

Bogdan Gabrys  
Robert J. Howlett  
Lakhmi C. Jain (Eds.)

LNAI 4253

# Knowledge-Based Intelligent Information and Engineering Systems

10th International Conference, KES 2006  
Bournemouth, UK, October 2006  
Proceedings, Part III

**3** Part III

 Springer

Lecture Notes in Artificial Intelligence 4253

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science



Bogdan Gabrys Robert J. Howlett  
Lakhmi C. Jain (Eds.)

# Knowledge-Based Intelligent Information and Engineering Systems

10th International Conference, KES 2006  
Bournemouth, UK, October 9-11, 2006  
Proceedings, Part III

## Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA  
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

## Volume Editors

Bogdan Gabrys  
Bournemouth University  
School of Design, Engineering and Computing  
Computational Intelligence Research Group  
Talbot Campus, Fern Barrow, Poole, GH12 5BB, UK  
E-mail: bgabrys@bournemouth.ac.uk

Robert J. Howlett  
University of Brighton  
School of Engineering  
Centre for SMART Systems, Brighton BN2 4GJ, UK  
E-mail: r.j.howlett@brighton.ac.uk

Lakhmi C. Jain  
University of South Australia  
School of Electrical and Information Engineering  
Knowledge-Based Intelligent Information and Engineering Systems Centre  
Adelaide, Mawson Lakes Campus, South Australia SA 5095, Australia  
E-mail: Lakhmi.Jain@unisa.edu.au

Library of Congress Control Number: 2006933827

CR Subject Classification (1998): I.2, H.4, H.3, J.1, H.5, K.6, K.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743  
ISBN-10 3-540-46542-1 Springer Berlin Heidelberg New York  
ISBN-13 978-3-540-46542-3 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2006  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper SPIN: 11893011 06/3142 5 4 3 2 1 0

## Preface

Delegates and friends, we are very pleased to extend to you the sincerest of welcomes to this, the 10th International Conference on Knowledge Based and Intelligent Information and Engineering Systems at the Bournemouth International Centre in Bournemouth, UK, brought to you by KES International.

This is a special KES conference, as it is the 10th in the series, and as such, it represents an occasion for celebration and an opportunity for reflection. The first KES conference was held in 1997 and was organised by the KES conference founder, Lakhmi Jain. In 1997, 1998 and 1999 the KES conferences were held in Adelaide, Australia. In 2000 the conference moved out of Australia to be held in Brighton, UK; in 2001 it was in Osaka, Japan; in 2002, Crema near Milan, Italy; in 2003, Oxford, UK; in 2004, Wellington, New Zealand; and in 2005, Melbourne, Australia. The next two conferences are planned to be in Italy and Croatia. Delegate numbers have grown from about 100 in 1997, to a regular figure in excess of 500. The conference attracts delegates from many different countries, in Europe, Australasia, the Pacific Rim, Asia and the Americas, and may truly be said to be 'International'. Formed in 2001, KES International has developed into a worldwide organisation that provides a professional community for researchers in the discipline of knowledge-based and intelligent engineering and information systems, and through this, opportunities for publication, networking and interaction. Published by IOS Press in the Netherlands, the KES Journal is organised by joint Editors-in-Chief R.J. Howlett and B. Gabrys. There are Associate Editors in the UK, the US, Poland, Australia, Japan, Germany and the Czech Republic. The Journal accepts academic papers from authors in many countries of the world and has approximately 600 subscribers in about 50 countries. KES produces a book series, also published by IOS Press, and there are plans for the development of focus groups, each with associated publications and symposia, and other ventures such as thematic summer schools.

The KES conference continues to be a major feature of the KES organisation, and KES 2006 looks set to continue the tradition of excellence in KES conferences. This year a policy decision was made not to seek to grow the conference beyond its current size, and to aim for about 450-500 papers based on their high quality. The papers for KES 2006 were either submitted to Invited Sessions, chaired and organised by respected experts in their fields, or to General Sessions, managed by Track Chairs and an extensive International Programme Committee. Whichever route they came through, all papers for KES 2006 were thoroughly reviewed. There were 1395 submissions for KES 2006 of which 480 were published, an acceptance rate of 34%.

Thanks are due to the very many people who have given their time and goodwill freely to make the conference a success.

We would like to thank the KES 2006 International Programme Committee who were essential in providing their reviews of the papers. We are very grateful for this service, without which the conference would not have been possible. We thank the high-profile keynote speakers for providing interesting and informed talks to catalyse subsequent discussions.

An important distinction of KES conferences over others is the Invited Session Programme. Invited Sessions give new and established researchers an opportunity to present a “mini-conference” of their own. By this means they can bring to public view a topic at the leading edge of intelligent systems. This mechanism for feeding new blood into the research is very valuable. For this reason we must thank the Invited Session Chairs who have contributed in this way.

The conference administrators, Maria Booth and Jo Sawyer at the Universities of Brighton and Bournemouth respectively, and the local committees, have all worked extremely hard to bring the conference to a high level of organisation, and our special thanks go to them too.

In some ways, the most important contributors to KES 2006 were the authors, presenters and delegates without whom the conference could not have taken place. So we thank them for their contributions.

In less than a decade, KES has grown from a small conference in a single country, to an international organisation involving over 1000 researchers drawn from universities and companies across the world. This is a remarkable achievement. KES is now in an excellent position to continue its mission to facilitate international research and co-operation in the area of applied intelligent systems, and to do this in new and innovative ways. We hope you all find KES 2006 a worthwhile, informative and enjoyable experience. We also hope that you will participate in some of the new activities we have planned for KES, for the next 10 years, to benefit the intelligent systems community.

August 2006

Bob Howlett  
Bogdan Gabrys  
Lakhmi Jain

# **KES 2006 Conference Organization**

## **Joint KES 2006 General Chairs**

Bogdan Gabrys  
Computational Intelligence Research Group  
School of Design, Engineering and Computing  
University of Bournemouth, UK

Robert J. Howlett  
Centre for SMART Systems  
School of Engineering  
University of Brighton, UK

## **Conference Founder and Honorary Programme Committee Chair**

Lakhmi C. Jain  
Knowledge-Based Intelligent Information and Engineering Systems Centre  
University of South Australia, Australia

## **Local Organising Committee (University of Brighton)**

Maria Booth, KES Administrator  
Shaun Lee, Simon Walters, Anna Howlett  
Nigel Shippam, PROSE Software Support  
Anthony Wood, Web Site Design

## **Local Organising Committee (University of Bournemouth)**

Jo Sawyer, KES 2006 Local Committee Coordinator  
Michael Haddrell, Mark Eastwood, Zoheir Sahel, Paul Rogers

## **International Programme Committee and KES 2006 Reviewers**

Abbass, Hussein	University of New South Wales, Australia
Abe, Akinori	ATR Intelligent Robotics and Communication Labs, Japan
Adachi, Yoshinori	Chubu University, Japan
Alpaslan, Ferda	Middle East Technical University, Turkey
Ang, Marcello	National University of Singapore, Singapore
Angelov, Plamen	Lancaster University, UK
Anguita, Davide	DIBE - University of Genoa, Italy
Anogeianakis, Giorgos	Aristotle University of Thessaloniki, Greece
Arotaritei, Dragos	Polytechnic Institute of Iasi, Romania

Arroyo-Figueroa, Gustavo	Electrical Research Institute of Mexico, Mexico
Asano, Akira	Hiroshima University, Japan
Asari, Vijayan	Old Dominion University, USA
Augusto, Juan	University of Ulster at Jordanstown, UK
Baba, Norio	Osaka Kyoiku University, Japan
Bajaj, Preeti	G.H. Rasoni College of Engineering, Nagpur, India
Bargiela, Andrzej	Nottingham Trent University, UK
Berthold, Michael	University of Konstanz, Germany
Berthouze, Nadia	University of Aizu, Japan
Binachi-Berthouze, Nadia	University of Aizu, Japan
Bod, Rens	University of St Andrews, UK
Bosc, Patrick	IRISA/ENSSAT, France
Bouvry, Pascal	University of Applied Sciences, Luxembourg
Brown, David	University of Portsmouth, UK
Burrell, Phillip	London South Bank University, UK
Cangelosi, Angelo	University of Plymouth, UK
Ceravolo, Paolo	University of Milan, Italy
Chakraborty, Basabi	Iwate Prefectural University, Japan
Chen, Yen-Wei	Ritsumeikan University, Japan
Chen-Burger, Yun-Heh	University of Edinburgh, UK
Chung, Jaehak	Inha University, Korea
Cios, Krzysztof	University of Colorado at Denver and Health Sciences Center, USA
Coello, Carlos A.	LANIA, Mexico
Coghill, George	University of Auckland, New Zealand
Corbett, Dan	SAIC, USA
Corchado, Emilio	University of Burgos, Spain
Correa da Silva, Flavio	University of Sao Paulo, Brazil
Cuzzocrea, Alfredo	University of Calabria, Italy
Damiani, Ernesto	University of Milan, Italy
Deep, Kusum	Indian Institute of Technology, Roorkee, India
Deng, Da	University of Otago, Dunedin, New Zealand
Dubey, Venky	University of Bournemouth, UK
Dubois, Didier	University Paul Sabatier, France
Duch, Wlodzislaw	Nicolaus Copernicus University, Poland
Eldin, Amr Ali	Delft University of Technology, The Netherlands
Far, Behrouz	University of Calgary, Canada
Finn, Anthony	DSTO, Australia
Flórez-López, Raquel	University of Leon, Spain
Fortino, Giancarlo	Università della Calabria, Italy
Fuchino, Tetsuo	Tokyo Institute of Technology, Japan
Fyfe, Colin	University of Paisley, UK
Gabrys, Bogdan	University of Bournemouth, UK

Galitsky, Boris	Birkbeck College University of London, UK
Ghosh, Ashish	Indian Statistical Institute, Kolkata, India
Girolami, Mark	University of Glasgow, UK
Gorodetski, Vladimir	St. Petersburg Institute of Informatics, Russia
Grana, Manuel	Universidad Pais Vasco, Spain
Grana Romay, Manuel	Universidad Pais Vasco, Spain
Grzech, Adam	Wroclaw University of Technology, Poland
Grzymala-Busse, Jerzy	University of Kansas, USA
Gu, Dongbing	University of Essex, UK
Håkansson, Anne	Uppsala University, Sweden
Hanh H., Phan	State University of New York, USA
Hansen, Lars Kai	Technical University of Denmark, Denmark
Harrison, Robert	The University of Sheffield, UK
Hasebrook, Joachim. P	University of Luebeck, Germany
Hatem, Ahriz	The Robert Gordon University, Aberdeen, UK
Hatzilygeroudis, Ioannis	University of Patras, Greece
Helic, Denis	Technical University of Graz, Austria
Hildebrand, Lars	University of Dortmund, Germany
Hirai, Yuzo	Institute of Information Sciences and Electronics, Japan
Hong, Tzung-Pei	National University of Kaohsiung, Taiwan
Honghai, Liu	The University of Aberdeen, UK
Hori, Satoshi	Institute of Technologists, Japan
Horio, Keiichi	Kyushu Institute of Technology, Japan
Howlett, Robert J.	University of Brighton, UK
Huaglory, Tianfield	Glasgow Caledonian University, UK
Illuminada, Baturone	University of Seville, Spain
Imada, Akira	Brest State Technical University, Belasus
Ishibuchi, Hisao	Osaka Prefecture University, Japan
Ishida, Yoshiteru	Toyohashi University of Technology, Japan
Ishida, Yoshiteru	Toyohashi University, Japan
Ishii, Naohiro	Aichi Institute of Technology, Japan
Jacquetnet, François	University of Saint-Etienne, France
Jadranka, Sunde	DSTO, Australia
Jain, Lakhmi C.	University of South Australia, Australia
Jarvis, Dennis	Agent Oriented Software Pty. Ltd., Australia
Jesse, Norbert	University of Dortmund, Germany
Kacprzyk, Janusz	Polish Academy of Sciences, Poland
Karacapilidis, Nikos	University of Patras, Greece
Karny, Miroslav	Institute of Information Theory and Automation, Czech Republic
Kasabov, Nik	Auckland University of Technology, New Zealand
Katarzyniak, Radoslaw	Wroclaw University of Technology, Poland

Kazuhiko, Tsuda	University of Tsukuba, Japan
Keskar, Avinash	Visvesvaraya National Institute of Technology, India
Kim, Dong-Hwa	Hanbat National University, Korea
Kim, Jong Tae	SungKyunKwan University, Republic of Korea
Kim, Sangkyun	Yonsei University, South Korea
Kim, Tai-hoon	Korea
Kittler, Josef	University of Surrey, UK
Kóczy, Tamás, László	Budapest University of Technology and Economics, Hungary
Koenig, Andreas	Technical University of Kaiserslautern, Germany
Kojiri, Tomoko	Nagoya University, Japan
Konar, Amit	Jadavpur University, India
Koshizen, Takamasa	Honda Research Institute Japan Co., Ltd., Japan
Kunifujii, Susumu	School of Knowledge Science, Japan
Kurgan, Lukasz	University of Alberta, Canada
Kusiak, Andrew	The University of Iowa, USA
Lanzi, Pier Luca	Politecnico di Milano, Italy
Lee, Dong Chun	Howon University, Korea
Lee, Geuk	Hannam Howon University, Korea
Lee, Hong Joo	Dankook University, South Korea
Lee, Hsuan-Shih	National Taiwan Ocean University, Taiwan
Lee, Raymond	Hong Kong Polytechnic University, Hong Kong, China
Liu, Yubao	Sun Yat-Sen University, China
Lovrek, Ignac	University of Zagreb, Croatia
Lu, Hongen	La Trobe University, Australia
Luccini, Marco	University of Pavia, Italy
Mackin, Kenneth J.	Tokyo University of Information Sciences, Japan
Main, J.	La Trobe University, Australia
Mandic, Danilo	Imperial College London, UK
Maojo, Victor	Universidad Politécnica de Madrid
Martin, Trevor	Bristol University, UK
Masulli, Francesco	University of Pisa, Italy
Mattila, Jorma	Lappeenranta University of Technology, Finland
Mazumdar, Jagannath	University of South Australia, USA
McKay, Bob	University of New South Wales, Australia
Mera, Kazuya	University of Hiroshima, Japan
Mesiar, Radko	STU Bratislava, Slovakia
Mira, Jose	ETS de Ingeniería Informática (UNED), Spain
Monekosso, Dorothy	University of Kingston, UK
Montani, Stefania	Università del Piemonte Orientale, Italy
Morch, Anders	University of Oslo, Norway
Munemori, Jun	Wakayama University, Japan



Munemorim Jun	Wakayama University, Japan
Murthy, Venu K.	RMIT University, Australia
Nakamatsu, Kazumi	University of Hyogo, Japan
Nakano, Ryohei	Nagoya Institute of Technology, Japan
Nakao, Zensho	University of Ryukyus, Japan
Nakashima, Tomoharu	Osaka Prefecture University, Japan
Narasimhan, Lakshmi	University of Newcastle, Australia
Nauck, Detlef	BT, UK
Navia-Vázquez, Angel	Univ. Carlos III de Madrid, Spain
Nayak, Richi	Queensland University of Technology, Brisbane, Australia
Neagu, Ciprian	University of Bradford, UK
Negoita, Mircea	KES, New Zealand
Nguyen, Ngoc Thanh	Wroclaw University of Technology, Poland
Nishida, Toyooki	University of Kyoto, Japan
Niskanen, Vesa A.	University of Helsinki, Finland
O'Connell, Robert	University of Missouri-Columbia, USA
Ong, Kok-Leong	Deakin University, Australia
Palade, Vasile	University of Oxford, UK
Palaniswami, Marimuthu	The University of Melbourne, Australia
Pant, Millie	BITS-Pilani, India
Papis, Costas	University of Piraeus, Greece
Paprzycki, Macin	Warsaw School of Social Psychology, Poland
Park, Gwi-Tae	Korea University, Korea
Pedrycz, Witold	University of Alberta, Canada
Peña-Reyes, Carlos-Andrés	Novartis Institutes for Biomedical Research, USA
Piedad, Brox	University of Seville, Spain
Polani, Daniel	University of Hertfordshire, UK
Popescu, Theodor	National Institute for Research and Development Informatics, Romania
Rajesh, R.	Bharathiar University, India
Reusch, Bernd	University of Dortmund, Germany
Rhee, Phill	Inha University, Korea
Rose, John	Ritsumeikan Asia Pacific University, Japan
Ruan, Da	The Belgian Nuclear Research Centre, Belgium
Ruta, Dymitr	BT Exact, UK
Rutkowski, Leszek	Technical University of Czestochowa, Poland
Sato-Ilic, Mika	University of Tsukuba, Japan
Sawada, Hideyuki	Kagawa University, Japan
Seiffert, Udo	Leibniz-Institute of Plant Genetics, Gatersleben, Germany
Semeraro, Giovanni	Università degli Studi di Bari, Italy
Sharma, Dharmendra	University of Canberra, Australia

Sirlantzis, Konstantinos	University of Kent, UK
Skabar, A.	La Trobe University, Australia
Sobecki, Janusz	Wroclaw University of Technology, Poland
Soo, Von-Wun	National University of Kaohsiung, Taiwan
Sordo, Margarita	Harvard Medical School, USA
Stumptner, Markus	University of South Australia, Australia
Stytz, Martin	Institute for Defense Analyses, USA
Suetake, Noriaki	Yamaguchi University, Japan
Sujitjorn, Sarawut	Suranaree University of Technology
Sun, Zhahao	University of Wollongong, Australia
Szczerbicki, Edward	University of Newcastle, Australia
Takahash, Masakazu	Simane University, Japan
Taki, Hirokazu	Wakayama University, Japan
Tanaka, Takushi	Fukuoka Institute of Technology, Japan
Tanaka-Yamawaki, Mieko	Tottori University, Japan
Teodorescu, Horia-Nicolai	Romanian Academy, Romania
Thalmann, Daniel	EPFL, Switzerland
Thatcher, Steven	University of South Australia, Australia
Tolk, Andreas	Virginia Modeling Analysis & Simulation center, USA
Torresen, Jim	University of Oslo, Norway
Treur, Jan	Vrije Universiteit Amsterdam, Netherlands
Turchetti, Claudio	Università Politecnica delle Marche, Italy
Tweedale, J.	DSTO, Australia
Uchino, Eiji	Yamaguchi University, Japan
Unland, Rainer	University of Duisburg-Essen, Germany
Verdegay, JoseLuis	University of Granada, Spain
Virvou, Maria	University of Piraeus, Greece
Walters, Simon	University of Brighton, UK
Wang, Dianhui	La Trobe University, Australia
Wang, Lipo	Nanyang Tech University, Singapore
Wang, Pei	Temple University, USA
Watada, Junzo	Waseda University, Japan
Watanabe, Keigo	Saga University, Japan
Watanabe, Toyohide	Nagoya University, Japan
Wermter, Stefan	University of Sunderland, UK
Wren, Gloria	Loyola College in Maryland, USA
Yamashita, Yoshiyuko	Tohoku University, Sedai, Japan
Yoo, Seong-Joon	Sejong University, Korea
Zahlmann, Gudrun	Siemens Medical Solutions; Med Projekt CTB, Germany
Zambarbieri, Daniela	University of Pavia, Italy
Zha, Xuan	NIST, USA

Zharkova, Valentina	Bradford University, Bradford, UK
Zhiwen, Yu	Northwestern Polytechnical University, China
Zurada, Jacek	University of Louisville, USA

## General Track Chairs

### Generic Intelligent Systems Topics

Track Title	Track Chair
Artificial Neural Networks and Connectionists Systems	Ryohei Nakano, Nagoya Institute of Technology, Japan
Fuzzy and Neuro-Fuzzy Systems	Detlef Nauck, BT, UK
Evolutionary Computation	Zensho Nakao, University of Ryukyus, Japan
Machine Learning and Classical AI	Mark Girolami, University of Glasgow, UK
Agent Systems	Ngoc Thanh Nguyen, Wroclaw University of Technology, Poland
Knowledge Based and Expert Systems	Anne Hakansson, Uppsala University, Sweden
Hybrid Intelligent Systems	Vasile Palade, Oxford University, UK
Miscellaneous Intelligent Algorithms	Honghai Liu, University of Portsmouth, UK

### Applications of Intelligent Systems

Track Title	Track Chair
Intelligent Vision and Image Processing	Tuan Pham, James Cook University, Australia
Intelligent Data Mining	Michael Berthold, University of Konstanz, Germany
Knowledge Management and Ontologies	Edward Szczerbicki, University of Newcastle, Australia
Web Intelligence, Multimedia, e-Learning and Teaching	Andreas Nuernberger, University of Magdeburg, Germany
Intelligent Signal Processing, Control and Robotics	Miroslav Karny, Academy of Science, Czech Republic
Other Intelligent Systems Applications	Anthony Finn, Defence Science & Technology Organisation, Australia

## Invited Session Chairs

Zhaohao Sun, University of Wollongong, Australia  
Gavin Finnie, Bond University, Queensland, Australia  
R.J. Howlett, University of Brighton, UK  
Naohiro Ishii, Aichi Institute of Technology, Japan  
Yuji Iwahori, Chubu University, Japan  
Sangkyun Kim, Yonsei University, South Korea  
Hong Joo Lee, Dankook University, South Korea  
Yen-Wei Chen, Ritsumeikan University, Japan  
Mika Sato-Ilic, University of Tsukuba, Japan  
Yoshiteru Ishida, Toyohashi University of Technology, Japan  
Dorothy Monekosso, Kingston University, UK  
Toyoaki Nishida, The University of Kyoto, Japan  
Ngoc Thanh Nguyen, Wroclaw University of Technology, Poland  
Rainer Unland, University of Duisburg-Essen, Germany  
Tsuda Kazuhiko, The University of Tsukuba, Japan  
Masakazu Takahash, Simane University, Japan  
Daniela Zambarbieriini, Università di Pavia, Italy  
Angelo Marco Luccini, Giunti Interactive Labs, Italy  
Baturone Iluminada, Institute de Microelectronica de Sevilla, University of Seville,  
Spain  
Brox Poedad, Institute de Microelectronica de Sevilla, University of Seville, Spain  
David Brown, University of Portsmouth, UK  
Bogdan Gabrys, University of Bournemouth, UK  
Davide Anguita, DIBE - University of Genoa, Italy  
P. Urlings, DSTO, Australia  
J. Tweedale, DSTO, Australia  
C. Sioutis, DSTO, Australia  
Gloria Wren, Loyola College in Maryland, USA  
Nikhil Ichalkaranje, UNISA, Australia  
Yoshiyuki Yamashita, Tohoku University, Sendai, Japan  
Tetsuo Fuchino, Tokyo Institute of Technology, Japan  
Hirokazu Taki, Wakayama University, Japan  
Satoshi Hori, Institute of Technologists, Japan  
Raymond Lee, Hong Kong Polytechnic University, China  
Tai-hoon Kim, Korea University of Technology and Education, Republic of Korea  
Kazumi Nakamatsu, University of Hyogo, Japan  
Hsuan-Shih Lee, National Taiwan Ocean University, Taiwan  
Ryohei Nakano, Nagoya Institute of Technology, Japan  
Kazumi Saito, NTT Communication Science Laboratories, Japan  
Giorgos Anogeianakis, Aristotle University of Thessaloniki, Greece  
Toyohide Watanabe, Nagoya University, Japan

Tomoko Kojiri, Nagoya University, Japan  
Naoto Mukai, Nagoya University, Japan  
Maria Virvou, University of Piraeus, Greece  
Yoshinori Adachi, Chubu University, Japan  
Nobuhiro Inuzuka, Nagoya Institute of Technology, Japan  
Jun Feng, Hohai University, China  
Ioannis Hatzilygeroudis, University of Patras, Greece  
Constantinos Koutsojannis, University of Patras, Greece  
Akinori Abe, ATR Intelligent Robotics & Communication Labs, Japan  
Shoji, ATR Intelligent Robotics & Communication Labs, Japan  
Ohsawa, ATR Intelligent Robotics & Communication Labs, Japan  
Phill Kyu Rhee, Inha University, Korea  
Rezaul Bashar, Inha University, Korea  
Jun Munemori, Wakayama University, Japan  
Takashi Yoshino, Wakayama University, Japan  
Takaya Yuizono, Shimane University, Japan  
Gwi-Tae Park, Korea University, South Korea  
Manuel Grana, Universidad Pais Vasco, Spain  
Richard Duro, Universidad de A Coruna, Spain  
Daniel Polani, University of Hertfordshire, UK  
Mikhail Prokopenko, CSIRO, Australia  
Dong Hwa Kim, Hanbat National University, Korea  
Vesa A. Niskanen, University of Helsinki, Finland  
Emilio Corchado, University of Burgos, Spain  
Hujun Yun, University of Manchester, UK  
Jaehak Chung, Inha University Korea, South Korea  
Da Deng, University of Otago, New Zealand  
Mengjie Zhang, University of Wellington, New Zealand  
Valentina Zharkova, University of Bradford, UK  
Jie Zhang, George Mason University, USA  
Richi Nayak, Queensland University of Technology, Australia  
Lakhmi Jain, University of South Australia, Australia  
Dong Chun Lee, Howon University, Korea  
Giovanni Semeraro, Università degli Studi di Bari, Italy  
Eugenio Di Sciascio, Politecnico di Bari, Italy  
Tommaso Di Noia, Politecnico di Bari, Italy  
Norio Baba, Osaka Kyoiku University, Japan  
Takumi Ichimura, Hiroshima City University, Japan  
Kazuya Mera, Hiroshima City University, Japan  
Janusz Sobecki, Wroclaw University of Technology, Poland  
Przemyslaw Kazienko, Wroclaw University of Technology, Poland  
Dariusz Król, Wroclaw University of Technology, Poland

Kenneth J. Mackin, Tokyo University of Information Sciences, Japan  
Susumu Kunifuji, Japan Advanced Institute of Science and Technology, Japan  
Motoki Miura, Japan Advanced Institute of Science and Technology, Japan  
Jong Tae Kim, SungKyunKwan University, Republic of Korea  
Junzo Watada, Waseda University, Japan  
Radoslaw Katarzyniak, Wroclaw University of Technology, Poland  
Geuk Lee, Hannam Howon University, Korea  
Il Seok Ko, Chungbuk Provincial Univ., Korea  
Ernesto Damiani, University of Milan, Italy  
Paolo Ceravolo, University of Milan, Italy  
Dharmendra Sharma, University of Canberra, Australia  
Bala Balachandran, University of Canberra, Australia  
Wanla Ma, University of Canberra, Australia  
Danilo P. Mandic, Imperial College London, UK  
Tomasz Rutkowski, RIKEN, Japan  
Toshihisa Tanaka, TUAT, Tokyo, Japan  
Martin Golz, Schmalkalden, Germany

## **Keynote Lectures**

Evolving Intelligent Systems: Methods and Applications

*Nikola Kasabov, Knowledge Engineering and Discovery Research Institute (KEDRI),  
Auckland University of Technology, New Zealand*

Feature Selection in Pattern Recognition

*Josef Kittler, University of Surrey, UK*

Ant Colony Optimisation

*Luca Maria Gambardella, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale  
(IDSIA), Switzerland*

Evolvable Genetic Robots in a Ubiquitous World

*Jong-Hwan Kim, Korea Advanced Institute of Science and Technology (KAIST),  
Republic of Korea*

Industrial Applications of Intelligent Systems

*Ben Azvine, BT, UK*

Towards Multimodal Computing: Extracting Information from Signal Nonlinearity  
and Determinism

*Danilo Mandic, Imperial College London, UK*

# Table of Contents – Part III

## Chance Discovery

Closed-Ended Questionnaire Data Analysis . . . . .	1
<i>Leuo-Hong Wang, Chao-Fu Hong, Chia-Ling Hsu</i>	
Strategies Emerging from Game Play: Experience from an Educational Game for Academic Career Design . . . . .	8
<i>Yoshikiyo Kato, Hiroko Shoji</i>	
Concept Sharing Through Interaction: The Effect of Information Visualization for Career Design . . . . .	16
<i>Hiroko Shoji, Kaori Komiya, Toshikazu Kato</i>	
What Should Be Abducible for Abductive Nursing Risk Management? . . . . .	22
<i>Akinori Abe, Hiromi Itoh Ozaku, Noriaki Kuwahara, Kiyoshi Kogure</i>	
The Repository Method for Chance Discovery in Financial Forecasting . . . . .	30
<i>Alma Lilia Garcia-Almanza, Edward P.K. Tsang</i>	
Emerging Novel Scenarios of New Product Design with Teamwork on Scenario Maps Using Pictorial KeyGraph . . . . .	38
<i>Kenichi Horie, Yukio Ohsawa</i>	
Creative Design by Bipartite KeyGraph Based Interactive Evolutionary Computation . . . . .	46
<i>Chao-Fu Hong, Hsiao-Fang Yang, Mu-Hua Lin, Geng-Sian Lin</i>	
Discovering Chances for Organizational Training Strategies from Intra-group Relations . . . . .	57
<i>Meera Patel, Ruediger Oehlmann</i>	
Trust, Ethics and Social Capital on the Internet: An Empirical Study Between Japan, USA and Singapore . . . . .	64
<i>Yumiko Nara</i>	

## Context Aware Evolvable and Adaptable Systems and Their Applications

Context-Aware Application System for Music Playing Services . . . . .	76
<i>Jae-Woo Chang, Yong-Ki Kim</i>	
Query Based Summarization Using Non-negative Matrix Factorization . . . . .	84
<i>Sun Park, Ju-Hong Lee, Chan-Min Ahn, Jun Sik Hong, Seok-Ju Chun</i>	
Intelligent Secure Web Service Using Context Information . . . . .	90
<i>Woochul Shin, Xun Li, Sang Bong Yoo</i>	
Study for Intelligent Guide System Using Soft Computing . . . . .	101
<i>Woo-kyung Choi, Sang-Hyung Ha, Seong-Joo Kim, Hong-Tae Jeon</i>	
Implementation of a FIR Filter on a Partial Reconfigurable Platform . . .	108
<i>Hanho Lee, Chang-Seok Choi</i>	
A Context Model for Ubiquitous Computing Applications . . . . .	116
<i>Md. Rezaul Bashar, Mi Young Nam, Phill Kyu Rhee</i>	
Adaptive Classifier Selection on Hierarchical Context Modeling for Robust Vision Systems . . . . .	124
<i>SongGuo Jin, Eun Sung Jung, Md. Rezaul Bashar, Mi Young Nam, Phill Kyu Rhee</i>	

## Advanced Groupware and Network Services

Wireless Interent Service of Visited Mobile ISP Subscriber on GPRS Network . . . . .	135
<i>Jeong-Hyun Park, Boo-Hyung Lee</i>	
An Exploratory Analysis on User Behavior Regularity in the Mobile Internet . . . . .	143
<i>Toshihiko Yamakami</i>	
Reliable Communication Methods for Mutual Complementary Networks . . . . .	150
<i>Takashi Kaneyama, Hiroshi Mineno, Takashi Furumura, Kunihiko Yamada, Tadanori Mizuno</i>	



Remote Plug and Play USB Devices for Mobile Terminals . . . . .	159
<i>Jungo Kuwahara, Hiroshi Mineno, Kiyoko Tanaka, Hideharu Suzuki, Norihiro Ishikawa, Tadanori Mizuno</i>	
SIP-Based Streaming Control Architecture for mobile Personal Area Networks . . . . .	166
<i>Hiroshi Mineno, Naoto Adachi, Tsutomu Eto, Kiyoko Tanaka, Hideharu Suzuki, Norihiro Ishikawa, Tadanori Mizuno</i>	
Development of Middleware GLIA for Connective Wide Collaborative Space with Networked I/O Device . . . . .	174
<i>Takaya Yuizono, Shinichi Nishimura, Jun Munemori</i>	
Development and Evaluation of No-Record Chat System Against Screen Capture . . . . .	181
<i>Takashi Yoshino, Takuma Matsuno</i>	
Home-Network of a Mutual Complement Communication System by Wired and Wireless . . . . .	189
<i>Kunihiro Yamada, Takashi Furumura, Yasuhiro Seno, Yukihisa Naoe, Kenichi Kitazawa, Toru Shimizu, Koji Yoshida, Masanori Kojima, Hiroshi Mineno, Tadanori Mizuno</i>	
Interaction Between Small Size Device and Large Screen in Public Space . . . . .	197
<i>Chunming Jin, Shin Takahashi, Jiro Tanaka</i>	
Development of a Medical Information Gathering System Using JMA Standard Health Insurance Claim System ORCA on IPv6 Network . . . . .	205
<i>Takashi Yoshino, Yoko Yukimura, Kunikazu Fujii, Yoshiki Kusumoto, Masayuki Irie</i>	
A Prototype of a Chat System Using Message Driven and Interactive Actions Character . . . . .	212
<i>Junko Itou, Kenji Hoshio, Jun Munemori</i>	

## Computational Methods for Intelligent Neuro-fuzzy Applications

Genetically Optimized Fuzzy Set-Based Polynomial Neural Networks Based on Information Granules with Aids of Symbolic Genetic Algorithms . . . . .	219
<i>Tae-Chon Ahn, Kyung-Won Jang, Seok-Beom Roh</i>	

Evaluation of the Distributed Fuzzy Contention Control for IEEE 802.11 Wireless LANs ..... 225  
*Young-Joong Kim, Jeong-On Lee, Myo-Taeg Lim*

Improved Genetic Algorithm-Based FSMC Design for Multi-nonlinear System ..... 233  
*Jung-Shik Kong, Jin-Geol Kim*

The Interactive Feature Selection Method Development for an ANN Based Emotion Recognition System ..... 241  
*Chang-Hyun Park, Kwee-Bo Sim*

Supervised IAFC Neural Network Based on the Fuzzification of Learning Vector Quantization ..... 248  
*Yong Soo Kim, Sang Wan Lee, Sukhoon Kang, Yong Sun Baik, Suntae Hwang, Zeunghnam Bien*

Walking Pattern Analysis of Humanoid Robot Using Support Vector Regression ..... 255  
*Dongwon Kim, Gwi-Tae Park*

**Intelligent Information Processing for Remote Sensing**

Low-Cost Stabilized Platform for Airborne Sensor Positioning ..... 263  
*Francisco J. González-Castaño, F. Gil-Castiñeira, J.M. Pousada-Carballo, P.S. Rodríguez-Hernández, J.C. Burguillo-Rial, I. Dosil-Outes*

Comparison of Squall Line Positioning Methods Using Radar Data ..... 269  
*Ka Yan Wong, Chi Lap Yip*

On Clustering Performance Indices for Multispectral Images ..... 277  
*Carmen Hernández, Josune Gallego, M. Teresa García-Sebastian, Manuel Graña*

Aerial Photo Image Retrieval Using Adaptive Image Classification ..... 284  
*Sung Wook Baik, Moon Seok Jeong, Ran Baik*

Integration of Spatial Information in Hyperspectral Imaging for Real Time Quality Control in an Andalusite Processing Line ..... 292  
*Abraham Prieto, F. Bellas, Fernando López-Peña, Richard J. Duro*

A Windowing/Pushbroom Hyperspectral Imager . . . . .	300
<i>Beatriz Couce, Xesus Prieto-Blanco, Carlos Montero-Orille, Raúl de la Fuente</i>	

## Evolutionary and Self-organising Sensors, Actuators and Processing Hardware

Genetic Programming of a Microcontrolled Water Bath Plant . . . . .	307
<i>Douglas Mota Dias, Marco Aurélio C. Pacheco, José Franco M. Amaral</i>	

Symbiotic Sensor Networks in Complex Underwater Terrains: A Simulation Framework . . . . .	315
<i>Vadim Gerasimov, Gerry Healy, Mikhail Prokopenko, Peter Wang, Astrid Zeman</i>	

Predicting Cluster Formation in Decentralized Sensor Grids . . . . .	324
<i>Astrid Zeman, Mikhail Prokopenko</i>	

Making Sense of the Sensory Data – Coordinate Systems by Hierarchical Decomposition . . . . .	333
<i>Attila Egri-Nagy, Chrystopher L. Nehaniv</i>	

Biologically-Inspired Visual-Motor Coordination Model in a Navigation Problem . . . . .	341
<i>Jacek Jelonek, Maciej Komosinski</i>	

A Self-organising Sensing System for Structural Health Management . . .	349
<i>Nigel Hoschke, Chris J. Lewis, Don C. Price, D. Andrew Scott, Graeme C. Edwards, Adam Batten</i>	

## Human Intelligent Technology

On Models in Fuzzy Propositional Logic . . . . .	358
<i>Jorma K. Mattila</i>	

Is Soft Computing in Technology and Medicine Human-Friendly? . . . . .	366
<i>Rudolf Seising, Jeremy Bradley</i>	

From Vague or Loose Concepts to Hazy and Fuzzy Sets – Human Understanding Versus Exact Science . . . . .	374
<i>Rudolf Seising, Jeremy Bradley</i>	

New Classifier Based on Fuzzy Level Set Subgrouping . . . . .	383
<i>Paavo Kukkurainen, Pasi Luukka</i>	

## Connectionist Models and Their Applications

Fast Shape Index Framework Based on Principle Component Analysis Using Edge Co-occurrence Matrix . . . . .	390
<i>Zhiping Xu, Yiping Zhong, Shiyong Zhang</i>	
Stock Index Modeling Using Hierarchical Radial Basis Function Networks . . . . .	398
<i>Yuehui Chen, Lizhi Peng, Ajith Abraham</i>	
Mathematical Formulation of a Type of Hierarchical Neurofuzzy System . . . . .	406
<i>Omar Sánchez, Sixto Romero, Francisco Moreno, Miguel A. Vélez</i>	
Hardware Support for Language Aware Information Mining . . . . .	415
<i>Michael Freeman, Thimal Jayasooriya</i>	
Measuring GNG Topology Preservation in Computer Vision Applications . . . . .	424
<i>José García Rodríguez, Francisco Flórez-Revueita, Juan Manuel García Chamizo</i>	
Outlier Resistant PCA Ensembles . . . . .	432
<i>Bogdan Gabrys, Bruno Baruque, Emilio Corchado</i>	

## Intelligent and Cognitive Communication Systems

Intelligent Channel Time Allocation in Simultaneously Operating Piconets Based on IEEE 802.15.3 MAC . . . . .	441
<i>Peng Xue, Peng Gong, Duk Kyung Kim</i>	
An Intelligent Synchronization Scheme for WLAN Convergence Systems . . . . .	449
<i>Sekchin Chang, Jaehak Chung</i>	
Aggressive Sub-channel Allocation Algorithm for Intelligent Transmission in Multi-user OFDMA System . . . . .	457
<i>SangJun Ko, Joo Heo, Yupeng Wang, KyungHi Chang</i>	
A Novel Spectrum Sharing Scheme Using Relay Station with Intelligent Reception . . . . .	465
<i>Jaehwan Kim, Junkyu Lee, Joonhyuk Kang</i>	
Intelligent Beamforming for Personal Area Network Systems . . . . .	473
<i>Younggun Ji, Seokhyun Kim, Hongwon Lee, Jaehak Chung</i>	

Multiband Radio Resource Allocation for Cognitive Radio Systems . . . . .	480
<i>Yangsoo Kwon, Jungwon Suh, Jaehak Chung, Joohee Kim</i>	

## **Innovations in Intelligent Agents and Applications**

Testing Online Navigation Recommendations in a Web Site . . . . .	487
<i>Juan D. Velásquez, Vasile Palade</i>	
An Overview of Agent Coordination and Cooperation . . . . .	497
<i>Angela Consoli, Jeff Tweedale, Lakhmi Jain</i>	
Innovations in Intelligent Agents and Web . . . . .	504
<i>Gloria Phillips-Wren, Lakhmi Jain</i>	

## **Intelligent Pattern Recognition and Classification in Astrophysical and Medical Images**

Robust Segmentation for Left Ventricle Based on Curve Evolution . . . . .	507
<i>Gang Yu, Yuxiang Yang, Peng Li, Zhengzhong Bian</i>	
Segmentation of Ovarian Ultrasound Images Using Cellular Neural Networks Trained by Support Vector Machines . . . . .	515
<i>Boris Cigale, Mitja Lenič, Damjan Zazula</i>	
Bayesian Decision Tree Averaging for the Probabilistic Interpretation of Solar Flare Occurrences . . . . .	523
<i>Vitaly Schetinin, Valentina Zharkova, Sergei Zharkov</i>	

## **Intelligent Multimedia Solutions and Security in the Next Generation Mobile Information Systems**

An Efficient 3-D Positioning Method from Satellite Synthetic Aperture Radar Images . . . . .	533
<i>Yeong-Sun Song, Hong-Gyoo Sohn, Choung-Hwan Park</i>	
Generalized Logit Model of Demand Systems for Energy Forecasting . . .	541
<i>Hong Sok Kim, Hoon Chang, Young- Kyun Lee</i>	
Secure Authentication Protocol in Mobile IPv6 Networks . . . . .	548
<i>Jung Doo Koo, Jung Sook Koo, Dong Chun Lee</i>	

Experiments and Experiences on the Relationship Between the Probe Vehicle Size and the Travel Time Collection Reliability ..... 556  
*Chungwon Lee, Seungjae Lee, Taehee Kim, Jeong Hyun Kim*

Conversion Scheme for Reducing Security Vulnerability in IPv4/ IPv6 Networks ..... 564  
*Do Hyeon Lee, Jeom Goo Kim*

Improved Location Management for Reducing Traffic Cost in 3G Mobile Networks ..... 572  
*Jae Young Koh*

**Engineered Applications of Semantic Web – SWEA**

A Semantic Web Portal for Semantic Annotation and Search ..... 580  
*Norberto Fernández-García, José M. Blázquez-del-Toro, Jesús Arias Fisteus, Luis Sánchez-Fernández*

A Semantic Portal for Fund Finding in the EU: Semantic Upgrade, Integration and Publication of Heterogeneous Legacy Data ..... 588  
*Jesús Barrasa Rodríguez, Oscar Corcho, Asunción Gómez-Pérez*

A Heuristic Approach to Semantic Web Services Classification ..... 598  
*Miguel Ángel Corella, Pablo Castells*

A RDF-Based Framework for User Profile Creation and Management ... 606  
*Ignazio Palmisano, Domenico Redavid, Luigi Iannone, Giovanni Semeraro, Marco Degemmis, Pasquale Lops, Oriana Licchelli*

Integrated Document Browsing and Data Acquisition for Building Large Ontologies ..... 614  
*Felix Weigel, Klaus U. Schulz, Levin Brunner, Eduardo Torres-Schumann*

A Workflow Modeling Framework Enhanced with Problem-Solving Knowledge ..... 623  
*Juan Carlos Vidal, Manuel Lama, Alberto Bugarín*

M-OntoMat-Annotizer: Image Annotation Linking Ontologies and Multimedia Low-Level Features ..... 633  
*Kosmas Petridis, Dionysios Anastasopoulos, Carsten Saathoff, Norman Timmermann, Yiannis Kompatsiaris, Steffen Staab*

## Intelligent Techniques in the Stock Market

On the Profitability of Scalping Strategies Based on Neural Networks . . .	641
<i>Marina Resta</i>	
Effect of Moving Averages in the Tickwise Tradings in the Stock Market . . . . .	647
<i>Felix Streichert, Mieko Tanaka-Yamawaki, Masayuki Iwata</i>	
A New Scheme for Interactive Multi-criteria Decision Making . . . . .	655
<i>Felix Streichert, Mieko Tanaka-Yamawaki</i>	
Utilization of NNs for Improving the Traditional Technical Analysis in the Financial Markets . . . . .	663
<i>Norio Baba</i>	

## Soft Computing Techniques and Their Applications

Use of Support Vector Machines: Synergism to Intelligent Humanoid Robot Walking Down on a Slope . . . . .	670
<i>Dongwon Kim, Gwi-Tae Park</i>	
Evolutionary Elementary Cooperative Strategy for Global Optimization . . . . .	677
<i>Crina Grosan, Ajith Abraham, Monica Chis, Tae-Gyu Chang</i>	
Human Hand Detection Using Evolutionary Computation for Gestures Recognition of a Partner Robot . . . . .	684
<i>Setsuo Hashimoto, Naoyuki Kubota, Fumio Kojima</i>	
A Simple 3D Edge Template for Pose Invariant Face Detection . . . . .	692
<i>Stephen Karungaru, Minoru Fukumi, Norio Akamatsu, Takuya Akashi</i>	
Emergence of Flocking Behavior Based on Reinforcement Learning . . . . .	699
<i>Koichiro Morihiro, Teiji Isokawa, Haruhiko Nishimura, Nobuyuki Matsui</i>	

## Human Computer Intelligent Systems

Real Time Head Nod and Shake Detection Using HMMs . . . . .	707
<i>Yeon Gu Kang, Hyun Jea Joo, Phill Kyu Rhee</i>	

Construction and Evaluation of Text-Dialog Corpus with Emotion Tags Focusing on Facial Expression in Comics . . . . .	715
<i>Masato Tokuhisa, Jin'ichi Murakami, Satoru Ikehara</i>	
Human Support Network System Using Friendly Robot . . . . .	725
<i>Masako Miyaji, Toru Yamaguchi, Eri Sato, Koji Kanagawa</i>	
Developing a Decision Support System for a Dove's Voice Competition . . . . .	733
<i>Chotirat Ann Ratanamahatana</i>	
Time Series Data Classification Using Recurrent Neural Network with Ensemble Learning . . . . .	742
<i>Shinichi Oeda, Ikusaburo Kurimoto, Takumi Ichimura</i>	
Expressed Emotion Calculation Method According to the User's Personality . . . . .	749
<i>Kazuya Mera, Takumi Ichimura</i>	
<b>Recommender Agents and Adaptive Web-Based Systems</b>	
Integrated Agent-Based Approach for Ontology-Driven Web Filtering . . .	758
<i>David Sánchez, David Isern, Antonio Moreno</i>	
A Constrained Spreading Activation Approach to Collaborative Filtering . . . . .	766
<i>Josephine Griffith, Colm O'Riordan, Humphrey Sorensen</i>	
Fuzzy Model for the Assessment of Operators' Work in a Cadastre Information System . . . . .	774
<i>Dariusz Król, Grzegorz Stanisław Kukla, Tadeusz Lasota, Bogdan Trawiński</i>	
Local Buffer as Source of Web Mining Data . . . . .	782
<i>Andrzej Siemiński</i>	
Lessons from the Application of Domain-Independent Data Mining System for Discovering Web User Access Patterns . . . . .	789
<i>Leszek Borzemski, Adam Druszcz</i>	
Application of Hybrid Recommendation in Web-Based Cooking Assistant . . . . .	797
<i>Janusz Sobecki, Emilia Babiak, Marta Stanina</i>	



Using Representation Choice Methods for a Medical Diagnosis Problem .....	805
<i>Kamila Aftarczuk, Adrianna Kozierekiewicz, Ngoc Thanh Nguyen</i>	

## Intelligent Data Analysis for Complex Environments

Construction of School Temperature Measurement System with Sensor Network .....	813
<i>Ayahiko Niimi, Masaaki Wada, Kei Ito, Osamu Konishi</i>	
Land Cover Classification from MODIS Satellite Data Using Probabilistically Optimal Ensemble of Artificial Neural Networks .....	820
<i>Kenneth J. Mackin, Eiji Nunohiro, Masanori Ohshiro, Kazuko Yamasaki</i>	

## Creativity Support Systems

Structure-Based Categorization of Programs to Enable Awareness About Programming Skills .....	827
<i>Kei Kato, Toyohide Watanabe</i>	
On-Screen Note Pad for Creative Activities .....	835
<i>Norikazu Iwamura, Kazuo Misue, Jiro Tanaka</i>	
IdeaCrepe: Creativity Support Tool with History Layers .....	843
<i>Nagayoshi Nakazono, Kazuo Misue, Jiro Tanaka</i>	
Personalized Voice Navigation System for Creative Working Environment .....	851
<i>Kaoru Tanaka, Susumu Kunifuji</i>	
An Editing and Displaying System of Olfactory Information for the Home Video .....	859
<i>Dong Wook Kim, Kazushi Nishimoto, Susumu Kunifuji</i>	
A Divergent-Style Learning Support Tool for English Learners Using a Thesaurus Diagram .....	867
<i>Chie Shimodaira, Hiroshi Shimodaira, Susumu Kunifuji</i>	

## Artificial Intelligence Applications in Power Electronics

Full Fuzzy-Logic-Based Vector Control for Permanent Magnet Synchronous Motors .....	875
<i>Jae-Sung Yu, Byoung-Kuk Lee, Chung-Yuen Won, Dong-Wook Yoo</i>	

Artificial Intelligent Application to Service Restoration Considering Load Balancing in Distribution Networks .....	883
<i>Sang-Yule Choi, Jae-Sang Cha, Myong-Chul Shin</i>	
Minimum Cost Operation Mode and Minimum Loss Operation Mode of Power System – Operation Mode Selection Based on Voltage Stability .....	893
<i>Sang-Joong Lee</i>	
Optimal Voltage and Reactive Power Control of Local Area Using Genetic Algorithm .....	900
<i>Hak-Man Kim, Jong-Yul Kim, Chang-Dae Yoon, Myong-Chul Shin, Tae-Kyoo Oh</i>	
Equivalent Electric Circuit Modeling of Differential Structures in PCB with Genetic Algorithm .....	907
<i>Jong Kang Park, Yong Ki Byun, Jong Tae Kim</i>	
Rule-Based Expert System for Designing DC-DC Converters .....	914
<i>Seok Min Yoon, Jong Tae Kim</i>	

## **Soft Computing Approaches to Management Engineering**

DNA-Based Evolutionary Algorithm for Cable Trench Problem .....	922
<i>Don Jyh-Fu Jeng, Ikno Kim, Junzo Watada</i>	
The Influences of R&D Expenditures on Knowledge-Based Economy in Taiwan .....	930
<i>Lily Lin</i>	
A Process Schedule Analyzing Model Based on Grid Environment .....	938
<i>Huey-Ming Lee, Tsang-Yean Lee, Mu-Hsiu Hsu</i>	
Fuzzy Set Theoretical Approach to the RGB Color Triangle .....	948
<i>Naotoshi Sugano</i>	
Spatial Equilibrium Model on Regional Analysis for the Trade Liberalization of Fluid Milk in Taiwan .....	956
<i>Lily Lin</i>	
Analysing the Density of Subgroups in Valued Relationships Based on DNA Computing .....	964
<i>Ikno Kim, Don Jyh-Fu Jeng, Junzo Watada</i>	

Structural Learning of Neural Networks for Forecasting Stock Prices . . . .	972
<i>Junzo Watada</i>	

Customer Experience Management Influencing on Human <i>Kansei</i> to MOT . . . . .	980
<i>Shin'ya Nagasawa</i>	

Getting Closer to the Consumer: The Digitization of Content Distribution . . . . .	988
<i>Peter Anshin, Hisao Shiizuka</i>	

## **Knowledge Processing in Intelligent and Cognitive Communicative Systems**

An Action Planning Module Based on Vague Knowledge Extracted from Past Experiences . . . . .	997
<i>Grzegorz Popiek</i>	

An Approach to Resolving Semantic Conflicts of Temporally-Vague Observations in Artificial Cognitive Agent . . . . .	1004
<i>Wojciech Lorkiewicz</i>	

Neural Network Approach for Learning of the World Structure by Cognitive Agents . . . . .	1012
<i>Agnieszka Pieczyńska, Jarosław Drapała</i>	

Towards a Computational Framework for Modeling Semantic Interactions in Large Multiagent Communities . . . . .	1020
<i>Krzysztof Juszczyszyn</i>	

Grounding Crisp and Fuzzy Ontological Concepts in Artificial Cognitive Agents . . . . .	1027
<i>Radosław Piotr Katarzyniak</i>	

## **Intelligent and Secure Digital Content Management**

A Design of Hybrid Mobile Multimedia Game Content . . . . .	1035
<i>Il Seok Ko, Yun Ji Na</i>	

Development of Oval Based Vulnerability Management Tool (OVMT) on a Distributed Network Environment . . . . .	1042
<i>Geuk Lee, Youngsup Kim, Sang Jo Youk</i>	

A Study on a Design of Efficient Electronic Commerce System . . . . . 1050  
*Yun Ji Na, Il Seok Ko, Jong Min Kwak*

A Scheduling Middleware for Data Intensive Applications on a Grid . . . . 1058  
*Moo-hun Lee, Jang-uk In, Eui-in Choi*

Development of a Monitoring Module for ITS (Intrusion  
Tolerant System) . . . . . 1068  
*Wankyung Kim, Wooyoung Soh, Hwangrae Kim, Jinsub Park*

Design of DRM-LMS Model in M-Learning Environment . . . . . 1075  
*Mingyun Kang, Seoksoo Kim, Gil-Cheol Park, Geuk Lee,  
Minwook Kil*

**Business Knowledge Modelling and Maintenance**

SBEAVER: A Tool for Modeling Business Vocabularies  
and Business Rules . . . . . 1083  
*Maurizio De Tommasi, Angelo Corallo*

A Semantic Recommender Engine Enabling an eTourism Scenario . . . . . 1092  
*Angelo Corallo, Gianluca Lorenzo, Gianluca Solazzo*

A Legal Architecture for Digital Ecosystems . . . . . 1102  
*Virginia Cisternino, Angelo Corallo, Gianluca Solazzo*

Evolutionary ANNs for Improving Accuracy and Efficiency  
in Document Classification Methods . . . . . 1111  
*Antonia Azzini, Paolo Ceravolo*

A Methodology for Determining the Creditability of Recommending  
Agents . . . . . 1119  
*Omar Khadeer Hussain, Elizabeth Chang,  
Farookh Khadeer Hussain, Tharam S. Dillon*

**Innovations in Intelligent Systems and Their  
Applications**

How to Solve a Multicriterion Problem for Which Pareto Dominance  
Relationship Cannot Be Applied? A Case Study from Medicine . . . . . 1128  
*Crina Grosan, Ajith Abraham, Stefan Tigan, Tae-Gyu Chang*

Interpretation of Group Measurements of Validation Data Using  
Fuzzy Techniques in an Object-Oriented Approach . . . . . 1136  
*Eduardo Mosqueira-Rey, Vicente Moret-Bonillo*

Accuracy of Neural Network Classifiers as a Property of the Size of the Data Set .....	1143
<i>Patricia S. Crowther, Robert J. Cox</i>	
Investigating Security in Multi-tree Based Technique in RFID Systems .....	1150
<i>Xu Huang, Dharmendra Sharma</i>	
An Improved ALOHA Algorithm for RFID Tag Identification .....	1157
<i>Xu Huang</i>	
A Dynamic Threshold Technique for XML Data Transmission on Networks .....	1163
<i>Xu Huang, Alexander Ridgewell, Dharmendra Sharma</i>	
Distributed Face Recognition: A Multiagent Approach .....	1168
<i>Girija Chetty, Dharmendra Sharma</i>	

## **Intelligent Agents and Their Applications**

Using Windows Printer Drivers for Solaris Applications – An Application of Multiagent System .....	1176
<i>Wanli Ma, Dat Tran, Dharmendra Sharma, Abhishek Mathur</i>	
A Novel Approach to Programming: Agent Based Software Engineering .....	1184
<i>Dharmendra Sharma, Wanli Ma, Dat Tran, Mary Anderson</i>	
Personalisation of Web Search: An Agent Based Approach .....	1192
<i>Gopinathan L. Ligon, M. Bala Balachandran, Dharmendra Sharma</i>	
Autonomy and Intelligence – Opportunistic Service Delivery in Mobile Computing .....	1201
<i>Jiangyan Chen, Michael J. O’Grady, Gregory M.P. O’Hare</i>	

## **Signal Processing Techniques for Knowledge Extraction and Information Fusion**

Acoustic Parameter Extraction from Occupied Rooms Utilizing Blind Source Separation .....	1208
<i>Yonggang Zhang, Jonathon A. Chambers, Paul Kendrick, Trevor J. Cox, Francis F. Li</i>	

An Online Method for Detecting Nonlinearity Within a Signal . . . . .	1216
<i>Beth Jelfs, Phebe Vayanos, Mo Chen, Su Lee Goh,</i>	
<i>Christos Boukis, Temujin Gautama, Tomasz Rutkowski,</i>	
<i>Tony Kuh, Danilo Mandic</i>	
Using Hierarchical Filters to Detect Sparseness in Unknown Channels . . . . .	1224
<i>Christos Boukis, Lazaros C. Polymenakos</i>	
Auditory Feedback for Brain Computer Interface Management – An EEG Data Sonification Approach . . . . .	1232
<i>Tomasz M. Rutkowski, Francois Vialatte, Andrzej Cichocki,</i>	
<i>Danilo P. Mandic, Allan Kardec Barros</i>	
Analysis of the Quasi-Brain-Death EEG Data Based on a Robust ICA Approach . . . . .	1240
<i>Jianting Cao</i>	
A Flexible Method for Envelope Estimation in Empirical Mode Decomposition. . . . .	1248
<i>Yoshikazu Washizawa, Toshihisa Tanaka, Danilo P. Mandic,</i>	
<i>Andrzej Cichocki</i>	
The Performance of LVQ Based Automatic Relevance Determination Applied to Spontaneous Biosignals . . . . .	1256
<i>Martin Golz, David Sommer</i>	
Alertness Assessment Using Data Fusion and Discrimination Ability of LVQ-Networks . . . . .	1264
<i>Udo Trutschel, David Sommer, Acacia Aguirre, Todd Dawson,</i>	
<i>Bill Sirois</i>	
Signal Reconstruction by Projection Filter with Preservation of Preferential Components. . . . .	1272
<i>Akira Hirabayashi, Takeshi Naito</i>	
Sensor Network Localization Using Least Squares Kernel Regression . . . .	1280
<i>Anthony Kuh, Chaopin Zhu, Danilo Mandic</i>	
<b>Author Index . . . . .</b>	<b>1289</b>

# Closed-Ended Questionnaire Data Analysis

Leuo-Hong Wang<sup>1</sup>, Chao-Fu Hong<sup>1</sup>, and Chia-Ling Hsu<sup>2</sup>

<sup>1</sup> Evolutionary Computation Laboratory,  
Department of Information Management, Aletheia University, Taiwan  
{wanglh, cfhong}@email.au.edu.tw

<sup>2</sup> Centre for Teacher Education, Tamkang University, Taiwan  
clhsu@mail.tku.edu.tw

**Abstract.** A KeyGraph-like algorithm, which incorporates the concept of structural importance with association rules mining, for analyzing closed-ended questionnaire data is presented in this paper. The proposed algorithm transforms the questionnaire data into a directed graph, and then applies association rules mining and clustering procedures, whose parameters are determined by gradient sensitivity analysis, as well as correlation analysis in turn to the graph. As a result, both statistically significant and other cryptic events are successfully unveiled. A questionnaire survey data from an instructional design application has been analyzed by the proposed algorithm. Comparing to the results of statistical methods, which elicited almost no information, the proposed algorithm successfully identified three cryptic events and provided five different strategies for designing instructional activities. The preliminary experimental results indicated that the algorithm works out for analyzing closed-ended questionnaire survey data.

## 1 Introduction

The questionnaire survey with closed-ended questions is one of the most common used tools for user information elicitation. The data collected are usually analyzed by various statistical techniques [1]. However, if no statistically significant events exist in the data, little information can be extracted by these statistical techniques.

Well-designed questionnaires, especially close-ended, rating scaled questionnaires, always try to capture the intended information as much as possible by carefully wording each question. Sentences in such questions therefore contain a certain keywords to appropriately represent the subjects of questions. Respondents convey how intensive they feel about these keywords at the same time when they answer questions to express their opinions. Hence, dealing with the collected data via the keywords viewpoint gives us an alternative way for eliciting information from survey data. If each keyword in the questions is, in other words, treated as an item and the relationships between these keywords can be appropriately defined, then data mining algorithms such as the a-priori algorithm for association rule mining [2] can be applied. As a result, the dependency

of keywords can be calculated. That implies the relationship between variables, which are observed by the questionnaire, is capable of being found.

Although dealing with data via the a-priori algorithm seems to work out for survey data just as good as other statistical techniques, only the most significant events would be identified. Eliciting little information is still a problem for such an analysis scheme. Therefore, we have designed a new algorithm integrating the concept of finding rare but structurally important events, which have been discussed by a handful of researchers [3][4][5][6] in *chance discovery* [7] discipline recently, with the a-priori algorithm for analyzing closed-ended questionnaires in this paper. Actually, the algorithm is similar to KeyGraph algorithm [4], except the measures of *support* and *confidence* used by association rules mining are introduced. In addition, a sensitivity analysis procedure which is applied to find the threshold values of these measures is also integrated with the process of identifying chances. With these modifications, the new algorithm can automatically identify chances and is suitable for questionnaire data analysis.

The rest of this paper is organized as follows. Section 2 presents the new algorithm to identify significant events and discover chances from closed-ended questionnaire survey data. Section 3 starts with introducing the instructional design application and then defines its structurally important events. The experimental results of applying our algorithm to the application are also shown in this section. We would summarize the paper in the final section.

## 2 The A-Priori Based Chance Discovery Algorithm

To deal with questionnaire data, we use the *support* and *confidence* measures which usually used by data mining applications to identify clusters in our algorithm. To investigate the structural importance of the survey data, a weighted, directed graph  $G = (V, A)$  where  $V$  is a set of nodes and  $A$  is a set of arcs between nodes must be constructed first. As mentioned in Section 1, a well-designed questionnaire is always carefully wording each question, so the set of nodes  $V$  for the graph  $G$  intuitively consists of every question of the questionnaire. However, question numbers but not keywords are used for representing the nodes in the following for simplicity. Meanwhile, each arc in the set  $A$  represents an association rule between two questions. Moreover, each node and each arc are assigned a weight. The weight of a node is the summation score given by all respondents. The weight of arc is defined as the co-occurrence frequency, which would be described later, of two adjacent questions. Once the graph is constructed, our algorithm can then be applied to identify the rare but structurally important nodes.

### 2.1 Preprocessing- Constructing a Weighted, Directed Graph

Assume a questionnaire consisting of  $N$  closed-ended, 1-to-5 scale rating questions is given to  $M$  respondents for evaluation. Then an  $M$  by  $N$  score matrix  $S = s_{mn}$ , where  $s_{mn}$  is the score of the  $n^{th}$  question given by the  $m^{th}$  respondent, can be filled. Each row of  $S$  is rewritten to a string record consisting of



repeatedly occurring question numbers. For example, if a row is (3, 5, 2), which means there are three questions and they got 3, 5 and 2 points respectively, then the derived string record is (1,1,1,2,2,2,2,2,3,3), which has three 1s, five 2s and two 3s. M string records are generated after rewriting. Once string records are ready, two more tasks must be fulfilled.

1. Count how many times each question number occurs in all M records. The occurring frequency vector  $F = (f(1), f(2) \dots, f(N))$ , where  $f(i)$  is the occurring frequency of the  $i^{th}$  question, is thus generated. This vector is used to calculate the co-occurrence matrix discussed later and the support of each question.
2. Generate an N by N co-occurrence matrix  $A = a_{ij}$ . In this matrix,  $a_{ij}$  is the co-occurring times of the  $i^{th}$  and  $j^{th}$  questions in all M records and is calculated by the following equation:

$$a_{ij} = \sum_{r=1}^M \min(f_r(i), f_r(j)) \quad (1)$$

where  $f_r(i)$  is the times of the  $i^{th}$  question occurring in the  $r^{th}$  record.

Consequently, the occurring frequency vector  $F$  and co-occurrence matrix  $A$  construct the weighted, directed graph for further processing.

## 2.2 Identifying the Rare but Structurally Important Nodes

The identification process consists of three steps:

1. **Extracting the high-support nodes.** The support measure of node is defined as:

$$support(i) = \frac{f(i)}{\text{the highest frequent that i can get}} \quad (2)$$

where  $f(i)$  is the occurring frequency of the  $i^{th}$  question. The highest frequency for a 1-to-5 scale rating question, evaluating by M respondents, is  $5 \cdot M$ . Hence, each question's support can be calculated by using Equation 2. Then the support measure of each question is sorted by descending order for applying gradient sensitivity analysis. The gradient sensitivity analysis facilitates the decision of threshold value. The nodes with high enough support, which is decided by sensitivity analysis, are therefore extracted.

2. **Connecting high-support nodes if their confidences to each other are high enough.** Based on Equations 1 and 2, the confidence measure of an arc from  $i^{th}$  question to  $j^{th}$  question is defined as follows:

$$confidence(i \rightarrow j) = \frac{a_{ij}}{support(i)} \quad (3)$$

The confidence values are computed for all pairs of high-support nodes. Then the node-pairs are sorted according to their confidence values. After applying

gradient sensitivity analysis to decide a reasonable threshold value, two nodes are connected if their both confidence values are higher than the threshold value. As a result, a refined graph consisting of high-support events and containing the information of high correlated events is generated.

3. **Calculating structural importance and emerging chances.** Each connected component in the refined graph is treated as a cluster. Then the correlation value of each node  $i$  not belonging to the high-support nodes and each cluster  $k$ ,  $correlation(i, cluster_k)$ , is calculated according to the following equation:

$$correlation(i, cluster_k) = \frac{\sum_{\forall j \in cluster_k} confidence(i \rightarrow j)}{|cluster_k|} \quad (4)$$

where  $|cluster_k|$  is the number of nodes of  $cluster_k$ , and  $confidence(i \rightarrow j)$  is just the confidence value defined in Equation 3. According to this equation, the correlation value of node  $i$  to the whole refined graph,  $correlation(i)$ , is defined as:

$$correlation(i) = \sum_{\forall cluster_k} correlation(i, cluster_k) \quad (5)$$

After all correlation values of non-high-support nodes are calculated, normalized and sorted, another gradient sensitivity analysis is applied to determine the threshold value of chances. Only nodes whose correlation values are larger than the threshold value are recognized as chances.

### 3 Experiments and Discussion

One questionnaire which was designed for the application of instructional design was used to examine the capability of our chance discovery algorithm. This questionnaire consists of 34 1-to-5 rating scale questions, which can be classified into four categories: attention, relevance, confidence and satisfaction. Actually, the questionnaire was designed according to the well-known ARCS (Attention, Relevance, Confidence and Satisfaction) model [8].

The survey data were collected from 56 respondents who enrolled in a general education course of TamKang University in 2005. Since the lecturer of the course made use of the power of the *open cyber classroom* (<http://mslin.ee.ntut.edu.tw/>) to assist her in accomplishing various instructional activities, a questionnaire survey examining the effectiveness of the course was therefore investigated when the course had been finished.

#### 3.1 Statistical Results

Originally, statistical techniques were applied to survey data in order to understand if any categories or questions were statistically superior to the others. Moreover, we also followed the data analysis process suggested in [9] to examine the questions whose average scores were the lowest ones. The next year's

instructional activities could be thus improved in the light of these statistical results. The average score and variance of each category is listed in Table 1. The questions with lowest three average scores are listed in Table 2.

**Table 1.** The statistical results of the survey

Categories	# of questions	summation	average	variance
<b>Attention</b>	8	31.018	3.877	0.035
<b>Confidence</b>	8	29.214	3.652	0.346
<b>Relevance</b>	9	32.643	3.627	0.448
<b>Satisfaction</b>	9	33.393	3.710	0.318

**Table 2.** The questions with lowest average scores. Other questions all got scores greater than 2.93. The question which got the highest average score 4.25 is the question: "accomplishable if working hard".

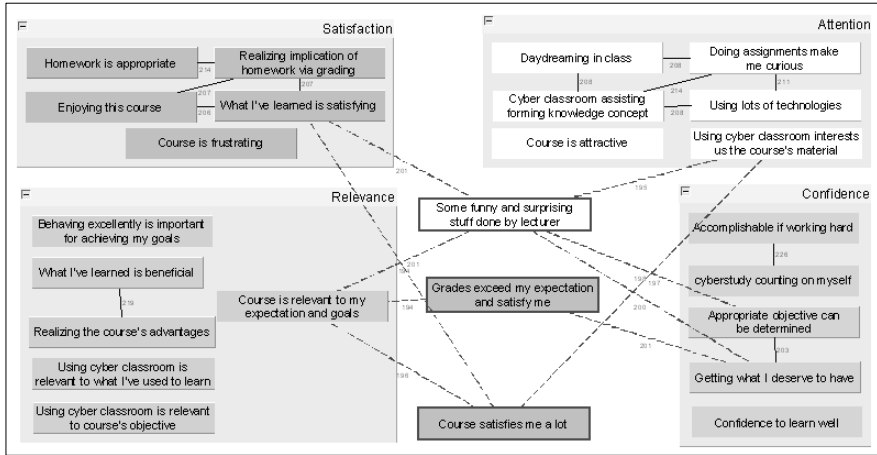
Categories	Questions	average	variance
<b>Attention</b>	Getting advantages from the course	1.95	0.71
<b>Satisfaction</b>	Getting passed only if I am working hard	2.29	0.83
<b>Confidence</b>	Grading is unpredictable	2.61	0.97

In Table 2, these three questions are identified as the weakest areas. Motivation enhancement strategies must be come up with to address these areas for the next year. However, the single-factor ANOVA indicates there are no differences between these four categories, and even between most of questions. In other words, little information has been obtained from these kinds of analysis. Thus, we developed the chance discovery algorithm described in Section 2 and applied it to analyze the same survey data.

### 3.2 Results of Applying the New Chance Discovery Algorithm

After applying the procedure of our algorithm, the refined graph was constructed and chances consequently were emerging, just as shown in Figure 1.

In Figure 1, there are 22 (out of 34) questions are emerging if the minimal support is 75% (it equals to 3.75 points). That means lots of questions got great scores from the respondents. The statistical results listed in Table 1 approve the fact as well. In addition, when the minimal confidence is 92.2% decided by gradient sensitivity analysis, the number of clusters is 13 (8 of them includes only one node). Two categories, the satisfaction and attention, are denser than the other two categories. That means instructional strategies concerning both the satisfaction and attention categories were impressive for comparatively more students in this course. Something interesting is statistical methods are impossible to sustain such a fact since they focus on explaining the behavior of the whole population but not individuals. As a result, we claim the strategies concerning the satisfaction and attention in this class are more successful. The key



**Fig. 1.** The emerging chances. The numbers over the links are calculated by Equation 1, which are the co-occurrence frequencies between two adjacent nodes.

points to improve for the next year thus focus on the other two categories. The suggestions include:

1. Design some activities that can effectively raise the confidence of students.
2. Create teaching material concerning the functionalities of the *open cyber classroom* in order to let students understand how cyber study facilitates learning.

The clusters consisting of significant association rules provide the inner-category explanation for the survey data. However, three inter-category chances (eight bridges totally) are also identified as shown in Figure 1. The support measures of these chances are not significant, but their correlation measures to clusters outperform to other nodes. In other words, relatively few respondents thought they had learned effectively from the viewpoint represented by these questions. More importantly, these three questions correlate closely with clusters in each category. Improving the effectiveness of categories from these chances should be acceptable even though these chances are obscure and risky. Hence, we suggest the following strategies for next year.

1. The lecturer should be more humorous and surprising. Because some students appreciate these funny or surprising stuff. As a result, both their learning motivation and satisfaction will increase.
2. Design some course-relating activities for additional grading bonuses. Because lots of students thought their grades were so unsatisfactory as to diminish their learning pleasure.
3. Design some cyber studying activities in order to attract and satisfy the major part of the students. The chance located at the lowest part of Figure 1 ("Course satisfies me a lot") indicates only a few students enjoyed

cyber studying as well as satisfying the material of this course. We hope their experience can be spread out to the other students if we design some interesting cyber studying activities.

## 4 Summary

A new chance discovery algorithm has been presented and applied to analyze questionnaire survey data in this paper. We follow the idea of *chances* to design the algorithm. Once the survey data transform into a form suitable for our algorithm, both statistically and structurally important events will be identified. According to an analysis on a real world survey data which were collected from a course in TamKang University, our algorithm successfully unveiled 5 clues for the lecturer of that course. In summary, the experimental results presented in this paper are preliminarily verify that our algorithm can provide an effective way to analysis closed-ended questionnaire data. More experiments will be examined in the near future.

## References

1. Sakai, T.: Questionnaire design, Marketing Research and Statistical Analysis. DrS-mart Press, Taipei (2004) (in Chinese).
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB Conference. (1994) 487–499
3. Nara, Y., Ohsawa, Y.: Application to Questionnaire Analysis. In: Chance Discovery. Springer Verlag (2003) 351–366
4. Ohsawa, Y., Benson, N., Yachida, M.: Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In: Proceedings of the Advances in Digital Libraries Conference. (1998)
5. Tamura, H., Washida, Y., Ohsawa, Y.: Emerging scenarios by using ddm: A case study for japanese comic marketing. In: Proceeding of KES 2004,. (2004) 847–854
6. Mizuno, M.: Chance Discovery for Consumers. In: Chance Discovery. Springer Verlag (2003) 367–382
7. Ohsawa, Y.: Modelling the Process of Chance Discovery. In: Chance Discovery. Springer Verlag (2003) 1–15
8. Keller, J., Kopp, T.: Application of the ARCS Model to Motivational Design. In: Instructional Theories in Action: Lessons Illustrating Selected Theories. Lawrence Erlbaum (1997) 289–320
9. Suzuki, K., Nishibuchi, A., Yamamoto, M., Keller, J.: Development and evaluation of website to check instructional design based on the arcs motivation model. Information and Systems in Education **2** (2004) 63–69

# Strategies Emerging from Game Play: Experience from an Educational Game for Academic Career Design

Yoshikiyo Kato<sup>1</sup> and Hiroko Shoji<sup>2</sup>

<sup>1</sup> Knowledge Creating Communication Research Center,  
National Institute of Information and Communications Technology  
4-2-1 Nukui-kitamachi, Koganei, Tokyo 184-8795, Japan  
ykato@nict.go.jp

<sup>2</sup> Department of Industrial and Systems Engineering, Chuo University  
1-13-27 Kasuga, Bunkyo, Tokyo 112-8551, Japan  
hiroko@indsys.chuo-u.ac.jp

**Abstract.** In this paper, we explore the possibility of chance discovery through game play. We describe an educational game for academic career design, and present strategic knowledge that players learned through game play. We discuss the role of game play in acquiring such knowledge in terms of chance discovery, and the validity of the acquired strategies in real-life career design.

## 1 Introduction

In this paper, we explore the possibility of chance discovery in game play. According to Abe [1], Ohsawa defines a chance as follows:

A chance is a new event/situation that can be conceived either as an opportunity or a risk.

In this sense, we consider the strategies that players acquire from playing educational career design game could be chances for them as they provoke rethinking of their own strategies for career design and making decisions in real-life career planning. Although the possibility of chance discovery through game play has not been explored much before, our experience tells us that there is a good *chance* that game play which enables player's chance discovery could be used to support her decision making process.

In the following section, we outline the chance discovery process we hypothesize. In Section 3, we describe an educational board game for academic career design, and strategies emerged from our experimental game play. In Section 4, we discuss the issues and future direction of applying game play for chance discovery. Finally, we conclude in Section 5.

## 2 Chance Discovery in Game Play

What role does 'game play' play in chance discovery process? To answer this question, we first review the definition of chance discovery from the abductive

point of view. Then, we present a hypothetical chance discovery process involving game play.

Abe defines chance from the abductive viewpoint as follows [1]:

Chance itself is a set of known facts, but it is unknown how to use them to explain an observation.

He characterizes rules that are abductively generated to explain *inexplicable* observations based on available facts as chance. In this formulation, we regard the game play participant's epistemic state regarding the reality as *facts*, and strategic knowledge applicable in reality as *observations*. In this framework, strategic knowledge can either emerge as a direct result of game play, or it can be a recognition of strategic knowledge in real life (usually in the form of moral or anecdotal events) by the participants. In case of recognition of the strategic knowledge, we assume that even if the participant has some kind of strategic knowledge before game play, she does not fully understand the implication of such knowledge, because of the lack of knowledge about the relationship between elements in reality pertaining to the strategy in concern. What is hampering the participants' understanding of the relationship between elements in reality is the complexity of the reality. They cannot readily see the relationship between facts. As a result, the construction of effective strategies in real-life is hindered.

Game play allows its participants to bridge the gap between *facts* in reality and *observations*, or strategic knowledge. According to abductive definition of chance discovery, we can say that such bridging process is indeed a kind of chance discovery. The process of chance discovery in game play is shown in Fig. 1.

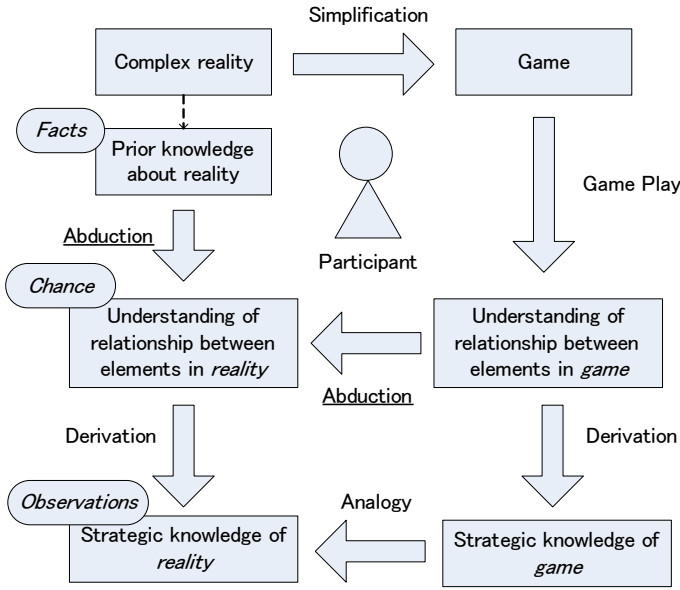
### 3 Experience with Educational Board Game for Academic Career Design

In this section, we describe an experience with an educational board game for academic career design, *Happy Academic Life 2006* (Fig. 2).

#### 3.1 Game Description

The object of the game is to be the first to achieve one of the seven goals, which represent stereotypical role models of academic career (Table 1). Each role model has different conditions to be satisfied in terms of various parameters (Table 2).

One of the key features of the game is time management. In each turn, a player has 600 hours to spend, which makes one turn corresponding to 3 months period of work. Drawing an analogy with financial accounting, 600 hours given for each turn are a player's revenue. The player has to spend fixed amount of time related to minimal duty of her position. Additionally, she can spend time on duties of her workplace (such as teaching classes, or being a member of departmental or university committee) or duties of academic societies (such as being a member of the program committee for a conference, or reviewing papers



**Fig. 1.** Chance discovery process in game play. Game play allows the participant to be aware of the relationship between elements in the complex reality, leading to acquisition of strategic knowledge in reality.

for a conference or a journal). Players also have to spend time on their students and postdocs. After deducting the time to be spent on the above-mentioned activities, remaining hours can be spent on doing research. Players always have a subject of study to work with. There is a specific amount of hours to be spent for each subject before one can submit a paper on the subject. After spending enough time on the subject, players draw an *accept/reject card* which tells the result of the review process: i.e. accept, conditional accept, inquiry, and reject. If accepted, the player places a chip representing the accepted paper on the research map (Fig. 2).

Each player starts her academic career as *Joshu* (an entry position in Japanese academic system, usually translated as “research associate”). In each turn, a player rolls a die, and move her piece the number of spaces showing on the die along the track. Each space on the track is marked with a label corresponding to five types of event cards<sup>1</sup> and *Wild*. Players stopped on a space with a label of one of the five types of event cards have to draw a card of the corresponding type. Players who stopped on a *Wild* can draw a card of any of the five types. Event cards provide users with opportunities and penalties. For example, chance cards include cards that announce an open position for a full professor. Players who qualify for the condition of the position can apply for it, and if successful be promoted to a professor.

<sup>1</sup> The five types of event cards are Chance, Private, *Gakkai* (“academic society”), *Gakunai* (“within university”), and *Shikin* (“research grant”).





**Table 1.** The seven goals and their conditions in Happy Academic Life 2006. To achieve the goal, a player have to be a full professor in addition to these conditions.

Goals	Conditions
Educator	To have 6 or more research-oriented graduates.
Research Director	To hire 3 postdocs or more at the same time.
Otium Cum Dignitate	To publish 12 papers or more without hiring a postdoc.
Politician	To score 10,000 workplace points.
Pundit	To publish at least one paper in all research field (A-F) and to score 10,000 connection points.
Prolific Author	To publish 10 papers or more, at or above level 3.
Outstanding Researcher	To publish 3 or more level 5 papers.

**Table 2.** Parameters of the career model in Happy Academic Life 2006

Parameters	Description
Funding Points	Represent a player's competence in earning research grants.
Workplace Points	Represent the status of a player in her workplace.
Connection Points	Represent the scale of a player's network of connections in the academic society.
Publication Quantity	The number of papers published by a player
Publication Quality	The level of papers published by a player, which is represented on the research map (Fig. 2) starting from level 1 up to 5.
Research-Oriented Graduates	The number of graduates who had motivation for doing research, whom a player has advised through out her career.
Postdocs	A player can hire a postdoc for every 5,000 funding points.

participants are asked to draw conclusions about the real world based on experience from game play. We have to consider four elements in generalization stage:

**Conclusions** Conclusions from the game experience.

**Game Data** The events happened during the game which support the conclusions.

**Judgment** Whether the conclusions are applicable in the real world.

**Life Data** Specific happenings in real life that either support or do not support the judgment.

In the following sections, we present two strategies that were identified in game plays we have conducted.

### Strategy 1: Never turn down an offer for a duty

*Conclusions* “Never turn down an offer for a duty.”

*Game Data* In one game, one of the participants chose *otium cum dignitate*<sup>2</sup> as his goal. He undertook all duties drawn from *Gakkai* or *Gakunai* cards. By coping

<sup>2</sup> A Latin word meaning “leisure with dignity”.

with all those duties, he gathered workplace points and connections points just enough to be promoted as a full professor. He was the first to achieve his goal in the game.

*Judgment* Not necessarily applicable in real life.

*Life Data* After the game, the participant said, “I was told by my teacher that I should never turn down an offer for a duty, and I have done so in my career.” Although the strategy worked for this particular case in game, we can easily imagine that blindly accepting any offers for duty will invariably lead to a very busy life, which is far from the spirits of *otium cum dignitate*. Other wise steps would be needed to achieve such goal in real life.

Although we have to take the words with caution, it still teaches us a lesson: avoiding duties as much as possible to concentrate merely on research is not necessarily a good strategy to achieve *otium cum dignitate* in real life.

## **Strategy 2: Enjoy the *Joshu* position as long as possible**

*Conclusions* “Enjoy the *Joshu* position as long as possible.”

*Game Data* In another game, one of the participants chose *otium cum dignitate* as his goal. He collected all the papers required for the goal condition (12 papers) while he was *Joshu*. After collecting the required papers, he was quick in being promoted to a full professor and was the first to achieve his goal.

*Judgment* Used to be applicable in real life, but not anymore.

*Life Data* *Joshu* used to be a tenured position in the past. However, in recent years many institutions have converted it into a fixed-term (usually 3 to 10 years) position, and one cannot stay as *Joshu* for a very long time as before.

## **4 Discussion**

In case of strategy 1, the participant had the strategic knowledge as a moral from his teacher beforehand. However, it is after the game play that he recognized his teacher’s words and related them to the strategy for academic career design. The game play made him aware of the importance of duties of the workplace and academic societies in achieving one’s goal in his academic career. We could say such awareness is the chance that game play provided.

In case of strategy 2, it is not clear whether the participant knew the strategy beforehand. However, he demonstrated the effectiveness of the strategy in the game, and we see that the strategy used to be effective in real life as well at least in the past. As a result of the game play, the participant will reinforce the belief that such strategy is effective, although it may not be effective anymore in real life. The discrepancies with the reality can be attributed directly to the difference between the model in the game and the reality. In the game, players can stay as *Joshu* as long as they want, while it is becoming not the case in

reality anymore. This case tells us that one has to be careful in choosing the model for the game for it to be effective in chance discovery.

Bedemeier et al. describes the effectiveness of frame games in providing “experiential learning about organizational politics, leadership, and decision making” [3]. Frame games are skeletal games that provide generic frameworks of games which can be adapted for a wide range of purposes. The authors discuss that more learning occurs when the participants involve in *redesigning* the frame game, where they construct another real-life game based on the original game. For example, nurses can construct a game modeling the medical world after playing the Academic Game (a type of frame game described in [3]), which models the world of academia. Although Happy Academic Life 2006 was not specifically developed as an instance of a frame game, one would easily see that it can be adapted to games in the fields other than academia, by redesigning the tracks, cards, and other elements of the game. Adding redesigning step to game play will shed light on the difference between game and reality, and allow participants to be aware of the limitation of the strategies in real life, which were effective in game.

In recent years, Gaming-Simulation has drawn attention as a pedagogical tool. It provides an environment for *experiential learning* [4]. Experiential learning is “the sort of learning undertaken by students who are given a chance to acquire and apply knowledge, skills and feelings in an immediate and relevant setting.” [5] Kolb created a model of experiential learning cycle consisting of four elements: i.e. 1) concrete experience, 2) observation and reflection, 3) forming abstract concepts, and 4) testing in new situations. In relation to our chance discovery process in game play, we see that game plays provide opportunities to complete the learning cycle where it is difficult in real life. People have little problem in going through step 1 and 2 of experiential learning cycle in real life. However, step 3 and 4 are not easy because of the complexity of real life, and testing is not possible in many cases. Through game play, participants discover chances to overcome the obstacles and can complete the experiential learning cycle.

Although game play is a powerful tool for education, it has some drawbacks because its learning process is basically a type of discovery learning [6]. In discovery learning, students are given a goal, but not a direct way to reach the goal, so that students have to find a way on their own to reach the goal. The strength of discovery learning is that it motivates students, they learn how to learn, and they learn more effectively than conventional methods. However, it has weakness such as missing core knowledge, or overly narrow studies.

Intelligent tutoring system can provide a facility to complement these weakness of discovery learning, and make the learning process effective [7]. Based on domain model, tutoring model, and student model, intelligent tutoring systems can monitor and guide the learning process. By observing events and behavior of the user, it can detect missing concepts to be learned, or misconceptions that the user has acquired, and take remedial action against them. An interesting direction of research is to see how such intelligent tutoring system can be employed to facilitate the chance discovery process in game play.

## 5 Conclusions

In this paper, we explored the possibility of chance discovery in game play. We described an educational game for academic career design, and the strategies emerged through game plays of it. We outlined the process of chance discovery which involves game play, and saw some evidence from experience that supports parts of our hypothesis. As we showed the possibility of chance discovery through game play, the next step would be to study the validity of the hypothetical chance discovery model of game play.

## Acknowledgments

Happy Academic Life 2006 was developed by Academic Life Club under the commemorative project of the 20th anniversary of the establishment of the Japanese Society for Artificial Intelligence. We would like to thank both the members of Academic Life Club and the Japanese Society for Artificial Intelligence for the support.

## References

1. Abe, A.: Chance discovery. <http://ultimavi.arc.net.my/ave/cd-j.html> (2006)
2. Stadskev, R.: Handbook of Simulation Gaming in Social Education. Part I: Textbook. Institute of Higher Education Research and Services, The University of Alabama, Tuscaloosa, AL (1974)
3. Bredemeier, M.E., Rotter, N.G., Stadskev, R.: “The academic game” as a frame game. *Journal of Experiential Learning and Simulation* **3** (1981) 73–83
4. Kolb, D.A.: *Experiential Learning: Experience as the Source of Learning and Development*. Financial Times Prentice Hall (1983)
5. Smith, M.K.: David A. Kolb on experiential learning. The encyclopedia of informal education, <http://www.infed.org/b-explrn.htm> (2001)
6. Baldwin, D.: Discovery learning in computer science. In: SIGCSE '96. (1996) 222–226
7. Siemer, J., Angelides, M.C.: Evaluating intelligent tutoring with gaming-simulations. In Alexopoulos, C., Kang, K., Lilegdon, W.R., Goldsman, D., eds.: *Proceedings of the 1995 Winter Simulation Conference*. (1995)

# Concept Sharing Through Interaction: The Effect of Information Visualization for Career Design

Hiroko Shoji, Kaori Komiya, and Toshikazu Kato

Faculty of Science and Engineering, Chuo University,  
1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan  
hiroko@indsys.chuo-u.ac.jp

**Abstract.** In this study, we have developed an interactive system called Mochi that effectively prompts users to visualize their concepts on 2D maps. We analyzed the process of concept sharing, in which two users initially had vague concept on career design and then gradually clarified it in the course of interaction. This paper discusses the effectiveness of the system in facilitating adjustment of communication gap between users in the domain of career designing.

## 1 Introduction

A number of tasks in our real society require concept sharing among plural persons involved, including concept designing of a new shop, advertisement creation, and upstream process of software development. Existing concept sharing support tools that have been proposed for these tasks include UML in the area of software engineering[1], the image scale for representing sensibility image[2], and groupware for process management[3][4], however, none of them have sufficient features for concept sharing. The effective support for concept sharing requires a mechanism to help the users involved build consensus and/or make a decision by enabling them to (i) clearly present each one's concept, (ii) understand each other's concept, and then (iii) share everyone's concept.

The authors are developing a system called Mochi (Mochi Object for Collaboration using Hyper Images) to support the collaboration that necessitates concept sharing. Mochi is a two-dimensional map for plural workers to represent and share their concepts. On this map, the user can represent their own concept using keywords, images, figures, and others. This paper applies Mochi to concept sharing in career designing as an example, and discusses its potential and challenges.

## 2 Mochi: A Concept Representation Support Tool

The authors are developing a system called Mochi (Mochi Object for Collaboration using Hyper Images) to support the collaboration that necessitates concept sharing. Mochi is a two-dimensional map for plural workers to represent and share their concepts(See Figure1). In order for them to do so, it must be able to represent:

- (1) Concept fragments, or subconcepts at various levels from vague one to concrete one

- (2) Relations between individual subconcepts represented
- (3) Consensus reached among coworkers on concept sharing

Mochi allows an individual user to represent their own concept freely on the map using keywords, images, figures, and others. An individual item such as a keyword, image, and figure that is displayed on the map is an element to represent a subconcept mentioned in (1). The positional relationship of those items arranged on the map represents relations between subconcepts mentioned in (2). In addition, when Mochi has led persons involved to understanding of each other's concepts and consensus on concept sharing, the degree and status of the consensus can be tagged on the map, allowing for the management of the progress of consensus building. This feature enables the representation of the consensus reached mentioned in (3).

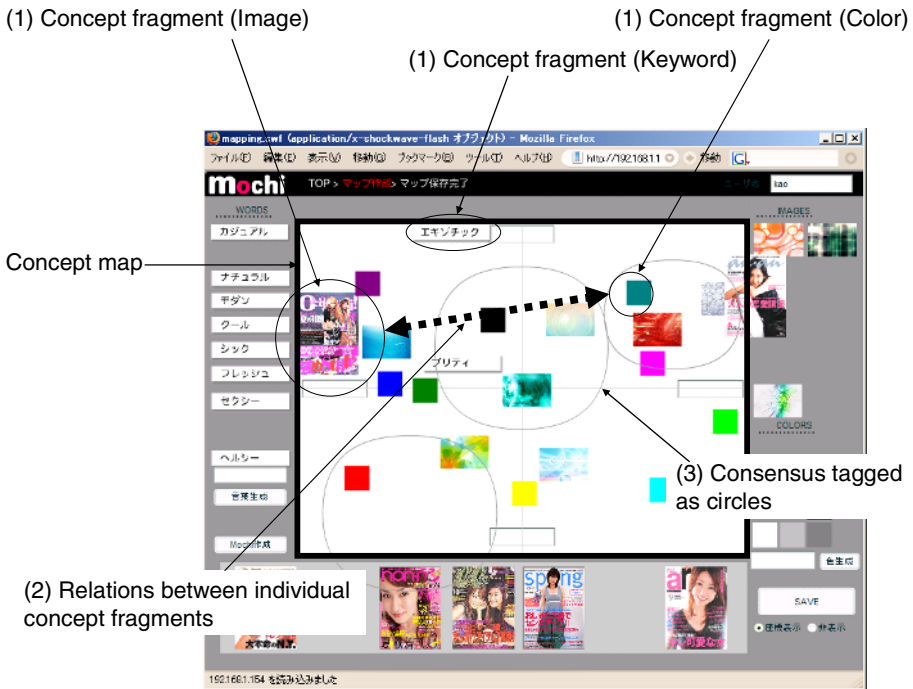


Fig. 1. Screenshot of Mochi System

### 3 Application of Mochi to Career Designing

Job hunters including university students must make clear their own concept of in what company and how they want to work[5]. The personnel staff of employing companies, on the other hand, must make clear their own concept of what company they are and what employees they need at the moment. Recently in Japan, the scarcity of employment for the younger generation is becoming a social problem. As one of its central causes, the perception gap between job offerers and job hunters has led to the

difficulty in finding employment. Therefore, if the both sides can clarify their own concepts, present them in visible form, and understand and share each other's concept, it can lead to support for job-hunting process. In addition, it can serve not only job-hunting but also career designing education after employment. Figure 2 shows an interaction scenario of concept sharing about career designing using Mochi.

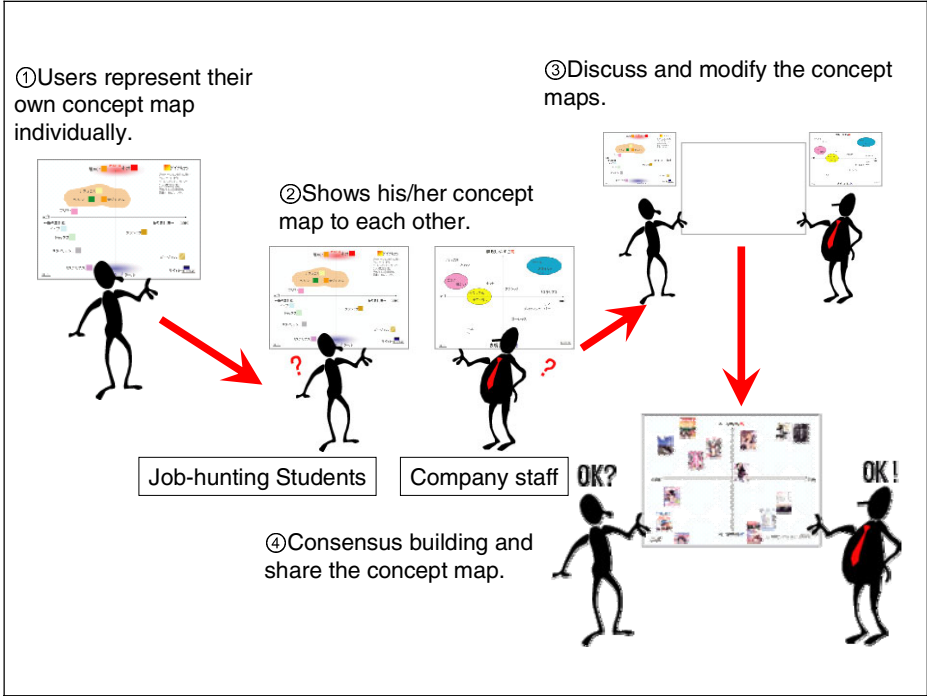


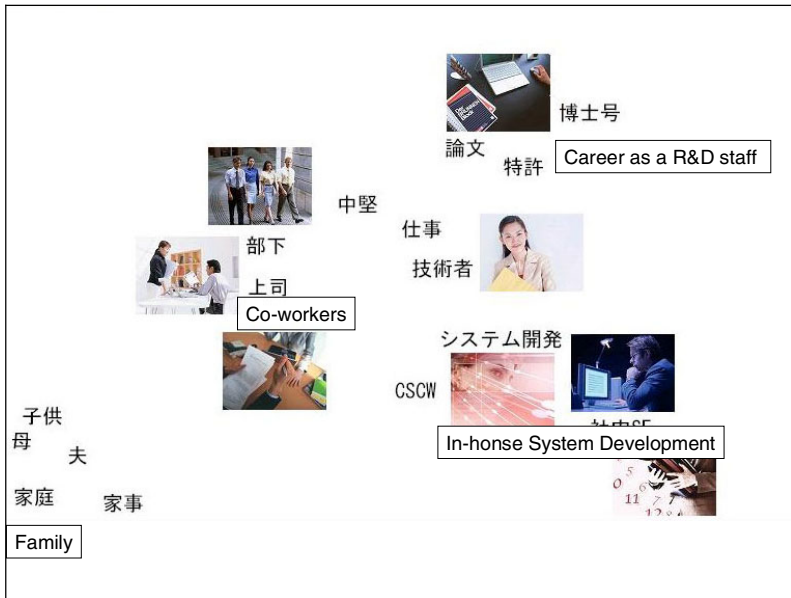
Fig. 2. Interaction scenario of concept sharing using Mochi

## 4 Experiments and Discussion

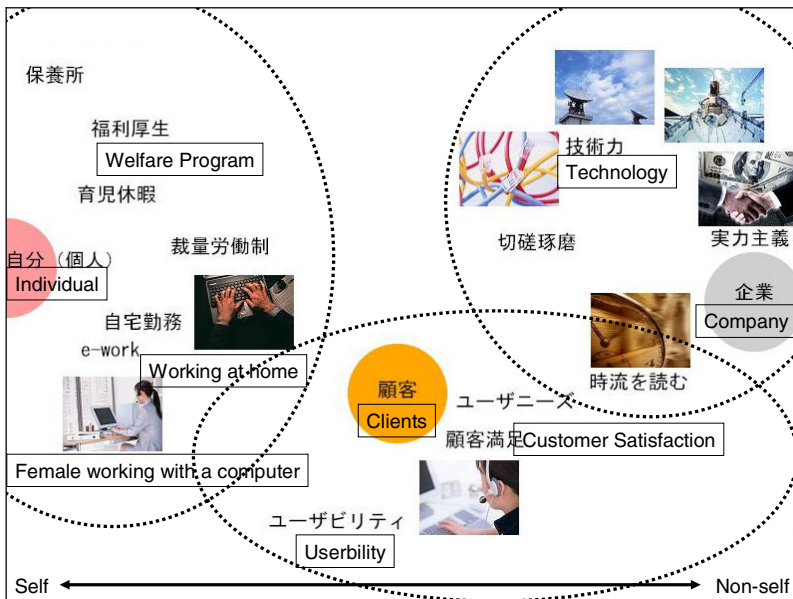
One experiment was conducted with four graduate students in job hunting as subjects, who were instructed to use Mochi to represent their "concept of the companies they want to work for". Additionally, another experiment was conducted with four employees of IT-related companies as subjects, who were instructed to use Mochi to represent their "image of young people whom their companies want to employ". Both experiments provided them with no set of keywords in advance, and allowed them to describe their concept in free words. The subjects were allowed to choose their favorite ones from more than 2000 images sorted out by category. Drawing capability for figures such as line and circle was also made available.

Figure 3 shows a concept map of "a desired company" created by a certain graduate student. A comment made by this student on this concept map is as follows:





(a) Initial map



(b) Final map

Fig. 3. A concept map of "a desired company" for a graduate student

"In any case, I regard highly that the company covers the various areas of activity and the staff in each of them are working actively. The reason for this lies in my image of a researcher I aim at. I want to be a researcher with multiple viewpoints, or in

other words, a researcher who does not stick to one area but rather covers several areas of expertise and also has a business point of view. I thought it is good for me to be such a one to work for a company that covers the various areas of activity and allows its employees to communicate with each other actively."

This student wishes to join a R&D division of an IT-related company, and is research-oriented, however, wants to build her career across several fields instead of delving into one thing. This student is a female and, thus, is very interested in the system of child-care leave and working at home as well. As for a vague concept that can not be clearly presented with keywords, on the other hand, she reinforced it with images. The images used for this purpose include ones easily understandable as concrete concepts, such as an image of a female working with a computer, and abstract ones. Arranged on the left side of the map is a concept mostly on herself, and on the right side is a concept of the company as a whole and its relationship with its customers, showing that she used the horizontal positional relationship on the map as a representation axis between the self and non-self (or surroundings).

All the four graduate students major in information systems engineering and wish to join a R&D division of an IT-related company. However, the comparison of the concept maps created by them has shown:

- The words and/or images used on the map vary considerably by person.
- The same image used on different maps has considerably different words arranged near it on each map.
- The same set of words used on different maps has a considerably different image arranged near it on each map.
- The implication of the positional relationship on the 2-D map varies considerably by person.

In presenting their concepts on the map, some students used a lot of words and few images and/or figures, others used only images. This shows that different persons have different concepts regardless of the similarity of their desired companies (there may be some debate over whether the difference lies in the concept itself or its representation method, however, at least the differences in concept map created were able to be identified). Furthermore, the comparison of concept maps of students and those of company employees has shown that the differences mentioned above appear more outstandingly for the employees. Not only the difference between students and employees in concept map is large, but also the difference among employees in concept map is larger than that among students. This result can be attributed to that the scarcity of work experience induces students to tend to have a general image of a company based on its brochure and others, whereas company employees have a concrete image that reflects their individual work experience.

## 5 Conclusion

This paper has described a system called Mochi that the authors have developed to support the collaboration that necessitates concept sharing. And then, it described a set of experiments conducted that applied Mochi to concept sharing in career designing as an example, and discussed its result. The experiments used only a few subjects,

however, interestingly have found out that there is a large difference among persons in concept of career design, and especially the difference between students and employees is outstanding. Any effective support for concept sharing using tools including Mochi is expected to overcome the perception gap between job offerers and job hunters and enable a more fruitful job matching for the both. Noticing the perception gap between one another can serve as a chance discovery[11] for a successful job matching. In this sense, Mochi may be useful as a chance discovery tool.

Unfortunately, the set of experiments conducted this time was limited to the representation of concepts by each student and employee and the verification of the difference between them, and was not able to follow up the process of concept sharing. As challenges of the future, the authors wish to apply Mochi in order to extend support to the concept sharing process and add necessary features to Mochi to make it more useful as a concept sharing support tool.

## References

1. Fowler, M., Scott, K., "UML Distilled: A Brief Guide to the Standard Object Modeling Language, 2nd Edition," Addison-Wesley, 1999.
2. Kobayashi, Y., "Color Image Scale, 2nd Edition," Kodansha, 2001. (In Japanese)
3. Microsoft Exchange, <http://www.microsoft.com/japan/exchange/>
4. Cybozu, <http://cybozu>
5. Shoji, H., "How can you discover yourselves during job-hunting?" KES2005 Conference Proceedings, Part I (LNAI3681), pp.1181-1185, Springer, 2005.
6. Shoji, H., Hori, K., Toward Improved Interface of Online Shopping System, IPSJ JOURNAL, Vol.42, No.6, pp.1387-1400, 2001 (in Japanese).
7. Kobayashi, K., Miraculous Four-line Diary Made Possible in Five Minutes A Day, Index Communications Corp., 2002 (in Japanese).
8. Omote, S., Magic of Diary: A Habit to Drastically Change Your Life, Sunmark Publishing, Inc., 2004 (in Japanese).
9. Schoen, D.A., The Reflective Practitioner: How Professional Think in Action, Basic Books, 1983.
10. Suwa, M., Purcell, T. and Gero, J., Macroscopic analysis of design processes based on a scheme for coding designers' cognitive actions, Design Studies Vol.19, No.4, pp.455-483, 1998.
11. Ohsawa, Y., McBurney, P. (Eds.), "Chance Discovery," Springer, 2003.

# What Should Be Abducible for Abductive Nursing Risk Management?

Akinori Abe, Hiromi Itoh Ozaku, Noriaki Kuwahara, and Kiyoshi Kogure

ATR Media Information Science Laboratories  
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan  
{ave, romi, kuwahara, kogure}@atr.jp

**Abstract.** In this paper, we analyze the hypothesis features of dynamic nursing risk management. In general, for risk management, static risk management is adopted. However, we cannot manage novel or rare accidents or incidents with general and static models. It is more important to conduct dynamic risk management where non-general or unfamiliar situations can be dealt with. We, therefore, propose an abductive model that achieves dynamic risk management where new hypothesis sets can be generated. To apply such a model to nursing risk management, we must consider types of newly generated hypotheses because sometimes newly generated hypotheses might cause accidents or incidents. We point out the preferable hypotheses features for nursing risk management.

## 1 Introduction

Despite recent high-quality nursing education and advanced medical treatments, the number of medical accidents due to nursing activities has not decreased. Instead, expectations about the safety and quality of nursing care have increased, and the range of nursing responsibilities has expanded. Therefore, to lower medical accidents, it is important to reduce nursing accidents and incidents to benefit both hospitals and patients. Medical risk management is one realistic solution to solve the problem, and many hospitals have introduced it. Currently medical risk management is based on a statistical model, which can be generated by inductive methodologies such as data mining. Around 80% of all accidents and incidents can be prevented by applying such inductive nursing risk management. However, if we only use inductive risk management, we cannot deal with novel or rare cases because all accidents or incidents cannot be known. Inductive (static) risk management cannot deal with novel or rare situations. We, therefore, think it would better to make risk models that dynamically perform risk management, which can be achieved by abduction. Risk management includes a concept — risk prediction. For computational prediction, abduction is the best selection. For dynamic risk management, we need to model nursing activities or human behaviour for errors. Based on Vincent's model, we previously proposed abduction-based nursing risk management [Abe et al. 2004] and extended the model to a scenario violation scheme [Abe et al. 2006]. Both models are quite flexible for conducting risk management, but currently they suffer certain

limitations when dealing with new error models. For instance, we can adopt CMS [Reiter and de Kleer 1987] or AAR [Abe 2000] for abduction, but for nursing risk management, abducted hypotheses cannot always be adopted as chances.

In this paper, we analyze both the features of hypotheses and suitable hypotheses for nursing risk management. Section 2 offers an overview of abductive nursing risk management. Section 3 discusses the features of abducible hypotheses in nursing risk management.

## 2 Abductive Nursing Risk Management

In previous papers, we pointed out the importance of dealing with possibly hidden (ignored or unconscious) events, factors, environmental elements, personal relationships, or matters likely to cause an unrecognized but serious accident in the future. Such factors can be regarded as chances. In [Abe 2003], we proposed an abductive framework for Chance Discovery [Ohsawa 2002] that can be achieved by abduction. In this section, we model risk management based on abduction. First, we briefly outline a pure abduction-based risk management model and then illustrate a scenario violation model for risk management.

### 2.1 Abduction Model

In [Abe et al. 2004], we formalized nursing risk management with an abductive framework. In cases where we know all possible hypotheses and their ideal observations<sup>1</sup>, we can detect malpractice beforehand because if someone selects a wrong hypothesis set or fails to generate a necessary hypothesis set, an ideal observation cannot be explained. When an ideal observation cannot be explained, an accident or incident occurs. By this mechanism, we can logically determine exactly where accidents or incidents might occur in advance. A simple logical framework for completing an activity is shown below (using the framework of Theorist [Poole et al. 1987]):

If

$$F \cup h_1 \not\models \textit{ideal\_observation}, \quad (1)$$

then find  $h_2$  satisfying (2) and (3).

$$F \cup h_2 \models \textit{ideal\_observation} \quad (2)$$

$$F \cup h_2 \not\models \square. \quad (3)$$

$$h_1, h_2 \in H, \quad (4)$$

where  $F$  is a set of facts that are always consistent and  $h_1$  and  $h_2$  are hypotheses that are not always consistent with the set of facts and other hypothesis sets.

---

<sup>1</sup> If we use a workflow sheet for nurses or an electronic medical recording system, we can determine ideal observations.

Hypotheses are generated (selected) from hypothesis base  $H$ .  $\square$  is an empty set. Therefore, the last formula means that  $F$  and  $h_2$  are consistent.

If we complete formula (2), the activity is successfully completed. On the other hand, if we cannot generate enough hypothesis sets to complete formula (2), certain problems will disturb the completion of the activity. Thus, beforehand we can determine the possibility of risk by abduction. That is, when we cannot explain an *ideal\_observation* with a current hypothesis set, a particular error might occur. If objective (*ideal\_observation*) cannot be explained, it cannot be completed. This situation is caused by particular accidents or incidents. This very simple logical that does not consider any effects of the generation order of hypotheses.

## 2.2 Scenario Violation Model

In a pure abduction model, we cannot deal with time information. For time information, Brusoni proposed a theoretical approach for temporal abduction [Brusoni et al. 1997], that shows abduction with absolute time information. However, we do not need to prepare strict models for temporal projection; instead we need to have a simple strategy to express a series of events.

For that, we introduced scenario in abduction and proposed a scenario violation model [Abe et al. 2006]. As shown in [Ohsawa et al. 2003], scenario is a *time series* of events under a coherent context. Accordingly, by introducing scenario, we can deal with time information in abduction. Thus, we introduced the effects of the generation order of hypotheses on the abduction model. In a scenario violation model, scenario violation means the possibility of error. This is a formalization of nursing risk management considering a series of events (time).

A simple logical model for checking a scenario violation is illustrated below. When all candidate scenarios are in a scenario base ( $SB$ ), risk determination inference can be achieved as follows:

$$s_i \in SB \quad (5)$$

$$s_i = \sum_{j(\text{in chronological order})} e_{ij}, \quad (6)$$

where  $s_i$  is a scenario and  $e_{ij}$  is an event.

As shown above, to avoid accidents or incidents (by completing an activity), it is necessary to reach a final goal. When reached, we observe a particular result. Accordingly, we can set an observation as a result from the final goal. Therefore, our aim is to explain observations with sets of scenarios. Thus when no accident or incident occurs, the following formulae are completed:

$$F \cup \sum_{i(\text{in chronological order})} O_i \models O \quad (7)$$

$$F \cup s_i \models O_i, \quad (8)$$

where  $F$  is background knowledge and  $O_i$  and  $O$  are observations (results of nursing activities).  $O_i$  can be regarded as a sub-observation of  $O$ . Of course, in some cases, we do not need to consider sub-observations.

Formulae (7) and (8) show abduction (hypothetical reasoning) that determines whether a scenario is completed. The most important difference from the hypothetical reasoning model is that formula (7) requires the verification of the chronological order of scenarios (hypotheses).

When

$$F \cup s_j \not\models O'_j, \quad (9)$$

$$F \cup s_j \models O_j, \quad (10)$$

$$O_j \neq O'_j, \quad (11)$$

and

$$F \cup \sum_{i(\text{in chronological order})} O_i \not\models O, \quad (12)$$

particular scenarios appear to be violated, indicating the possibility of an error. The possibility of accidents or incidents occurring can logically be determined (explained) by abduction before they occur.

In this formalization, a scenario can be regarded as a structured and ordered hypothesis. In addition, each event can also be regarded as an ordered hypothesis.

### 3 Features of Abducible Hypotheses

We proposed abductive nursing risk management to achieve dynamic risk management. However, the current formalization is still based on hypothetical reasoning where hypotheses are generated (selected) from a hypothesis base. In this section, we discuss the features of hypotheses to be generated in actual dynamic risk management.

#### 3.1 Abduction Model

In abduction models, part or all necessary hypotheses are previously prepared as an hypothesis base. We can extend hypothetical reasoning by introducing a mechanism of CMS [Reiter and de Kleer 1987] where missing hypotheses can logically be generated (created). For instance, consider the following case ( $h_1$  is another necessary hypothesis set):

$$F \cup h_2 \models \text{injection}(\text{Diamox}), \quad (13)$$

$$h_2 = \text{Diamox} \vee h_1. \quad (14)$$

Even if the hypothesis base does not include *Diamox* as an hypothesis, if we have the following *fact*, we can apply CMS to generate *Diamox* as a missing hypothesis:

$$\text{injection}(X) :- \text{content}(X) \wedge \text{distilled\_water} \wedge \text{give\_injection}. \quad (15)$$

In fact, CMS can logically generate missing hypotheses, but its limitation is that it can only generate clauses, that is, a minimal conjunction of known or to be known terms or their negations. For instance, in the above example, *Diamox* is not a known hypothesis, but it can be known from observation. Then *Diamox* can be abduced. We presume that a fact base must be complete. We believe that after a complete model of nursing activities is obtained from data in the E-nightingale project [Kuwahara et al. 2004], if we consult with electronic medical recording systems, we can presume the completeness of the fact base.

However, it would be better to generate completely unknown knowledge during abduction, which is “real abduction.”

### 3.2 Abductive Analogical Reasoning

For “real abduction,” in [Abe 2000], we proposed Abductive Analogical Reasoning (AAR) that logically and analogically generates missing hypotheses. Its generation mechanism is similar to CMS’s. Structures of generated knowledge sets are analogous to the known knowledge sets. In the framework of AAR, not completely unknown but rather unknown hypotheses can be generated. The inference mechanism is briefly illustrated as follows (for notations, see [Abe 2000]):

When

$$\Sigma \not\models O, \quad (O \text{ cannot only be explained by } \Sigma.) \quad (16)$$

$\Sigma$  (background knowledge) lacks a certain set of clauses to explain  $O$ . Consequently, AAR returns a set of minimal clauses  $S$  such that

$$\Sigma \models S \vee O, \quad (17)$$

$$\neg S \notin \Sigma. \quad (18)$$

The result is the same as CMS’s. This is not always a guaranteed hypothesis set. To guarantee the hypothesis set, we introduced analogical mapping from known knowledge sets.

$$S \mapsto S', \quad (S' \text{ is analogically transformed from } S.) \quad (19)$$

$$\neg S' \in \Sigma, \quad (20)$$

$$S' \mapsto S'', \quad (21)$$

$$\Sigma \models S'' \vee O, \quad (22)$$

$$\neg S'' \notin \Sigma. \quad (23)$$

$O$  is then explained by  $\neg S''$  as an hypotheses set. Thus we can generate a new hypothesis set that is logically abduced whose structure is similar to authorized (well-known) knowledge sets.



### 3.3 Features of Abducible Hypotheses in Nursing Risk Management

We introduced AAR to Chance Discovery [Abe 2003] where we defined two types of chances. The first suggests unseen or unknown events as chance, and the second suggests known events as chance by generating new rules. In both types of chances, abduction and analogical mapping play a significant role. The role of abduction is the discovery and suggestion of chance, and the role of analogical mapping is the adjustment and confirmation of chance. AAR works well in usual situations, but in the case of nursing risk management, such alternatives cannot always be applied. Actually, we cannot adopt optional medicine as an hypothesis. Sometimes, nurses mistakenly give a similarly named medicine that causes an accident or incident. Even if the effectiveness of the medicine is similar, it might cause a problem. Thus, for a medicine, we should abduce the same medicine or one that has identical effectiveness to remove nursing accidents or incidents. For other factors, the situation is identical. Thus we cannot always apply the first framework to determine nursing accidents or incidents. Instead, we can adopt the second framework.

In the second type of chance discovery, we refer to the structure of a knowledge set that can be regarded as a scenario for generating an hypothesis set. In [Abe et al. 2006], we introduced the concept of a scenario to express flow and time in nursing activities. For a scenario violation model, the main aim is to determine the violation of a scenario, for which we need to prepare all possible scenarios. We proposed to utilize nursing practice manuals provided by hospitals and nursing academies to build a nursing scenario base. Even if we automatically generate a nursing scenario base by referring to those materials, it is still difficult to compile a perfect one. As pointed out in [Abe et al. 2006], in nursing scenarios, not all but just part of the scenario order is important. For similar activities, the important part of a scenario is almost identical. For instance, a necessary medicine must be dissolved before an injection. Thus for similar activities, there should be common unchangeable scenarios. We can refer to such scenarios to determine scenario violations even if we do not know the complete scenario of the activity.

### 3.4 Ontology for Nursing Risk Management

For analogical mapping, we need to prepare dictionaries that show similarities between multiple scenarios. A thesaurus can be applied to such problems. Of course it can be partially applied to nursing risk management based on scenario violation. However, a thesaurus usually gives linguistic similarities. We need similarities for actual activities. That is, we need to prepare a dictionary that can provide similarities for actual nursing activities. For that, we are currently building an ontology that deals with nursing activities [Abe et al. 2005]. We are also trying to build a set of nursing corpora [Ozaku et al. 2005] and to extract nursing workflow patterns (scenario) by analyzing transcribed nursing dialogues [Ozaku et al. 2006a, Ozaku et al. 2006b]. We manually add the tags of nursing tasks to the transcribed nursing dialogues (Table 1 (Private information is modified.)). The types of tags are determined by referring to authorized

job categories provided as Classification of Nursing Practices [CNP 2005] and Nursing Practice Classification Table [NPCT 2004]. They include such labels as “conference (18-106)” and “intravenous infusion (13-63-6A0502).” After adding such tags, we can build an ontology that can be applied to AAR-based nursing risk management.

**Table 1.** Labeled dialogues from nurses

Time	dialogue	Job Category
11:01:00	I'm going to a short conference (meeting or handover).	18-106 conference
11:20:48	The short conference is finished.	18-106 conference
11:28:11	I'm going to prepare a drip infusion for Abe-san.	13-63-6A0502 intravenous infusion
11:32:01	I have finished preparing the drip for Abe-san.	13-63-6A0502 intravenous infusion

## 4 Conclusions

For dynamic risk management, we need to deal with an incomplete knowledge base that lacks knowledge. To supplement the missing knowledge, we can apply abduction, and we proposed abduction-based nursing risk management. In this paper, we analyzed the features of hypotheses for nursing risk management. The results are as follows:

- Hypotheses that suggest unseen or unknown events as chance  
An exact hypothesis set is necessary to conduct nursing risk management.
- Hypotheses that suggest known events as chance by generating new rules  
Similarly structured scenarios can be referred to for conducting nursing risk management.

As shown in this paper, for nursing activities, we cannot always freely adopt alternatives as new hypotheses. Even if analogically correct, an accident or incident might occur. Thus we should abduce the same element or one that has the same effectiveness for removing nursing accidents or incidents by preparing a specialized knowledge base that can be applied to such problems as shown above. We need to prepare proper categorization of nursing tasks for knowledge. Thus we need to prepare an ontology that deals with the categorization of nursing tasks, which are now building by collecting nursing activities from hospitals.

## Acknowledgments

This research was supported in part by the National Institute of Information and Communications Technology (NICT).

## References

- [Abe 2000] Abe A.: Abductive Analogical Reasoning, *Systems and Computers in Japan*, Vol. 31, No. 1, pp. 11–19 (2000)
- [Abe 2003] Abe A.: The Role of Abduction in Chance Discovery, *New Generation Computing*, Vol. 21, No. 1, pp. 61–71 (2003)
- [Abe et al. 2004] Abe A., Kogure K., and Hagita N.: Determination of A Chance in Nursing Risk Management, *Proc. of ECAI2004 Workshop on Chance Discovery*, pp. 222–231 (2004)
- [Abe et al. 2005] Abe A., Sagara K., Ozaku H.I. Kuwahara N., and Kogure K.: On Building of Nursing Ontology, *JCMI25*, 3-D-2-4 (2005) (In Japanese)
- [Abe et al. 2006] Abe A., Ozaku H. I., Kuwahara N., and Kogure K.: Scenario Violation in Nursing Activities — Nursing Risk Management from the viewpoint of Chance Discovery, *Soft Computing Journal*, Springer (2006) to appear
- [Brusoni et al. 1997] Brusoni V., Console L., Terenziani P., and Dupré D.T.: An Efficient Algorithm for Temporal Abduction. *Proc. 5th Congress of the Italian Association for Artificial Intelligence*, pp. 195–206 (1997)
- [CNP 2005] Japan Academy of Nursing Science eds.: Classification of Nursing Practice, *Japan Academy of Nursing Science* (2005) (in Japanese).
- [NPCT 2004] Japan Nursing Association eds.: Nursing Practice Classification Table, *Japan Nursing Association* (2004) (in Japanese).
- [Kuwahara et al. 2004] Kuwahara N. et al.: Ubiquitous and Wearable Sensing for Monitoring Nurses' Activities, *Proc. SCI2004*, Vol. VII, pp. 281–285 (2004)
- [Ohsawa 2002] Ohsawa Y.: Chance Discovery for Making Decisions in Complex Real World, *New Generation Computing*, Vol. 20, No. 2, pp. 143–163 (2002)
- [Ohsawa et al. 2003] Ohsawa Y., Okazaki N., and Matsumura N.: A Scenario Development on Hepatics B and C, *Technical Report of JSAI*, SIG-KBS-A301, pp. 177–182 (2003)
- [Ozaku et al. 2005] Ozaku H.I., Sagara K., Naya F., Kuwahara N., Abe A., and Kogure K.: Building Dialogue Corpora for Nursing Activity Analysis, *Proc. of LINC-2005 Workshop*, pp. 41–48 (2005)
- [Ozaku et al. 2006a] Ozaku H.I., Abe A., Sagara K., Kuwahara N., and Kogure K.: A Task Analysis of Nursing Activities Using Spoken Corpora, *Advances in Natural Language Processing, Research in Computing Science*, Vol. 18, pp. 125–136 (2006)
- [Ozaku et al. 2006b] Ozaku H.I., Sagara K., Kuwahara N., Abe A., and Kogure K.: Nursing Spoken Corpora for Understanding Nursing Assignments, *Proc. of NI2006*, pp. 481–485 (2006)
- [Poole et al. 1987] Poole D., Goebel R., and Aleliunas R.: Theorist: A Logical Reasoning System for Defaults and Diagnosis, *The Knowledge Frontier: Essays in the Representation of Knowledge (Cercone N.J., McCalla G. Eds.)*, pp. 331–352, Springer-Verlag (1987)
- [Reiter and de Kleer 1987] Reiter R. and de Kleer J.: Foundation of assumption-based truth maintenance systems: Preliminary report. *Proc. of AAAI87*, pp.183–188 (1987)
- [Vincent et al. 1993] Vincent C., Ennis M., and Audley R. J. eds.: *Medical Accidents*, Oxford University Press (1993)

# The Repository Method for Chance Discovery in Financial Forecasting

Alma Lilia Garcia-Almanza and Edward P.K. Tsang

Department of Computer Science  
University of Essex  
Wivenhoe Park, Colchester, CO4 3SQ, UK  
algarc@essex.ac.uk, edward@essex.ac.uk

**Abstract.** The aim of this work is to forecast future opportunities in financial stock markets, in particular, we focus our attention on situations where positive instances are rare, which falls into the domain of Chance Discovery. Machine learning classifiers extend the past experiences into the future. However the imbalance between positive and negative cases poses a serious challenge to machine learning techniques. Because it favours negative classifications, which has a high chance of being correct due to the nature of the data. Genetic Algorithms have the ability to create multiple solutions for a single problem. To exploit this feature we propose to analyse the decision trees created by Genetic Programming. The objective is to extract and collect different rules that classify the positive cases. It lets model the rare instances in different ways, increasing the possibility of identifying similar cases in the future. To illustrate our approach, it was applied to predict investment opportunities with very high returns. From experiment results we showed that the Repository Method can consistently improve both the recall and the precision.

## 1 Introduction

Financial forecasting is one of the important areas in computational finance [1]. Based on Genetic Programming (GP) [2] and aided by constraints [3], EDDIE is a machine learning tool for financial forecasting [4,1]. However, in some situations, the number of profitable investment opportunities is extremely small, this occurs, for example, in finding arbitrage opportunities [5]. The interest in finding rare opportunities motivates our research in chance discovery [6,7].

Machine learning classifiers, like other forecasting techniques, extend the past experiences into the future. However, the imbalance between positive and negative cases poses a serious challenge to machine learning. Specifically GP, which is a evolutionary technique, has limitations to deal with imbalanced data sets. Because it favours negative classifications, which has a high chance of being correct due to the nature of the data. In imbalanced data sets the classifier performance must not be measured only by the *accuracy*<sup>1</sup> [8,9]. A common measure

---

<sup>1</sup> Accuracy is the proportion of the total number of predictions that were correctly predicted.

for a classifier performance, in imbalanced classes, is the geometric mean of the product of *precision*<sup>2</sup> and *recall*<sup>3</sup>[9].

The objective of the Repository Method (RM) is to increase the recall in the classification without sacrificing the precision (i.e. without substantially increase the total number of false positives). Genetic algorithms are able to produce multiple solutions for a single problem. However, the standard procedure is to choose only the best individual of the evolution as the optimal solution of the problem and discard the rest of the population. A GP process spends a lot of computational resources evolving entire populations for many generations. For this reason we presume that the remaining individuals could contain useful information that is not necessarily considered in the best individual. We propose to analyse the decision trees in a wider part of the population and in different stages of the evolutionary process. The idea behind this approach is to collect different rules that model the rare cases in diverse ways. The over-learning produced by this method attempts to compensate the lack of positive cases in the data set. This work is illustrated with a data set composed by closing prices from the London Financial Market.

The remainder of this paper is organized as follows: Section 2 contains an overview of the problem that illustrates our method; Section 3 presents our approach, while Section 4 describes the experiments to test our method. Section 5 presents the experiment results. Finally, Section 6 summaries the conclusions.

## 2 Problem Description

To illustrate our method, it was applied to a problem for discovering classification rules in a financial stock data set. The goal is to create a classifier to predict future movements in the stock price. This problem has been addressed previously by Tsang *et al.* [10,4,5]. Every case in the dataset is composed by a *signal* and a set of attributes or *independent variables*. The signal indicates the opportunities for *buying* or *not buying* and *selling* or *not selling*. The signal is calculated looking ahead in a future horizon of  $n$  units of time, trying to detect an increase or decrease of at least  $r\%$ . The independent variables are composed by financial indicators derived from financial technical analysis. Technical analysis is used in financial markets to analyse the price behaviour of stocks. This is mainly based on historical prices and volume trends [11].

## 3 Repository Method

The objective of this approach is to compile different rules that model the positive cases in the training data set. Since the number of positive examples is very small, it is important to gather all available information about them. RM analyses a wider part of the population to collect useful rules. This analysis is

---

<sup>2</sup> Precision is the proportion of the predicted positive cases that were correct.

<sup>3</sup> Recall is the proportion of positive cases that were correctly identified.

**Table 1.** Discriminator Grammar

G	→ <Root>
<Root>	→ "If-then-else", <Conjunction>   <Condition>, "Class", "No Class"
<Conditional>	→ <Operation>, <Variable>, <Threshold>   <Variable>
<Conjunction>	→ "and"   "or", <Conjunction>   <Conditional>, <Conjunction>   <Conditional>
<Operation>	→ "<", ">"
<Variable>	→ Variable <sub>1</sub>   Variable <sub>2</sub>   ... Variable <sub>n</sub>
<Threshold>	→ Real number

extended to different stages of the evolutionary process. The selection is based on the performance and novelty of the rule. Decision tree analysis and rule collection has been previously addressed by Quinlan [12].

RM involves the following steps: 1) Rule extraction 2) Rule simplification 3) New rule detection. The above procedures will be explained in the following sections.

### 3.1 Rule Extraction

Rule extraction involves the analysis of decision trees in order to delimit their rules. For this reason decision trees are generated and evolved using Discriminator Grammar (DG), see Table 1. This grammar<sup>4</sup> produces trees that classify or not a single class, Figure 1 illustrates a decision tree that was created using DG.

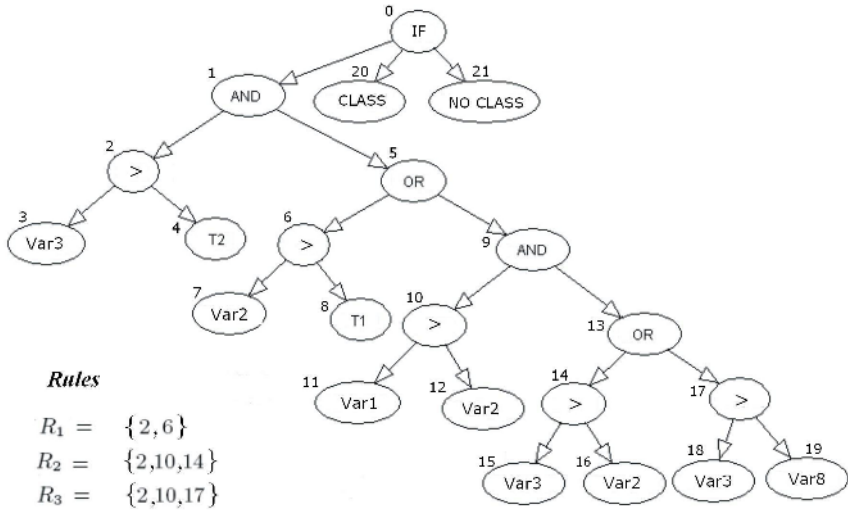
To extract the tree rules, let  $T$  be a tree with syntax DG, it means that  $T$  is composed by rules and it can be expressed as the union of its rules such as  $T = (R_1 \cup R_2 \cup \dots \cup R_n)$  where  $R_i$  is a rule and  $n$  is the total number of rules in  $T$ . A rule  $R_k$  is a minimum set of conditions that satisfy the tree  $T$ , to satisfy  $R_k$  every condition in the rule has to be satisfied. Rule extraction is concerned with the discovery of all minimal sets of conditions that satisfy  $T$  ( see Figure 1). Once a rule  $R_k \in T$  has been extracted this is individually evaluated against the training data. If the precision of  $R_k$  achieves a predefined Precision Threshold (PT), then  $R_k$  is considered for the next step (rule simplification), otherwise  $R_k$  is discarded.

### 3.2 Rule Simplification

The procedure to identify new rules involves the comparison between  $R_k$  and the rules in repository. However noisy and redundant conditions are an impediment to make an effective rule comparison. For this reason rule simplification is an essential part of RM.

Rule simplification is a hard task, specially in latest stages of the evolutionary process, due to decision trees generated by GP tend to grow and accumulate

<sup>4</sup> The term grammar refers to a representation concerned with the syntactic components and the rules that specify it [13].



**Fig. 1.** The figure shows a decision tree generated by DG and its rules. Every rule is composed by a set of conditions that are represented by the number of the *conditional node* (i.e. node with syntax <Conditional> in DG).

introns [14,15,16,17]. To simplify rules we have defined two types of conditions: *hard conditions* and *flexible conditions*. A hard condition is a comparison of two variables (e.g.  $var_1 < var_2$ ). A flexible condition is the equation between a variable and a threshold (e.g.  $var_1 < .8$ ). When two flexible conditions have the same variable and operator they are defined as *similar conditions* (e.g.  $var_1 < 3$  and  $var_1 < 2$  are similar conditions). Conditions were divided, in hard and flexible, because the conditions that compare thresholds could be difficult to differentiate (e.g.  $var_1 < .08998$  and  $var_1 < .08991$ ). In addition similar conditions can be simplified (e.g.  $var_1 < 5$  and  $var_1 < 10$  can be replaced by  $var_1 < 5$ ). Rule simplification details were omitted in this paper due to lack of space. However, a detailed explanation is available in [18].

To simplify rules, the redundant and noisy conditions have to be removed and flexible conditions have to be simplified. Let  $R_k = \{c_1, c_2 \dots c_n\}$  be the set of conditions in  $R_k$ .

- If  $c_1, c_2 \in R_k$  are hard conditions and  $c_1 = c_2$  then  $R_k = R_k - c_2$
- If  $c_1, c_2 \in R_k$  are flexible conditions and  $c_1$  and  $c_2$  are similar conditions then  $c_1$  and  $c_2$  are simplified using the simplification table in [18]. (e.g. conditions  $var_1 < 12$  and  $var_1 < 10$  are similar and they can be replaced by  $var_1 < 10$ )
- If  $c_i \in R_k$  and  $Performance(R_k) = Performance(R_k - c_i)$  then  $R_k = R_k - c_i$

### 3.3 New Rule Detection

Once the rule  $R_k$  has been simplified, we are able to determine if  $R_k$  is different from the rules in the repository. Let  $Rep = \{R_i\}$  be the set of rules in the

repository. Let  $R_i$  be a *hard rule* if  $R_i$  is comprised exclusively of hard conditions. Let  $R_i$  be a *flexible rule* if it has at least one flexible condition.  $R_k$  and  $R_i$  are *similar rules* if they have the same hard conditions and similar flexible conditions. The following procedure determines if  $R_k$  is added or not to the rule repository.

- If  $R_k$  is a hard rule and  $R_k \notin Rep$  then  $Rep = Rep \cup R_k$
- If  $R_k$  is a flexible rule and  $\exists R_i \in Rep$  such as  $R_k$  and  $R_i$  are similar rules and  $Fitness(R_k) > Fitness(R_i)$  then  $Rep = (Rep - R_i) \cup R_k$
- If  $R_k$  is a flexible rule and there is not a  $R_i \in Rep$  such as  $R_k$  and  $R_i$  are similar rules then  $Rep = Rep \cup R_n$

## 4 Experiments Description

To test our approach a series of experiments was conducted. The performance was measured in terms of the recall, precision and accuracy. A population of 1,000 individuals was evolved using a standard GP. Every ten generations the entire population was saved, let's call them  $P_{10}, P_{20}, \dots, P_{100}$ . Subsequently RM analysed these populations and compiled the useful rules. RM accumulated all the useful rules during the entire process. The experiment was tested using different values for the precision threshold (PT = 60%, 70%, 80%). This process was repeated twenty times, the results of the experiment were grouped and averaged by generation and PT. Table 2 presents the GP parameters used to evolve the populations.

### 4.1 Training Data Description

The data sets to train and test the GP in the experiment came from the London stock market. Every data set contains 890 records each from Barclays stock (from March, 1998 to January, 2005). The attributes of each record are composed by indicators derived from financial technical analysis; these were calculated on the

**Table 2.** Summary of Parameters

Parameter	Value
Population size	1,000
Initialization method	Growth
Generations	100
Crossover Rate	0.8
Mutation Rate	0.05
Selection	Tournament (size 2)
Elitism	Size 1
Control bloat growing	Tarpeian method, 50% of trees whose largest branch exceed 6 nodes are penalized with 20% of the fitness for each node that surpassed the largest branch allowed.
Fitness Function	$\sqrt{Recall \cdot Precision}$



basis of the daily *closing price*<sup>5</sup>, volume and some financial indices as the FTSE<sup>6</sup>. We looked for *selling* opportunities of 15% in ten days. The number of positives cases is 39 in the training data set and 23 in the testing. The opportunities are naturally grouped in clusters, close to the peak to predict.

## 5 Main Results

This section documents the results obtained by applying RM to the set of populations described in the previous section. All figures and tables given in this section denote averaged results from series of twenty test runs. All the results were obtained using the testing data set. Table 3 shows the recall, precision and accuracy of the best individual (according to the fitness function defined in Table 2) in each generation. The same measures (recall, precision and accuracy) were reported for RM using PT =60%,70%,80%.

**Table 3.** Recall, Precision and Accuracy of a standard GP and RM using PT = 60%, 70%, 80%. Averaged results from series of twenty runs.

Gen	RECALL				PRECISION				ACCURACY			
	GP	Repository method			GP	Repository method			GP	Repository method		
		PT=60%	70%	80%		PT=60%	70%	80%		PT=60%	70%	80%
10	1%	1%	0%	0%	1%	1%	0%	0%	97%	97%	97%	97%
20	0%	4%	3%	2%	0%	2%	2%	1%	97%	95%	95%	96%
30	13%	11%	6%	3%	2%	5%	3%	2%	90%	93%	93%	95%
40	10%	21%	16%	10%	4%	6%	5%	5%	92%	91%	91%	93%
50	13%	30%	21%	12%	6%	7%	6%	5%	93%	89%	90%	92%
60	8%	39%	28%	19%	7%	9%	8%	7%	94%	88%	89%	91%
70	16%	44%	35%	23%	5%	9%	8%	7%	88%	87%	88%	90%
80	10%	44%	39%	29%	4%	9%	8%	8%	93%	87%	87%	89%
90	6%	46%	40%	32%	6%	9%	8%	8%	94%	86%	87%	88%
100	21%	47%	43%	34%	3%	9%	9%	8%	82%	85%	86%	87%

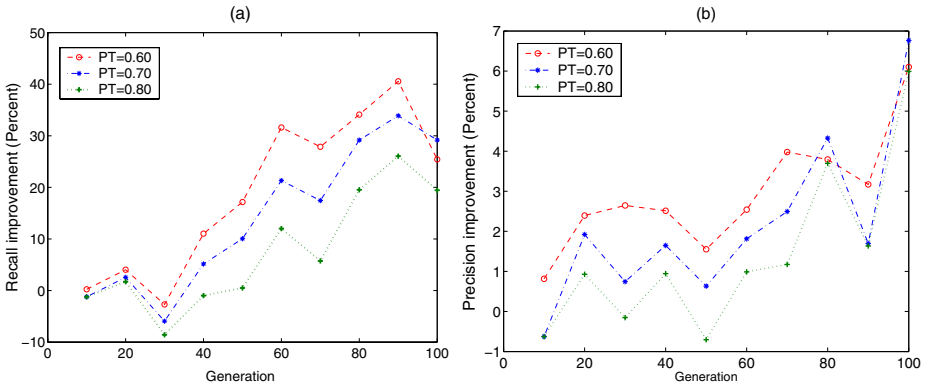
From Table 3, it can be seen that GP recall fluctuates from generation to generation. It is because a single individual only can hold a small set of rules. Thus when the best individual is tested with a different data set it is unable to have a consistent behavior. In contrast the recall obtained by RM increases consistently with the generations. Table 3 shows that, except for earliest generations, RM can obtain better recall than the best GP tree. The low performance in earliest generations is probably due to the fact that most of the trees were not too far from being random. However when the evolutionary process advances, it tends to generate more and better rules, in consequence RM performance

<sup>5</sup> The settled price at which a traded instrument is last traded at on a particular trading day.

<sup>6</sup> An index of 100 large capitalization companies stock on the London Stock Exchange, also known as "Footsie.

improves. As can be seen in Table 3, precision obtained by RM shows a high level of consistency. In the other hand the precision of the best individual fluctuates, due to the same reasons given for recall fluctuation.

The improvement in recall and precision is paid, in some cases, by decrease in accuracy, as it can be seen in generation 90. This is due to the evolution pressure discourages positive classifications in GP, since they have a small chance of being correct (a standard feature in chance discovery). Experiments show that RM is able to pick out rules that together classify more positive cases. In addition, most of the extra positive classifications were correctly made. This is reflected in both increase in recall and precision. But, since more errors were made (not as much, in proportion, as the correct classifications), the overall accuracy has been decreased. Given that our goal is to improve recall and precision, this is an acceptable price to pay.



**Fig. 2.** RM improvement, (a) Recall improvement (b) Precision improvement

## 6 Conclusions

The objective of RM is to mine the knowledge acquired by the evolutionary process to compile more features from the training data. RM compiles rules from different individuals and stages of the evolutionary process. Our experimental results showed that by combining rules from different trees, we can classify more positive cases. RM outperformed the best tree generated by GP, improving the precision and recall. Our approach is a general method, and therefore, results should not be limited to financial forecasting. It should be useful for classification problems where chances are rare, i.e. the data set is imbalanced. Therefore, the Repository Method is a promising general tool for chance discovery.

## Acknowledgement

Authors would like to thank reviewers for their helpful comments. The first author thanks to (CONACyT) to support her studies at the University of Essex.

## References

1. E. P. Tsang and S. Martinez-Jaramillo, "Computational finance," in *IEEE Computational Intelligence Society Newsletter*, 2004, pp. 3–8.
2. J. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, Massachusetts: The MIT Press, 1992.
3. J. Li, *A genetic programming based tool for financial forecasting*. Colchester CO4 3SQ, UK: PhD Thesis, University of Essex, 2001.
4. E. P. Tsang, P. Yung, and J. Li, "Eddie-automation, a decision support tool for financial forecasting" in *Journal of Decision Support Systems, Special Issue on Data Mining for Financial Decision Making*, ser. 4, vol. 37, 2004.
5. E. P. Tsang, S. Markose, and H. Er, "Chance discovery in stock index option and future arbitrage," in *New Mathematics and Natural Computation, World Scientific*, ser. 3, vol. 1, 2005, pp. 435–447.
6. A. Abe and Y. Ohsawa, "Special issue on chance discovery," in *New Generation Computing*, ser. 1, vol. 21. Berlin: Springer and Tokyo: Ohmsha, November 2002, pp. 1–2.
7. Y. Ohsawa and P. McBurney, *Chance discovery*. Springer, 2003.
8. F. J. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *In Proc. Fifteenth Intl. Conf. Machine Learning*, W. Madison, Ed., 1998, pp. 445–553. [Online]. Available: [citeseer.ist.psu.edu/provost97case.html](http://citeseer.ist.psu.edu/provost97case.html)
9. M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," in *Machine Learning*, vol. 30. 195–215, 1998.
10. E. P. Tsang, J. Li, and J. Butler, "Eddie beats the bookies," in *International Journal of Software, Practice and Experience*, ser. 10, vol. 28. Wiley, August 1998, pp. 1033–1043.
11. W. F. Sharpe, G. J. Alexander, and J. V. Bailey, *Investments*. Upper Saddle River, New Jersey 07458: Prentice-Hall International, Inc, 1995.
12. J. R. Quinlan., "Rule induction with statistical data" in *Journal of the operational research Society*, 38, 1987, pp. 347–352
13. N. Chomsky, *Aspects of the theory of syntax*. Cambridge M.I.T. Press, 1965.
14. P. Angeline, "Genetic Programming and Emergent Intelligence," in *Advances in Genetic Programming*, K. E. Kinnear, Jr., Ed. MIT Press, 1994, ch. 4, pp. 75–98.
15. P. Nordin, F. Francone, and W. Banzhaf, "Explicitly Defined Introns and Destructive Crossover in Genetic Programming," in *Proceedings of the Workshop on Genetic Programming: From Theory to Real-World Applications*, J. P. Rosca, Ed., Tahoe City, California, USA, 9 July 1995, pp. 6–22.
16. T. Soule and J. A. Foster, "Code size and depth flows in genetic programming," in *Genetic Programming 1997: Proceeding of the Second Annual Conference*, Eds. Morgan Kaufmann, 1997, pp. 313–320.
17. W. B. Langdon, "Quadratic bloat in genetic programming," in *Proceedings of the Genetic and evolutionary Computation Conference*, 2000, pp. 451–458.
18. A. L. Garcia-Almanza, "Technical report, rule simplification," <http://privatewww.essex.ac.uk/~algarca/documents/Rule-simplification.doc>.

# Emerging Novel Scenarios of New Product Design with Teamwork on Scenario Maps Using Pictorial KeyGraph

Kenichi Horie and Yukio Ohsawa

Department of Quantum Engineering and Systems Science,  
Graduate School of Engineering  
The University of Tokyo

7-3-1 Hongo Bunkyo-Ku, Tokyo 113-8656

Japan

kenhorie@eastcom.ne.jp, Ohsawa@q.t.u-tokyo.ac.jp

**Abstract.** We developed a method of teamwork for products design in real manufacturing company, where Scenario Maps using Pictorial KeyGraph assist creating novel scenarios of new product design. In Pictorial KeyGraph, photographs of physical objects corresponding to incomprehensible items in given data are embedded to the visual result of KeyGraph applied to their business report. In their communications with Pictorial KeyGraph, novel and practical scenarios of new products design were extracted, and 5 new patents have been applied. We found evidences that the team members tend to combine established concepts via rare words in creative designing.

## 1 Introduction

In manufacturing companies, it is expected for the members to develop new products according to customers` requirements. For this purpose, they create scenarios of features of forthcoming products. Here a scenario means a story in which a designed product/service is used. In this study, “Chance Discovery Double Helix Loop Process” [1] as Chance Discovery method was adopted, which interact objective data on computer with subjective data such as experience, inspiration, etc. created by human in order to discover a chance [2]. In the process, tools for visualizing the relation among events/items based on data, such as KeyGraph [3, 4], has been introduced. By looking at the diagram, team members are supposed to understand the meaningful sequence of events, by connecting closely located items. However, a critical problem has been remaining if they intend to apply the method to design and development. That is, they are born and bred in different contexts but they should talk for developing a new product. For example, engineers and designers are graduated from engineering school, and marketers or managers are from business or management schools. Thus, the vocabulary gap among these people causes a deadlock, i.e., the creative ideas of developers and designers can not contribute to the corporate decision. In turn, the proposals from marketers do not move designers or developers. Even if there are technical sales people good at talking both with designers and customers, their words may not easy for marketing people or management staffs, due to the vocabulary gaps. In this paper, we propose and validate a method to aid the cross-disciplinary communication to achieve a decision of a product design.

## 2 Summaries of the Problem and the Solution Method

Here we can point out a critical issue: The system to be produced here is a complex machine, to be dealt with by highly trained technicians in CCD surface inspection system manufacturing companies. The team members are skilled engineers, as well as the designers, developing technicians, and the technical sales staffs. Although technical sales people collected and brought customers' reports, those reports were written in specific engineering terms. Even though the reports included call reports and test reports, no new and practical ideas for product innovations have been extracted after all.

To this problem, we developed the following procedure. Here the tool is KeyGraph, with a new function to embed pictures of real entities corresponding to items (i.e., words) in the customers report. This function is used in 2-2).

[The Design Communication with Pictorial KeyGraph]

- 1) Visualized the textual dataset  $D$ , obtained by putting customer reports in one, by KeyGraph.
- 2) Do communication about considerable scenarios with looking at the graph shown in 1). This communication goes as:
  - 2-1) Each participant presents a scenario about the product behaviors.
  - 2-2) if there is an item on KeyGraph, which is hard to understand, then, a participant may request other participants to embed a picture of the corresponding entity.
  - 2-3) The moderator may control participants so that the utterances do not be biased to a few members.
- 3) After a fixed time length of communication, the moderator stops and select the most feasible scenarios from the criteria of the cost of development and the expected sales.

We expected that the simple revision with showing pictures for unknown words realizes a significant breakthrough. The advantage is three fold: First, the vocabulary gap is filled by the pictures. Second, the deeper level gap that is the difference in the concerns of team mates is also filled, by visualizing the interestingness of the most uncertain components of the diagram. It has been pointed out that uncertain information triggers a favorable design process [5], but this stands only if the collaborators share a going concern with the development. Third, the pictures are easier than words for the user to imagine the real scenes of product behaviors.

This evaluation is purely subjective and uncontrolled, in that we see the effects of the presented method on human's creativity in the real process of design, not on any precision/accuracy measures. This is an ideal way of evaluation in this study, because our goal is to resolve the communication gap caused by the participants' difference in their expertise. By solving this severe human-factor problem, we find significant creativity apparently caused by the pictures on Pictorial KeyGraph.

## 3 KeyGraph and Pictorial KeyGraph

A "chance" is defined as event or situation with significant impact on human decision making. The "discovery" of a chance is to become aware of and to explain the significance of a chance, especially if the chance is rare and its significance has been unnoticed. [6]

The conceptual model of “Chance Discovery Double Helix Loop Process” is developed based on the fusion of human’s process and computer’s process, both of which approach spirally and connect helically in the process of chance discovery. [7] It is defined as to detect, understand, and use events that a significant for a decision. KeyGraph is employed as a tool for event map [2] visualization prevalent in the process. KeyGraph presents a two dimensional undirected graph. (Figure1.) In KeyGraph, Black node is appeared with a word, which is higher frequent occurrence in the text data. Black node itself shows event, which is growing or dying fragment of an ontological entity in human recognition. Moreover, Black nodes are connected with links if frequent co-occurrence of these words is higher in the same sentence. The cluster of Black nodes and links as a graph structure is named as *Island*. The *Island* is a stable structure with well-established familiar events, corresponding to an ontological entity in human recognition. On the other hand, Red node is appeared with a word, which is rare frequent occurrence in the text data. The Red node, which is connected with links among Black nodes in the *island*, if frequent co-occurrence with these Black nodes is higher in the same sentence, is named as *Bridge*

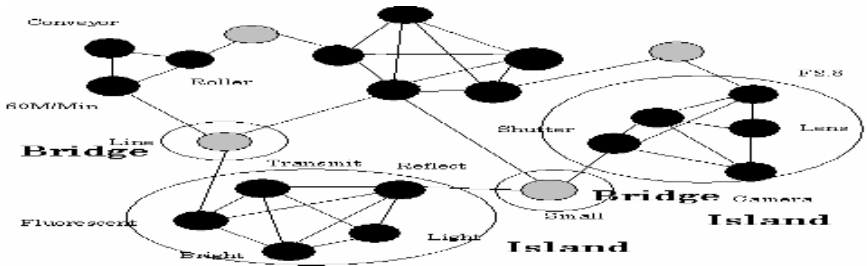


Fig. 1. Sample of KeyGraph

Pictorial KeyGraph succeeds the method to embed realistic images onto the graph, after the experience of the textile company. Here, we introduce Pictorial KeyGraph, where nodes are replaced by pictures of corresponding real entities in data set *D*. This replacement has been executed by user’s dragging picture icons, from the PC desktop to the output window of KeyGraph (See Fig.2).

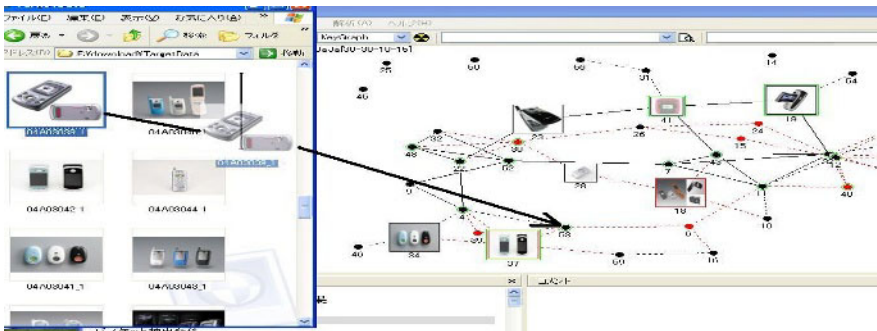


Fig. 2. The usage of Pictorial KeyGraph (on our original software named Polaris [8])

## 4 Application of KeyGraph to Product End –User Reports

### 4.1 Preliminary Study and Tasks

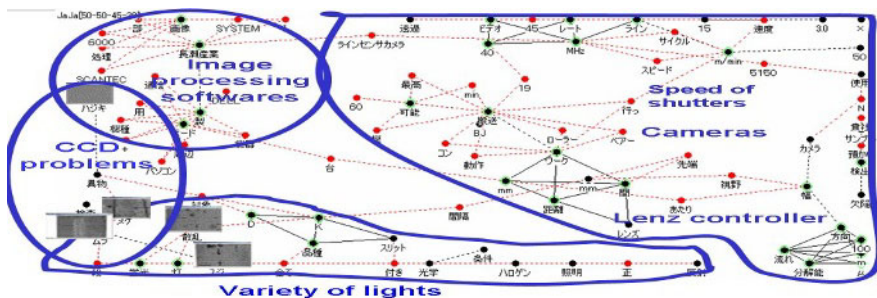
We executed preliminary study based on customer call reports and test reports. The reports were written in free text format by technical sales people, on new functions and new products related to linear CCD surface inspection system. The aim of these reports was to detect defects (scratches on web film surface), as required by customers, on their products by the inspection system. It was, however, neither possible for all of technical sales people to interpret nor to create scenarios from KeyGraph, after processing all customer call reports and test reports by KeyGraph working on the Polaris platform.[8] After this preliminary study, we found three major tasks to be settled. Firstly, 20 groups of defects, which include 64 defect categories, had some names given by the customers. However, the meanings of those names were neither identified commonly by technical sales people, nor by the managing executive of technical sales section. Secondly, the large majority of these names came to be located as red nodes in KeyGraph, i.e., as bridges. Lastly, the meanings of the defect names were ambiguous: Same names were assigned for seemingly different defects, in the customer call reports and test reports.

### 4.2 Application of Pictorial KeyGraph to Redesigning CCD Surface Inspection Systems

We executed the following procedure.

1. Prepare photographs of all defects for 20 groups, of 64 categories, and identify the names used for these defects among the subject customers.
2. Create a graph, with Pictorial KeyGraph, embedding above photographs of defects to nodes for corresponding names (mainly red nodes on KeyGraph). See Figure 3.
3. Separate customer call reports and test reports into each customer basis.

On the way of this procedure, it came out to be much easier to identify names of defects commonly among them, and understand the relation with defects as bridges, among islands corresponding to topics on camera, image processing software, lighting, etc. The vocabulary gap among them are fulfilled and the common opinions of subjects are agreed



**Fig. 3.** A result of Pictorial KeyGraph for the reports from customers, i.e., the end users of CCD surface inspection system

### 4.3 Experimental Conditions and Process

In this study, one (1) sales manager managing the sales of the system, two (2) experienced technical sales people with more than 10 years experience for technical sales, and three (3) inexperienced technical sales people with experience less than 3 years, were chosen, 6 subjects in total. This number may look small, but is the maximum number of experts we can collect. To these subjects, 16 pictorial KeyGraphs (Fig 3.) were shown one by one. Each scenario was created through group discussion among subjects. That is, they uttered scenarios, and threw objections if any to presented scenarios, for each KeyGraph. And, when all subject participants in the group agreed with a scenario, the scenario was recorded. Thus, a number of scenarios came out as a data set in time series.

## 5 Experimental Results

### 5.1 Classification of Extracted Scenarios

During the discussion, 104 scenarios were obtained sequentially. These 104 scenarios could be classified into 85 about the present design and 19 about future designs, respectively, of CCD surface inspector.

For example, the following are two of the 19 scenarios for the future designs.

1. *Use Line Sensor Camera N5150Hs with 50mm lens, of which the resolution and the video rate are 100 micron toward width and length and 40 MHz. This is for inspecting unevenness, line, sputter and drop out, etc., which our customers require.*
2. Use fluorescent lamp in regular reflection and Halogen lamp with slit in transmission, by changing the width of slit. This change should be done according to the category of the defect.

### 5.2 The Roles of Bridges (Red Nodes)

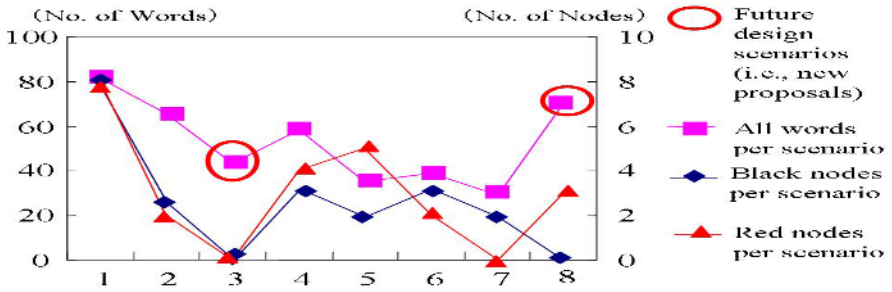
The analysis of temporal variance among the number of characters in each scenario and of words for the black/red nodes contained in each scenario, among those extracted from the 16 graphs obtained by Pictorial KeyGraph, was executed. The following features were common to 10 (of the 16) graphs, that were all the graphs from which the 19 future-design scenarios were extracted (100%), according to the curves in Figure 4.

1. The scenarios mostly correspond to present designs. Then, just before the appearance of a future-design scenario, the number of characters per scenario decreases.
2. The number of words for red nodes per scenario decreases once, a few scenarios before the appearance of a future-design scenario.

In addition, we observed the following feature in 8 of the 10 graphs (80%).

3. The number of words for red nodes per scenario increases, just on/after extracting a future proposal.





**Fig. 4.** The temporal variance of the numbers of words and red/black nodes, and the timing of the appearance of future-design scenarios

Based on these observations, we hypothesized that new scenarios emerge from the combination of established islands, via the awareness of bridging red nodes. If this hypothesis stands, it gives us useful implication: For example, we can improve the way of use of KeyGraph as follows: First show the black nodes and black links, i.e., only the islands, until the user thinks over for creating new scenarios. Then, show the red nodes and red links which may play a role as a bridge for expressing new ideas about the future scenarios.

In order to validate this hypothesis, we investigated the data of presented scenarios into more details. First, we counted the number of words corresponding to red nodes per present-design scenario, and per future-design scenario. As a result:

1. The number of present-design scenarios containing red nodes, appearing one scenario before a future-design: 16 (of 19) scenarios
2. The number of present-design scenarios *not* containing red nodes, appearing one scenario before a future-design: 3 scenarios
3. The number of future-design scenarios containing red nodes: 10 scenarios

Thus, we can say that a present-design scenario, presented when the participants are close to creating a future-design scenario, tends to be created referring to the words for red nodes. When the participants go into the phase to present future scenarios, more than half of scenarios are created referring to red nodes.

Then, we checked the topics of future-design scenarios and present-design scenarios, appearing just before and after the appearance of future-design scenarios. Here a topic means the theme discussed, corresponding to the component of the product they considered to improve. For example, “about camera in the CCD surface inspection system” “about image processing software,” and so on, was the topics.

The results were as follows:

1. The topics of the future-design scenarios were the same as of present-design scenarios just *before* the future-design: 12 (of 19) scenarios.
2. The topics of the future-design scenarios were the same as of present-design scenarios just *after* the future-design: 2 (of 12: the other 7 future-design scenarios appeared at the last of discussion, so no scenarios after the 7 could be counted) scenarios.

The same topic has been discussed by examinees mainly just before “Future proposal”. But the topic was changed suddenly to another one on or just after “Future proposal”

## 6 Conclusion

We developed a method of teamwork for a product design in a real manufacturing company, where Pictorial KeyGraph aids in the creative consensus of team mates. In Pictorial KeyGraph, photographs of real entities corresponding to incomprehensible items in given data are embedded to the visual result of KeyGraph applied to their business reports, by users with tacit and explicit expertise in the real world of business on the way of the communication for designing a product. In their communications, novel and practical scenarios of product behaviors were extracted, and 5 new patents have been applied based on one of 11 new design scenarios, i.e., “*Develop the marking system and marking Ink to draw marks near by defects such as Scratch, Chink, Pustule, Black dot, White dapple and foreign body after detecting them on the surface of film.*”

The CCD inspector developed and sold from this company is the current most well accepted by users (customers), even though the company had been suffering from slow pace of inventions.

Behind this success, we found evidences that the team members tend to combine established concepts via rare words in creative designing. Conceptually, such a mechanism of creativity has been considered in the literature [9], and has been applied to realizing creative communication environment [10, 11]. However, this paper still presents a new finding. That is, the visualization of KeyGraph should be bottom-up in 5.2: Show the islands first, until the user thinks over for creating new scenarios and then show the bridges which may aid in presenting new ideas about the future scenarios. This finding presents supportive evidence to the instructive studies in design communication, where good questions make the trigger to good designs [11]. The uncertain information as words in the red nodes has been said to be helpful to creative design (see [5]), but presents even a stronger hint when given at the better timing, i.e., when the designer is asking for a new hint for creating a design.

## References

- [1] Ohsawa, Y., Modeling the Process of Chance Discovery, Ohsawa, Y. and McBurney P., Eds, *Chance Discovery*, Springer Verlag pp.2-15 (2003)
- [2] Y. Ohsawa, and P. McBurney eds., *Chance discovery (Advanced information processing)*. Springer-Verlag, 2003.
- [3] Ohsawa, Y., Benson E. N., and Yachida, M., KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor, Proc. Advanced Digital Library Conference (IEEE ADL'98), pp.2-18 (1998)
- [4] Kushiro, N., and Ohsawa, A Scenario Acquisition Method with Multi-Dimensional Hearing and Hierarchical Accommodation Process, *New Mathematics and Natural Computation*, in Vol.1, No.4, 2006.

- [5] Gaver W.W., et al, Ambiguity as a Resource for Design, in Proceedings of Computer Human Interactions, 2003.
- [6] Ohsawa Y.: Chance Discovery for Making Decisions in Complex Real World, New Generation Computing, Vol.20, No.2, pp.143-163(2002)
- [7] Y. Ohsawa and Y. Nara, "Decision process modeling across Internet and real world by double helical model of chance discovery," New generation computing, vol. 21, 2003, pp. 109-121.
- [8] Okazaki, N. and Ohsawa, Y.: "Polaris: An Integrated Data Miner for Chance Discovery" In proceedings of The Third International Workshop on Chance Discovery and Its Management, Crete, Greece (2003)
- [9] Goldberg, David E. The Design of Innovation: Lessons from and for Competent Genetic Algorithms Boston, MA: Kluwer Academic Publishers, 2002.
- [10] Eris, O., Effective Inquiry for Innovative Engineering Design, Kluwer Academic Publisher, 2004.
- [11] Fruchter, R. et al, Knowledge reuse through chance discovery from an enterprise design-build enterprise data store, New Mathematics and Natural Computation Vol.3 pp.393-406, 2005.

# Creative Design by Bipartite KeyGraph Based Interactive Evolutionary Computation

Chao-Fu Hong<sup>1</sup>, Hsiao-Fang Yang<sup>2</sup>, Mu-Hua Lin<sup>2</sup>, and Geng-Sian Lin<sup>3</sup>

<sup>1</sup>Department of Information Management  
Aletheia University, Taipei County, 251, Taiwan  
cfhong@email.au.edu.tw

<sup>2</sup>Department of Management Information Systems  
National Chengchi University Taipei, 116, Taiwan  
jimmy52@ms35.hinet.net, fa925710@email.au.edu.tw

<sup>3</sup>Graduate School of Management Sciences,  
Altheia University, Taipei County, 251, Taiwan  
aa912219@email.au.edu.tw

**Abstract.** Kotler and Trias De Bes (2003) at Lateral Marketing said that in customer's designing process the creativity as a kind of lateral transmitting. It meant that the customer would be stimulated by new need to discover a new concept, which could be merged into his design. Therefore, in designing process how to help the designer quickly designed a creative product that was the important problem. The other was that in interactive creative design the designer had to face the fatigue problem. In this paper, we developed a Bipartite KeyGraph based Interactive Evolutionary Computation (BiKIEC), which could collect the interactive data and ran the KeyGraph analysis to find the key components (chance1). And then the bipartite analysis was used to discover the chance2. Finally, the chance3 was the probability for entering the shot-cut in Small World. The BiKIEC emerged some creative components for helping designer designed the creative product. After analyzing the designer interactive data, we found that the product was created by chance mechanism, which was quickly accepted by designer in his design process. Furthermore, the questionnaire results also indicated that the BiKIEC could significantly help the designer to design his favorite product. Therefore, the BiKIEC was a useful tool for helping the designer discovered his creative chance in interactive design process.

## 1 Introduction

The enterprise always adopted the different strategy, one strategy of the Porter (1980) value chain, to earn the maximum profit. It had to correctly understand what the customer need was and then include his need into the product design. Therefore, before the designer entering the design process, the analyst had to correctly segment the market for capturing the target customer and it was very important for enterprise. Although, the strategy was used to segment the market and capture the niche market for narrowing down the market size and increasing the classifying precision. In practical, when the market size was too small it would hardly earn enough profit for enterprise. Kotler and Trias De Bes at 2003 announced the lateral marketing: using the

multi-values reconfirmed the customer need to create the new market as the new service or business for enterprise enlarging the market size to increase the profit. Unfortunately, the Kotler and Trias De Bes only proposed a conceptual model, but they did not develop the operating system. Here we tried to include the laterally transmitting and vertical transmitting into the dynamic interactive evolutionary computation (IEC) to develop a creatively operating system.

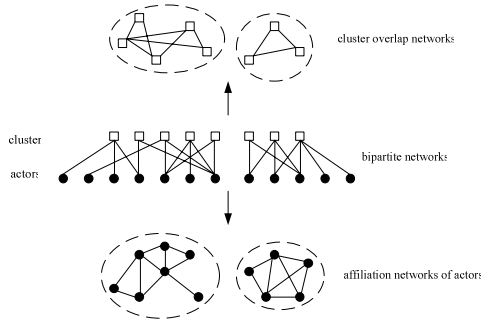
The IEC had been help the police to recognize the criminal face (Caldwell and Johnston, 1991), besides, it broke the limitation of EC which needed the fitness function to run the evolutionary computation. Because the IEC only depended on the designer's value to select the better chromosomes, in order to propagate its genes to the offspring. Therefore, in IEC the interactive process was not only supplied the stimulating information for designer to drive him discovering what did his need, but also supplied the choosing and recombining power for designer to make creative product. Consequently the creative designing problem became how to help the designer emerging a new creativity or recombining the creative concept into the making product process. The other is the fatigue problem, in IEC if the chromosome have  $j$  genes and each gene have  $k$  alleles, there will have  $k^j$  assembling patterns and its population is about 6-12, it means the probability for designer find the best pattern is  $\frac{1}{k^j}$ ; the

interactive process becomes a heavy load for designer. Find a good solution in the limited times is an important problem for IEC (Nishino, Takagi, Cho and Utsumiya, 2001) (Ohsaki and Ingu, 1998) (Takagi, 2001).

Fortunately, Ohsawa and McBurney (2003) proposed the Chance Discovery: Firstly, analyzed all contexts by terms presenting frequency and terms link to build the KeyGraph. Secondly, according to terms frequency and link's frequency discovered the important weak keyterms as the chance to predict the future. This process may help the IEC to discover the chance and to filter the good surviving components to assemble the creative product. But it still has a problem: the IEC's population size is too small and the evolutionary times must limit in few generations. These constrains make the IEC hardly collecting enough data to discover the chance.

Here we tried to adopt the concept of bipartite network to overcome the small data problem. In social relationship, the people connected each other as connecting by friend network and the company organizational network to format the complex social relationship became a kind of affiliation network, as in Fig.1. The relationships between the actors and movies were presented (Watts, 2003), some actors had played a movie and an actor had played some movies, too. This network was called the bipartite network. In addition, according to the bipartite network the affiliation networks of actors and cluster overlap networks were able to format and every actor had his living network (affiliation network of actors). An actor could start from his affiliation network and depended on the bipartite network to discover his new cluster overlap network, and then arrive to the new affiliation network as the chance for him to start his new life. Now the creative problem was as how to build the meaningful bipartite network.

As above discussion, some important mechanisms were found: one was the choosing mechanism which was not only according to KeyGraph method discovered the



**Fig. 1.** Affiliation networks (Watts, 2003)

important keyterms and weak keyterms to limit the space of the affiliation network, but also started from the affiliation network to run the lateral transmitting mechanism by bipartite network. And then the recombination mechanism based on the above process assembled the creative product as the shot-cut in Small World for designing the creative product.

## 2 Methodology of BiKIEC Model

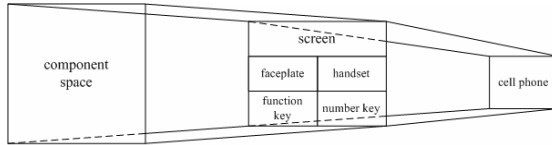
As above discussion, in interactive design how to create the meaningful bipartite network became an important problem. In this paper, the brainstorm and value-focused thinking (VFT) was adopted to define the bipartite network and format the BiKIEC model as in following section.

### 2.1 Value Focus Thinking and Designing Space

Keeney (1992) announced VFT method to solve the decision problem. First, the brainstorming meeting was used to collect the customer’s data for building the complete problem space and defining the objective space. Second, the analyst according to the problem space induced the customers’ value networks and created whole context space. In this paper our researching problem was what cell phone was the customer wants to buy. We celebrated three brainstorm meetings and every brainstorm meeting has about twenty customers attended. In the meetings, the customers’ value network and their objectives were collected and used to descript the content space. Finally, according to these results the cell phone’s elements was designed and its chromosome was shown in table 1, too. In brainstorm meeting most of the customers very cared about the faceplate of cell phone and in cell phone’s marketing many types of faceplate were presented, too. These phenomenons told us the faceplate was a very important factor in cell phone market. Therefore, 16 bit genetic length was used to present the faceplate of cell phone. All designing space of cell phone was  $(2^6 \times 2^2 \times 2^2 \times 2^2 \times 2^2)$  16384.

**Table 1.** The cell phone's chromosome

items	faceplate	handset	screen	function key	number key
size	1...64	1...4	1...4	1...4	1...4



**Fig. 2.** The cell phone's VFT

### 2.2 Lateral Transmission and Chance

After the brainstorm meeting the large designing space was defined, but in large content space it was not suite for the customer directly chose his favorite cell phone or using the IGA to design the creative cell phone. In this section we proposed a new method to break through this bottleneck. First, our system only supplied six products in every generation for customer. The customer depended on his need to estimate every cell phone and picked up some his favorite components as following operative graph.



**Fig. 3.** The interface of BiKIEC

Second, two kinds of the data were collected from every cell phones' population: one was all cell phones' score, what was given by designer and the other data was from the check boxes' data that was the customer discovering chance for making the creative product.

In KeyGraph method, the elements' frequencies were used to find the important key components and important weak key components in every generation. But in IGA, every generation only six products were supplied for customer to calculate the weight of elements. Here, the Eq.1 was used to calculate every component's weight and sorted these elements by weight. If the component's weight was under the average of weight, it would be removed.

Bipartite KeyGraph (BiK model)

In IGA, the new products (offspring) were generated by the genetic operation and elite product for customer to choose his favorite product in every generation. But in bipartite KeyGraph the components' weight were defined by the components' score divided it appearing times. The weight was higher than average weight, which was calculated by Eq.1, and  $a_c$  was called the high weight term.

$$a_c = \frac{\sum_{i=1}^p C_c}{T_c} \quad (1)$$

The score  $C_c$  was given by the designer for  $c$  component,  $T_c$  was the  $c$  component appearing time, and  $p$  was the population size.

Besides, the product's component was as a terms, the product was as a sentence, and all sentences (products) in one generation as an article. The terms' link was counted in every sentences, and all articles' sentences' links were calculated by Eq.2 to discover which terms was supported by other terms, and if the term was belonged to the high weight cluster (attribute) it would be the important keyterms ( $key(a)$ ).

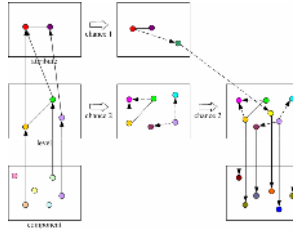
$$\begin{aligned} Assoc(a_i, a_j) &= \sum_{s \in D} \min(|a_i|_s, |a_j|_s) \\ key(a) &= 1 - \prod_{g \in G} \left[ 1 - \frac{base(a, g)}{neighbors(g)} \right] \\ base(a, g) &= \sum_{s \in D} |a|_s |g - c|_s \\ neighbors &= \sum_{s \in D} \sum_{a \in s} |a|_s |g - a|_s \\ \text{if } a \in g, \quad &|g - a|_s = 0 \\ \text{if } a \notin g, \quad &|g - a|_s = |g|_s \end{aligned} \quad (2)$$

The  $s$  was the sentence, the  $g$  was the cluster,  $D$  was the article and  $G$  was the all clusters. Here our model had one constrain, in the same cluster the components link's frequency was not counted.

As above discussion, the VFT model had been constructed by all fundamental objectives and the KeyGraph (Eq.2) was used to analyze the interactive data to find the important key components and high weight components. And then these components were formatted many clusters. Of course, if the weak weight component could connect with many clusters, there would be called the important weak component. And these important weak components were called the chance, too. In BiKIEC model, the bipartite network could emerge high potential cluster (chance1) and the KeyGraph emerged high potential components (chance2); there were shown in Fig.4.

Here we proposed two kind of merging chance methods: one was the important key components and high weight components to decide the good cluster (attributes), and in good cluster (attributes) the never presented components were the chance2 in Fig.4. At KeyGraph level: on the red cluster (attribute) only one component was found and on blue cluster (attribute) could find two components, there belonged to the short distance lateral transmitting. These two kind components belonged to the  $a$  asset.





**Fig. 4.** The diagram of bipartite KeyGraph process

### 2.3 Another Long Lateral Transmission the Chance3(Small World)

Our model was not like the traditional IGA, it had choosing mechanism. The customer could pick up his preferable components by check box from 6 products. It meant this operation helped the customer could acquire a new idea to design a creative product. This chance was like the concept of shot-cut in Small World but did not like the merging chance as chance1 and chance2. Therefore, the Eq.3 was used to discover the creative direction by himself to change the designing direction.

$$b_c = \frac{\sum_{i=1}^p p_c}{T_p} \quad (3)$$

The  $b$  was the probability for entering the shot-cut in Small World. The  $p_c$  was the times of picked up  $c$  component and the  $T_p$  was the times of all components were picked up.

### 2.4 The Recombination Mechanism

In the choosing mechanism, the designer could depend on three kind of chance mechanisms to get the asset  $a$  and  $b$ . And then the  $a$  connected with  $b$  to make a creative product  $ab$ , this value would be higher than the any  $a$  or any  $b$ .

$$CBIEC = a_t + b_t \quad (4)$$

The BiKIEC model in designing the creative cell phone was as following: according to Eq.2 found the key components  $a_{t-1}$  ( $t$  is generation), and then the bipartite KeyGraph was used to extend the  $a_{t-1}$  to  $a_t$  and discovered the chance as shown in Fig.5. The lateral transmitting (chance2) was extending in same cluster, but the long lateral transmitting (chance1) could extend to other cluster. The  $b_t$  was decided by picked up components (Eq.3). The recombining process could assemble the creative products ( $b_t + a_t$ ) and there would be supplied in next generation for customer.

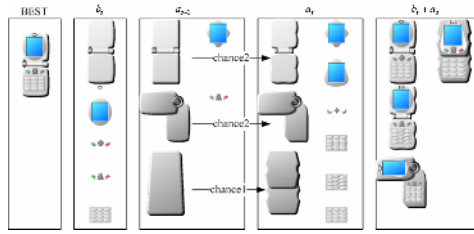


Fig. 5. According to the Chance process to assemble the creative product

### 3 Experiment Results and Discussion

The purpose of this experiment was that recombined IEC with chance to induce the customer found his favorite product. In addition, before developing this system we had surveyed the cell phone’s market and celebrated three brainstorm meetings to discover the value network. The cell phone was defined by five parts as the faceplate, handset, screen, function key and number key.

Therefore in our experiment the faceplate had 4 types and each type had 4 level, the handset had 4 level, the function’s key had 4 level and the number’s key had 4 level. And the variables of IGA are shown on Table2.

Table 2. Parameter’s design

Items	Context
Coding	Binary
Population size	6
Crossover rate	0.8
Mutation rate	0.01

We had celebrated two test to evidence the BiKIEC in creative ability was better than the IGA. The first test we invited the professors and college’s students, the secondary test we invited the high school’s teachers. These schools are all in north Taiwan. We had collected 40 samples and the recovery rate was 77.5% (valid data divided all samples).

#### 3.1 Used a Case to Discuss the Performance of Chance Process

Here we used the evolutionary data and the graph method to building the evolutionary KeyGraph. On the first generation all chromosomes were generated by random, and the Eq.1 was used to calculate components’ weight to decide the important key components and the components’ link to draw the KeyGraph. Even though, in every generation only 6 data were collected, the important key components and the components’ links were discovered and shown in Fig.6a.

The system analyzed the components’ link for finding an important weak key component 70 as the chance for customer. Furthermore, the important key clusters and important weak clusters could be extended by bipartite network for finding the potential cluster, and then assigned a component as the chance2. Then the customer was stimulated by these new creative products and continued to estimate the new products again. In Fig.7a, the graph of secondary generation was drawn by the

evolutionary data for us to check the component, what were suggested by all chance processes and to check these components how efficiency induced the customer to discover his favorite product. In Fig.9a the chance2 had suggested 14, 79, 89, and 67 to assemble the new product for customer to investigate what components would be accepted. The chance1 had suggested a new cluster for customer, too. The experimental results indicated that the components were selected by chance1, which were not accepted by customer. In Fig. 10a the chance1 (component 70) was still not observed that it became an important key component. But in this generation the chance2 suggested components 69 and 78 were observed that there became the important key components and there also satisfied the customer's need (value network). But suggested by chance1 component still had no chance accepted by customer. Finally, the customer had made his favorite product and stopped the designing process.

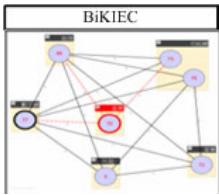


Fig. 6a. 1's generation

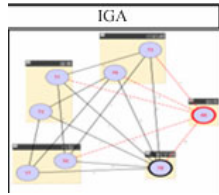


Fig. 6b. 1's generation

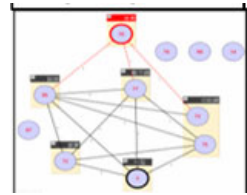


Fig. 7a. 2's generation

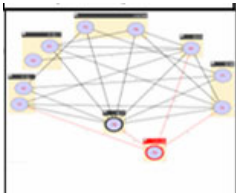


Fig. 7b. 2's generation

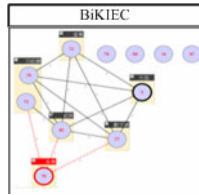


Fig. 8a. 3's generation

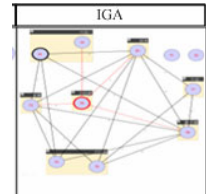


Fig. 8b. 3's generation

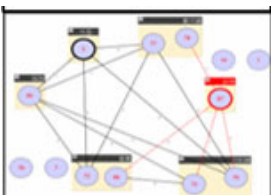


Fig. 9a. 4's generation

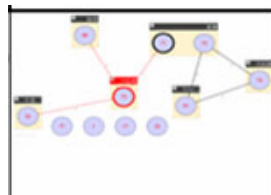


Fig. 9b. 4's generation

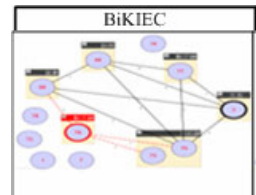


Fig. 10a. 5's generation

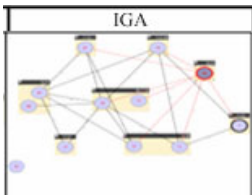


Fig. 10b. 5's generation

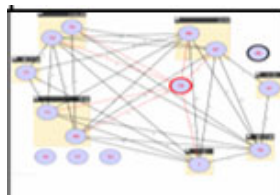
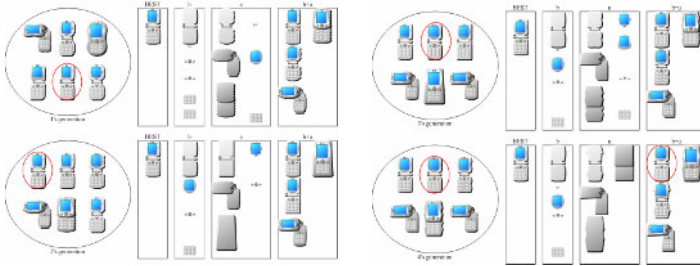


Fig. 11. 9's generation

But the IGA’s data was presented in Fig. 6b to Fig.11, the path of evolution was not stable. After 9<sup>th</sup> generation the customer still could not find what was he wanted cell phone and he had faced the fatigue problem. He stopped the designing process.

From this experimental result some interesting things were found: first (chance1), the KeyGraph was the useful method for extracting the important key components and important weak key term, in which these components could help the IGA focused on the important space to find the best component. Second (chance2), the bipartite network help the IGA filtered out the invalid space of the component to increase the sampling probability of the weak important components. This sampling mechanism could overcome the random sampling in IGA to increase the evolutionary efficiency. This sampling mechanism agreed the long distance clusters’ components could link each other as the component 74 and 75 in function key space. Third, the chance3 as component 78 suddenly appeared on 4<sup>th</sup> generation to help the IGA entering the better cluster (evolutionary path). This phenomenon was as the shot-cut in Small World.

In this case the chance1 did not apply its power on evolutionary design, but chance2 and chance3 could quickly discover the customer’s favorite value and re-combined the creative product. In addition, the BiKIEC according to the KeyGraph, bipartite KeyGraph and chance3 emerged some creative product for customer. The customer could feel our model did not only keep multi-direction in evolutionary computation, but also could quickly discover his favorite components to create the product. He could enjoy the design work on IGA environment.



**Fig. 12.** Bipartite KeyGraph evolutionary process

Furthermore, our system supplied the selecting power and indirectly recombining power for customer to make his favorite product. Therefore, comparing the chance product *ab* with the last time best chromosome, the result also was present in table 4 and the chance *ab* was significantly better than the last time best chromosome (8.11 vs 6.64). This result evidenced the bipartite KeyGraph could discover what was the customer need to demonstrate that our model could jump to the favorite creative space as the customer wanted.

**Table 4.** T test between *ab* and *b*

Method	N	Mean	Sig.
<i>ab</i>	15	8.11	0.008<0.05
the best chromosome	12	6.64	

### 3.2 Investigate the Chance in Designer's Mind

In this section we wanted to directly evidence one thing: the customer could feel that the Chance mechanism and recombination were the useful methods for assembling a creative product. The t-test was used to compare the speed of convergence of BiKIEC and IGA, there were not significant difference, but the BiKIEC was still faster than IGA.

**Table 5.** T test of evolution times

		N	Mean	Sig. (2-tailed)
Evolutionary times	IGA	16	10.00	0.308
	BiKIEC	15	6.67	

The table 6 was the questionnaire result. The chance3 operation could help him choosing a creative *b* and entering another cluster as the short-cut in Small World to design the favorite *ab*. The result was that in the 95% confidence about 79.7% designer agreed it.

**Table 6.** T test of question

Question	Test Value = 3.986
	Sig. (2-tailed)
The chance3 operation could help him create the favorite cell phone	0.047

**Table 7.** T test of questionnaire

Question		Method	N	Mean	Sig. (2-tailed)
6	Cell phone is your need	IGA	16	3.69	0.088
		BiKIEC	15	4.07	
7	It is a creative cell phone	IGA	16	3.56	0.127
		BiKIEC	15	4.13	
9	Do you like	IGA	16	3.56	0.090
		BiKIEC	15	4.00	
11	It worth to buy	IGA	16	3.88	0.841
		BiKIEC	15	3.93	
14	Easily operate	IGA	16	3.63	0.112
		BiKIEC	15	4.00	
16	Next generation your feasible cell phone will be disappeared	IGA	16	2.50	0.061
		BiKIEC	15	1.93	
17	System know what you need	IGA	16	3.81	0.242
		BiKIEC	15	4.13	
18	The cell phone usually is not your want	IGA	16	2.63	0.103
		BiKIEC	15	2.07	

The IGA did not have the chance process or did not supply the recombining power for customer to generate the favorite new product for him. But the BiKIEC had three kind chance process to search the favorite components and supplied the indirect

power for customer to recombine the  $a$  with  $b$  as the creative  $ab$ . Therefore at question 16 we could evidence the chance1 and chance2 could help the customer found what his favorite cell phone was, and it was supported. The question 17 and 18 asked customer that the  $ab$  was as same as his favorite design in his mind and the results were almost supported. The above analysis results could evidence that the bipartite KeyGraph and chance processes were the useful method to help the customer designing his favorite product. The question 6 and 7 asked the customer: is this a creative product. All were supported. Finally, the question 9 and 11 were used to investigate the customer wanted to purchase this product. The customer had high intention to buy it. These analysis results supported the bipartite KeyGraph and chance process were useful method to help the customer to accomplish his design.

## 4 Conclusion

In this study we wanted to develop a good creative designing system and broke through the IGA's fatigue problem, therefore, we defined three kinds of lateral transmitting (chance) for customer to assemble his creative product in BiKIEC model. In case study, it was not only according to the interactive data for calculating the component's weight and its associable relation to build the KeyGraph, but these data were also used to build the bipartite KeyGraph for discovering the creative chance and emerging the chance in the design process. Beside, the results of questionnaire also supported that the bipartite KeyGraph and chance mechanism were the useful method in customer's design.

## References

1. Caldwell, C., and Johnston, V. S. (1991), "Tracking a criminal suspect through 'face-space' with a genetic algorithm." Proceedings of the fourth international conference on genetic algorithms, San Francisco, CA:Morgan Kauffman Publishers Inc, pp. 416-421.
2. Keeney, R. L. (1992), Value focused thinking: A path to creative decision making. Cambridge, MA: Harvard University Press.
3. Kotler, P., and Trias De Bes, F. (2003), Lateral marketing: New techniques for finding breakthrough ideas. John Wiley & Sons Inc.
4. Nishino, H., Takagi, H., Cho, S., and Utsumiya, K. (2001), "A 3d modeling system for creative design." The 15<sup>th</sup> international conference on information networking, Beppu, Japan, IEEE Press, pp. 479-486.
5. Ohsaki, M., Takagi, H., and Ingu, T. (1998), "Methods to reduce the human burden of interactive evolutionary computation." Asia fuzzy system symposium, Masan, Korea, IEEE Press, pp. 495-500.
6. Ohsawa, Y., and McBurney, P. (2003), Chance Discovery, New York: Springer Verlag.
7. Porter, M. (1980), Competitive Strategy, The Free Press, New York, 1980.
8. Takagi, H. (2001), "Interactive evolutionary computation: fusion of the capabilities of ec optimization and human evaluation." Proceeding of the IEEE, IEEE Press, pp. 1275-1296.
9. Watts, D. J. (2003), Six degrees: The science of a connected age, London:W.W.Norton & Company Ltd.

# Discovering Chances for Organizational Training Strategies from Intra-group Relations

Meera Patel and Ruediger Oehlmann

Kingston University,  
Faculty of Computing, Information Systems and Mathematics,  
Penrhyn Road, Kingston upon Thames, KT1 2EE, UK  
R.Oehlmann@Kingston.ac.uk

**Abstract.** Chance discovery has been described as a process of identifying opportunities or risks for future decision making that are referred to as chances. We argue that in the area of organizational training such chances are widely ignored, because either standard off-the-shelf training strategies or ad-hoc strategies are used. In contrast, this paper describes a case-study that uses a two-stage approach, which is tailored to the specific training situation. The case study involves the development of a training scheme to improve customer relationship management (CRM). In the first stage a study is conducted that analyzes relevant groups of the organization based on diagrammatic self-descriptions and descriptions of intra-group relations. In the second stage, chances are identified from the diagrams.

## 1 Introduction

The design of organizational training schemes is a difficult area. They are important, because if they are successful, they may contribute to an increase in the organization's revenue. But they are also very costly. Therefore, if poorly designed, they may have the opposite effect. A contributing blockage in achieving the intended goals may be seen in staff's lack of acceptance and cooperation. A reason for this is that the majority of organizational training schemes are using off-the-shelf approaches that cannot appreciate the internal situation of a particular organization or workgroup. Furthermore, do not address the problems in an individual workgroup.

In the approach described below, we have utilized the concepts of chance discovery to design a training scheme that is tailored to the needs of a particular organization. However, the methodology is generally applicable. Chance Discovery aims at identifying and managing rare, but significant events, such as potential risks or opportunities, in some domain or application. These events are termed *chances*. [4]. Therefore in the context of organizational training, the objective is to identify chances in an organization for the development of successful training strategies.

Our method begins with an investigation of the intra-group relations for each group involved and the inter-group relations as well. In the course of the investigation, each group member makes his or her views of the relations to other group members explicit by producing diagrams that characterize these relationships. The diagrams are based on the Social Diagrammatic Language (SDL) that provides graphic elements

for describing various aspects of an inter-personal relation [3]. All the produced diagrams are then analyzed with respect to the training issue at hand. Weak and strong characteristics of the group situation are considered as potential chances for new training strategies. A team of experts assesses these potential chances and considers two areas. The first area involves behavioral changes in the group as a whole and also of single group members. The underlying assumption is that weaknesses in an organization also can be related to behavioral problems and wrong attitudes. The second area involves factual knowledge of the domain in which training is intended to take place. Often the diagrams indicate lack of knowledge or skills in particular domain aspects. Discussions within the expert team and considerations of both areas then lead to an agreed set of training strategies.

This method has been evaluated in a case-study of a high-tech golf driving range. This is a leisure facility for practicing golf techniques in a way similar to ten-pin-bowling. The domain for which the company was seeking training support was customer relationship management (CRM). Due to space restrictions, this paper cannot describe the entire case study, which involved 14 members of staff and some 100 diagrams<sup>1</sup>. Therefore, the next section will describe the investigation method but will reduce the results to a few examples. Section 3 will then use these examples and relate them to group work and CRM. Based on this analysis, Section 4 will outline the main features of the proposed training scheme and the final section will discuss the approach.

## **2 An Investigation on Intra-group Relations**

### **2.1 Method**

The study involved two groups. The customer service group had 10 members and the management group had 4 members. Below we will report some of the results from the customer service group. The 10 members of the customer service group had an age range between 18 and 41 years (means 29.5 years). Each group member received an instruction document that described all the graphic elements of SDL. In addition, the document contained a detailed example of a relationship that was described in SDL. In a first phase the group members were trained in producing SDL diagrams. This training included demonstrations and exercises. In a second phase, the group members had to generate their own SDL graphics, which described their views of their relationships with other group members. All diagrams were produced at the same day and analyzed by an expert team that consisted of a psychologist, an educationalist and a customer relationship expert.

### **2.2 Results**

The results were derived from the SDL diagrams, which the subjects generated. An example of a single diagram is given in Fig 1. Here is expressed that Subject RI considers his relationship with Subject CA positively as professional. She thinks that the negative aspects of the relationship are that CA is unhelpful and overfriendly. RI

---

<sup>1</sup> A detailed technical report with the complete results can be obtained from the authors.



considers herself as socially highly accepted and believes that on a joy-distress scale from 1 to 5, the relationship is positioned about in the middle. While RI used elements of self-perception and emotions, she did not use elements to describe conflict resolutions, although such elements are also supported in SDL. However other subjects have done so. The results could be summarized by listing opposite feature pairs, such as helpful vs. unhelpful. The feature pairs in turn could be divided into three categories. Category I contained feature pairs, where both components actually appeared in the graphics. Category II contained those pairs where only one component appeared in the graphic and the opposite was never mentioned. For instance, only the component bossy of the pair bossy vs. obedient appeared in the graphics. Category III contained pairs of features which did not appear at all in the graphics. However these pairs were expected by the experts based on their knowledge of customer relationship management. It was particularly Category III that gave rise to the introduction of factual knowledge into the training scheme.

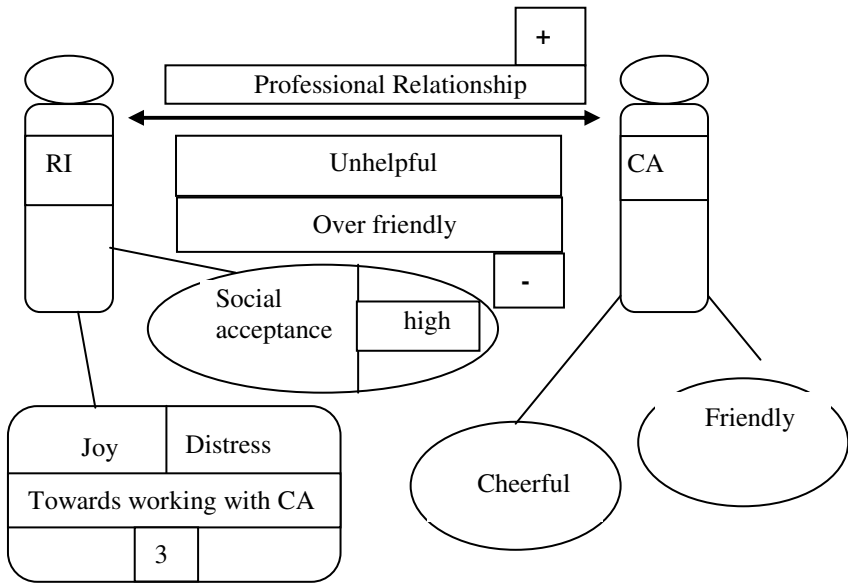


Fig. 1. SDL Example

### 3 Chances for Training Strategies in Group Work and CRM

The results indicated that some relations were considered as *helpful* and others as *unhelpful*. They also showed that some relations were qualified as *bossy* but that no relationship was characterized by *obedience*. These results already suggest that there are problems with interactions within the group and not just problems with customer interactions. However, these problems will also affect relations between staff and customers. For instance, it is unlikely that a member of staff, who is unhelpful towards his or her colleagues, will be helpful to customers.

Therefore these results represent potential chances for a type of training that is suitable to change intra-group behavior. This can be achieved with a non-interventional approach [1]. Typically in non-interventional group meetings, initially the group leader does not direct or guide the group. Therefore *facilitator* is a more appropriate term for this function [7]. In a non-interventional setting the three main phases of *dependency*, *counter-dependency* and *interdependency* can be distinguished. All these phases are necessary to achieve a mature group that is capable of self-disclosure. This in turn is necessary to accomplish behavioral changes.

In the *dependency* phase, the facilitator sits back and allows the group members to find their own direction. This reveals the instinctive unconscious reactions of group members much more clearly than in a more structured approach. It also will allow the group to begin the difficult process of taking responsibility for itself. The *counter-dependency* phase is characterized by the airing of differences and the open acknowledgement of conflict. The open aggression, which sometimes occurs, may be directed towards the facilitator or towards each other. This phase is important because it helps the group members to be open about their less positive reactions to one another. This openness is necessary to achieve the kind of meaningful, honest communication that groups, when they function well, should provide. It is also a requirement to enter the third phase of *interdependency*. In this phase, the group begins to have confidence in its own ability to do things without the facilitator's 'permission' or instruction. They also take responsibilities for each other.

Even in group oriented societies and the more in western societies, it is typical that people shy away from sharing very personal experiences with a group of strangers and these are all people who do not belong to one's family. However, groups who share the same development and history of the three phases outlined above are willing to offer such self-disclosure and respond appropriately [8]. As the group evolves to such a level of maturity, it is likely to be the differences rather than the similarities, which lead to learning and behavioral change. The entire process can be made more concrete and focused by utilizing the conceptual pairs that have been derived from the social diagrams. For instance discussion of *helpful vs. unhelpful* or of *bossiness*, may lift important issues into consciousness and may lead to a change in behavior. In fact all the results from the group model need to be addressed in the group sessions.

Although a number of group members were characterized as "good with customers", the results showed that many dimensions of customer relationship management, which experts consider as relevant, did not at all appear in the diagrams. This in particular applies to many characteristics of good communication with customers. For instance, the results showed that pairs such as *consistent vs. inconsistent* communication or *attentive vs. defensive* behavior towards customers were not a concern of the customer service group.<sup>2</sup> Such deficits can be considered as potential chances for new training strategies. However they cannot simply be corrected by non-interventional group sessions, because in order for the group to discuss these issues even in a controversial fashion, it first needs to become aware of them. Therefore in addition to the non-interventional style, a second training style is required, that presents factual knowledge in a very interventional way.

The next section will describe how this contradiction can be resolved in an integrated training scheme.

---

<sup>2</sup> These requirements have been discussed from a CRM perspective in [2].

## 4 A Structure for an Organizational Training Scheme

The proposed training scheme integrates non-interventional and interventional training strategies. These are organized in three phases.

**Initial Phase.** An initial phase is purely non-interventional. At the beginning, interventions and the presentation of factual knowledge make not much sense, because the willingness among group members to accept interventions is very low. In this phase, the focus is on chances that are derived from the intra-group analysis. For instance, the considerable difference in the group members' willingness to help one another needs to be addressed. Furthermore, a consensus needs to be achieved that helping one another is a basic requirement for working successfully as a group. For this purpose, the group needs to address the question of what prohibits them to lend support to one another.

**Intermediate Phase.** The degree to which the group matures determines how much factual knowledge can be introduced. Typically in this later stage, presentations of factual knowledge and discussions will alternate. In particular, the knowledge that in the group analysis has been identified as missing should be introduced in this phase. At the end of this second stage, the group should be able to relate the external requirements of CRM to the achieved consensus about how the group members work as a group and they should recognize that their collaboration as a group forms the basis from which they can address the CRM requirements.

**Final Phase.** In the third phase, additional strategies such as role play, brainstorming or planned external experiential situations can be introduced. The selection of these strategies should again be based on chances that have been identified in the group analysis. For instance, noticing that the group considers the task of customer relationship management only from their staff perspective, rather than from additional perspectives such as those of the customer or the management, may lead to the introduction of role play, where group members take the role of a manager or a customer. In such role play scenarios it is important to ensure that the group member who takes the role also has a strong motivation to succeed in that role. The analysis has also shown that group members saw themselves in a position of executing directives of the management without being able to identify themselves with such orders. Such a view can in part be addressed by brainstorming sessions with the objective to improve the CRM framework that is currently in place [6]. This type of session would allow the group members to feel to be involved in the overall operations, which could lead to a higher degree of identification with the organization. On the other hand, management may gain additional creative input. Experiential learning situations can be introduced for instance by letting the group compete against another group or giving the group a task that can only be accomplished, if all the group members work together.

## 5 Conclusion

We began with a criticism of current off-the-shelf approaches in organizational training, in that such approaches do not consider the particular internal situation of a given

organization. Clearly the method described in this paper does not have this limitation. Although cast in a generally applicable framework, it focuses on the particular social relations in an organization. It is the identification of strength and weaknesses in these relations that provide chances for changes in the organizational behavior.

From the non-interventional group sessions, the approach moves to a mix of non-intervention and intervention. With the increasing maturity of the group, the non-interventional components are more and more replaced by interventional strategies. However, it is important to note that it is not sufficient to have a non-interventional group session and then to move on to traditional training forms. The non-interventional phases need to be closely integrated with interventional phases. In fact, it may be necessary to return to purely non-interventional sessions, if a presentation of new material indicates problems within the group.

The same applies to the final phase, although it can be expected that the group is now sufficiently mature to benefit from the strategies of this phase. These strategies require group members to work together and to take new perspectives they may not have thought of before. If there are still any unresolved relationship issues lingering, these strategies will fail. This is the reason why these strategies are only considered in the final phase. Their purpose is to reinforce and strengthen what has been already achieved. They are a tool to reach collaboration where there was no collaboration before.

The strategies in all three phases will only succeed if they address chances that have been identified in the current group situation. This is why the method is superior to previous standard approaches to organizational training.

Group-based chance discovery has been considered within the double-helix framework [5]. Here two strands of the helix represent the processing of object and subject data. Object data are obtained from the domain under investigation and subject data are obtained from the group discussions of the analysis of object data. Our approach can be matched with this framework, when we accept that the domain in question is the work group. Then the object data are the SDL diagrams and the subject data are derived from the discussions of the team that designs the training scheme. However currently these discussions have not formally been analyzed. Nevertheless, via the training scheme, they affect the generation of future object data in the group sessions.

Generally the group members commented very positively about the SDL-based method. For two reasons, they found it easier to describe their relations in terms of SDL diagrams than in the form of verbal protocols. First, SDL diagrams were produced only at certain points in time rather than in parallel to other activities. Second, they found the constraints that were imposed by the limited set of graphical elements, a useful guide as opposed to the free expression required in verbal protocols. Nevertheless the *comment* feature in the diagrams enabled subjects to express views beyond the diagram notation. Of course, a questionnaire would provide even more guidance, if closed questions are used. But closed questions would not allow the same freedom in structural expression. In addition, questionnaires involve the framing problem, in that answers may be given depending on the formulation in the question. This does not apply to the SDL approach. In comparison to verbal protocols, SDL provides structural information, which in verbal protocols is hidden. A disadvantage of SDL is that it is typically used at certain point in time and therefore only provides snapshots. However, this is not so relevant in the area of social relationships, because these

relations do not change so frequently. Moreover the SDL graphics enabled the group members to visualize their relationships. Therefore with respect to constrains on expressiveness, our approach can be positioned between verbal protocols and questionnaires. With respect to visualization, the approach is superior to both methods.

## References

1. Crago, H.: Couple, Family and Group Work: First Steps in Interpersonal Intervention, Open University Press, Maidenhead, UK, 2005.
2. Irons, K. The World of Superservice, Addison-Wesley, Reading, MA, 1997.
3. Oehlmann, R.: The Function of Harmony and Trust in Collaborative Chance Discovery, *New Mathematics and Natural Computation*, 2, No. 1 (2006), pp. 69-84.
4. Ohsawa, Y.: Modeling the Process of Chance Discovery, in Ohsawa, Y. and McBurney, P. (eds.), *Chance Discovery*, Springer, Berlin, 2003.
5. Ohsawa, Y., Nara, Y.: Decision Process Modeling across Internet and Real World by Double Helical Model of Chance Discovery, *New Generation Computing*, 21 (2003), pp. 109-121.
6. Rawlinson, J.: *Creative Thinking and Brainstorming*. Gower Publishing, Aldershot, UK, 1986.
7. Rogers, C.: *On Encounter Groups*, Harper & Row, New York.
8. Yalom, *The Theory and Practice of Group Psychotherapy*, 4<sup>th</sup> ed., Basic Books, New York, 1995.

# **Trust, Ethics and Social Capital on the Internet: An Empirical Study Between Japan, USA and Singapore**

Yumiko Nara

The University of the Air  
narayumi@u-air.ac.jp

**Abstract.** Social capital is becoming increasingly important in the knowledge society. Most studies of the phenomena that are considered as social capital have, in the tradition of Robert Putnam's writings mainly focused on what can be called social capital in civil society in the real world - outside the virtual world, i.e. the Internet community. This paper tries to clarify the significance of trust as the element of social capital that relates ethics of the Internet community to theoretical and empirical approaches. Results indicate that studies on trust related to the Internet can be positioned in two dimensions – “studies on the system trust (especially, trust in the system infrastructure) – studies on the personality trust” and “studies with risk management approach – studies with trust management approach.” Generalized trust has significant relationships with reciprocity, human interaction and cooperation on the Internet. However ethics does not simply correlate with trust.

## **1 Introduction**

Trust plays an extremely important role as a lubricant of social relationships including politics and economic activities. This has been consistently argued by a lot of researchers in various fields of the social science such as economics (Akerlof, 1970; Frank, 1988), sociology (Simmel, 1900; Barber, 1983; Coleman, 1988; Luhmann, 1973; Giddens, 1990), politics (Hardin, 1992; Putnam, 1993), psychology (Rotter, 1980; Yamagishi&Yamagishi, 1994), and anthropology (Gambetta, 1988). Especially since 1990s, the trust research has remarkably developed in various fields of social science, while social situations and relations have been unstable as fluidizing in politics and economy.

In this paper, the author is aiming at the following three points related to trust focusing on the social space -the Internet, i.e. the integrated system with the social relationships. 1) The peculiarity and the pattern of the former research related to trust and the Internet including social capital are considered. 2) The elements of trust in the web community should be examined with survey data focusing on personality trust. 3) The relationship between trust and ethics in the web community is examined in detail.

## **2 The Peculiarities and Patterns of the Approach to Trust and the Internet**

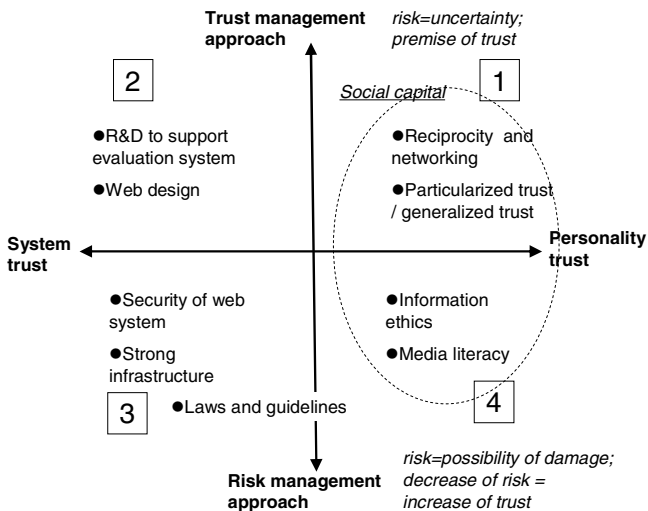
### **2.1 Concept of Trust**

Trust that has been mainly treated in the area of social sciences seems to have the following three peculiarities.

1. Trust is conceptualized on the premise that one's situation is in certain social uncertainty, i.e. it lacks certain information. It means trust is a concept connected to risk. In the situations that have no social uncertainty, there is no necessity to think whether the other party can be trustworthy or not. This idea distinguishes trust from "assurance" (= situation with no social uncertainty) (Yamagishi & Yoshikai, 2005).
2. Trust is divided into "personality trust" and "system trust." Both Luhmann and Giddens tried to grasp trust by two patterns of "personality trust (persoenliches Vertrauen)" and "system trust (Systemvertrauen)." Personality trust is the trust based on the observation and the experience to concrete "person". Society has been so complicated that it becomes impossible to cope with such complexity only by personality trust. Then, the system trust is needed. We trust the system that is institutionalized so that we cope with the complexity of society.
3. Trust is generated through the commitment to person(s) by person(s). People try to cope with the social uncertainty because trust contains both possibilities – the possibility to be betrayed and to gain profit when his/her trust pays off.

Since 1990's the Internet has been spreading rapidly, the research of the trust related to the Internet has been performed. Various studies are conducted not only in the fields of the social science but also in the standpoints of natural science like engineering.

Here the studies on trust related to the Internet are grouped according to the following two dimensions– "studies on system trust (especially, trust of system infrastructures) – studies on personality trust" and "studies with risk management approach – studies with trust management approach" (Figure 1). Common recognition among these four quadrants is that trust should be more important in the web community than in the daily life spaces, because on the Internet everyone can communicate with others easily with high levels of anonymity and invisibility. That means the Internet space has a higher uncertainty than the real world.



**Fig. 1.** Approach to Trust on the Internet

## **2.2 System Trust (Especially Trust in System Infrastructures) and Personality Trust: From the Viewpoint of Risk Management <The Third and Fourth Quadrants>**

As for the Internet, it is set up on the social telecommunication infrastructure. In this sense, an information environment such as the Internet doesn't consist without the trust in the system infrastructure. From this viewpoint a considerable number of studies have been performed since the approval of the Internet; they are mainly conducted in the field of engineering. The main concerns are security countermeasures of the telecommunication network, strong web systems against accidents, countermeasures against unlawful computer access, etc. The mission of this research is to improve the reliability of the telecommunications system as the infrastructure. In other word, they grasp the Internet with the possibility to cause damage, so decreasing such damage may bring security and increase of trust. It means this type of approach recognizes risk as the possibility of loss/damage, and aims to minimize damage (Okamoto, 2001; Kawashima, 2001; Sugino, 2003; Yamamoto, 1996 etc). There are other studies from the standpoints of institutions like laws (eg. copyright law), politics and guidelines with the mission to improve system trust of the Internet.

The research on personality trust in the Internet has been conducted in the area of social sciences, such as information ethics, sociology, social psychology and so on. Trust functions as a certain kind of system management mechanism. Trust is also a concept that includes ethical and moral elements. Standing on this feature, studies in the areas of information ethics and information education have been developed, there is recognition that each user is a subject with responsibilities to generate trust on the Internet, so that it is necessary to deal with others morally (even though she /he is not identified). There is an awareness of the issues that users should not be victim/assailant on the Internet. In this quadrant, trust is discussed relates to the necessity to decrease risk as the possibility of the damage. Moreover, the research from the aspect with "improving trust = reducing risk" has been performed in the field of sociology (Endo, 2004), which indicates that internet literacy should include the ability and technology to construct an appropriate relation and cooperation with others in the Internet space (though counterparts' faces cannot be seen).

## **2.3 System Trust and Personality Trust on the Internet: From the Viewpoint of Trust Management <The First and Second Quadrant>**

Trust and the risk have been addressed in the research described above with the intention to decrease damage on the Internet. From another viewpoint, i.e. intending to further develop the possibility of the Internet as a social device, various researchers have approached trust (system trust and personality trust) in the web community.

Technical system development to support trust formation in the web community has been realized in the engineering area. For example, systems that improve information reliability by addressing user reliability and information reliability, systems that calculate trust of information based on a "web of trust", and a method of trust evaluation with reputation systems have been developed (Nomura, 2005; Usui, 2002; Endo, 2003; Tomobe, 2005).

Some studies that treat trust on the Internet from the viewpoint of social capital have been published in the area of social science. Social capital is the concept that was



introduced in the 1970's. Sociology, politics, and economics, etc. have been mainly paying attention to it; especially the conceptualization by Coleman (1988) and the development by Putnam (1993) are well known. The meaning of social capital is that it makes the influence on society, economy, and politics to be "capital" which differs from other capitals with the aspect of relations between subjects, which cannot be seen. Trust is generally considered to be part of a larger concept of "social capital." Social capital has been defined to include trust, norms of reciprocity, and networks of civic engagement (Putnam, 1993).

Trust can be divided into particularized trust and generalized trust. Generalized trust is the perception that most people are part of your moral community (Uslaner, 2003). The difference between generalized and particularized trust is similar to the distinction that Putnam (2000) drew between "bonding" and "bridging" social capital. Yamagishi & Yamagishi (1994) formulated this distinction: generalized trust is trust in people who are different from yourself. Particularized trust is faith in people who are like oneself. Particularized trust can be formulated by her/his confidence that her/his counterparts are trustworthy though actual experiences of communication and interaction with them (the concerns and ascriptions are shared among them). On the other hand, generalized trust is formed without sharing any actual experiences.

The generalized trust becomes important in the Internet space. Mistrusters view dealing with strangers as taking big risks. Trusters see expanding their horizons as great chances. Generalized trust is useful in modern knowledge society, especially in the Internet space to interact with others whom she/he doesn't know, and to obtain an advantage mutually through the trust. On the other hand, particularized trust towards others with the same background has an advantage in the semi-closed communities with similar ascription of members. It is supposed that using the Internet with trust and reciprocity would achieve the formation of human networks and the sharing the resource. As for the use of the Internet, some studies have discussed and examined how the use of the Internet influences the formation of social capital (Uslaner, 2003; Miyata, 2004; Norris, 2003). Miyata (2004) tries to demonstrate that social activities in online communities increase social capital with keywords; trust and reciprocity. The study examined whether social capital offers benefits to individuals as well as the collective. The results suggest that the mobilization of social resources embedded in social networks in online communities increases psychological well-being and satisfaction in their decision-making. These results also show the possibility that activities in web space may lead to empowerment and social solution. Furthermore, research has referred to the possibilities of the information network, which generates mutual trust and the commitment from the standpoint of organization theory, emphasizing the importance of the trust management as a requirement for the paradigm shift from the web community to the knowledge community (Syozugawa, 1999; Atsujii, 2002).

### **3 The Structure of Personality Trust (Generalized Trust) on the Internet**

#### **3.1 Framework of Study**

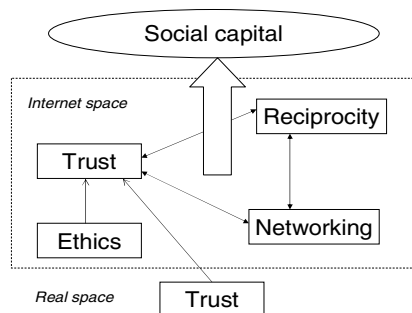
This section aims at the first and fourth quadrants of Fig. 1, i.e. the aspect of personality trust as a basis of the web community. A questionnaire-based social survey was

conducted in Japan, USA and Singapore; the main variables were generalized trust, reciprocity and human interaction in the web space, and information ethics and consciousness/behavior on the Internet.

The significance and possibility of trust focusing on the Internet as a new community has been clarified by using data of a questionnaire-based survey between Japan and the US (Nara, 2005); the findings suggest that people's positive attitude of taking information will influence their conditions of trust as a receiver (trust-examining of objective data, feeling of others' trustworthiness as the basis of subjective data) and as a sender (trust-generating of objective data, trust-generating of the basis for subjective data). Based on the former perspective, in this paper, the status of general trust should be examined, as well as the relationships with reciprocity, human interaction and cooperation in web communities from the viewpoint of social capital.

Furthermore, the relationship between trust and ethics is considered. Moralistic trust is a moral commandment to treat people as if they were trustworthy. The central idea behind moralistic trust is the belief that most people share our fundamental moral values (Fukuyama, 1995). Moralistic trust is based upon "some sort of belief in goodwill of the other" (Yamagishi & Yamagishi, 1994), and it is a value that rests on an optimistic view of the world and one's ability to control it. So it is supposed that the person with a high generalized trust has a concrete ethical and good consciousness and behavior on the Internet.

The analytical framework of the research is shown in Figure 2. Indexes of the main variables are shown below. Among these indexes, generalized trust is based on the pre-standardized scales, the others were made by the author specific for this study.



**Fig. 2.** Analytical framework of the research

### Generalized Trust on the Internet

Question: What do you think about the Internet?

1. Strongly agree    2. Agree    3. Somewhat agree    4. Somewhat disagree  
5. Disagree    6. Strongly disagree

- On the Internet, most people are basically honest.
- On the Internet, I tend to trust people.
- On the Internet, most people are basically innocent and kind.
- On the Internet, most people trust others.
- On the Internet, most people are trustworthy.

### Information Ethics Consciousness

Question: What do you think of the following conduct done on the Internet? Please choose the most suitable number for each item.

1. Very problematic      2. Problematic      3. Somewhat problematic  
4. Somewhat unproblematic      5. Unproblematic      6. Not problematic at all

- Circulating incorrect information knowingly
- Reading another person's e-mail without telling him/her
- Looking at pornographic web sites on a computer at work or at school
- Giving your password to another person
- Accessing a computer that you are not authorized to access
- Tampering with data on another person's computer through the Internet
- Sending a computer virus deliberately
- Circulating somebody's name and phone number to a large number of unidentified people without telling him/her
- Using pictures or texts on somebody's web site for your own web site without telling him/her (excluding cases in which permission has been obtained)
- Publicizing pornographic pictures or texts on your freely-accessible web site
- Speaking ill of others in a chat room or electronic bulletin board

### Information Ethics Behavior

Question: How often have you done the following things on the Internet? Please choose the most suitable number for each item.

1. very often      2. often      3. sometimes      4. rarely      5. never  
(Eleven items are same as those of "Information ethics consciousness.")

### Cooperation to Generate Trust in the Web Community

Question: Regarding your attitude in using the Internet, how much is the following description applicable to you? Choose the most suitable one.

- I try cooperating with others on the Internet, so that the Internet society functions smoothly.
- On the Internet, I actively seek to build cooperative relationships with others when they seem trustworthy.

### Reciprocity in the Internet Community

Question: Regarding your attitude in using the Internet, how much is the following description applicable to you? Choose the most suitable one.

- On the Internet, I think users are all in a give-and-take relationship.

### Human Interaction on the Internet

Question: How often do the following things happen when you interact with others on the Internet?

1. very often      2. often      3. sometimes  
4. rarely      5. never

- They tell you about themselves.
- They tell you about their personal trouble.
- You tell them about yourself.
- You tell them about your personal trouble.

### 3.2 Outline of Survey

The subjects of the survey were male and female Internet users, 20 to 39 years old from all parts of their countries. They were picked randomly from panels composed by survey facilities ([J] NOS list, [US] Greenfield Online list). Surveys were conducted in February and March of 2005, with slight differences in the survey period among two countries ([J] 2005 Feb7-28 and Mar11-12, [US] Feb9-20). For Japan, the questionnaire was sent and returned via ordinary mail; for the US, subjects logged on to a questionnaire website with a log-in-name and password. The sample size was 2412 (with 1175 usable samples) for Japan, 2461 (with 551 usable samples) for the US. Investigation implementation organizations were [J] Nippon Research Center (NRC) and [US] Taylor Nelson Sofres Intersearch (TNS).

Basic attributes of respondents are as follows; Gender: female 68.0% and male 32.0% in Japan, 50.6% and 49.4 % in US. Age: between 20-29 years old 47.0 % and between 30-39 years old 53.0% in Japan (average 30.37 years old), 48.1% and 51.9% in US (average 29.70 years old).

### 3.3 The Relationships Between Generalized Trust and Some Variables

In this section, the relationships among some variables, such as generalized trust in the web community, recognition of reciprocity in web, human interaction on the Internet as networking tendency and cooperation tendency, are examined.

Table 1 shows the frequency distribution of generalized trust on the Internet, and table 2 and 3 the result of the ANOVA among the three nations ([J], [US]and [S] in table mean Japan, USA and Singapore). The total score of generalized trust of each respondent, which is used in table 2 and 3 was calculated by adding the code numbers for the answer category for each question. So smaller the total score becomes, the higher the level of generalized trust becomes. These indicate that the level of

**Table 1.** Frequency distribution of generalized trust on the Internet

		(%)						
On the internet		1. Strongly agree	2. Agree	3. Somewhat agree	4. Somewhat disagree	5. Disagree	6. Strongly disagree	Total
Most people are basically honest	[J]	0.34	3.78	16.85	38.87	27.09	13.07	100.0
	[US]	1.27	7.08	35.75	25.59	17.06	13.25	100.0
	[S]	0.54	5.06	23.10	35.74	24.19	11.37	100.0
I tend to trust people	[J]	0.42	4.39	23.67	33.82	24.27	13.43	100.0
	[US]	1.63	4.36	25.77	29.58	22.87	15.79	100.0
	[S]	1.44	3.97	21.84	31.41	27.44	13.90	100.0
Most people are basically innocent and kind	[J]	0.34	1.98	24.44	36.92	23.58	12.74	100.0
	[US]	1.64	5.44	31.58	27.95	19.96	13.43	100.0
	[S]	1.08	2.53	19.31	32.13	29.61	15.34	100.0
Most people trust other	[J]	0.17	1.55	20.70	39.95	26.20	11.43	100.0
	[US]	1.63	7.08	35.93	28.32	17.42	9.62	100.0
	[S]	0.90	3.79	24.37	32.67	23.83	14.44	100.0
Most people are trustworthy	[J]	0.01	0.77	13.23	37.54	30.67	17.78	100.0
	[US]	2.00	4.17	30.31	30.31	19.42	13.79	100.0
	[S]	0.54	2.71	18.05	33.94	26.89	17.87	100.0

**Table 2.** Result of ANOVA (oneway) relation between generalized trust and nation

	sum of squares (SS)	d.f.	mean square (MS)	F
SS between	902.621	2	451.3105418	21.270***
SS within	47931.080	2259	21.21783093	
	48833.701	2261		

\*\*\* p&lt;.001

**Table 3.** Follow-up test (Tukey's HSD) relation between generalized trust and nation

	n	mean	S.D.	multivariate comparison
Japan	1157	21.406	4.1786	(USA)***
USA	551	19.884	5.149	(Japan)*** (Singapore)***
Singapore	554	21.215	4.878	(USA)***

\*\*\* p&lt;.001

**Table 4.** Correlation coefficient among some variables (Pearson's R)

Japan (N=1113~1163)	generalized trust on the Internet	reciprocity	interaction	cooperation	generalized trust in the real space	Information ethics consciousness	Information ethics behavior
generalized trust on the Internet	1.000	0.287***	0.157***	0.246***	0.477***	-0.002	0.003
reciprocity		1.000	0.158***	0.570***	0.178***	-0.055	0.036
interaction			1.000	0.276***	0.064*	-0.048	0.114***
cooperation				1.000	0.137***	0.057	0.060*
generalized trust in the real space					1.000	0.074*	-0.013
Information ethics consciousness						1.000	-0.183***
Information ethics behavior							1.000
USA (N=551)	generalized trust on the Internet	reciprocity	interaction	cooperation	generalized trust in the real space	Information ethics consciousness	Information ethics behavior
generalized trust on the Internet	1.000	0.223***	0.048*	0.210***	0.465***	-0.128**	0.166***
reciprocity		1.000	0.12**	0.653***	0.189***	0.072	0.045
interaction			1.000	0.274***	-0.016	-0.086*	0.317***
cooperation				1.000	0.200***	0.133**	0.007
generalized trust in the real space					1.000	0.037	-0.01
Information ethics consciousness						1.000	-0.224***
Information ethics behavior							1.000
Singapore (N=554)	generalized trust on the Internet	reciprocity	interaction	cooperation	generalized trust in the real space	Information ethics consciousness	Information ethics behavior
generalized trust on the Internet	1.000	0.390***	0.216***	0.352***	0.557***	-0.132**	0.189***
reciprocity		1.000	0.215***	0.630***	0.354***	-0.019	0.088*
interaction			1.000	0.206***	0.115**	-0.112**	0.298***
cooperation				1.000	0.259***	0.057	0.075
generalized trust in the real space					1.000	-0.130**	0.045
Information ethics consciousness						1.000	-0.252***
Information ethics behavior							1.000

\*\*\* p&lt;.001 \*\* p&lt;.01 \* p&lt;.05

generalized trust of Americans is the highest among the three. The result of the ANOVA was highly significant ( $p<.001$ ). Yamagishi & Yamagishi (1994) conducted surveys and found that Japanese rank higher on particularized trust and Americans on generalized trust in the daily life. In this study almost the same result was obtained even in the online community. That means the American people are used to social uncertainty since they have faced high migration. Singapore is a multiethnic nation as

well similar to the United States. However, more than 90% of the population consist of Chinese and Malay, and people's inclination to accept the others' decision is comparatively strong because the centralization tendency is higher in Singapore than in the US. Precisely the US society is characterized by a culture that allows putting trust in the good sense and the judgment of the common citizens and various opinions. This background can explain these results.

Table 4 shows the results of relationships among some variables related to generalized trust with correlation coefficient (Person's R). Table 4 indicates that trusters recognize high reciprocity in the web community, and interact with others often. This tendency has correlations with high cooperation to contribute to the Internet, it shows the possibility to formulate social capital of the online communities.

As for the relationships between generalized trust and information ethics consciousness/behavior, unexpected results are obtained. There is no significant correlation about Japanese data, yet in the US and Singapore there are results with negative correlations, i.e., the person with high trust (low score means high trust) tends to have immoral consciousness (high score means immoral consciousness) and immoral behavior (low score means immoral behavior). What does it indicate? Assumption for this problem will be discussed in the next section (3.4).

### **3.4 Relationship Between Generalized Trust and Ethics: From the Viewpoint of Rational Egoist Tendency**

There are studies by Rotter (1980) and Kosugi & Yamagishi (1998) based on psychological experiment in the real space that provided the significant results – contrary to the common sense idea that trustful people are gullible and easily believe whatever other people may say, the idea was not necessarily valid. Trusters are sensitive to their counterparts, the situation and the information, so that rational judgments depend on counterparts, situation and information. The same tendency would be observed in the Internet, too. Therefore with this another hypothesis, the author examines the relationships between trust and ethics depending on one's level of "rational egoist tendency." Rational egoist tendency is an aspect of human behavior that examines whether external sanctions would be imposed. If such sanctions are not expected, the behavior will be immorally and egoistically. Since the internet space has high invisibility and anonymity, rational egoists easily behave immorally.

How does rational egoism influence the relationship between trust and ethics? The rational tendency and the egoist tendency were measured separately, using six-step scales which ranged from "applies very well" to "doesn't apply at all." Rationality here means instrumental rationality, namely choosing rational means for the end. The author used "To attain a goal, I assess the current situation carefully before acting" and two other questions to measure this tendency. Egoism is defined here as self-centeredness in relation to others and the society. Three questions were used for this tendency, including "I would like to live my life as it suits me even if that means that other people have to suffer".

Based on these measures, the respondents were divided into four groups for each country, to obtain correlation coefficient between generalized trust and ethics in each group.

- Group 1: High Egoism, High Rationality  
 Group 2: High Egoism, Low Rationality  
 Group 3: Low Egoism, High Rationality  
 Group 4: Low Egoism, Low Rationality

The results of the correlation between generalized trust in the online communities and information ethics are shown in table 5. It is observed that relationships among these variables show different tendencies depending on the rational egoist tendency. In group 4 (low egoism, low rationality), trusters tend to have good moral consciousness and action. On the other hand, in group 2 (high egoism, low rationality), even though she/he has a high level of trust, the egoist tendency is supposed to supersede and the participant behaves immorally and egoistically. Others excluding herself/himself in the web communities are honest and trustworthy – this would be a recognition of counterparts and situation by rational egoists, which makes them behave immorally. This tendency is remarkable in the US and Singapore according to table 5.

**Table 5.** Correlation coefficient between generated trust on the Internet and ethics consciousness/behavior in four groups of rational egoist tendency (Pearson's R)

Japan	rational egoist tendency		generalized trust in the real space	Information ethics consciousness	Information ethics behavior
Grp.1 (N=321~331)	egoist tendency: high rationality : high	generalized trust on the Internet	0.467***	0.059	-0.113*
Grp.2 (N=310~317)	egoist tendency: high rationality : low	generalized trust on the Internet	0.474***	-0.066	0.038
Grp.3 (N=242~248)	egoist tendency: low rationality : high	generalized trust on the Internet	0.472***	-0.050	-0.078
Grp.4 (N=232~239)	egoist tendency: low rationality : low	generalized trust on the Internet	0.516***	0.092	0.129*
USA	rational egoist tendency		generalized trust in the real space	Information ethics consciousness	Information ethics behavior
Grp.1 (N=115)	egoist tendency: high rationality : high	generalized trust on the Internet	0.579***	-0.094	0.279**
Grp.2 (N=167)	egoist tendency: high rationality : low	generalized trust on the Internet	0.336***	-0.188*	0.217**
Grp.3 (N=157)	egoist tendency: low rationality : high	generalized trust on the Internet	0.473***	-0.144	0.132
Grp.4 (N=112)	egoist tendency: low rationality : low	generalized trust on the Internet	0.534***	-0.006	-0.212*
Singapore	rational egoist tendency		generalized trust in the real space	Information ethics consciousness	Information ethics behavior
Grp.1 (N=157)	egoist tendency: high rationality : high	generalized trust on the Internet	0.619***	-0.063	0.142
Grp.2 (N=154)	egoist tendency: high rationality : low	generalized trust on the Internet	0.482***	-0.244**	0.296***
Grp.3 (N=132)	egoist tendency: low rationality : high	generalized trust on the Internet	0.638***	-0.165	0.104
Grp.4 (N=111)	egoist tendency: low rationality : low	generalized trust on the Internet	0.400***	0.045	0.146

\*\*\* p<.001 \*\* p<.01 \* p<.05

## 4 Conclusion

This study has tried to make clear the type of former studies on trust and Internet, and based on these categories, examined the significance and possibility of trust related to

social capital on the Internet with survey data. Results indicate that studies on trust related to the Internet are categorized according to two dimensions – “studies on the system trust (especially, trust in the system infrastructure) – studies on the personality trust” and “studies with risk management approach – studies with trust management approach.” Among them the concept of social trust should be more important in the web community because of its higher level of uncertainty rather than the real world. Generalized trust has significant relationships with reciprocity, human interaction and cooperation on the Internet. However ethics does not simply correlate with trust. It is clarified that a trustor does not always behave morally in the process of forming the web community. The features of the Internet space change consciousness and the actual behavior of an individual. Yet a trustor actually helps and cooperates each other. We have to design web communities based on this result.

## Acknowledgement

This study was supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (B).

## References

- [1] Akerlof, G. A., 1970, The Market for “Lemon”: Qualitative Uncertainty and the Market Mechanism, *Quarterly Journal of Economics*, **84**, pp.488-500
- [2] Atsuji, S., Shyi, S. & Hirayama, Y., 2002, Trust Management in Knowledge Community, *Journal of the Japan Society for Management Information*, **11**(2), 31-41
- [3] Barber, B., 1983, *The Logic and Limit of Trust*, Rutgers University Press.
- [4] Coleman, J.S., 1988, Social capital in the creation of human capital, *American Journal of Sociology*, **94** Supplement, pp.95-120
- [5] Endo, H., & Noto, M., 2003, Information Recommendation System by World-of-Mouth Communication Model, *Technical Report of the Institute of Electronics, Information and Communication Engineers*, **103**(78), pp.13-18
- [6] Frank, R.H., 1988, *Passions within Reason*, Norton.
- [7] Fukuyama, F., 1995, *Trust: The Social Virtues and the Creation of Prosperity*, New York: Simon & Schuster
- [8] Gambetta, D., 1988, Mafia: The Price of Distrust, *Trust: Making and Breaking Cooperative Relations*, Basil Blackwell.
- [9] Hardin, R., 1992, The Street-level Epistemology of Trust, *Politics and Society*, **21**, pp.505-529
- [10] Kawashima, H. & Mukaidono, M., 2001, Proposal of Reliability Enhancement Activities on the Network (No.1 & No.2), *Journal of Reliability Engineering Association of Japan*, **23**(4 & 5), pp.314-325 & pp.437-445
- [11] Kosugi, M. & Yamagishi, T., 1998, General Trust and Judgment of Trustworthiness, the Japanese Journal of Psychology, **69**(5), pp.349-357
- [12] Giddens, A., 1990, *The Consequences of Modernity*, Stanford: Stanford University Press.
- [12] Loury, G., 1987, Why should we care about group inequality?, *Social Philosophy and Policy*, **5**, pp. 49-271
- [13] Luhmann, N., 1973, *Vertrauen : ein Mechanismus der Reduktion sozialer Komplexität*, Stuttgart: Ferdinand Enke Verlag.



- [14] Miyata, K., 2004, The Internet as Medium of Social Networking: The Formation of Social Capital in Online Communities and Effects on Micro and Macro Level, *Cognitive Studies*, **11**(3), 182-196
- [15] Nara, Y., 2005, Trust and Information-taking Behavior in the Web Community, *Proceedings of the 9th International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES2005)PartIV*, Springer pp.839-847
- [16] Nara, Y., 2004, Ethics on the Internet: A Comparative Study of Japan, the United States, and Singapore, *Community in the Digital Age: Philosophy and Practice*, Rowman & Littlefield
- [17] Nomura, T., Katayama, t., et al., 2005, Potential Social Network Search Mechanism for Driving Knowledge Sharing, *IPSJ Transaction on Databases*, **46**(SIG 8), pp.72-81
- [18] Norris, P., Social Capital and Information and Communication Technologies: Widening or Reinforcing Social Networks?, 2003, Trust in the Knowledge Society, International Forum on Social Capital for Economic Revival, Economic and Social Research Institute, Cabinet Office, Japan
- [19] Putnam, R.D., 1993, *Making democracy work: Civic traditions in modern Italy*, Princeton, NJ: Princeton University Press
- [20] Rotter, J., 1967, A new scale for the measurement of interpersonal trust, *Journal of Personality*, **35**, pp.651-665
- [21] Rotter, J., 1980, Interpersonal Trust, Trustworthiness and Gullibility, *American Psychologist*, **35**, pp.1-7
- [22] Syozugawa, Y. & Yoshida, H., 1999, A Study of How Information Technology and Communications Affect Individual Relationships and Corporate Organizations: The Possibility for Mutual Trust and Commitment, *Journal of the Japan Association for Social Informatics*, **11**, 120-134
- [23] Sugino, T., 2003, Security of Information Network and Risk management, *Operations Research*, **48**(7), pp.29-34, 2003
- [24] Tomomobe, H. & Matsuo, Y., et al., 2005, Extraction and Utilization of Human Social Network toward Semantic Web, *Journal of Information Processing Society of Japan*, **46**(6), pp.1470-1479
- [25] Uslaner, E.M., 2003, Trust in the Knowledge Society, International Forum on Social Capital for Economic Revival, Economic and Social Research Institute, Cabinet Office, Japan
- [26] Usui, Y., Takahashi, H., et al., 2002, Study on the Effect of Reputation System as Trust Evaluation in Network Community, *Technical Report of the Institute of Electronics, Information and Communication Engineers*, **102**(505), pp.19-24
- [27] Yamagishi, T., Cook, K.S. & Watabe, M., 1998, Uncertainty, Trust and Commitment Formation in the US and Japan, *American Journal of Sociology*, **104**(1), pp.165-194
- [28] Yamagishi, T. & Yamagishi, M., 1994, Trust and Commitment in the US and Japan, *Motivation and Emotion*, **18**, pp.129-166
- [29] Yamagishi, T. & Yoshikai, N., 2005, Towards Trustable Information Society, *Journal of the Institute of Electronics, Information and Communication Engineers*, **88**(1), 54-56
- [30] Yamamoto, T., Sakai, R. & Kasahara, M., 1996, A Fair Exchange System on Anonymous Channel, *Technical Report of the Institute of Electronics, Information and Communication Engineers*, **96**(365), pp.9-13

# Context-Aware Application System for Music Playing Services\*

Jae-Woo Chang and Yong-Ki Kim

Dept. of Computer Engineering  
Center for Advanced Image and Information Technology  
Chonbuk National University, Chonju, Chonbuk 561-756, South Korea  
jwchang@chonbuk.ac.kr, ykkim@dblab.chonbuk.ac.kr

**Abstract.** Context-awareness is a technology to facilitate information acquisition and execution by supporting interoperability between users and devices based on users' context. In this paper, we design a middleware and a context server for dealing with context-aware application system in ubiquitous computing. The middleware plays an important role in recognizing a moving node with mobility as well as in executing an appropriate execution module according to context. In addition, the context server functions as a manager that efficiently stores context information, such as user's current status, physical environment, and resources of a computing system. Using them, we implement our application system which provides a music playing service based on context. It is shown to take below two seconds that our application system can detect a user's context and start playing music according to the context.

## 1 Introduction

Mark Wieser at Xerox Palo Alto Research Center described ubiquitous computing as being about interconnected hardware and software that are so ubiquitous that no one notices their presence [1]. An effective software infrastructure for running ubiquitous applications must be capable of finding, adapting, and delivering the appropriate applications to the user's computing environment based on the user's context [2]. Thus, context-aware application systems determine which user tasks are most relevant to a user in a particular context. They may be determined based on history, preferences, or other knowledge of the user's behavior, as well as the environmental conditions. The context-awareness can facilitate information acquisition and execution by supporting interoperability between users and devices based on users' context, in a variety of applications including location-based services and Telematics. In this paper, we design middleware and context server components for context-aware application systems. The middleware plays an important role in recognizing a moving node with mobility as well as in executing an appropriate execution module according to context. In addition, the context server functions as a manager that efficiently stores

---

\* This work is financially supported by the Ministry of Education and Human Resources Development (MOE), the Ministry of Commerce, Industry and Energy (MOCIE) and the Ministry of Labor (MOLAB) through the fostering project of the Lab of Excellency.

context information, such as user's current status, physical environment and resources of a computing system. Using them, we implement our application system which provides a music playing service based on context.

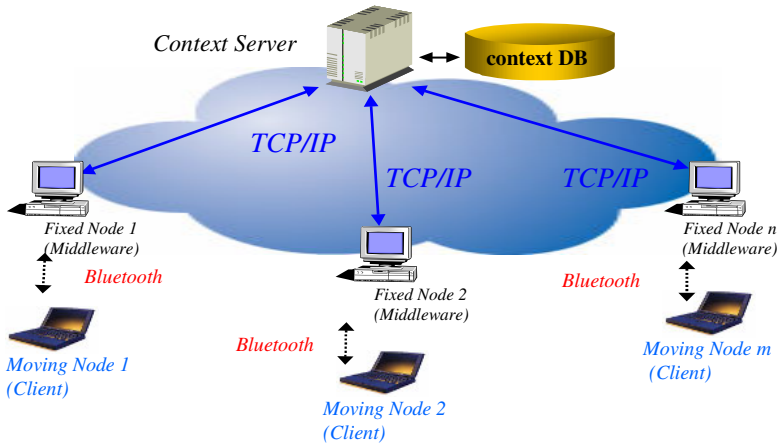
## 2 Related Work

Context-aware application systems determine which user tasks are most relevant to a user in a particular context. They may be determined based on history, on preferences, or other knowledge of the user's behavior, as well as the environmental conditions. In this section, we discuss the typical context-aware application systems. First, INRIA in France [3] proposed a general infrastructure based on contextual objects to design adaptive distributed information systems in order to keep the level of the delivered service despite environmental variations. The contextual objects (COs) were mainly motivated by the inadequacy of current paradigms for context-aware systems. The use of COs does not complicate a lot of development of an application, which may be developed as a collection of COs. The value of a particular object used in a context-dependent application is automatically updated by the adaptive framework, independently from the application. Secondly, AT&T Laboratories Cambridge in U.K [4] presented a platform for context-aware computing which enables applications to follow mobile users as they move around a building. The platform is particularly suitable for richly equipped, networked environments. Users are required to carry a small sensor tag, which identifies them to the system and locates them accurately in three dimensions. Thirdly, Arizona State Univ. [5] presented Reconfigurable Context-Sensitive Middleware (RCSM), which made use of the contextual data of a device and its surrounding environment to initiate and manage ad hoc communication with other devices. The RCSM provided core middleware services by using dedicated reconfigurable FPGA (Field Programmable Gate Arrays), a context-based reflection and adaptation triggering mechanism, and an object request broker that are context-sensitive and invokes remote objects based on contextual and environmental factors, thereby facilitating autonomous exchange of information. Finally, Lancaster Univ. in U.K [6] presented a comprehensive description of the GUIDE project which has been developed to provide city visitors with a hand-held context-aware tourist guide. The development of GUIDE has involved: capturing a real set of application requirements, investigating the properties of a cell-based wireless communications technology in a built-up environment and deploying a network based on this technology around the city, designing and populating an information model to represent attractions and key buildings within the city, prototyping the development of a distributed application running across portable GUIDE units and stationary cell-servers.

## 3 Middleware and Context Server for Context-Awareness

Context is any information that can be used to characterize the situation of any entity [7]. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves. We design an overall architecture of context-adaptive computing system for supporting various context-aware application services by combining the advantage of the INRIA

work [3] with that of the AT&T work [4]. The system is divided into three components, context server, fixed node (middleware), and moving node (client). First, the context server functions as inserting remote object and context information into object and context database, respectively, and retrieving them from the databases. Secondly, a fixed node functions as a middleware, which can find, insert, and execute a remote object for context awareness. Finally, a moving object communicates with a fixed node and executes a predefined built-in program according to the context information acquired from a middleware. The context server communicates with a fixed node by using a network based on TCP/IP, while a moving object communicates with a fixed node using Bluetooth wireless communication [8]. The proposed context-aware computing system has a couple of powerful features. First, our middleware can define context objects describing context information as well as can keep track of a user's current location. Secondly, our context server can store context objects and their values depending on a variety of contexts as well as can manage users' current locations being acquired from a set of fixed nodes by using spatial indexing. Finally, our client can provide users with adaptive application services based on the context objects. Figure 1 shows the overall architecture for supporting various context-aware application services.



**Fig. 1.** Overall architecture for supporting context-aware application services

Our middleware for context-aware application services consists of three layers, such as detection/monitoring layer, context-awareness layer, and application layer. First, the detection/monitoring layer serves to monitor the locations of remote objects, network status, and computing resources, i.e., CPU usage, memory usage, bandwidth, and event information related with devices. Secondly, the context-awareness layer serves as a middleware which is an essential part for handling context-aware application services. It can be divided into five managers, such as script processor, remote object manager, context manager, context selection manager, communication proxy manager. The script processor analyzes the content of context-aware definition script and executes its specified action. The remote object manager manages a data structure

for all the context objects used in application programs. The context manager manages context and environmental information including user preference, user location, etc. The context selection manager chooses the most appropriate context information under the current situation. The communication proxy manager serves to communicate with the context server and temporarily reserve data for retransmission in case of failure. Finally, the application layer provides functions to develop various context-aware applications using the application programming interface (API) of the middleware while it is executed independently of the middleware.

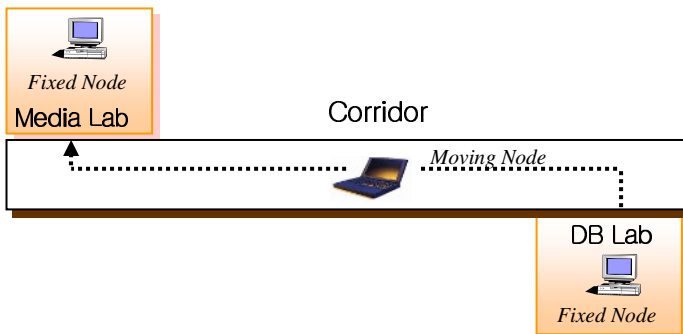
The procedure to execute proper services based on context using our middleware has three steps. First, a moving node broadcasts a connection request signal so as to connect with a fixed node by using Bluetooth. The fixed node covering an interesting area analyzes the signal and accepts the connection with the corresponding fixed node. When the connection between them is established, the moving node delivers its own information to the fixed node. Secondly, when remote objects are found, the fixed node delivers their information to the context server through a network using TCP/IP. The server stores into the context database the information delivered from the fixed node. Finally, the server searches the fixed node's context which is the most suitable with current situation from the context database. The fixed node executes predefined programs based on the context information and notifies a moving node of services being executed. Then, the moving node executes the service being requested by the fixed node. Because a moving node maintains all the addresses of fixed nodes and makes a connection to a fixed node, a moving node communicates with the fixed node periodically once a connection between them is established, and determines whether or not the connection between them should hold.

Because a context server is required to store and retrieve remote objects and context information extracted from them, we design a context server to efficiently manage both the remote object and the context information using a commercial DBMS. The context server analyzes and executes the content of packets delivered from the middleware. It is divided into four managers, such as communication manager (CM), packet analysis manager (PAM), context/object manager (COM), and SQL query manager (SQM). The CM serves as communicate between the server and the middleware. The CM delivers to PAM packets transferred from the middleware so as to parse them, and it delivers to the middleware result packets made from the server. The PAM parses the packets from the CM and determines what action wants to be done currently. Based on the parsing, the PAM calls the most proper functions in the COM. The COM translates into SQL statements the content delivered from the PAM and delivers the SQL statements to SQM to execute them. The application programming interface (API) for the COM is ContextDefine, ContextDestroy, ContextInsertTuple, ContextDeleteTuple, ContextSearch, ContextSearchTuples, and ContextCustom. The SQM executes the SQL statements from the COM by using the DBMS and delivers the result to the middleware via the CM. The API for the SQM is sql-Reader and sql-NonQuery.

## 4 Context-Aware Music Playing Application System

Using the middleware and the context server, we develop our context-aware music playing application system under Redhat Linux 7.3(kernel version 2.4.20) with 866

MHz Pentium-III CPU and 64 MB main memory. We make use of GCC 2.95.4 as a compiler and affix 2.0.2 as a Bluetooth device driver. The Bluetooth device follows the specification of Version1.1/Class1 and makes a connection to PCs using USB interfaces [9]. We also use MySQL DBMS as a commercial DBMS because we can reduce the developing time, compared with using a storage system, and we can increase the reliability of developed systems. In our music playing application system, when a user belonging to a moving node approaches to a fixed node, the fixed node plays a music with the user's preference according to the user's location. In general, each user has a list of his music with his(her) preference and moreover has different lists of popular music depending on time, i.e., morning time, noon time, and night time. Thus, when a user, which is listing to his popular music in the area of the fixed node 1, moves to the area of the fixed node 2 (Figure 1), the music stops playing in the area of the fixed node 1 while it starts playing in the area of the fixed node 2. The fixed node differentiates a user from another user and plays his(her) preferred music depending on the current time by considering the time when a user enters into the area of the fixed node. The record of a database in context server for our music playing application has six attributes, such as User\_ID, User\_Name, Location, Music\_M, Music\_A, and Music\_E. The User\_ID serves as a primary key to identify a user uniquely. The User\_Name means a user name and the Location means a user's current location which can be changed whenever a fixed node finds the location of a moving object. Finally the Music\_M, the Music\_A, and the Music\_E represent his(her) popular music file in the morning time, the noon time, and the night time, respectively.



**Fig. 2.** Testing environment for our application system

To determine whether or not our application system works well, we test it by adopting the scenario used in Cricket [10], one of the MIT Oxygen project. We test its execution in the following three cases; the first is when a user covered by a moving node approaches to a fixed node or move apart from it, the second is when two different users approaches to a fixed node, and final case is when a user approaches to a fixed node at different times. Among them, because the first case is the most general case, we will explain the first case in more detail. For our testing environment, we locate two fixed nodes in Database laboratory (DB Lab) and Media communication laboratory (Media Lab) of Chonbuk National University, as shown in Figure 2, where

the fixed node can detect a moving node by using Bluetooth wireless communication. There is a corridor between DB Lab and Media Lab and its distance is about 60 meter. We test its execution in case when a user having a moving node moves from DB Lab to Media Lab or in a reverse direction.

Figure 3 shows testing in case when a user having a moving node is approaching to a fixed node. First, the fixed node receives a user name from the moving node as the moving node receives a user name from the moving node as the moving node (①). Secondly, we determine whether the information of the user has already been stored into a server or not. If does, we search the music file belonging to the user in a given time and downloads the music file from the server (②). Finally, we play the downloaded music file by using a MP3 music player. On the contrary, testing in case when a user having a moving node is moving apart from to a fixed node is so little different. First, when the middleware detect that a user is too far from the fixed node to communicate with it, we output an error message and stop the process playing the music. Finally, we remove the music player process. In a short, when a user is approaching to a fixed node, the music belonging to the user is playing while when a user is moving apart from the fixed node, the music stops playing.

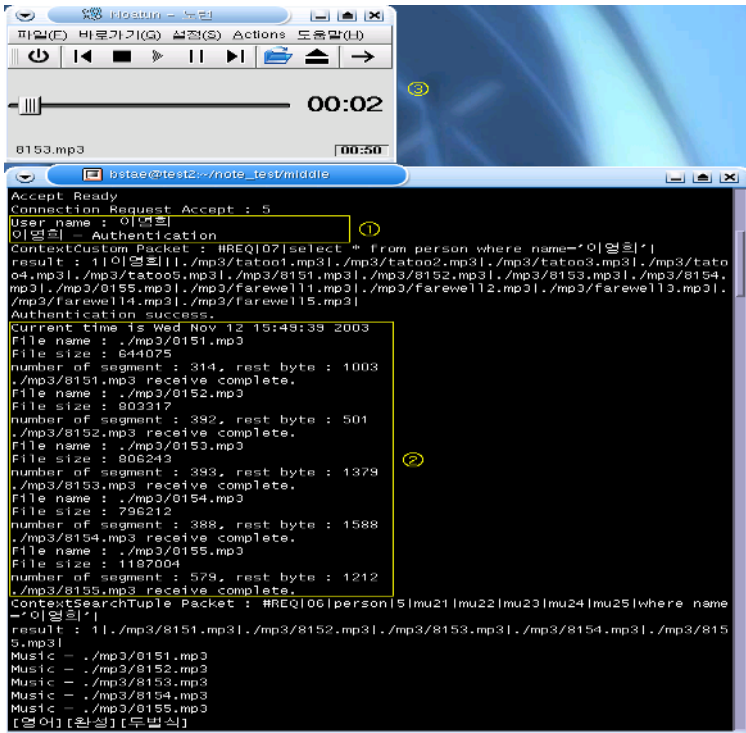


Fig. 3. Testing in case when a user is approaching to a fixed node

For the performance analysis of our application system, we measure the average times by adopting a boundary detection of beacons used in Cricket [10]. First, as a moving node is approaching to a fixed node, it takes 1.34 second to make a connection

between them. The time means the one to detect the connection of a moving node by middleware when a moving node enters into the communication boundary of a fixed node. The time mainly depends on the specification of Bluetooth wireless communication. Secondly, it takes 0.5 second to start a corresponding service after making the connection between them. The time means the one to search the profile of the corresponding user and to call the module playing music by the middleware. Here the searching time for a user's profile is dependant on the packet transfer time of TCP/IP and on the DBMS performance of the context server. The calling time for the music playing module means the one for loading it by an operating system (OS), which is affected by the available memory of the OS kernel and the speed of a hard disk. Finally, as a moving node is moving apart from a fixed node, it takes 1.45 second to make a disconnection between them. The time means the one to detect the disconnection of a moving node by the middleware. The time is relatively long due to the kernel's release of socket resources because the kernel tries to communicate with the moving node even though the moving node goes out of the communication boundary of a fixed node. When the kernel is connected or disconnected with the moving object, it can be considered very reasonable that the middleware sets the time limit to be two seconds. Thus, if it takes over two seconds for the middleware to make a connection with a moving node and detect context from it, a user may consider the situation as a fault.

## 5 Conclusions and Future Work

In this paper, we designed both a middleware and a context server for dealing with context-aware application system in ubiquitous computing. The designed middleware played an important role in recognizing a moving node with mobility by using a Bluetooth wireless communication as well as in executing an appropriate execution module according to the context acquired from a context server. The designed context server functioned as a manager that efficiently stores into the database server context information, such as user's current status, physical environment, and resources of a computing system. Using the middleware and the context server, we implemented our application system which provides a music playing service based on context. It was shown to take below 2 seconds that our application system could detect a user's context and start playing music according to the context. As future works, it is required to study on an inference system to acquire new context information from the existing context information.

## References

1. M. Weiser, "Some Computer Science Issues in Ubiquitous Computing", *Communication of the ACM*, Vol 36(7), pp. 75-84, 1993.
2. G. Banavar, A. Bernstein, "Issues and challenges in ubiquitous computing: Software infrastructure and design challenges for ubiquitous computing applications", *Communication of ACM*, Vol 45(12), pp. 92-96, 2002.
3. P. Couderc, A. M. Kermarrec, "Improving Level of Service for Mobile Users Using Context-Awareness", *Proc. of 18th IEEE Symposium on Reliable Distributed Systems*, pp. 24-33, 1999.



4. A. Harter, A. Hopper, P. Steggles, A. Ward, P. Webster, "The anatomy of a Context-aware application", *Wireless Networks* Vol. 8, Issue 2/3, pp. 187-197, 2002.
5. S. S. Yau and F. Karim, "Context-sensitive Middleware for Real-time Software in Ubiquitous Computing Environments", *Proc. of 4th IEEE Symposium on Object-oriented Real-time Distributed Computing*, pp.163-170, 2001.
6. K. Cheverst, N. Davies, K. Mitchell, A. Friday, "Experiences of developing and deploying a context-aware tourist guide: the GUIDE project", *Proceedings of the sixth annual international conference on Mobile computing and networking*, pp. 20-31, 2000.
7. A. K. Dey, "Understanding and Using Context", *Personal and Ubiquitous Computing Journal*, Vol. 5, No. 1, pp. 4-7, 2001.
8. Bluetooth Version 1.1 Profile, <http://www.bluetooth.com>.
9. Affix: Bluetooth Protocol Stack for Linux, <http://affix.sourceforge.net>.
10. N. B. Priyantha, A. Chakraborty, and H. Balakrishnan, "The Cricket Location Support System", *6th ACM/IEEE Int'l Conf. on Mobile Computing and Networking(MOBICOM)*, pp. 32-43, 2000.

# Query Based Summarization Using Non-negative Matrix Factorization

Sun Park<sup>1</sup>, Ju-Hong Lee<sup>1,\*</sup>, Chan-Min Ahn<sup>1</sup>, Jun Sik Hong<sup>2</sup>, and Seok-Ju Chun<sup>3</sup>

<sup>1</sup>Department of Computer Science and Information Engineering, Inha University, Korea  
{sunpark, ahnchl}@datamining.inha.ac.kr,  
juhong@inha.ac.kr

<sup>2</sup>Department of Electronic Engineering, Youngdong University, Korea  
jnshong@youngdong.ac.kr

<sup>3</sup>Department of Computer Education, Seoul National University of Education, Korea  
chunsj@snue.ac.kr

**Abstract.** Query based document summaries are important in document retrieval system to show the concise relevance of documents retrieved to a query. This paper proposes a novel method using the Non-negative Matrix Factorization (NMF) to extract the query relevant sentences from documents for query based summaries. The proposed method doesn't need the training phase using training data comprising queries and query specific documents. And it exactly summarizes documents for the given query by using semantic features and semantic variables without complex processing like transformation of documents to graphs because the NMF have a great power to naturally extract semantic features representing the inherent structure of a document.

## 1 Introduction

Query based summarization of text document becomes important with increasing the amount of information available on the internet. Text document summaries can be either generic summaries or query based summaries. A generic summary presents an overall sense of the documents' contents. A query based summary presents the contents of the document that are related to the user's query. Query based summarization are tailored to the requirements of a particular user. The summary takes into account some representation of the user's interests, which can range from user model to profiles recording subject area terms or even a specific query containing terms that are deemed expressive of a user's information need [10, 11].

The recent studies for query based summarization are as follows: Berger and Mittal proposed a method that uses frequently-asked question (FAQ) for training a query-relevant summarization. Each frequently-asked question document is comprised of questions and answers about specific topic [2].

Bosma proposed a method using the Rhetorical Structure of the document. It transforms a document into a weighted graph, in which each vertex represents a sentence and the weight of an edge represents the distance between the two sentences. It is difficult to apply the Rhetorical Structure Theory to multimodal documents without extensive

---

\* Corresponding author.

modifications [3]. Varadarajan and Hristidis proposed a method to create query specific summaries by adding structure graph to documents by extracting associations between their fragments [15]. Mani and Bloedorn used graphs to formalize relations between sentences inside a document for multi-document summarization [12].

Sakurai and Utsumi proposed a method that generates the core part of the summary from the most relevant document to a query, and then the additional part of the summary, which elaborates on the topic, from the other documents. Their method has a beneficial effect on long summaries. But its performance is not satisfactory for the specific task [13].

Saggion used the content reduction which is a process of sentence elimination. The content reduction method removes sentences from a pool of candidate sentences until the desired compression is achieved [14].

The NMF can find a parts representation of the data because non-negative constraints are compatible with the intuitive notions of combining parts to form a whole, which is how the NMF learns a parts-based representation [8, 9].

In this paper, we propose a novel method that makes query-based summaries by extracting sentences using the similarity between query and Non-negative Semantic Feature vectors obtained from the NMF.

The proposed method in this paper has the following advantages: First, it is an unsupervised text summarization method that doesn't require the training data comprising queries and query specific documents. Second, the NMF have a great power to naturally extract semantic features representing the inherent structure of a document. By virtue of the power of the NMF, it also can select sentences that are highly relevant to a given query because it can chooses the sentences related to the query relevant semantic features that well represent the structure of a document. Third, it can be applied to the query based summarization for multi-documents.

The rest of the paper is organized as follows: Section 2 describes the proposed method and section 3 shows the experimental results. We conclude the paper in section 4 with future researches.

## 2 Query Based Summarization

In this section, we propose a method that creates query-based summaries by selecting sentences using the NMF. The proposed method consists of the preprocessing step and the summarization step.

### 2.1 Preprocessing

In the preprocessing step, after a given document is decomposed into individual sentences, we remove all stopwords and perform words stemming. Then we construct the weighted term-frequency vector for each sentence in document by Equation (1) [1, 4, 5].

Let  $T_i = [t_{1i}, t_{2i}, \dots, t_{ni}]^T$  be the term-frequency vector of sentence  $i$ , where elements  $t_{ji}$  denotes the frequency in which term  $j$  occurs in sentence  $i$ . Let  $A$  be  $m \times n$  weighted terms by sentences matrix, where  $m$  is the number of terms and  $n$  is the number of sentences in a document. Let element  $A_{ji}$  be the weighted term-frequency of term  $j$  in sentence  $i$ .

$$A_{ji} = L(j, i) \cdot G(j, i) \quad (1)$$

Where  $L(j, i)$  is the local weighting for term  $j$  in sentence  $i$ , and  $G(j, i)$  is the global weighting for term  $j$  in the whole documents. That is,

$$L(j, i) = t_{ji} \quad (2)$$

$$G(j, i) = \log(N/n(j)) \quad (3)$$

Where  $N$  is the total number of sentences in the document, and  $n(j)$  is the number of sentences that contain term  $j$ .

## 2.2 Query Based Summarization by NMF

In the summarization step, sentences are selected by using the NMF.

We perform the NMF on  $\mathbf{A}$  to obtain the Non-negative Semantic Feature Matrix  $\mathbf{W}$  and Non-negative Semantic Variable matrix  $\mathbf{H}$  such that:

$$\mathbf{A} \approx \mathbf{W}\mathbf{H} \quad (4)$$

Here  $\mathbf{W}$  is an  $m \times r$  matrix and  $\mathbf{H}$  is an  $r \times n$  matrix. Usually  $r$  is chosen to be smaller than  $n$  or  $m$ , so that the total sizes of  $\mathbf{W}$  and  $\mathbf{H}$  are smaller than that of the original matrix  $\mathbf{A}$ . This results in a compressed version of the original data matrix. We keep updating  $\mathbf{W}$  and  $\mathbf{H}$  until  $\|\mathbf{A} - \mathbf{W}\mathbf{H}\|^2$  converges under the predefined threshold. The update rules are as follows [8, 9]:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T A)_{\alpha\mu}}{(W^T W H)_{\alpha\mu}} \quad (5)$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(A H^T)_{i\alpha}}{(W H H^T)_{i\alpha}} \quad (6)$$

A column vector corresponding to  $j$ 'th sentence  $A_j$  can be represented as a linear combination of semantic feature vectors  $W_{.l}$  and semantic variable  $H_{lj}$  as follows:

$$A_{.j} = \sum_{l=1}^r H_{lj} W_{.l} \quad (7)$$

Where  $W_{.l}$  is the  $l$ 'th column vector of  $\mathbf{W}$ .

The powers of the two non-negative matrices  $\mathbf{W}$  and  $\mathbf{H}$  are described as follows: All semantic variables ( $H_{lj}$ ) are used to represent each sentence.  $\mathbf{W}$  and  $\mathbf{H}$  are represented sparsely. Intuitively, it make more sense for each sentence to be associated with some small subset of a large array of topics ( $W_{.l}$ ), rather than just one topic or all the topics. In each semantic feature ( $W_{.l}$ ), the NMF has grouped together semantically related terms. In addition to grouping semantically related terms together into

semantic features, the NMF uses context to differentiate between multiple meanings of the same term [8].

To evaluate the degree of similarity of the semantic feature vector  $W_l$  with regard to the query  $\vec{q}_j$  as the correlation between the vector  $W_l$  and  $\vec{q}_j$ . This correlation can be quantified, for instance, by the cosine of the angle between these two vectors [1, 4, 5]. That is,

$$\begin{aligned} \text{sim}(W_l, \vec{q}_j) &= \frac{W_l \cdot \vec{q}_j}{|W_l| \times |\vec{q}_j|} \\ &= \frac{\sum_{i=1}^m w_{i,l} \times q_{i,j}}{\sqrt{\sum_{i=1}^m w_{i,l}^2} \times \sqrt{\sum_{i=1}^m q_{i,j}^2}} \end{aligned} \quad (8)$$

Where  $|W_l|$  and  $|\vec{q}_j|$  are the norms of the semantic feature vector and query vectors.

We propose the following query based summarization method:

1. Decompose the document  $D$  into individual sentences, and let  $k$  be the number of sentences for summarization.
2. Perform the stopwords removal and words stemming operations.
3. Construct the weighted terms by sentences matrix  $A$  using Equation (1).
4. Perform the NMF on the matrix  $A$  to obtain the matrix  $W$  and the matrix  $H$  using Equation (5) and (6).
5. Select a column vector  $W_p$  of matrix  $W$  whose similarity to the query is the largest using Equation (8).
6. Select the sentence corresponding to the largest index value of the row vector  $H_p$ , and include it in the summary.
7. If the number of selected sentences reaches the predefined number  $k$ , then stop the algorithm. Otherwise go to step 5 to find the next most similar column vector excluding  $W_p$ .

In step 5, the fact that the similarity between  $W_p$  and the query is largest means the  $p$ 'th semantic feature vector  $W_p$  is the most relevant feature to the query. In step 6, it select the sentences that has the largest weight for the most relevant semantic feature.

### 3 Experimental Results

As an experimental data, we used Yahoo Korea News [6]. We gave 5 queries to retrieve news documents from Yahoo Korea News. The retrieved news documents are preprocessed using HAM (Hangul Analysis Module) which is a Korean language analysis tool based on Morpheme analyzer [7]. The evaluator was employed to manually create the query based summaries for the retrieved Yahoo Korea news documents. Table 1 provides the particulars of the evaluation data corpus.

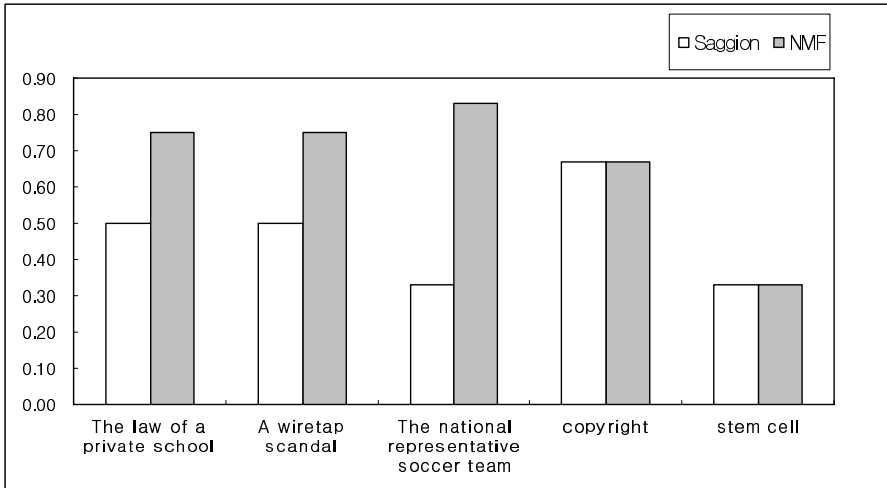
**Table 1.** Particulars of the Evaluation Data Corpus (Yahoo Korea News)

Document attributes	values
Number of docs	50
Number of docs with more than 10 sentences	42
Avg sentences / document	16
Min sentences / document	2
Max sentences / document	29

We used the precision ( $P$ ) to compare the performances of the two summarization methods, Saggion's method[14] and our method. We modified a Saggion's method for the experimental environment. Let  $S_{man}$ ,  $S_{sum}$  be the set of sentences selected by the human evaluators, and the summarizer, respectively. The standard definition of the precision is defined as follows [1, 4, 5]:

$$P = \frac{|S_{man} \cap S_{sum}|}{|S_{sum}|} \quad (9)$$

The evaluation results are shown in figure 1.

**Fig. 1.** Evaluation Results

Experimental results show that the proposed method surpassed the Saggion's method. This is because the NMF have the great power to grasp the innate structure of a document like human's cognition process [8].

## 4 Conclusions

In this paper, we proposed a novel method that makes query based summaries by extracting sentences using the NMF. The proposed method has the following

advantages: it doesn't require the training data comprising queries and query specific documents. By virtue of the power of the NMF that have a great power to naturally grasp the innate structure of a document. It can select sentences that are highly relevant to a given query. It also can be used to make the summaries for multi-documents.

In future work, we have a plan to evaluate our method on various term weighting schemes. And we will study the relation between Non-negative Semantic Feature Matrix  $W$  and Non-negative Semantic Variable Matrix  $H$  for performance elevation of the summarization.

**Acknowledgment.** This work was supported by the Brain Korea 21 Project in 2006.

## References

1. Baeza-Yaters, R., Ribeiro-Neto, B.: *Modern Information Retrieval*, ACM Press (1999)
2. Berger, A., Mittal, V. O.: Query-Relevant Summarization using FAQs. In *Proceeding of the 38<sup>th</sup> Annual Meeting on Association for Computational Linguistics (ACL'00)*, (2000)
3. Bosma, W.: Query-based Summarization using Rhetorical Structure Theory. In *Proceeding of the 15<sup>th</sup> Meeting computational Linguistics in the Netherlands (CLIN'04)*, (2004)
4. Chakrabarti, S.: *mining the web: Discovering Knowledge from Hypertext Data*, Morgan Kaufmann. 67-74 (2003)
5. Frankes, W. B., Baeza-Yates, R.: *Information Retrieval : Data Structure & Algorithms*, Prentice-Hall (1992)
6. [Http://kr.news.yahoo.com](http://kr.news.yahoo.com) (2005)
7. Kang, S. S.: *Information Retrieval and Morpheme Analysis*. HongReung Science Publishing Co. (2002)
8. Lee, D. D. and Seung, H. S.: Learning the parts of objects by non-negative matrix factorization, *Nature*, 401:788-791 (1999)
9. Lee, D. D. and Seung, H. S.: Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13, pages 556-562 (2001)
10. Mani, I.: *Automatic Summarization*. John Benjamins Publishing Company (2001)
11. Mani, I., Maybury, M. T.: *Advances in automatic text summarization*. The MIT Press (1999)
12. Mani, I., Bloedorn, E.: Multidocument summarization by graph search and matching. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI'97)*, (1997)
13. Sakurai, T., Utsumi, A.: Query-based Multidocument Summarization for Information Retrieval. In *Proceeding of the Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Summarization Workshop (NTCIR'04)*, (2004)
14. Sansion, H.: Topic-based Summarization at DUC 2005. In *Proceedings of the Document Understanding Conference 2005 (DUC'05)*, (2005)
15. Varadarajan, R., Hristidis, V.: Structure-Based Query-Specific Document Summarization. In *Proceeding of the ACM Fourteenth Conference on Information and Knowledge Management (CIKM'05)*, (2005)

# Intelligent Secure Web Service Using Context Information\*

Woochul Shin, Xun Li, and Sang Bong Yoo

School of Computer Science, Inha University  
Incheon, Korea

Fax: 82-32-874-1435

{woochul, inhaalgm}@dbssi.inha.ac.kr, syoo@inha.ac.kr

**Abstract.** Web service is one of the important components of the new computing paradigm, which uses the Web technologies for building distributed systems. This paper presents a design and implementation of a secured GIS Web Service for mobile devices. As many mobile devices are equipped with GPS (Global Positioning System), it is required to handle the position information more effectively. We have extended the proxy program in the client device to actively send the context information to the server. Based on the context information the server determines the optimal service mode to a particular client. Because the Web service interfaces open network domain, proper security model should be incorporated for classified information. A working example of location-based secured GIS Web Service is presented. By using Web Service standards and XML messages we can achieve the maximal interoperability for heterogeneous mobile devices.

## 1 Introduction

Recently, with the rapid development of the information systems and mobile communication technology, major computing power has now evolved from personal computer to being embedded in chips in many surrounding objects. In realizing this kind of ubiquitous computing environment, computers should automatically identify information and the environments of users in order to execute tasks more effectively. Because the Web service interfaces open network domain, proper security model should be incorporated for classified information.

The definition of context is not the same in various fields. In this paper we focus on the mobile terminals that are equipped with GPS (Global Positioning System) like PDAs. The information about user's location, moving direction, operating system, and screen size represents the context of mobile devices. The information obtained from a user can be classified into software, hardware and geographic information. For most standard client developers, programming to collect context information are not trivial because they include many system variables and GPS APIs. In this paper, we present a novel concept of active proxy, which not only helps the user to develop the client system efficiently but also collects the context information from mobile devices.

---

\* This work was supported by Inha University.



Most GIS systems are developed in some specific service environments; it is quite difficult to exchange information in different systems. For example PLIS (Productivity Layout Information System) and LMIS (Land Management Information System) and other systems use a specific database and development environment. Therefore, it is hard to directly exchange information each other. Being based on Web technologies, Web Service is promising technology for system integration. It is especially adequate for mobile environments because it employs simple XML messages. A Web Service interface and the message structure have been designed and implemented for location-based GIS service.

This paper is organized as follows. Section 2 discusses related work. Section 3 describes structure and handling of context-aware information for the location-based GIS Web Service. In Section 4, implementation of the Secured Web Service is described. A working example of the Secured Web Service is also presented. Section 5 concludes the paper and summarizes the contributions of this paper.

## 2 Related Work

Researches on context-awareness technologies have been emerged in the field of natural language processing. However, as the more devices are controlled by embedded computers, it is required to handle the context information more effectively [3, 5, 6, 7, 8, 12].

Context Toolkit (CT) is developed by Georgia Tech to facilitate building context-aware applications with standard components, such as context widgets, context aggregators and context interpreters [1]. CT utilizes these three components to offer appropriate services, and uses necessary context information; however the systems are not widely used. University of Arizona research relating to context-awareness technology is accomplished in the RCSM project [2]. The university had designed a kind of a high-efficiency description language based context interface.

However, common characteristics of the researches in this field are they usually require a large number of sensors [13, 14]. The GIS Web Service presented in this paper, the sensor that users contain is only GPS. The application areas that we are targeting are location-based service to mobile devices such as PDA and mobile phones. Because there are various platforms for those mobile devices, it is important task to collect context information correctly and provide services suitable for users [4, 10].

In this paper the proposed standard Web Service technologies are used to realize the Active Proxy for Web Service [9, 11]. Using the Active Proxy, programmers are not required to consider OS system variables and hardware properties. The Active Proxy will be embedded in user's terminal and provide the context information to the server.

## 3 Modeling Context Information in Mobile Devices

### 3.1 Data Types Including Context Information

Context Information in mobile devices comes from GPS and OS. Because we provide location-based service, the data from GPS are important. Standard GPS system provides the following location data.

- Universal Time
- Latitude ddmm.mm
- N or S (North or South)
- Longitude ddmm.mm
- E or W (East or West)
- GPS Quality Indicator
  - 0 - fix not available
  - 1 - GPS fix
  - 2 - Differential GPS fix
- Number of satellites in view2
- Antenna Altitude above/below mean-sea-level
- WGS-84 Geoidal separation, the difference between the WGS-84 earth ellipsoid and mean-sea-level
- Units of antenna altitude, meters
- Client speed
- Checksum
- User Password
- User Security Level
- Active Proxy Identification Key

Other context data come from the operating system variables. The following variables are defined in Windows CE.

- public SIZE sizeScreenSize; // Screen Size
- public String szComputerName; // Computer Name
- public String szUserDomainName; // Domain Name
- public String szUserName; // User Name
- public String szLocalIP; // IP
- public String szOSType; // OS Type
- public String szUserPassword // User Password
- public String szSecurityLevel // User Security Level
- public String szActiveProxyID // Active Proxy ID
- public OS\_VERSION OsVersion; // OS Version
- public Int32 dwSystemLanguageID; // System Language
- public bool bCanPlaySound; // Sound Availability
- public Int32 dwMonitorCount; // Monitor count
- public Int32 dwMemorySize; // Memory Size
- public Int32 dwTickCount; // Running Time
- public DATE\_TIME dtSystemTime; // System Time
- public Int32 dwNetworkSpeed; // Network Speed
- public String szProcessModel; // Process Model

The context data listed above will be collected by the Active Proxy automatically and sent to the Web Service Server.

### 3.2 Handling Context Information

The context information received from a user's terminal will be classified into two categories, i.e., software and hardware. The platform of the web service is based on XML. There are various kinds of operating systems in the market. There are large differences between Linux and Windows systems, in terms of naming objects, memory processing, and allocating hard disk space. Moreover, different versions of operation systems from the same corporation have different services. This section outlines the handling of context information.

The server decides the size of the map to be displayed on the client's screen based on the size of the user's screen. For example, large maps are delivered to large monitors those are usually used for desktop computers. On the other hand, smaller maps are delivered to devices with small monitors such as PDAs to be more effective. Considering the security, the system should check the validity of users in terms of computer-name, user-location-name and user-name. The system also traces the location of the clients through their local IPs and provides proper services to them by considering the trajectories.

Directions of moving to a inquired destination can be informed by means of a voice information. The server check the client system if it equipped with a sound card or not. It also checks the network speed in order to determine the optimal voice quality and transmission time. The server also verifies the language of the client system by its system language ID. Proper voice messages should be delivered depending on the system language of the client.

Multiple monitors on the client enable the server to provide some additional services. For instance, if a client has two monitors; one shows a map of overall area and the other shows map zoomed up the target area searched by the users. The server also determines the computing capability of the client by means of the memory-size and CPU. If the memory-size is small or the processing speed of CPU is low, it will be difficult to transmit large information from server to client. The server should decrease image resolution automatically. The client's system time are used for determining the time difference between the server and the client.

## 4 Implementation of GIS Web Service

### 4.1 Development Environment

In this research, the overall systems are developed on Microsoft platforms. Detail information of the development environment is summarized in Table 1.

**Table 1.** Development Environment of Secured GIS Web Service

<b>Operating System</b>	Microsoft Windows XP (SP2)
	Microsoft Windows CE.NET
<b>Web Server</b>	Microsoft IIS Server 5.1
<b>Develop Framework</b>	Microsoft .NET Framework 1.1
<b>Develop Platform</b>	Microsoft Visual Studio.NET 2003
<b>Voice Generation</b>	TTS Voiceware
<b>GPS Interface</b>	NMEA0183

The function of the Web Service server is to collect context information from the client and generate the information to be responded to the client. To be interoperable, it is required to use WSDL and SOAP to provide information to the client for development and service. The server is mainly divided into three parts: network interface, user authentication service and DB interface module. The features of this Web Service can be summarized as follows:

- For the sake of internationalization, Unicode encoding is employed.
- Standard GPS interface is used.
- Client and server are synchronized.
- The maps are fitted to client's monitor automatically.
- Web Service standards, i.e., UDDI, WSDL, and SOAP are employed
- User and messages are authenticated before served.
- Security modules are incorporated with the Web Service architecture.

### 4.2 Active Proxy for Client Development

In order to provide proper service, the server should collect the context information from clients. One approach would be define the variables for the required context information and make the users to provide their own information to the server. However, it is not trivial for average programmers to make programs to collect all the context information such as position, speed, OS type, screen size, memory size, network speed, so on. In this work we have extended the functionality of proxy to solve this problem.

A proxy exists in the form of window.dll on client device. The proxy has two main functions: one is to provide a web service interface; the other is to communicate with the server according to user's request. We have added one more function to the proxy, which is to collect context information of the client. The overall structure of Web Service system is depicted in Figure 1.

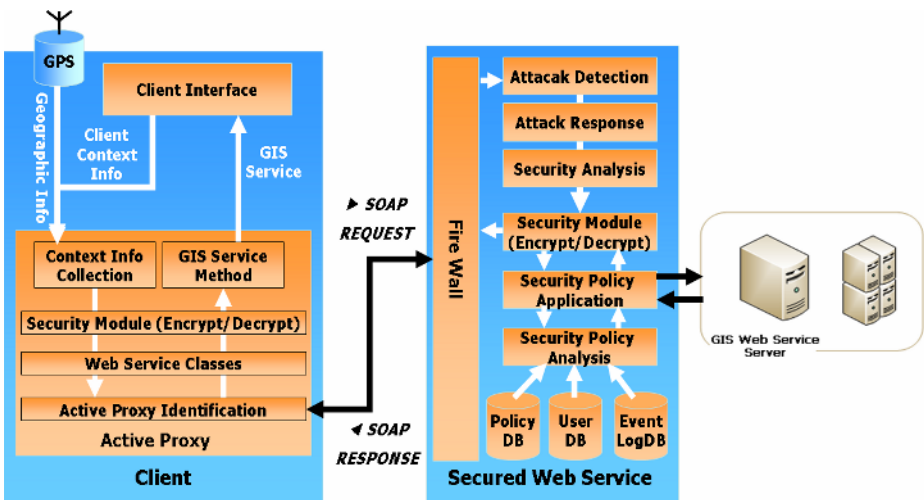


Fig. 1. Structure of Secured GIS Web Service using active proxy library

The active proxy connects to the GIS Web Service and provides context data to the server. The client program can just refer the active proxy library and use the operations and types. A simple declaration of the web service object using dll library (it is named ISLAB.dll in this work) can be done as follows.

```
// declares an object for the GIS Web Service by ISLAB
private ISLABGIS.refGWS.ISLABGISWebService GisWebService;
// declares a service object by the GIS Web Service
private ISLABGIS.refGWS.GIS_SERVICE_DATA GisServiceData;
// declares an object of client information
private ISLABGIS.refGWS.CLIENT_INFO ClientInfo;
// declares an object of geographic data
private ISLABGIS.refGWS.CLIENT_GEOGRAPHY_DATA CGData;
```

Using these objects, the active proxy can directly call the Web Service operations. A simple example is shown below.

```
// sends the client's context information to server
GisWebService.GetClientInfo(ClientInfo);
// receives the response from server
GisServiceData = GisWebService.GISService(CGData);
```

The ISLAB.dll library includes the function to collect the context information of a client system as follows.

```
// collects the context information of client system
Public CLIENT_INFO GetClientInfo()
```

This method returns the context information that will be sent to the server. This library is compiled in a .NET CLI environment; therefore it cannot be used in other environments, such as a java and MFC. For such platforms, different versions of libraries need to be developed.

### 4.3 Web Service Interface Design

The overall structure of operations provided by the GIS Web Service is presented in Figure 2. First the system verifies system performance of client device and network status using the CheckSystemStatus() operation. It decides if the Web Service can be provided properly. It then waits a service request from the client. Receiving a request, the system checks the validity of user by UserAuthentication(). After user authentication, the GetClientinfo() method receives context information of the client through the active proxy. The response to the client's request will be provided to the users via GIS-Service() operation. The VerifyData() operation checks the integrity of service data.

### 4.4 A Working Example of Web Service

The procedure for users to use the secured Web Service is as follows.

1. Web application developer searches the Web Service that he/she want to use. Proper Web Services can be located using UDDI.
2. Download the WSDL files that describe the Web Services.

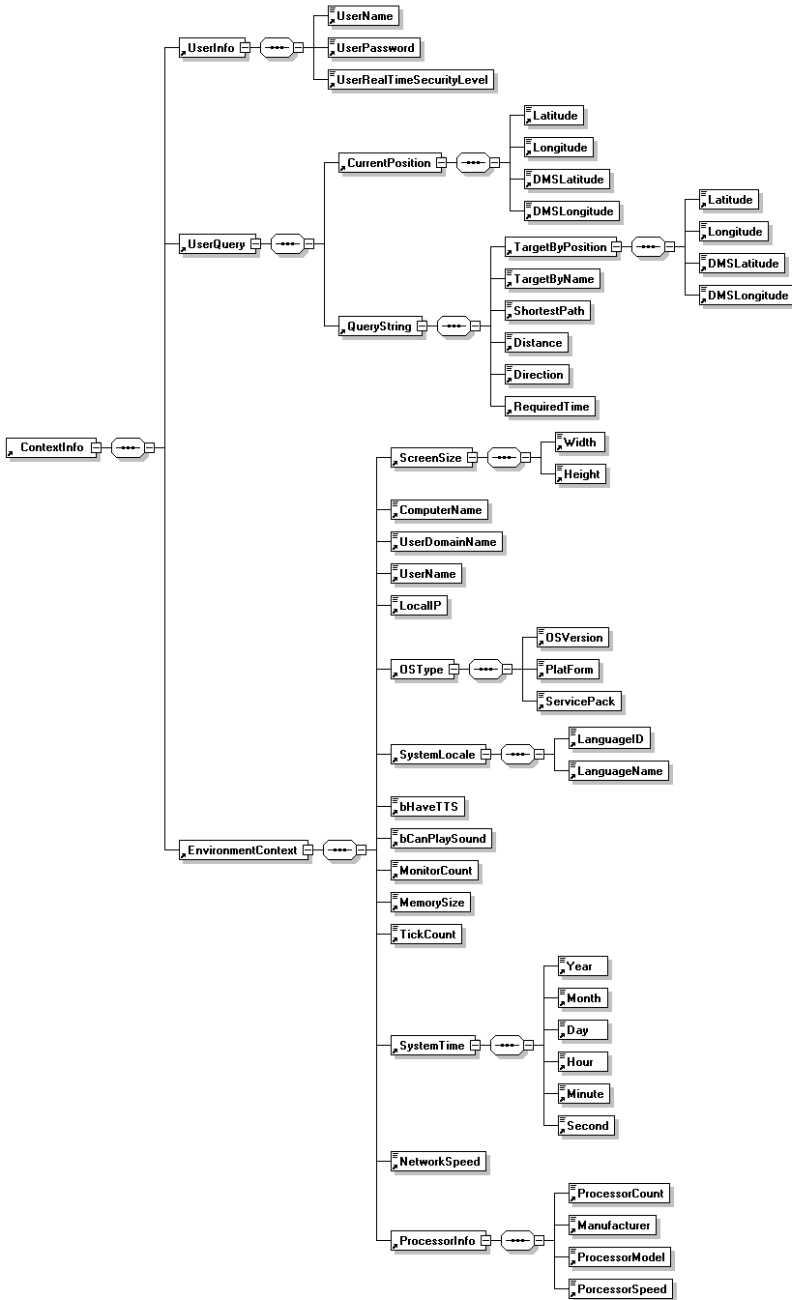
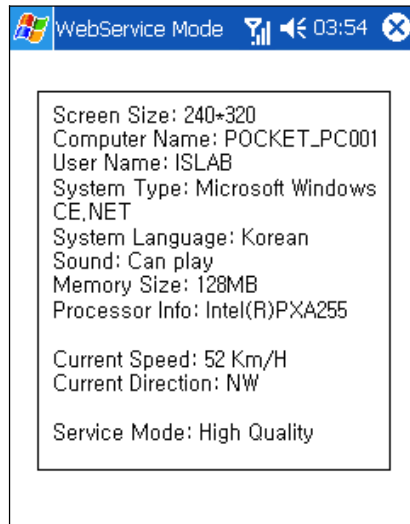


Fig. 2. Interface structure for Secured GIS Web Service

3. In a .NET environment, using the trace function of WSDL an active proxy can be created.
4. The active proxy collects context information of the client. Then it sends the value to the server using SOAP transmitted over HTTP.
5. The client and the server communicate each other according to the requests from the client.

When a client program connects the server, it will be requested to login to the system. If login succeeds, the server will begin to analyze the received context information as defined in the previous section. Some of the basic context information will be displayed on the client's monitor as shown in Figure 3.



**Fig. 3.** Context information displayed on client's monitor

```

for $a in document("textshape.txt")//TextShape
let $b := min(sp_distance($a/obj, here))
where contains($a/tex(), Object )and text_kind eq '50'
return <POI> {
    for $c in document("textshape.xml")//TextShape
    where contains($a/tex(), Object ) and sp_distance($a/obj, here) <= $b
and
    text_kind eq '50'
return {
    <Label> $c/text() </Label>
    <Location> $c/obj</Location>
    <Distance> sp_distance($c/obj,point(12.23323232,
88.123334323 )) </Distance> }
</POI>

```

**Fig. 4.** Template to search for the nearest object

Using the “Path” menu, the user can search his/her destination. For example, when the user searches for the nearest university, the server will search it and respond the result to the user. For this query, a template written in XQuery language (see Figure 4) will be executed.

The search result is shown in Figure 5. The user’s current position is denoted by A and the destination is marked by B. The small arrow near A indicates the moving direction of the user. The shortest path from A to B would be the dotted path. However, considering the moving direction of the user the server recommends the solid path from A to B. This is one example of using the user’s context information. The lower right-hand side of Figure 5 is marked by C. The detail geographical information of this area is blocked because it includes military facilities that are classified as confidential information.

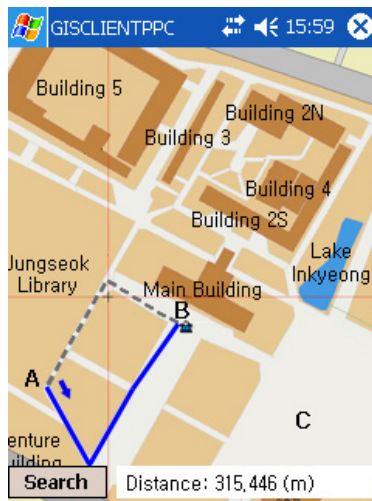


Fig. 5. Map showing the path from the current position A to the destination B

## 5 Conclusion

Location-based GIS service has been designed and implemented as a secured Web Service. Because many of mobile devices as PDA are equipped with GPS, the server can be informed the location of the clients more easily. In order to provide more appropriate service to each client, we need to analyze the context information from the client’s device. The context information of mobile devices can be classified into three categories, i.e., position information, S/W information, and H/W information. Most of the context information can be acquired from the GPS string and OS system variables. Microsoft proxy in client devices has been extended in order to actively send the context information to the server. Using the active proxy client program, developers have no burden of handling GPS string and system variables. A working example of Location-based GIS System is also presented. The contribution of this paper can be summarized as follows.



- The context model of mobile devices has been defined. It includes position information, S/W information, and H/W information.
- Microsoft proxy has been extended in order to actively send the context information to the server.
- Interface structure for Location-based GIS Web Service have been designed and implemented. Messages structures to be used in this service are also designed in XML.
- Message structures for in XML, which can improve the interoperability of the service.
- Security modules are incorporated with the Web Service architecture for proper handling of confidential information.

Because the proposed system follows all the standard technologies of Web Service, in can interoperable most of mobile devices. The proposed system can be easily extended to almost all mobile and location-based systems such as POI systems, tracking systems, ubiquitous systems, and distributed control and management systems. Some user-friendly designed GUI tools could make the system more effective and the active proxy currently based on .Net environment needs to be extended to other platforms such as J2EE.

## References

1. G. Abowd and E. Mynatt, "Charting Past, Present, and Future Research in Ubiquitous Computing," *ACM Transactions on Computer-Human Interaction*, Vol. 7, No. 1, pp 29-58, March 2000.
2. J. Altmann et al., "Context-awareness on mobile devices - the hydrogen approach," *Proceedings of the 36th Annual Hawaii International Conference on 6-9*, pp.10 Jan. 2003.
3. S. Berger, H. Schulzrinne, S. Sidiroglou, and X. Wu, "Ubiquitous Computing Using SIP," *NOSDAV'03*, PP.82-89, June 2003.
4. P. Brezillon, "Using context for supporting users efficiently," *Proceedings of the 36th Annual Hawaii International Conference on 6-9* pp.9 Jan. 2003.
5. R. Hall and H. Cervantes, "Gravity: Supporting Dynamically Available Services in Client-Side Applications," poster paper at *ESEC/FSE 2003*.
6. R. Hauch, A. Miller, R. Cardwell, "Information Intelligence: Metadata for Information Discovery, Access, and Integration," *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 793-798, June 2005.
7. Jeffrey Hightower and Gaetano Borriello, "Location Systems for Ubiquitous Computing," *IEEE Computer*, Volume 34, Issue 8, pp.57-66, August 2001.
8. Eija Kaasinen, "User needs for location-aware mobile services," *Personal and Ubiquitous Computing*, Volume 7, Issue 1, pp. 70-79, May 2003.
9. F. Leymann, "Web Services: Distributed Applications without Limits - An Outline," *Procds. of Database Systems for Business, Technology and Web*, 2003.
10. Jian Li, Ruofeng Tong, MinTang, and Jinxiang Dong, "Web Service-based distributed feature library" *Computer Supported Cooperative Work in Design*, 2004. *Proceedings. The 8th International Conference on Volume 1*, pp. 753 – 758, May 2004.
11. M. Moschgath, J. Hahner, and R. Reinema, "Sm@rtLibrary - An Infrastructure for Ubiquitous Technologies and Applications," *Proceedings of Distributed Computing Systems Workshop 2001*, pp. 208-213, April 2001.

12. M. P. Papazoglou, "Web Services and Business Transactions," *World Wide Web Journal*, vol. 6, March 2003.
13. Tak-Goa Tsuei and Chin-Yang Sung, "Ubiquitous Information Services with JAIN Platform," *Mobile Networks and Applications*, Volume 8, Issue 6, pp. 655-662, December 2003.
14. J. Yang and M.P. Papazoglou, "Service Components for Managing the Life-Cycle of Service Compositions," *Information Systems*, Vol. 29, Issue 2, pp. 97-125, April 2004.

# Study for Intelligent Guide System Using Soft Computing

Woo-kyung Choi, Sang-Hyung Ha, Seong-Joo Kim, and Hong-Tae Jeon

School of Electronic and Electrical Engineering, Chung-Ang University,  
Dongjak-Gu, Seoul, 156-756, Korea  
chwk001@wm.cau.ac.kr

**Abstract.** GPS navigation system has been begun to install to the car since the 1990's. The early system was road guide but it is giving much serviceableness to user because various functions are added by the development of various techniques. However the growth of the most important guide thing of navigation system is yet not conspicuous. In this paper, intelligent guide system that infers information of various recommended road and can guide suitable road to personal tendency was proposed. By using fuzzy logic, it updates user's driving tendency at regular intervals and infers road state. Also path breakaway inference system that learns user's movement path and can secure personal security was proposed by using GPS information.

**Keywords:** Road guide system, Personal security, Fuzzy logic, Neural network.

## 1 Introduction

Modern society expanded to information society based on industrial society. Much technologies with expansion of information society has developed and it has used to human's convenience. Currently, one of technology that gives much assistance in human's life is car navigation system[1][2]. Many cars are installing navigation system and its market is more and more extending. These phenomenon is because navigation provides various functions as well as road guide[3][4]. These functions are because of development of communications network and operation ability of many data that are used to navigation such as CPU. These phenomenon will be more and more expanded and will develop continuously with concept of telematics[5][6]. Present navigation system has been decided by map engine composed from one-side algorithm of company and user's a driving pattern or a mental state has been disregarded. Also developer certainly needs to add individual tendency or character because a lot of electronic equipment is growing with thinking of personalization, characteristic and intelligence. Also personal security is becoming much issue because the society is complicated by fast industrialization and the information age. So we constituted hierarchical fuzzy structure that infer personal driving tendency by using fuzzy logic to accomplish path recommendation method in this paper[7]. And path condition is inferred by fuzzy logic using the information of recommended paths. By comparison of outputs, we proposed intelligent guide system that can infer

suitable path to user. Also personal security system that can inform personal situation using user’s position information was proposed. It separates pattern of movement paths or tracks and it learns user’s life path through user’s position information[8].

## 2 Structure of Intelligent Guide System

System that used in the paper consists of road guide system and personal security system(Fig 1). In the paper, road guide system is composed of module that performs driving pattern inference and road state inference. Personal security system is formed into path learning module and breakaway inference module. The system does easily correction and addition of algorithm because each part is composed by module.

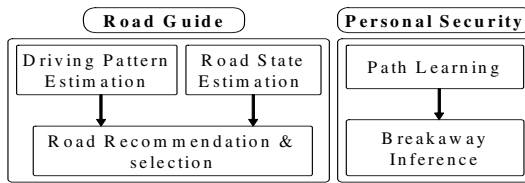


Fig. 1. Structure of intelligent guide system

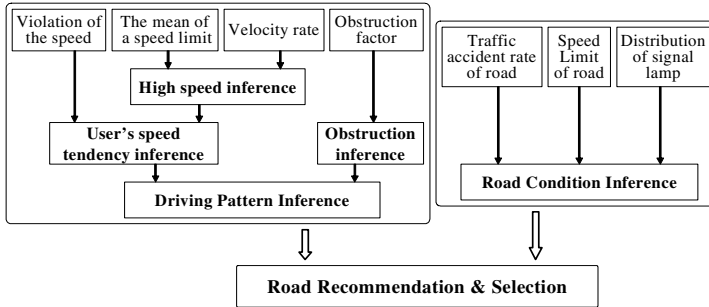


Fig. 2. Structure of Road Guide System

Driving pattern inference modules is composed of speed tendency inference module and obstruction factor inference module. Driving pattern inference module outputs result by using structure of hierarchical fuzzy logic. Road condition inference module gets output from real road environment and geographical information. Because driving pattern continuously change, it is not fixed system but updating system according to pattern attitude of driver. Path learning module consists of algorithm of movement path data collection using GPS, path data scale algorithm and path learning algorithm using neural network. Path data gets movement path from a certain distance and uses vector computation method.

### 3 Algorithm of Intelligent Guide System

The proposed algorithm is composed of driving pattern inference, road condition inference, path learning inference and path breakaway inference.

Driving pattern inference algorithm used user's moving information and map information to learn user's real driving pattern. Input value is the moving time, the speed limit, the movement distance, weight about time, obstruction factor and distribution of signal lamp. Violation of the speed(eq.1), the mean of a speed limit and velocity rate(eq.2), the mean of a speed limit(eq.2), velocity rate(eq.3) and obstruction factor(eq.4) was showed to the next part.  $WT$  is weight about time.  $MS$  is mean of speed.  $Dis$  is Distance.  $OT$  is obstacle number.

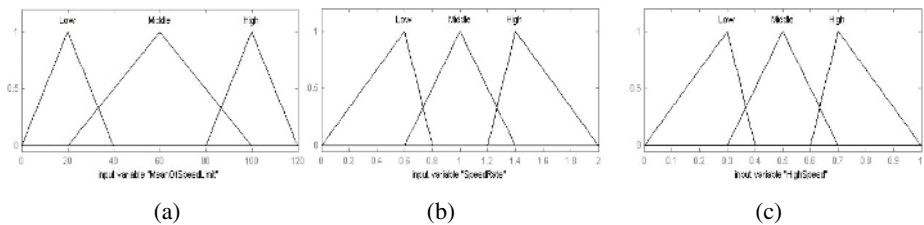
$$Vis(n) = \frac{MS(n)}{MSL(n) + WT_1} \quad (1)$$

$$MSL(n) = \frac{MSL(n-1) \times Dis(n-1) + \Delta MSL_{n-1}^n \times \Delta Dis_{n-1}^n}{Dis(n)} \quad (2)$$

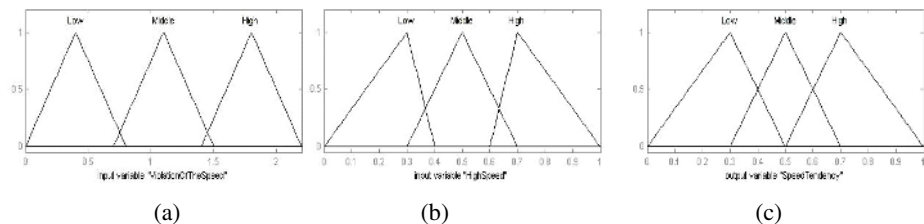
$$VL(n) = \frac{MS(n)}{MSL(n) + WT_2} \quad (3)$$

$$OF(n) = \frac{ON}{Dis(n)} \times 600 \quad (4)$$

Fuzzy reasoning consists of premise and consequent. The premise of the high speed inference is the mean of a speed limit and the velocity rate(Fig. 3). The speed tendency acquires output by the high speed and the violation of speed(Fig. 4). The premise of user's driving pattern reasoning uses the obstruction factor and the speed tendency(Fig 5).



**Fig. 3.** Membership function that used to high speed reasoning: (a) Premise-Speed limit, (b) Premise-The velocity rate, (c) Consequent-High speed



**Fig. 4.** Membership function that used to user's speed tendency reasoning: (a) Premise-High speed, (b) Premise-The violation of speed, (c) Consequent-Speed tendency

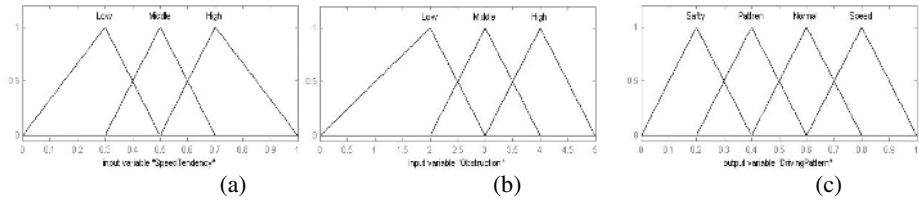


Fig. 5. Membership function that used to user's driving pattern reasoning: (a) Premise-Speed tendency, (b) Premise-Obstruction factor, (c) Consequent-User's driving pattern

Road condition inference module reasons the geographical information of road recommended. The traffic accident rate, the speed limit and the distribution of signal lamp is calculated by information of the traffic accident, the speed limit, the movement distance and the signal lamp number. Road situation is gained by fuzzy reasoning.

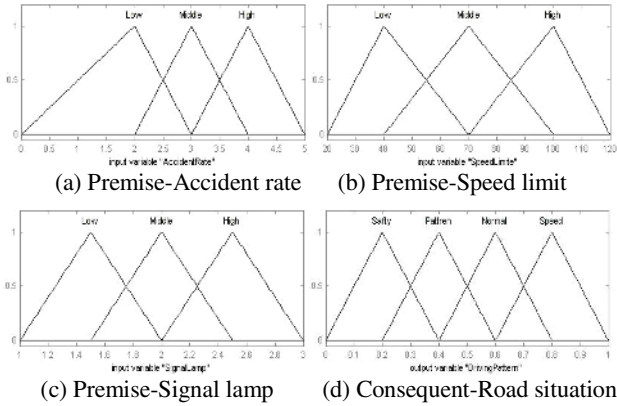


Fig. 6. Membership function that used to road situation reasoning

Paths are learned to get safety from the abduction or the accident by chase of user's driving path. First, data of path learning divided into two types. It is path and non-path. Collection of learning data gains from data variation about distance and uses vector computation method.

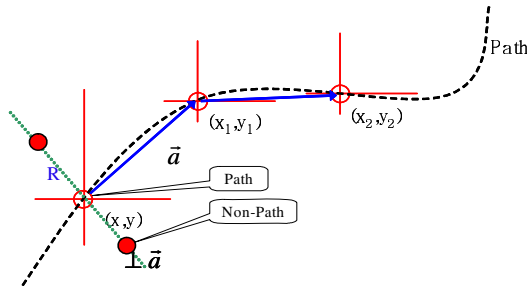


Fig. 7. Gain method of learning data

The personal security system calculates  $\vec{a}$  and computes coordinates  $(a_1, b_1)$  that distant of two point is 'R' in  $\perp \vec{a}$ . The calculated data did scale from '-1' to '1'. Path learning used neural network and back-propagation algorithm[9][10]. Output is path, non-path and undecided.

After learning movement path, the personal security system reasons path break-away by using user's position information. Range of path breakaway reasoning is described in equation 5.

$$\begin{cases} \text{If } O_i \geq 0.7, O_i \text{ is classified to ON} \\ \text{If } O_i \leq 0.3, O_i \text{ is classified to OFF} \\ \text{Otherwise, } O_i \text{ is "undecided"} \end{cases} \quad (5)$$

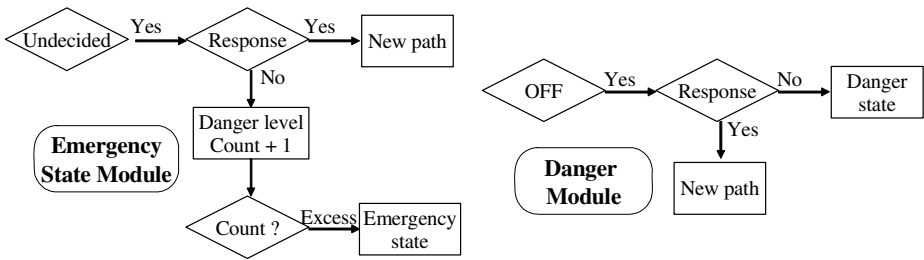


Fig. 8. Flow chart of emergency module and danger module

The path breakaway inference is composed of module of emergency condition and danger condition. The goal of emergency module transmits position information to administrator as user intention for assistance request of other people. When user is not response about breakaway degree and approval, the danger module transmits warning. If user was no response until the last warning, this module considers that user is at danger condition and it has function that informs information and condition to the police.

## 4 Simulation and Result

We performed simulation that applies method proposed by programming. Inputs used information that can easily get to navigation. This system set up data of virtual map for simulation and we set necessary information in the map. Simulation device used PDA(Personal digital assistant) that is portable.

We measured real driving tendency information of user that has various driving pattern and used to simulation. The next process set up virtual map and is process that gain user's driving pattern. This simulator can output driving pattern after it determines and moves path that can get pattern value. User's tendency agreed with simulation result.

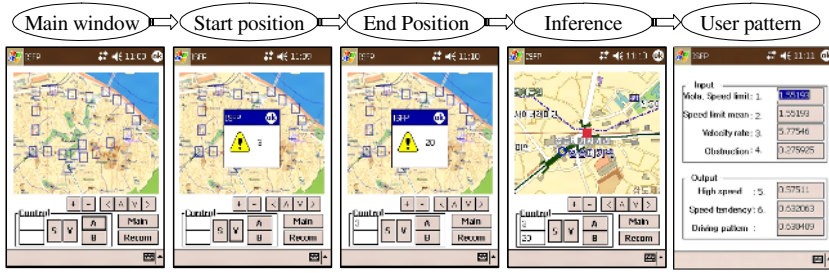


Fig. 9. Simulation for driving pattern inference

If user selects information of a start point and an arrival point, this simulation recommends suitable roads. This simulator evaluates each road from information of recommended roads. It arranges roads as similar ranking of user's pattern from comparison driving pattern value with road condition value. It recommends the last road, after user selects the one of recommended roads. The last road is suitable road at user pattern or road the time required is little.

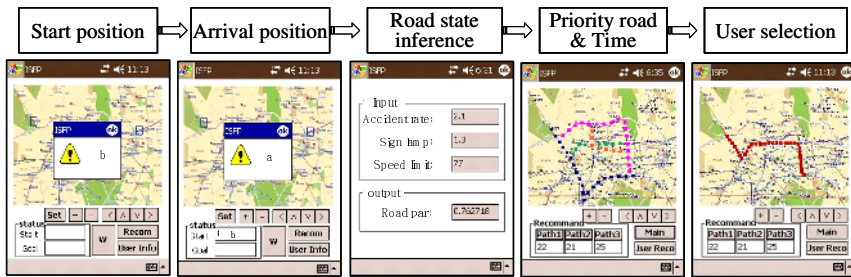


Fig. 10. Simulation of road condition and road recommendation

The data that receive from GPS was learned by neural network through pre-process and data scale. We simulated the road learning, breakaway inference, new path learning and add function of new road using virtual map in PDA.

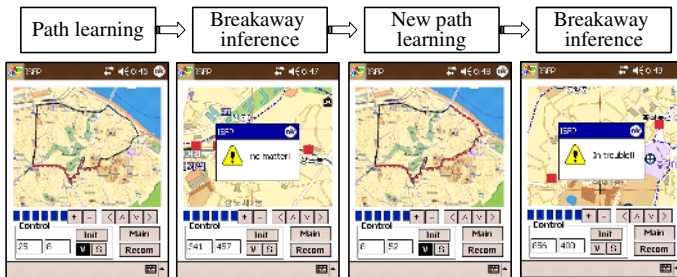


Fig. 11. Simulation for learning and breakaway inference



## 5 Conclusion

In this paper, we proposed intelligent guide system that can recommend new road and can guarantee personal security. Important factor of this paper is recommendation of suitable road that considers the character of driver, the times of driving and the condition of road. Also this study extended range of road selection by recommendation of various roads to user. It learn path of user and can accomplish personal security by reasoning of breakaway. In this paper, proposed algorithm can easily apply to real navigation. Also this system can easily add inputs that need to reasoning process because of modules structure. Addition of input is possible according to development of every technique and this can make more elaborate system. Specially, if car information of telematics is used to this guide system, intelligent guide system can be more necessary to user.

The next work will become study of various algorithms that is necessary at driving. Also it will be study of intelligent guide system that help more drivers by supply of a mental condition, car information, traffic information and weather information.

## References

1. K. Daniel Wong, and Donald C. Cox: Two-state pattern-recognition handoffs for corner-turning situations. *IEEE Trans.* pp. March (2001) 589-594.
2. Peralta, J. O. and de Peralta, "Security PIDS with phtysical sensors, real-time pattern recognition, and continuous patrol," *IEEE Trans.*, pp. 340-346, Nov. 2002.
3. Trahanias, P. E. and Venetsanopoulos, A. N.: Vector directional filter-a new class of multichannel image processing filters. *IEEE Trans.* pp. Oct. (1993) 528-534
4. Ranganathan, N., Vihaykrishnan, N. and Bhavanishankar, N.: A linear array processor with dynamic frequency clocking for image processing applications. *IEEE Trans.* Aug. (1998) 435-445
5. Takeuchi H., Mawatari M. and Tamura M., Shirokane T.: Digital signal processing for home VCR circuitry. *IEEE Trans.* Aug. (1989) 429-435
6. M. Sugeno, M. Nishida: Fuzzy control of model car. *Fuzzy Sets Syst.*, vol. 16. (1985) 103-113, 1985
7. T. Tagaki and M. Sugeno: Fuzzy identification of System and Its Application to Modeling and Control. *IEEE Trans. Syst. Man Cybern*, vol. SMC-15. (1985) 116-132
8. H. J. Ra: Tracking Methods of User Position for Privacy Problems in Location Based Service. *Journal of Korea Fuzzy Logic and Intelligent Systems Society*, Vol. 14. No. 7. (2004) 865-870
9. Chin-Teng Lin and Ya-Ching Lu: A neural fuzzy system with fuzzy supervised learning. *IEEE Trans.* Oct. (1996) 744-763
10. Simon Haykin: *Neural Networks-A comprehensive Foundation* 2nd edition. Prentice Hall. (1999)

# Implementation of a FIR Filter on a Partial Reconfigurable Platform

Hanho Lee and Chang-Seok Choi

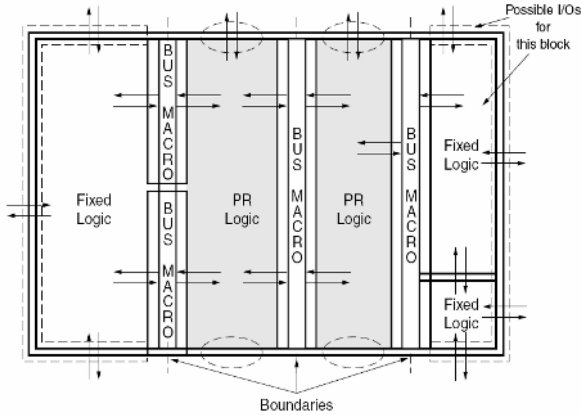
School of Information and Communication Engineering  
Inha University, Incheon, 402-751, Korea  
hhlee@inha.ac.kr

**Abstract.** This paper presents our implemented, synthesized and tested on demand and partial reconfiguration approaches for FIR filters using Xilinx Virtex FPGAs. Our scope is to implement a low-power, area-efficient autonomously reconfigurable digital signal processing architecture that is tailored for the realization of arbitrary response FIR filters on Xilinx Virtex FPGAs. The implementation of design addresses area efficiency and flexibility allowing dynamically inserting and/or removing the partial modules to implement the partial reconfigurable FIR filters with various taps. This partial reconfigurable FIR filter design shows the configuration time improvement, good area efficiency and flexibility by using the dynamic partial reconfiguration method.

## 1 Introduction

The possibility to perform dynamic partial hardware reconfiguration of FPGAs increases their flexibility and ability of run-time adaptation. This feature is provided by Xilinx Virtex II FPGAs, which can be dynamically reprogrammed using the so called ICAP interface. FPGAs provide an array of logic cells that can be configured to perform a given functionality by means of a configuration bitstream. Many of FPGA systems can only be statically configured. Static reconfiguration means to completely configure the device before system execution. If a new reconfiguration is required, it is necessary to stop system execution and reconfigure the device it over again. Some FPGAs allow performing partial reconfiguration, where a reduced bitstream reconfigures only a given subset of internal components. Dynamic Partial Reconfiguration (DPR) allows the part of FPGA device be modified while the rest of the device (or system) continues to operate and unaffected by the reprogramming [1].

This paper presents a partially reconfigurable FIR filter design that targets to meet all the objectives (low-power consumption, autonomous adaptability/reconfigurability, fault-tolerance, etc.) on the FPGA. The FIR filters are special kind of digital filters and have a wide applicability because it has a good characteristic such as linear phase and stability. They are employed in the majority digital signal processing (DSP) based electronic systems. The emergence of demanding applications (image, audio/video processing and coding, sensor filtering, etc.) in terms of power, speed, performance, system compatibility and reusability make it imperative to design the reconfigurable architectures. However, FIR filters may need a large number of coefficients to obtain



**Fig. 1.** Design layout with two reconfigurable modules [3]

the desired specification. This results in the large number of area (slice) for FPGA design. Therefore, there are certain disadvantages associated with run-time reconfigurable design of higher order tap FIR filters using conventional FPGA design techniques. One of the major disadvantages is the so called reconfigurable overhead, which is the time spent for reconfiguration. This depends on the reconfigurable device and the method of reconfiguration. The partial reconfiguration technique can be used in this case since the various taps FIR filters have so many similarities in their structure. Therefore, partial reconfiguration addresses the reduced reconfiguration overhead, coefficient flexibility and area efficiency for higher order FIR filters.

## 2 Module-Based Partial Reconfiguration

Module-based partial reconfiguration was proposed by Xilinx [3][4]. And now many researchers have been proposed many partial reconfiguration methods (JBits, PARBIT, etc) [1][2]. But these methods are difficult to apply real applications because these methods reconfigure the gate-level based. However module-based partial reconfiguration technique can reconfigure the system-level based.

The modular design flow allows the designer to split the whole system into modules. For each module, the designer generates a configuration bitstream starting from an HDL description and going through the synthesis, mapping, placement, and routing procedures, independently of other modules [3]. The modular design flow consists of Modular Design Entry/Synthesis and Modular Design Implementation steps. Modular Design Entry/Synthesis step must be done for top-level design and the modules. Top-level design is designed by the team leader and consists of 'black box' for each sub-modules and 'wiring' for interconnection of each sub-modules.

Modular Design Implementation step comprises following three phase: 1) Initial budget phase, 2) Active module implementation, 3) Final assembly. Module-based partial reconfiguration method is a special case of modular design [3]. This method can

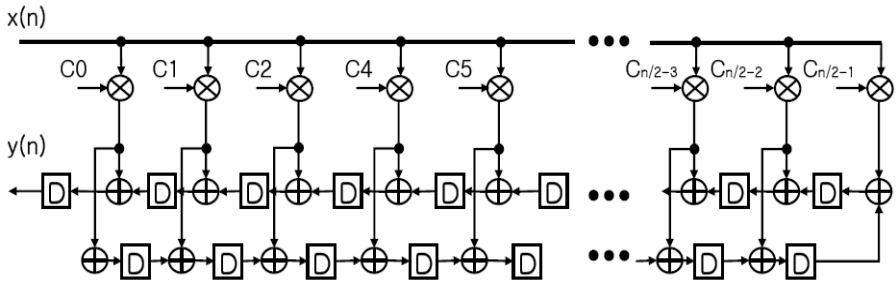
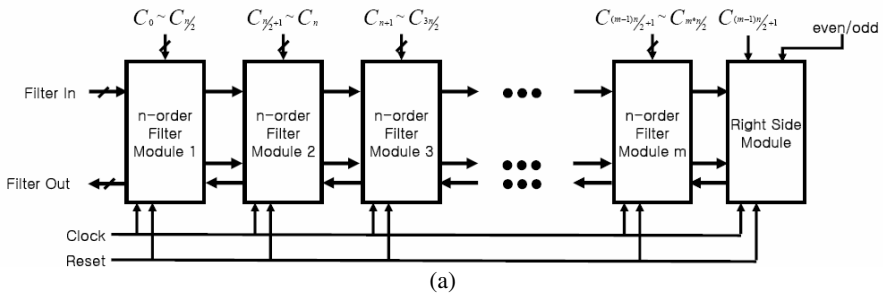
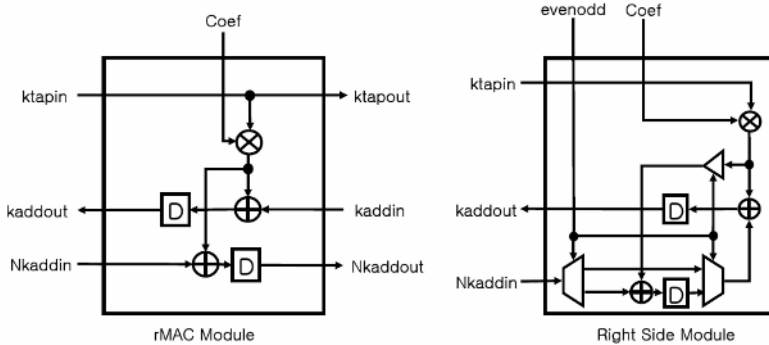


Fig. 2. *n*-tap symmetric transposed FIR filter



(a)



(b)

Fig. 3. Block diagram of (a) partial reconfigurable  $m \times n$  order FIR filter, (b) reconfigurable multiply-accumulate (rMAC) modules

reconfigure only a given subset of internal components during device is activating. A complete initial bitstreams are generated for each reconfigurable module. Fig.1 shows the design layout using partial reconfigurable module. Hardwired bus macros must be included in design for guarantee that each time partial reconfiguration is performed routing channels between modules remain unchanged, avoiding contentions inside the FPGA and keeping correct inter-module connections.

### 3 Reconfigurable FIR Filter Design

The FIR filter computes an output from a set of input samples. The set of input samples is multiplied by a set of coefficients and then added together to produce the output as shown in Fig. 2. Implementation of FIR filters can be undertaken in either hardware or software [5]. A software implementation will require sequential execution of the filter functions. Hardware implementation of FIR filters allows the filter functions to be executed in a parallel manner, which makes improved filter processing speed possible but is less flexible for changes. Thus, reconfigurable FIR filter offers both the flexibility of computer software, and the ability to construct custom high performance computing circuits.

Fig. 3 shows the partial reconfigurable  $m \times n$  order FIR filter, which consists of  $m$   $n$  order filter modules and right side module. These FIR filter is consisted of  $m$  filter modules, which connected by bus macros on FPGA. And each filter module consists of  $n/2$  reconfigurable multiply-accumulate (rMAC) unit, which includes the serial-to-parallel register to get coefficient inputs in serial.

### 4 Implementation

On adaptive systems, a limiting factor for the overall system performance is often the speed of which the system is able to adapt to perform a certain task. This section describes the implementation method of 20-tap FIR filter, which is reconfigured partially from 12-tap FIR filter. The whole system is implemented on a Xilinx Virtex2p30 FPGA device [6].

#### 4.1 HDL Coding and Synthesis

This step is composed to following two phase:

- Top module design:

In this phase, designer must consider each sub-module interconnection, area assignment and bus macro assignment.

- Reconfigurable sub-module design:

This phase is same to traditional HDL design method. But designer must consider input and output assign rule for partial reconfiguration.

#### 4.2 Module-Based Design

- Initial Budget:

In this phase, the team leader assigns top-level constraints to the top-level design. Top-level constraint needs to area constraint and bus-macro assignment. This step is as sequence of top module design. In this step, designer must do bus macro manual setting, sub module area constraint by using floorplanner and top module IOB assignment. Bus macro is limited by target size. Through equation (1), designer can estimate maximum usable bus macro.

$$\text{MaxBus} = 4 * \text{rowCLB} \quad (1)$$

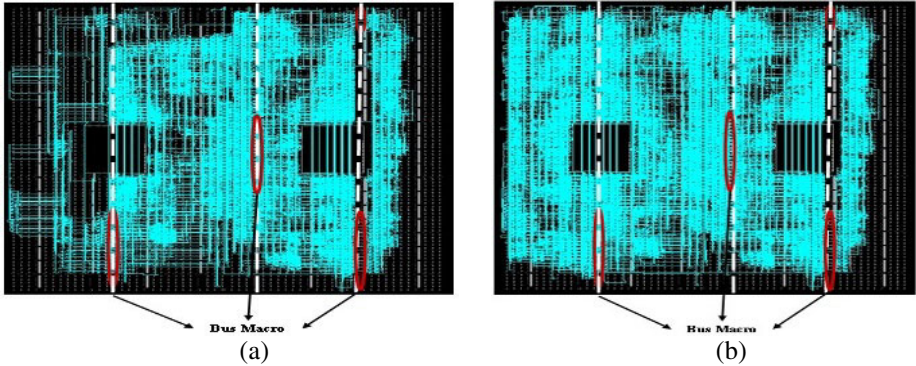


Fig. 4. PAR map of (a) 12-tap and (b) 20-tap FIR filter using DPR

If designer needs area optimization, optimized area can be estimated in a synthesis step. An optimized width equation is described by

$$x = \left[ \frac{\text{slice}}{4 * \text{row}} + 1 \right] \quad (2)$$

where slice is a maximum slice number estimated in a synthesis step and row is a target row size.

- Active module implementation:

In this phase, the team members implement the reconfigurable modules. That is, partially reconfigurable sub-modules are generated by top module and .ucf file. Each sub-module generates a partial bitstream during this step. Fig. 4 shows a post-PAR (placement and routing) diagram. Through  $n$ -order filter module1 is reconfigured to bypass module and module2 is reconfigured to 4-tap module on 12-tap FIR filter while other module is processing, 20-tap FIR filter is composed by partial reconfiguration of module1 showing Fig. 4 (b).

- Final module assemble:

In the phase, the team leader assembles and implements the top-level design using each sub-modules and generates top-module bitstream. That is, designer assembles on system from partially generated modules. All partial modules generated in active module implementation step are combined to the top-level module.

## 5 Experiment and Result

The partial reconfiguration of symmetric transposed FIR filters was implemented on Xilinx Virtex-II FPGA device using test environment shown in Fig. 5 [7]-[9]. XUPV2P FPGA test board and Agilent logic analyzer were used for board level verification. And configuration bitstream download is operated by Xilinx Platform Cable USB and IMPACT. For dynamic partial reconfiguration experiment, the partial reconfigurable module1 and module2 were reconfigured to bypass module and 4-tap two rMAC modules respectively while other areas of modules remain operational. For

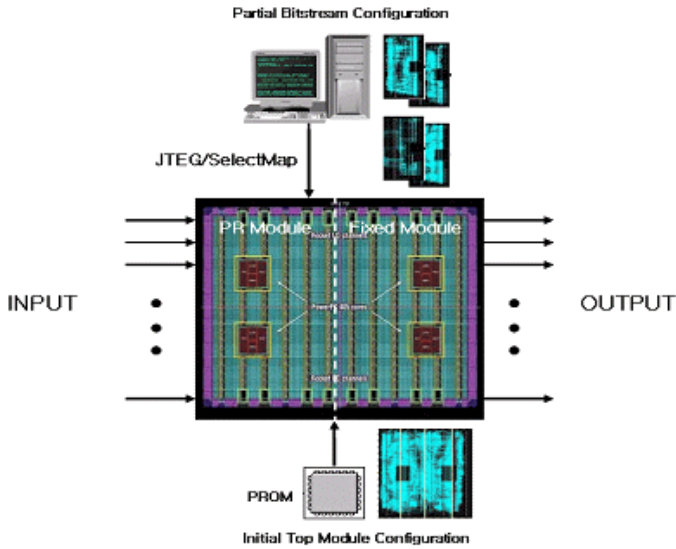


Fig. 5. Test Environment.

verification, we have performed following two methods. First, 12-tap and 20-tap FIR filters before/after partial reconfiguration have been simulated to verify the output results on FPGA test board using Xilinx ChipScope Pro Analyzer. Second, each module has been assigned by identification number such as bypass=00, 2-tap=01, 4-tap=10, 6-tap=11, and then during the partial reconfiguration process the waveform of logic analyzer shows the change of identification number to verify the partial reconfiguration of FIR filter.

Fig. 6 shows the board test result to verify the partial reconfiguration and measure the configuration time. Fig. 6(a) shows the board test result for FIR filter using DPR technique. Because most of modules are operating except reconfigured module, module identification number is changed continuously. After completing DPR, the waveform shows the output change from 3D(111101) to 31(110001). This result

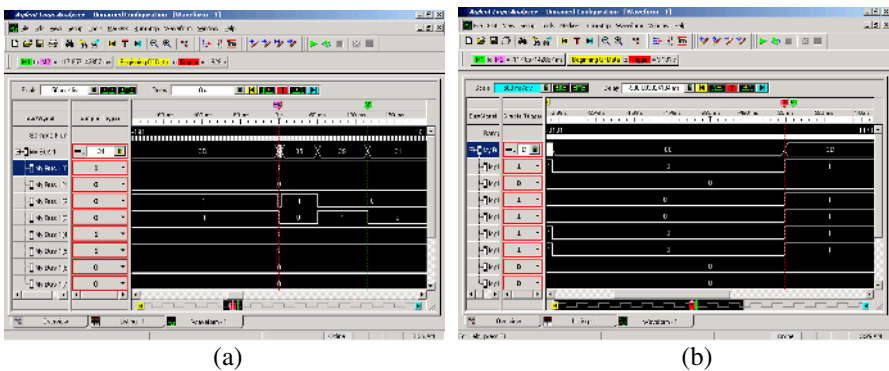


Fig. 6. Board test result of (a) partial reconfiguration, and (b) full reconfiguration

shows that module2 is reconfigured partially from 6-tap three rMAC modules to bypass module. And measured reconfiguration time shows about 112.5 *ms*. Otherwise, Fig. 6(b) shows the board test result in which the full reconfiguration is processed after FPGA reset. Measure reconfiguration time is about 3.05 s. Thus the reconfiguration time of DPR FIR filter is reduced about 1/30 compared to full reconfiguration of FIR filter.

**Table 1.** FPGA device utilization for several FIR filters

	<b>GF</b>	<b>MBF</b>	<b>DPR</b>
<b>Slice</b>	3,058	5,349	4,733
<b>LUT</b>	5,980	9,669	8,427
<b>Equivalent Gate</b>	N/A	76,024	68,063

*GF: General Symmetric FIR Filter*

*MBF: Multiplexer Based Reconfigurable FIR Filter*

*DPR: Reconfigurable symmetric transposed FIR filter using DPR*

*EG: Equivalent Gate Count*

For performance comparison, we have implemented general symmetric FIR filter (GF) using variable multipliers, multiplexer based reconfigurable FIR filter (MBF) and reconfigurable symmetric transposed FIR filter using DPR (DPR). Table 1 shows the utilization of slice, LUT and equivalent gate count after technology mapping. The reconfigurable FIR filter using DPR can save about 11.5% slice compared to the multiplexer based reconfigurable FIR filter, which can be reconfigured to various FIR filter using multiplexer. Compared to the general symmetric FIR filter (GF), the slice number in reconfigurable FIR filter using DPR method was increased about 54% because of adding bus macro, serial-to-parallel register and a little controller. But if we want to change 2, 4, 6-tap rMAC modules in general symmetric FIR filter, the full reconfiguration must be needed and required long reconfiguration time. However, reconfigurable FIR filter using DPR method requires the partial reconfiguration of about 1,461 slices out of total 4,733 slices for 2-tap one rMAC module or 6-tap three rMAC module, which adds flexibility allowing dynamically inserting and/or removing the coefficient taps.

## 6 Conclusion

This paper discusses a partial reconfigurable FIR filter design approach using dynamic partial reconfiguration. This approach has area efficiency, flexibility and fast configuration time allowing dynamically inserting and/or removing the part of modules. The proposed reconfigurable FIR filter design method produces a reduction in hardware cost compared to multiplexer-based reconfigurable FIR filter and allows performing fast partial reconfiguration, where a reduced bitstream reconfigures only a given subset of internal components. In the future, self-reconfigurable hardware



platform using microcontroller unit and configuration memory will be promising solution for automatic partial reconfiguration of digital circuit in the run-time environment.

## Acknowledgement

This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program.

## References

- [1] D. Mesquita, F. Moraes, J. Palma, L. Moller, N. Calazanas, "Remote and Partial Reconfiguration of FPGAs: tools and trends," International Parallel and Distributed Processing Symposium, April 2003.
- [2] A. K. Raghavan, P. Sutton, "JPG-A partial bitstream generation tool to support partial reconfiguration in Virtex FPGAs," Proc. of the International Parallel and Distributed Processing Symposium, 2002.
- [3] Xilinx Inc., "XAPP 290: Two flows for Partial Reconfiguration: Module Based or Difference Based," www.xilinx.com, Sept 2004.
- [4] Xilinx Inc., "Development System Reference Guide," www.xilinx.com.
- [5] Uwe Meyer-Baese, *Digital Signal Processing with Field Programmable Gate Arrays*, Springer, 2001.
- [6] Hyuk Kim, *Real Xilinx FPGA World*, Ant Media, Oct. 2003.
- [7] Xilinx Inc., "Virtex configuration architecture advanced user's guide," Oct. 2004.
- [8] Philippe Brutel, "Managing Partial Dynamic Reconfiguration in Virtex-II Pro FPGAs," Xcell Journal, Xilinx, Fall 2004.
- [9] Xilinx Inc., "Xilinx University Program Virtex-II Pro Development System Hardware Reference Manual," March 2005.

# A Context Model for Ubiquitous Computing Applications

Md. Rezaul Bashar, Nam Mi Young, and Phill Kyu Rhee

Intelligent Technology Laboratory, Dept. of Computer Science & Engineering  
Inha University, Incheon 402-751, Republic of Korea

{bashar, rera}@im.inha.ac.kr, pkrhee@inha.ac.kr

**Abstract.** Now these days, much research has been carried out on context-aware computing that is an important field in the area of ubiquitous computing, which offers a pervasive vision to implement a smart system by connecting computers, sensors and other peripherals in wired or unwired fashion. The main focus of this paper is on context modeling to design a real-time face recognition system for ubiquitous computing. In this research, a real-time framework with the combining concepts of context-awareness and genetic algorithm referred as real-time genetic algorithm (RGA) is proposed that meets the characteristics of context model and developments of a ubiquitous application. This framework is implemented on a real-time environment and a recognizable success is notified.

## 1 Introduction

Ubiquitous Computing [1][2][3], Uicom in the shortest form integrates computation into the environment, rather than having computers which are distinct objects. Pervasive computing is the another term for ubiquitous computing. Computer Scientists hope that embedding computation into the environment would enable people to move around and interact with computers more naturally than they currently do. This system dynamically adapts to the needs of the user and to the current operational context. Thus, ubiquitous system would exhibit a high level of adaptive behavior in order to respond to a wide range of contexts and stimuli that may be difficult to define a priori in a comprehensive manner. In addition, the context and input stimuli experienced by adaptive software applications and the effectiveness of the adaptive behavior they exhibit are highly dependent on the user's involved and the environmental context. One of the goals of ubiquitous computing is to enable devices to sense changes in their environment and to automatically adapt and act on these changes based on user needs and preferences.

Context [9] is information that can be used to characterize the situation of a real world entity, basically location and identity. Due to the rapid development of sensors, it is easy to capture an entity to sense and use context. Context-awareness [6, 9] is the system that takes the advantages of context. To make understandable about a context-awareness system, it is important to make a context model [9]. In this research, a context model is proposed and related experiments establish the developed concepts.

## 2 A Context Model

A context model represents a designer's understandings and feelings on context that is the organization of the physical world's data into logical structure, the key idea to

implement in an artificial way, and a foundation for the implementation of context-awareness function. Objects are the basic elements of a model and these objects are described with logical or physical entities and relations between them. The proposed comprehensive model (shown in Fig.1) is composed with four layers i.e, sensor layer, core layer ,context layer, and application layer. Each layer uses its own data of different type from its previous level.

(a) Sensor layer: These data are produced from the environment through a set of sensors. The output set of sensor layer is denoted by

$$Sd = (S_{t,1}, S_{t,2}, \dots, S_{t,n}), Sd_{t,n} = (Sen_n, t, h_n)$$

Where,  $Sd_{t,n}$  denotes the sensor data of sensor  $i$  in time  $t$ ,  $Sen_n$  denotes the value of  $n$ th sensor and  $h_n$ , the hidden parameter, is a value between 0 to 1 to model uncertain characteristic of context. The hidden value is assigned to each kind of sensors.

(b) Core layer: As this is the basic part of this system because the performance of the entire system depends on this, it is referred as core layer. Data from Sensor layer are treated as input to this core layer. The output of core dataset is defined as

$$Cd = (Cd_{t,1}, Cd_{t,2}, \dots, Cd_{t,m}), Cd_{t,i} = (\tau_{i,t}, t, h_i).$$

Where,  $Cd_{t,i}$  denotes a semantic piece called core data at time  $t$ . Each sensor corresponds to a core data.  $\tau$  denotes an assertion that is to retrieve a fact from the sensor data and the fact cannot be divided to more trivial parts, such as Person (Kim) and Element (Spectacles). Sometimes AND, OR and NOT operations are performed to construct more complicated facts.

(c) Context Layer: In this layer, context situation is identified based on core data. It makes cluster for recognition and makes sense to the software system. Context situation is defined as

$$S = (S_1, S_2, \dots, S_p), S_i = (Cd_i, Ser_i, p_i).$$

Where,  $Cd_i$  denotes the core data, which constitutes the situation,  $Ser_i$  denotes the services for the  $Cd_i$ , and  $p_i$  stands for priority for each context situation that is predefined.

(d) Application Layer: This layer determines the action to be performed on the basis of context situation and services. The application to be executed is determined as

$$A = (A_1, A_2, \dots, A_n), A_i = (S_i, Mod_i, trig_i)$$

Where  $S_i$  is the context situation,  $Mod_i$  is the program module to take action, and  $trig_i$  is the triggering for the application.

The design of the comprehensive context model is shown in Figure 2, where three databases are used for element collections, context ontology and actions. Element collection is proposed because of the system's real-time application, as if there new element is detected, it is added to the database.

Context ontology deposits the context category for a new incoming context and keeps the track of context. As a result, every context is at least one entry in this table. It is used for real-time adaptations while the sensor detects system changes with time, contexts are classified and make a entry into the table.

Action database, an off-line database, stores the probable actions that are specified by this system. A look-up table implements this to reduce the execution time.

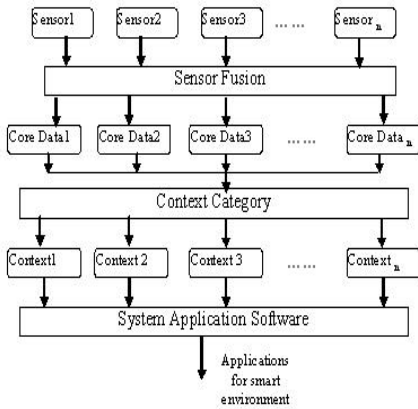


Fig. 1. A context model

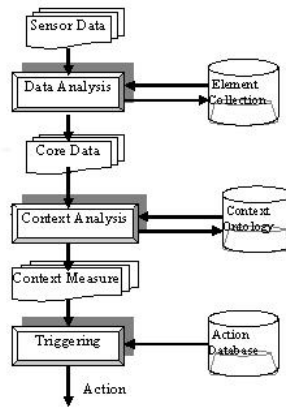


Fig. 2. Design paradigm of the conceptual context model

### 3 Proposed Experimental Environment

The proposed real-time framework of the genetic algorithm is suitable for adaptive, ubiquitous and pervasive applications with the capability of real time adaptation called real-time genetic algorithm (RGA). The designed real-time genetic algorithm operates in two modes: the evolutionary mode and the action mode. In the evolutionary mode, it accumulates its knowledge by exploring its application environments, while in the action mode; it performs its designated task using the accumulated knowledge from the evolutionary mode. The evolutionary mode is either online or offline adaptation. For offline adaptation, environmental context is categorized according to some predefined characteristics (here, illumination) and genetic algorithm is used for learning. For online adaptation, when a new context is encountered, it directly interacts with the action mode. Whenever an application environment changes, the system accumulates and stores environmental context knowledge in terms of context category and its corresponding action.

#### 3.1 Environmental Context-Awareness

Environmental context-awareness is carried out using environmental context data that is defined as any observable and relevant attributes and its interaction with other entities and/or surrounding environment at an instance of time [6].

For identifying and categorize environmental context data, Fuzzy Adaptive Resonance Theory (FART), a variation of first generation ART [ 6,7] algorithm is adopted. First ART, named ART1, works with binary inputs, while FART is a synthesis of the ART algorithm and Fuzzy operators that (FART) allows both binary and continuous input patterns [6]. The image space of object instance with varying illuminations must be clustered properly so that the location error can be minimized. However, the classification of images under varying illumination is very subjective and ambiguous. Thus, FART method, which shows robustness in subjective and ambiguous applications in order to achieve optimal illumination context clustering, is preferred for adaptation.

The performance of clustering is improved by observing previously clustered data repeatedly [6].

### 3.2 Proposed Real-Time Genetic Algorithm (RGA)

FART has its capability for incremental learning that introduces clustering for real-time system in a dynamic environment. For real-time learning, As with the usual work for separating environmental context, FART looks for an unknown type of cluster, if it finds, it makes a new cluster as shown in Fig.3.

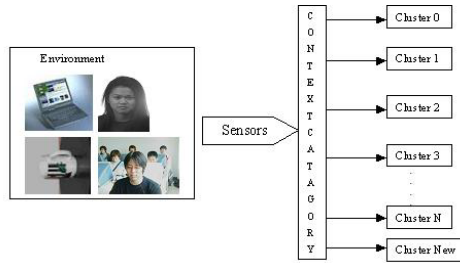


Fig. 3. On-line learning using FART

RGA consists Context-Category Module (CCM), Evolution Control Module (ECM), Adaptive Action Module (AAM), and Context Knowledge Base (CKB) as shown in Fig 4. CCM identifies a current context using environmental context data. AAM consists of one or more action primitives, which can be heterogeneous, homogeneous, or hybrid operational entities. For example, the action primitives of a pattern classifier are divided into preprocessing, feature representation, class decision, and post processing primitives. ECM searches for the best combining structure of action primitives for an identified context. The structures of optimal actions are stored in the CKB with the corresponding context expression.

Initially, the system accumulates the knowledge through off-line evolution (Fig 4. a) to the CKB that guarantees optimal performance for individual identified context. The CKB stores the expressions of identifiable contexts and their matched actions that will be performed by the ECM. The matched action can be decided by either experimental trial-and-error or some automating procedures. In the operation time, the context expression is determined from the derived context representation, where the derived context is decided from the context data.

The adaptive task is carried out using the knowledge of the CKB evolved in the evolutionary mode and then action mode is performed. For on-line evolution, when a new context data is found, it generates a new category and updates the CKB as shown in Fig 4.b. The detail process for RGA is as follows:

- Step1. Environmental contexts are clustered by CCM using FART.
- Step2. The action configuration structure of the AAM is encoded as the chromosome and the fitness of GA is decided.

Step3. GA decided the most effective subset of action configurations using the associated training action data.  
 Step4. The chromosome with their associated contexts (either trained or new context) is stored at CKB.

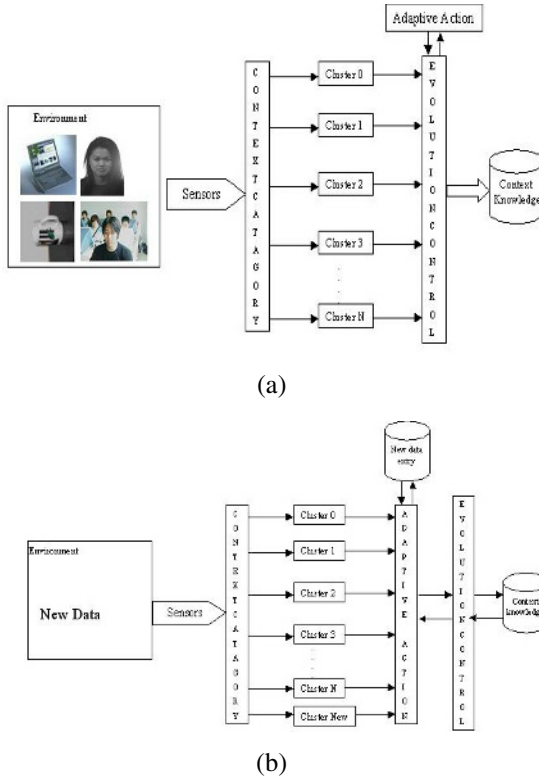


Fig. 4. Block diagram of proposed RGA (a) off-line evolution, and (b) on-line evolution

## 4 Design Example

The proposed framework is applied in the field of visual information processing i.e face recognition. Face images with different illumination are preferred for this experiment due to its spatial boundaries so that it is easy to distinguish among the environmental contexts. In this research, FART constructs clusters according to variation of illumination as shown in Fig5. The AAM of RGA consists of three stages: preprocessing, feature extraction and classification [8]. The action primitives in the preprocessing steps are histogram equalization, contrast stretching and retinex [8]. The action primitives in the feature extraction stage are PCA and Gabor representation [8] and finally cosine distance measurement is concerned for classification.



Fig. 5. Example of face images clustered by FART on different illumination

## 5 Experimental Results

In this experiment, FERET face image dataset with its normal illumination *fafb* and bad illumination *fafc* are used for making artificial environmental contexts. Both real-time and non real-time GA based face recognition experiments are carried out. FART has constructed 9 types of cluster for non-RGA and 13 types for RGA and hybrid vectorization method is applied for clustering. Fig.6 shows the result for non-RGA technique.

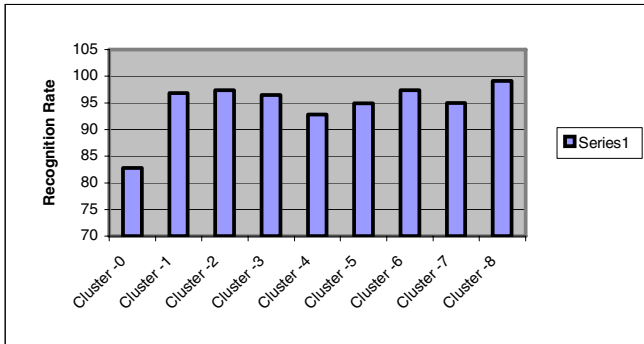


Fig. 6. Performance of face recognition with non-RGA

Fig. 7 describes the recognition rate as well as recognition ratio between Real-time and non real-time GA. Initially the system has accumulated knowledge from the environmental context through offline evolution and it produces more than 96% accuracy, however, when a lot of context categories are present, it takes comparatively more time for evolution, as a result the recognition rate decreases. Gathering knowledge from offline evolution, the on-line evolution starts and for some times it achieves better performance than previous offline system. After some times, as the number of contexts increases, the recognition rate decreases, while the evolution is finished, it receives the highest recognition rate. In this figure, off-line evolution is shown up to

position A (0-6 cluster), on-line evolution starts from cluster 6 and it produces upward recognition rate up to cluster 9. From cluster 9, it decreases its accuracy up to cluster 12. And finally, at the finishing point of the evolution, it produces maximum recognition rate due to less number of context.

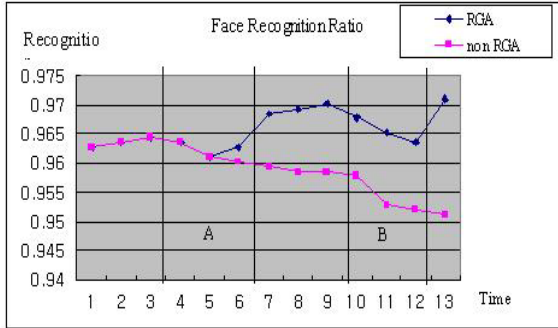


Fig. 7. Face recognition rate for RGA with respect to time

## 6 Conclusion

This paper proposes a comprehensive model for context-awareness to make a ubiquitous environment. In this research, a four-layered context aware model is introduced and this concept is applied to a robust face recognition scheme using real-time genetic algorithm. The important feature of the proposed model is the context data i.e to make categorize the environmental data and illustrated experiments bring out this categorization very successfully. Although multiple sensors are proposed, experiments are carried out on a single sensor and our future experiments will contribute on multiple sensors and sensor fusion. The developed system produces highly robust and real-time face recognition system on different illumination categorized images with the technique of adaptation by real-time genetic algorithm. This research also establishes a new concept for adaptive real-time genetic algorithm that reduces the execution time of traditional genetic algorithm with higher performance and this makes a notable contribution for ubiquitous computing.

## References

1. H. Liu et. al.: Illumination Compensation and Feedback of Illumination Feature in Face Detection. Proc. International Conferences on Information-technology and Information-net, Beijing, vol. 3,(2001) pp.444-449.
2. Jinho Lee, etc.: A Bilinear Illumination Model for Robust Face Recognition. 10<sup>th</sup> IEEE International conference on Computer Vision (ICCV'05), 2005.
3. Marios Savvides et.al.: Corefaces- Robust Shift Invariant PCA based Correlation Filter for Illumination Tolerant Face Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), 2004.
4. Laiyun, et. al.: Face Relighting for Face Recognition Under Generic Illumination. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'04), 2004.



5. Haitao Wang, Stan Z. Li, et. al.: Illumination Modeling and Normalization for Face Image. IEEE International workshop on Analysis and Modeling of Faces and Gestures (AMFG'2003), 2003.
6. Phill Kyu Rhee, et, al.: Context-Aware Evolvable System Framework for Environment Identifying Systems. Knowledge-Based Intelligent Information and Engineering Systems, KES2005.
7. B.D. Ripley.: Pattern Recognition and Neural Networks. Cambridge University Press, 1997.
8. Mi Young, et. al.: Hybrid Filter Fusion for Robust Visual Information Processing. Knowledge-Based Intelligent Information and Engineering Systems, KES2005.
9. Jie Sun, and ZhaoHui Wu.: A Comprehensive Context Model for next generation ubiquitous computing applications. IEEE conference on Embedded and Real-Time Computing Systems and Applications (RTCSA'05), 2005.

# Adaptive Classifier Selection on Hierarchical Context Modeling for Robust Vision Systems

SongGuo Jin, Eun Sung Jung, Md. Rezaul Bashar, Mi Young Nam,  
and Phill Kyu Rhee

Dept. of Computer Science & Engineering., Inha University  
253, Yong-Hyun Dong, Nam-Gu  
Incheon, South Korea

{sgkim, eunsung, bashar, rera}@im.inha.ac.kr, pkrhee@inha.ac.kr

**Abstract.** This paper proposes a hierarchical image context based adaptable classifier ensemble for efficient visual information processing under uneven illumination environments. In the proposed method, classifier ensemble is constructed in two stages: i) it distinguishes the illumination context of input image in terms of hierarchical context modeling and ii) constructs classifier ensemble using the genetic algorithm (GA). It stores its experiences in terms of the illumination context hierarchical manner and derives artificial chromosome so that the context knowledge can be accumulated and used for identification purpose. The proposed method operates in two modes: the learning mode and the action mode. It can improve its performance incrementally using GA in the learning mode. Once sufficient context knowledge is accumulated, the method can operate in real-time. The proposed method has been evaluated in the area of face recognition. The superiority of the proposed method has been shown using international face database FERET.

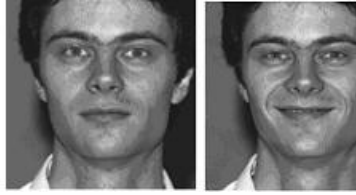
**Keywords:** context awareness, face recognition, classifier ensemble, evolvable classifier selection, hierarchical context modeling, and genetic algorithm.

## 1 Introduction

The aim of this paper is to improve the accuracy of vision system using hierarchical context modeling and adaptive classifier ensemble. The difficulty of object recognition can be understood by observing Fig. 1 where the same person looks different under varying environments. Efficient interpretation or understanding of images in vision systems frequently relies on some priori knowledge. Several vision systems using prior knowledge have been proposed in the past. Most of them require knowledge bases specifically designed for the application domain, and designing such knowledge bases requires very time consuming process.

In this paper, we focus on the variation of illumination in input images referred as image context that affect the performance of object recognition. However, the proposed method can be readily applied or extended to other kinds of variations. It constructs the most effective structure of classifier system for individual illumination environmental ontology, called hierarchical context modeling, using unsupervised learning method.

A general classifier system for object recognition can be thought to having multiple stages. Each stage consists of several competitive action primitives. The action primitives are basic functional elements and the same functional elements with different behavior by changing parameters, threshold, etc. are treated as different action primitives. Simple model for object recognition can be divided into three stages: preprocessing, feature representation, and class decision. The preprocessing action primitives are histogram equalization, end-in contrast stretching, Retnix, Hormomorphic filter, etc. The feature representation action primitives are PCA [22], FLD [22], Gabor wavelet [11], etc. The class decision action primitives are Bayesian classifier, neural network, SVM, etc. General Classifier selection scheme can be thought to select action primitive and associated parameters in each stage, and produces an efficient classifier system for a given operational environment.



**Fig 1.** The face images of the same person with different facial expressions

The framework of classifier ensemble can be formalized as follows. Let  $PP = \{pp_1, pp_2, \dots, pp_s\}$  be a set of preprocessing action primitives. Let  $FR = \{fr_1, fr_2, \dots, fr_t\}$  be a set of feature representation action primitives. Let  $CD = \{cd_1, cd_2, \dots, cd_n\}$  be a set of class decision action primitives. Assume that there is  $k$  possible classes  $= \{\omega_1, \omega_2, \dots, \omega_k\}$ . Let  $R^n$  be an input space. The input of each classifier system is represented by input vector  $x \in R^n$ , i.e.  $x = \{x_1, x_2, \dots, x_n\}^T$ . Individual classifier system assigns class label from  $\Omega$  to an input vector  $x$ . That is,  $CLS_i : R^n \rightarrow \Omega$  with  $i = 1, \dots, c$ . The set of all possible classifier systems are expressed as follows:

$$CLS = PP \times FR \times CD \quad (1)$$

Total  $k = s \times t \times u$  classifier systems can be produced. For example, a classifier can be denoted as follows.

$$CLS_i = [pp_j, fr_k, cd_l] \quad (2)$$

The output of classifier  $CLS_i(x)$  can be represented as a vector in the following.

$$CLS_i(x) = [O_{i,1}(x), O_{i,2}(x), \dots, O_{i,c}(x)], \quad (3)$$

where  $O_{i,j}(x)$  is the output derived from  $CLS_i$  using the input vector  $x$ .

In this paper, we present a novel classifier selection method by introducing the novel concept of hierarchical image context model to achieve high efficient object recognition. It distinguishes the illumination variations of input image using unsupervised learning method repeatedly and derives a hierarchical illumination image category, called hierarchical context modeling. It constructs a classifier system for each illumination category for effective exploration of the GA search space of various classifier systems. Classifier system structure is encoded in terms of artificial chromosome, called action reconfiguration chromosome. GA is used to explore a most effective classifier system structure for each identified data context category. The proposed method adopts the novel strategy of context knowledge accumulation. The knowledge of an individual context category and its associated chromosomes of effective classifiers are stored in the context knowledge base. Similar research can be found by [7]. Once sufficient context knowledge is accumulated, the method can react to such variations in real-time.

## **2 Hierarchical Context Modeling and Adaptive Classifier Selection**

In this session, we will discuss about context knowledge modeling that is derived from context data using unsupervised learning method. Context data is defined as any observable and relevant attributes that affect system behavior, and its interaction with other entities and/or surrounding environment at an instance of time [15]. We limit the context data as the variation of images due to illumination change. It can be, however, extended to various configurations, computing resource availability, dynamic task requirement, application condition, application related environment, etc. [15].

### **2.1 Hierarchical Context Modeling and Adaptive Classifier Selection**

In this research, Context-awareness is carried out based on the hierarchical context modeling and identification of context data. Input context data need to be identified (context identification), and used to validate a most effective classifier for a given action data. Thus, context data should be modeled in association with input action data as much as possible. Since there is no direct way to find context modeling, we used unsupervised learning method repeatedly for each context. The resulting context cluster hierarchy is called hierarchical context model. The proposed method controls the classifier system selection (reconfiguration) based on the identified data context category. Context modeling clusters context data set into several data context categories. Context modeling can be performed by an unsupervised learning algorithm such as SOM, Fuzzy Art, K-means etc. [16, 17]. Context identification is to determine the context category of a given context data [17]. Context identification can be carried out by employing a normal pattern classification method such as NN, K-nn, SVM, etc.

### **2.2 Context Knowledge Accumulation Using GA**

Context knowledge describes a trigger of system action (combined classifier system output) in association with an identified context stored in the context knowledge base

over a period of time [8]. Initially, the proposed method learns and accumulates the knowledge of context-action configuration chromosome associations, and stores them in the CKB. The knowledge of context-action association denotes that of a most effective classifier system for an identified context. The AM configures the action configuration chromosome and our method decides the fitness of the GA.

The proposed method adopts the novel strategy of context knowledge accumulation. The knowledge of an individual context category and its associated chromosomes of effective classifier systems are stored in the context knowledge base. In addition, once the context knowledge is constructed, the system can react to changing environments at run-time [17].

### 2.3 Adaptive Classifier Selection Using Hierarchical Context Modeling

The proposed Evolvable Classifier Selection (ECS) method uses two types of data: context data and action data as inputs. The action data, denoted by  $\mathbf{x}$ , is a normal data being processed. The context data, denoted by  $\mathbf{y}$ , is used to identify a data context of  $x$ , the normal input. The proposed method controls classifier selection based on the identified data context. Action data itself can be used as context data. We assume that the context data can be modeled in association with the input action data. We need identify a data context (category) firstly, and select a best classifier system based on the data category. The classifier selection is formalized as follows.

$$ECS(x, y) = CKO(y)(CLS_1(x), CLS_2(x), \dots, CLS_k(x)), \quad (4)$$

where,  $CKO$  is a context-aware knowledge operator.  $CKO$  selects the best classifier from total  $k$  classifier systems. The implementation of  $CKO$  can be done by some learning method and hierarchical context knowledge base. We use a single input image as both context and action data in this paper. The proposed method consists of the context identification module (CIM), the action control module (ACM), the action module (AM), the evolution control module (ECM), and the context knowledge base (CKB) (see Fig. 2).

The proposed method tries to distinguish the data characteristics of input image (data contexts) and selects a classifier system accordingly using the genetic algorithm (GA). It stores its experiences in terms of the data context category and the artificial chromosome, called action configuration chromosome, so that the context knowledge can be accumulated and used later. Each chromosome represents the encoding of the structure of an optimal AM for corresponding data context category. Data context is identified by the CIM. The ACM searches for a best combining structure of action primitives (i.e. classifier system) for an identified data context. The action configuration chromosomes of optimal actions are stored in the CKB with the corresponding data context category. Our method evolves itself by accumulating the knowledge which classifier system guarantees an optimal performance for each identified data context. The ECM manages the evolution. It operates in two modes: the evolution mode and action mode, and the details are described in the following sub section.

### 2.4 The Action Mode

In the action mode, data context is identified by the CIM. The ACM searches the action configuration chromosome for the identified data context in the CKB (see Fig. 2). The AM is reconfigured by the ACM, if necessary. Then, the reconfigured AM performs its task using the action data, and produces the response of the ECS. Whenever the ACM identifies that the data context is changed, the system reconfigures the AM. If our method measures the performance being fell down below a predefined criterion, our method activates the evolution mode, or it may evolve the system periodically.

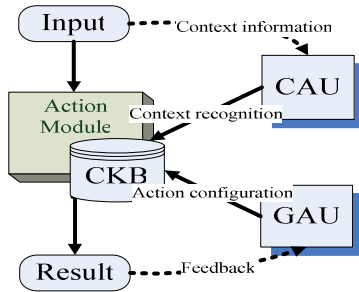


Fig. 2. Block diagram of the proposed architecture for object recognition

### 2.5 The Evolution Mode

In the evolution mode, the training data are clustered into data context categories by the CIM (see Fig. 2). The training data of each data context category are used to accumulate the knowledge of the action configuration chromosome of the AM. The evolution process is controlled by the ECM, and the details will be discussed in the next session. The determined action configuration chromosome with its corresponding data context category is stored in the CKB.

## 3 Object Recognition Using ECS Method

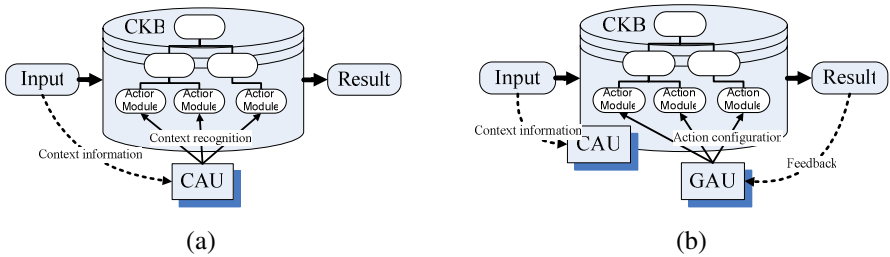
The proposed ECS method has been tested in the area of object recognition. We employ the ECS strategy where the face recognition system structure is allowed to evolve in accordance with changing quality of input image data, i.e. data context. The data sets Inha, FERRET, and Yale were used in our experiments.

### 3.1 The Design of ECS Based Face Recognition

The AM of ECS based face recognition consists of three stages: the preprocessing, feature representation, and class decision. Preprocessing is performed for providing stable quality images as much as possible for face recognition. The action primitives employed for preprocessing stage here is the histogram equalization, the Retnix, and the end-in contrast stretching [17]. We adopt Gabor vectors [17] with different weight values of individual fiducial points as the action primitives of feature representation.

For the simplicity, we adopt non-parametric classification method k-nn's with different threshold values as the action primitives of the class decision stage. The architecture of face recognition using the proposed method is shown in Fig. 3.

In hierarchical context based face recognition, the input images are used as the action data as well as the context data. We assume that the training set of input face image is provided. In the action mode, and the data context category of input face image is identified by the CIM. If the data context category is the same as the previous one, the ACM activates the AM. Otherwise, the ACM gets the action reconfiguration chromosome from the CKB, and the AM reconfigures itself using the information in the chromosome. In other words, a classifier system that is optimized for the data context of the input image is selected (or reconfigured). Finally, the AM produce the recognition result using the action data (input face image).



**Fig. 3.** Two modes of hierarchical context based face recognition: (a) the flow diagram of the action mode, and (b) the flow diagram of the evolution mode.

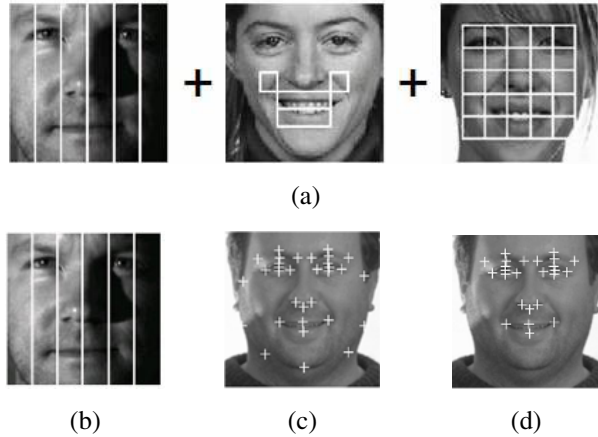
In the evolution mode, the CIM clusters (models) face data images into several data context categories. The details of the data context modeling and identification will be discussed in the next sub-session. Each cluster denotes one data context category, and the ACM generates the corresponding action reconfiguration chromosome for each data context using the face images in each cluster, respectively.

### 3.2 Hierarchical Illumination Context Modeling

We used face image of  $128 \times 128$  spatial resolution and 256 gray levels as input context data  $\mathbf{y}$  here (see the session 3.1). Three methods of derived context feature were investigated here as shown in Fig. 4. In the vertical scanning method, the input context data, i.e.  $128 \times 128$  face image is reduced into  $6 \times 6$  images, and the reduced image is scanned in the vertical direction first (from top to bottom) and from left to right. Then, we generated dcf's [17] 1-D vector with 36 elements. The horizontal scan is carried out in a similar way except that the first scanning direction is horizontal. In the hybrid scan, three different scans are concatenated into 1-D dcf vector with 36 elements.

Next step, dcf's is clustered into several context categories in order to assign a data context category to each face images. Several types of unsupervised learning methods have been investigated for constructing the context model. However, SOM [22] is selected to be the most promising algorithm. SOM can be used to create an intuitive model of the important concepts contained in information [15, 16]. After a sufficient

number of input vectors have been presented, network connection weights specify clusters, the point density function of which tends to approximate the probability density function of the input vectors. In addition, the connection weights will be organized such that topologically close nodes are sensitive to inputs that are similar. Fig. 5 shows hierarchical context model generated by the hybrid scan where images are categorized according to environmental situation.



**Fig. 4.** Facial geometry representation by hybrid scanning (a), vertical scanning (b), coordinates of facial feature points (c) and facial feature points excepting face edges (d)



**Fig. 5.** Hierarchical context modeling using hybrid scanning

## 4 Experimental Results

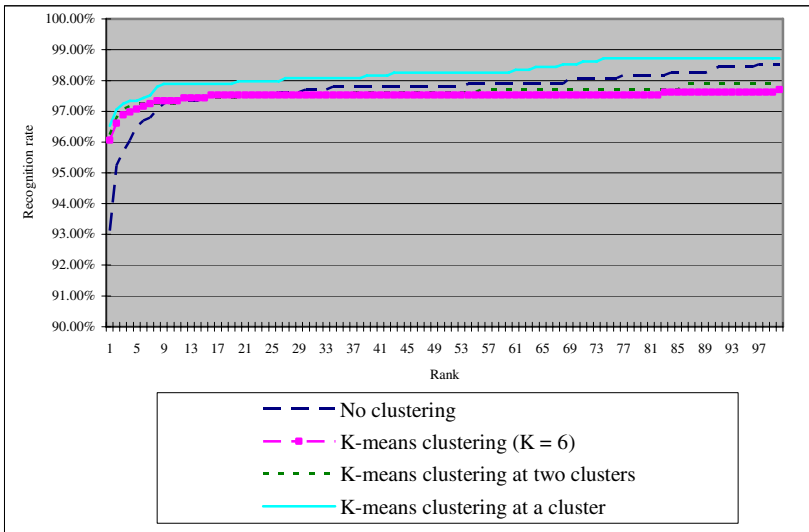
The feasibility of the proposed hierarchical context modeling method has been tested in the area of face recognition using FERET [20] data sets. We used 2418 images from 1209 persons in FERET data set.



First, we clustered the images into context models by hierarchical clustering using K-means algorithm. Extracted vectors using hybrid scanning are feed into K-means clustering. Second, we evolve classifier systems for individual context models. Genetic algorithm evaluates the best classifier system for each context model using images in the corresponding cluster.

**Table 1.** Face recognition results on clustered data after GA adaptation. Face images are hierarchically clustered using K-means clustering repeatedly.

	Recognition rate (Total number of images)			
	No clustering	K-means clustering (K = 6)	K-means clustering at two clusters	K-means clustering at a cluster
Cluster-0	<b>93.13</b>	96.09% (228)		
Cluster-1		97.45% (157)		
Cluster-2		96.94% (131)		
Cluster-3		95.27% (148)	96.68%(87)	
			94.56%(61)	
Cluster-4		99.36% (157)		
Cluster-5		92.62% (217)	96.15%(130)	
			90.8%(87)	90.91%(22)
				92.30%(65)
<b>Total</b>		<b>96.05%</b>	<b>96.23%</b>	<b>96.54%</b>



**Fig. 6.** The CMC curve shows the performance on clustered data after GA adaptation

In the process of clustering, K-means clustering is done repeatedly. Some clusters constructed by K-means algorithm are clustered by K-means algorithm again. Table 1 shows the performance of successful face recognition using hierarchical K-means clustering and adaptation. At first, whole test images are clustered to 6 clusters. Next, on some clusters, K-means clustering is done again. And this process is repeated. In the event, that enables hierarchical context modeling. As you can see in those tables, we can achieve better performance by hierarchical context modeling after adaptation using genetic algorithm. Fig. 6 is showing the CMC curve of the adaptable face recognition for each case of clustering.

In Table 2, one can find that the proposed method can achieve the highest performance. The superiority comes from the flexibility of our method since it has the capability of hierarchical ontology based context-awareness and adaptation.

**Table 2.** Performance comparison of the proposed system comparing with other approaches

Algorithm/Method	Recognition rate
arl_cor	0.827
arl_ef	0.797
Ef_hist_dev_anm	0.774
Ef_hist_dev_l1	0.772
ef_hist_dev_l2	0.716
ef_hist_dev_md	0.741
ef_hist_dev_ml1	0.733
ef_hist_dev_ml2	0.772
excalibur	0.794
mit_mar_95	0.834
mit_sep_96	0.948
umd_mar_97	0.962
usc_mar_97	0.95
Proposed method	0.9654

## 5 Conclusion

In this paper, a novel method of classifier combination using hierarchical image context-awareness is proposed and applied to object recognition problem. The proposed method tries to distinguish its input data context and evolves the classifier combination structure accordingly by Genetic algorithm (GA). It stores its experiences in terms of the data context category and the evolved artificial chromosome so that the evolutionary knowledge can be used later. The main difference of the proposed classifier selection method from other methods is that it can select classifiers in accordance with the identified context. In addition, once the context knowledge is constructed, the system can react to changing environments at run-time.

The proposed method has been evaluated in the area of face recognition. Data context-awareness, modeling and identification of input data as data context categories, is

carried out using K-means algorithm. The face data context can be decided based on the image attributes such as, light direction, contrast, brightness, spectral composition, etc. The proposed scheme can optimize itself to a given data in real-time by using the identified data context and previously derived chromosome. The proposed method is tested using four datasets: Inha, FERET, and Yale. Its performance is evaluated through extensive experiments to be superior to those of most popular methods, especially in each cluster.

## References

- [1] M. Potzsch, N. Kruger, and C. Von der Malsburg, Improving Object recognition by Transforming Gabor Filter responses, *Network: Computation in Neural Systems*, vol.7, no.2 (1996) 341-347.
- [2] C. Liu and H. Wechsler, Evolutionary Pursuit and Its Application to Face recognition, *IEEE Trans. on PAMI*, vol. 22, no. 6 ( 2000) 570-582.
- [3] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, From Few to Many: Illumination Cone Models for face recognition under Variable Lighting and Pose, *IEEE Trans. on PAMI*, vol. 23 no. 6 (2001) 643-660.
- [4] P. Phillips, The FERET database and evaluation procedure for face recognition algorithms, *Image and Vision Computing*, vol.16, no.5 (1999) 295-306.
- [5] Kuncheva L.I. and L.C. Jain, Designing classifier fusion systems by genetic algorithms, *IEEE Transactions on Evolutionary Computation*, vol.4, no.4 (2000) 327-336.
- [6] Sang-Woon Kim; Oommen, B.J., "On using prototype reduction schemes and classifier fusion strategies to optimize kernel-based nonlinear subspace methods," *Pattern Analysis and Machine Intelligence*, *IEEE Trans.* vol. 27, (2005) 455-460.
- [7] Ludmila I. Kuncheva, Switching Between Selection and Fusion in Combining Classifiers: An Experiment, *IEEE Transactions on Systems, Man, and Cybernetics - part B: cybernetics*, vol.32, no.2 (2002) 146-156.
- [8] C. Liu and H. Wechsler, Gabor Feature Classifier for Face Recognition, *Computer Vision 8th IEEE International Conference*, (2001) 270-275.
- [9] L. Kuncheva, K Andreeva, DREAM: a shell-like software system for medical data analysis and decision support, *Computer Methods and Programs in Biomedicine*, vol.40, no.2 (1993) 73-81.
- [10] Y. Liang, H. Kato, M. Taya, and T. Mori, Infinitesimal approach to the crystallography of Martensitic transformation: Application to Ni-Ti , *Scripta Materia*, vol. 43 (2000) 535-540
- [11] In Ja Jeon, Ki Sang Kwon, and Phill Kyu Rhee, Optimal Gabor Encoding Scheme for Face Recognition Using Genetic Algorithm, *KES 2004*, (2004) 227-236.
- [12] X. Wang and X. Tang, Random Sampling LDA for Face Recognition, in *Proceedings of CVPR*, vol.2 (2004) II-259- II-265.
- [13] X. Wang and X. Tang, A Unified Framework for Subspace Face Recognition, *IEEE Trans. on PAMI*, vol. 26, no. 9 ( 2004) 1222- 1228.
- [14] H. Kim, D. Kim, and S. Y. Bang, Face Recognition Using LDA Mixture Model, in *Proceedings of ICPR*, (2002) 486-489.
- [15] S. Yau, F. Karim, Y. Wang, B. Wang, and S. Gupta, Reconfigurable Context-Sensitive Middleware for Pervasive Computing, *IEEE Pervasive Computing*, 1(3), July-September 2002, (2002)33-40.

- [16] M.Y. Nam and P.K. Rhee, A Novel Image Preprocessing by Evolvable Neural Network, LNAI3214, vol.3, (2004) 843-854.
- [17] M.Y. Nam and P.K. Rhee : An Efficient Face Recognition for Variant Illumination Condition, ISPACS2005, vol.1 (2004) 111-115.
- [18] Shafer, S., Brumitt, B., and Meyers, B., The EasyLiving Intelligent Environment System, CHI Workshop on Research Directions in Situated Computing, (2000).
- [19] <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>
- [20] P. Phillips, The FERET database and evaluation procedure for face recognition algorithms, Image and Vision Computing, vol.16, no.5 (1999) 295-306.
- [21] <http://www.nist.gov/>
- [22] Nam Mi Young and Phill Kyu Rhee: Adaptive classifier combination for visual information processing using data context-awareness. Advances in Intelligent Data Analysis VI, September 2005.

# Wireless Internet Service of Visited Mobile ISP Subscriber on GPRS Network

Jeong-Hyun Park<sup>1</sup> and Boo-Hyung Lee<sup>2</sup>

<sup>1</sup> Postal Technology Research Center,  
Electronics and Telecommunications Research Institute (ETRI)  
# 161 Kajong-Dong, Yusong-Ku, Daejeon, 305-700, Korea  
jh-park@etri.re.kr

<sup>2</sup> Department of Computer Science and Engineering, Kongju University  
# 275, BooDae-Dong, CheonAnn, ChungNam, 330-717, Korea  
bh11998@kongju.ac.kr

**Abstract.** This paper shows wireless Internet access of visited ISP subscriber using dynamic IP in home ISP based on GPRS network. There is core network testbed and protocol stack, information flow, operation, and parameters of signaling messages for wireless Internet access of visiting ISP subscriber on GPRS network in this paper. There are also some implementation ideas of wireless Internet access system for remote mobile subscriber, simulation results of packets with IP-in-IP and without IP-in-IP between GGSN and ISP web server on our core network testbed in this paper.

## 1 Introduction

The combination of both developments, the growth of the Internet and the success of mobile networks, suggests that the next trend will be an increasing demand for mobile access to Internet applications. It is therefore increasingly important that mobile radio networks support these applications in an efficient manner. Thus, mobile radio systems currently under development include support for packet data services. The most widely deployed standard for second-generation mobile radio networks is the Global System for Mobile Communications (GSM) [3]. Networks based on this standard will be extended in the near future with the General Packet Radio Service (GPRS) [1][2], which provides data rates up to 384 kb/s (2 Mb/s).

When discussions about GPRS started in the early 1990s, applications such as road transport telematics and financial services driving the demand. The high costs for circuit-switched GSM connections prevented the widespread use of mobile data transmission for such services. In recent years, however, end-user applications such as Web browsing and email are becoming increasingly popular; therefore, the Internet has dominated the standardization of GPRS. Internet applications are predicted to contribute the largest share of the expected traffic volume.

In brief, GPRS can be described as a service providing optimized access to the Internet, while reusing to a large degree existing GSM infrastructure. Advanced mobile terminals using multiple slots will offer more convenient and faster Internet access than today's technology. The GPRS concept allows volume-oriented charging, which permits users to have cheap, permanent connections to the Internet.

General Packet Radio Service (GPRS) is being defined by the European Telecommunications Standards Institute (ETSI) to provide packet data service using Global System for Mobile Communications (GSM) cellular networks. As impressively demonstrated by the Internet, packet-switched networks make more efficient use of the resources for bursty data applications and provide more flexibility in general. Fig. 1 depicts the development system architecture for wireless Internet.

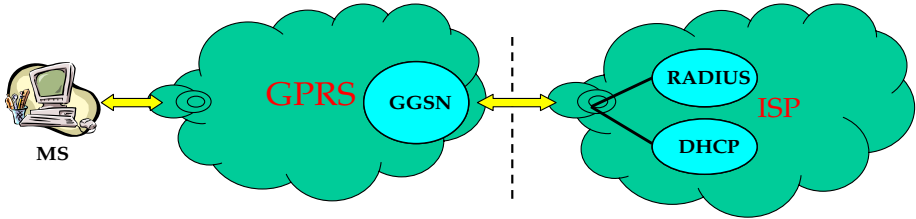


Fig. 1. Network Architecture

The Gateway GPRS Support Node (GGSN) is the gateway node between an external packet data network (IP) or packet-switched data network (X.25/X.75) and the GPRS core network. In the case of an external IP network, the GGSN is seen as an ordinary IP router serving all IP addresses of the mobile stations (MSs). This node may include firewall and packet-filtering mechanisms. Additionally, its task is to assign the correct SGSN for a mobile station depending on the location of the MS.

This paper describes the core network testbed and operation scenario for wireless internet service of mobile ISP subscriber on GPRS network in section 2. There are also messages between GGSN and ISP, and GGSN and mobile IP for wireless internet service of visited mobile ISP subscriber, and overall simulation results of packets with IP-in-IP and without IP-in-IP between GGSN and ISP web server on our core network testbed in section 3, and finally, conclusions are drawn.

## 2 Wireless Internet Access of Visited Mobile ISP Subscriber on GPRS Network

### 2.1 Testbed

To gain experience and iterate on our design, we have been implementing wireless Internet access model in a testbed. This currently consists of a GGSN, GPRS Support Node (SGSN), mobile IP (MIP) included Foreign Agent (FA) and Home Agent (HA), and ISP network included Remote Authentication Dialing in User Service (RADIUS), Dynamic Host Configuration Protocol (DHCP)/DHCP relay, Point-to-Point Protocol (PPP), Layer 2 Tunneling Protocol (L2TP), and Web server. We have Remote Access Node (RAN) simulator which can be supported UTRAN (UMTS (Universal Mobile Telecommunications System) Terrestrial Radio Access Network) with multimedia mobile terminal. Fig. 2 depicts the core network testbed which is implemented. We verify the roaming service of remote mobile subscriber through SGSN and GGSN



## 2.2 Operation

We now describe operation scenario of mobile IP based access for wireless Internet access of visited mobile ISP subscriber based on GPRS in this section.

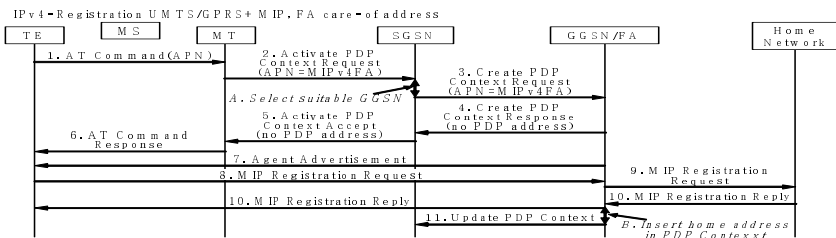
Fig. 3 shows signalling flow of wireless Internet access case of remote mobile ISP subscriber based on GPRS using mobile IP. A way to allow users to roam from one environment to another, between fixed and mobile, between public and private as well as between different public systems is to use Mobile IP. Mobile IP (MIP) is a mobility management protocol developed by IETF. The Mobile IP Foreign Agent (FA) is located in the Core Network in the GGSN. MIP also uses a Home Agent (HA) [8] which may or may not be located in a GSM/UMTS network. The interface between the GGSN and the FA will probably not be standardized as the GGSN/FA is considered being one integrated node. The mapping between these two is a matter of implementation. Each FA must be configured with at least one care-of address. In addition a FA must maintain a list that combines IP addresses with tunnel endpoint identification (TEIDs) of all the visiting MSs that have registered with the FA. IP packets destined for the MS are intercepted by the HA and tunneled to the MS's care-of address, i.e. the FA. The FA de-tunnels the packets and forwards the packets to the MS. Mobile IP related signalling between the MS and the FA is done in the user plane. MIP registration messages are sent with UDP. Address allocation - at PDP context activation no IP address is allocated to the MS indicated by 0.0.0.0. in the "Requested PDP Address" field. If the MS does not have a static IP address which it could register with the HA, it will acquire a dynamic IP address from the HA. After completion of the PDP activation the SGSN is informed of the assigned IP address by means of the GGSN initiated PDP Context Modification Procedure. A signalling scheme is described below. The PS attach procedures have been omitted for clarity in fig. 3.

In case of mobile IP based access,

- (1) The AT command carries parameters that the MT needs to request the PDP Context Activation. The important parameter here, is the APN (Access Point Name). The AT command is followed by a setup of the PPP connection between the MT and the TE, which are not included in the fig. 3.
- (2) The MT sends the "Activate PDP Context Request" to the SGSN. The message includes various parameters of which the "APN" (Access Point Name) and the "Requested PDP Address" are of interest here. The TE/MT may use APN to select a reference point to a certain external network or to select a service. APN is a logical name referring to the external packet data network or to a service that the subscriber wishes to connect to. The "Requested PDP Address" should be omitted for all MS's using Mobile IP. This is done irrespective of if the MT has a permanently assigned Mobile IP address from its Mobile IP home network, a previously assigned dynamic home address from its Mobile IP home network or if it wishes the Mobile IP home network to allocate a "new" dynamic home address. The SGSN will base the choice of GGSN based on the APN that is given by the MS.
- (3) The SGSN requests the selected GGSN to set up a PDP Context for the MS. The PDP address and APN fields are the same as in the "Activate PDP Context Request" message.



- (4) A Create PDP Context Response is sent from the GGSN/FA to the SGSN. If the creation of PDP Context was successful, some parameters will be returned to the SGSN, if not, an error code will be returned. If the GGSN has been configured, by the operator, to use a Foreign Agent for the requested APN, the PDP address returned by the GGSN shall be set to 0.0.0.0. indicating that the PDP address shall be reset by the MS with a Home Agent after the PDP context activation procedure.
- (5) The Activate PDP Context Accept message is sent by the SGSN to the MS and contains similar information as the Create PDP Context Response message.
- (6) The MT sends an AT response back to the TE to confirm that the PDP context activation has been done.
- (7) The Agent Advertisement [8] is an ICMP (Internet Control Message Protocol) Router Advertisement message with a mobility agent advertisement extension. The latter part contains parameters of the FA that the mobile node needs, among those are one or more care-of addresses that the FA offers. This message should be sent, in the Packet Domain user plane, as an IP limited broadcast message, i.e. destination address 255.255.255.255, however only on the TEID for the requesting MS to avoid broadcast over the radio interface.
- (8) The Mobile IP Registration Request is sent from the mobile node to the GGSN/FA across the Packet Domain backbone as user traffic. The mobile node includes its (permanent) home address as a parameter. Alternatively, it can request a temporary address assigned by the home network by sending 0.0.0.0 as its home address, and include the Network Access Identifier (NAI) in a Mobile-Node-NAI Extension.
- (9) The FA forwards the Mobile IP Registration Request to the home network of the mobile node, where a home agent (HA) processes it. Meanwhile, the GGSN/FA needs to store the home address of the mobile node or the NAI and the local link address of the MS, i.e. the TEID (Tunnel Endpoint ID).
- (10) The Registration Reply is sent from the home network to the FA, which extracts the information it needs and forwards the message to the mobile node in the Packet Domain user plane. As the FA/GGSN knows the TEID and the NAI or home address, it can pass it on to the correct MS. The GGSN/FA extracts the home address from the Mobile IP Registration Reply message and updates its GGSN PDP Context.
- (11) The GGSN triggers a "GGSN initiated PDP Context modification procedure" in order to update the PDP address in the SGSN.



**Fig. 3.** Signalling Flow for Mobile IP Based Access



### 3.2 Performance

As mentioned in previous section, we implemented GGSN and SGSN on Solaris, and FA, HA and AAA on Linux, ISP servers on Solaris and Linux. We also implemented performance tool called loadbox which can generate packets with some options such as packet length and interval, and measure delay, packet loss, and throughput at IP layer. For synchronization between GGSN and ISP, and GGSN and FA, we prepared network time protocol server. We observed the performance of IP-in-IP of test packet between GGSN and ISP web server, and GGSN and FA. For simulation, we stimulated test packets with varying length (1054 Byte x 10, 100, 1000) and varying interval (10ms, 100ms, 1000ms). The following are simulation results of delay, packet loss, and throughput with IP-in-IP and without IP-in-IP on the testbed.

- 1) Delay: In case of packet with IP-in-IP, delay time is about 30ms, and in case of packet without IP-in-IP, it's about 20ms. If test packet size increase with fast interval, we can see that time delay of packets with IP-in-IP is more than the time delay of packets without IP-in-IP. But the increased interval of time delay is minor. Also the time delay of the packets is almost the same situations between GGSN and FA, and GGSN and ISP web server.
- 2) Lost Packet: Both of packets with IP-in-IP and packets without IP-in-IP, lost packets rate is the same.
- 3) Throughput: Both of packets with IP-in-IP and packets without IP-in-IP at reliability level 95 % with significant level +/- 5, throughput is the same. Normally, we can see that throughput is about 78 Mb/s with 8 packets loss on our core network testbed.

**Table 1.** Performance Test Result

	Without IP in IP	With IP in IP	Remark
Delay	20ms	30ms	Increasing Packet Size : $\text{Delay}_{\text{with IP-in-IP}} > \text{Delay}_{\text{without IP-in-IP}}$
Loss	Same rate		
Throughput	Same [78 Mb/s with 8 packets loss]		Reliability level 95 % with significant level +/- 5

## 4 Concluding Remark

If the advent of the commercial Internet is the engine behind the new post-industrial revolution, then "wireless Internet" will surely accelerate innovation in this economy. These days many people use the term wireless Internet to indicate wire access to web services and content. However, the architecture, protocols, services and wireless technologies that constitute wireless Internet are still under consideration and a subject of great debate. There are a number of companies, standards bodies, and industrial for a vying to define future wireless Internet technology. The end result is that operators are faced with a large and confusing array of choices on how best to build next

generation mobile networks. Each technology has its pros and cons. For example, the IETF mobile IP protocol represents a simple and scalable global mobility solution but lacks support for fast handoff control, real-time location tracking, authentication and distributed policy management found in cellular networks today. In contrast, the International Mobile Telecommunications 2000 (IMT-2000) mobile network system offers support for seamless mobility, paging, and service quality but are built on complex and costly connection-oriented networking infrastructure that lacks the Internet flexibility, scalability, and cost effectiveness found in IP networks. Wireless Internet should be capable of combining the strengths of both approaches without inheriting their weaknesses.

Actually, we have studied how we apply GPRS network model to Internet as next generation wireless network model. This article showed wireless Internet access model for remote mobile ISP subscriber based on GPRS using IP, L2TP and mobile IP. As previously seen in this paper, there is wireless Internet access network model, and operation scenarios based on GPRS and mobile IP. For wireless Internet access model of remote mobile ISP subscriber based on GPRS, we defined messages and parameters between GGSN and mobile IP, and GGSN and ISP network. We also implemented SGSN, GGSN, FA, and HA for simulation. We simulated GPRS functions, Internet services, and performance of IP-in-IP between GGSN and FA, and GGSN and ISP web server using our core network testbed included GPRS and MIP components, and RAN simulator.

## References

1. 3GPP, "GPRS Service Description, Stage 2", 3G TS 23.060 version 3.3.0, March 2000.
2. 3GPP, "GPRS Service Description, Stage 1", 3G TS 22.060 version 3.3.0, March 2000.
3. 3GPP, "Combined GSM and Mobile IP Mobility Handling in UMTS IP CN", 3G TR23.923 version 3.0.0, May 2000.
4. R. Droms, "Dynamic Host Configuration Protocol (DHCP)", RFC 2131, March 1997.
5. C. Rigney, S. Willens, A. Rubens, and W. Simpson, "Remote Authentication Dial In User Service (RADIUS)", RFC 2865, June 2000.
6. W. Townsley, A. Valencia, A. Rubens, G. Zorn, G. Pall, and B. Palter, "Layer Two Tunneling Protocol (L2TP)", RFC 2661, August 1999.
7. William Allen Simpson, "The Point-to-Point Protocol (PPP)", RFC1661, July 1994.
8. C. Perkins, "IP Mobility Support", RFC2002, Oct. 1996.
9. C. Perkins, "IP Encapsulation within IP", RFC2003, Oct. 1996.

# An Exploratory Analysis on User Behavior Regularity in the Mobile Internet

Toshihiko Yamakami

ACCESS, 2-8-16 Sarugaku-cho, Chiyoda-ku, Tokyo, Japan  
yam@access.co.jp

**Abstract.** The ever-changing nature of the mobile Internet contributes to the difficulties encountered when experts try to identify the user behavior characteristics. Using thin channels with so-called 24-hour 365-day always on nature, it is crucial to understand regularity of user access in the mobile Internet. It is leveraged by the mobile Internet-specific features like user identifiers provided by wireless carriers. The author attempts to identify the easy-gone mobile Internet users from regularity dimension using a long-term user log with user identifiers. The author proposes an interval probability comparison method to predict the user behavior in the next month. The experiment from the mobile clickstream data shows the positive effect of the proposed method.

## 1 Introduction

With the emergence of increasingly powerful multimedia-capable mobile handset, the mobile Internet technologies have made their way into any time, any place computing with a wide spectrum of information sharing and e-commerce. This leverages the increasing demand to analyze mobile Internet behaviors in both of research and industrial communities. The key aspects of the mobile Internet include a limited small display size. In order to occupy this premium place, the service providers have to catch the shifts of user demands in a timely manner. Other aspects include an always-on feature that reflects the end user real life. Many past analysis techniques can be applied to some of the mobile Internet behavior analysis, however, the mobile Internet-specific aspects are still to be explored to cope with increasing challenges in the mobile Internet. One of the important sources of user behavior analysis is mobile clickstream.. A mobile clickstream is a record of a user's activity on the mobile Internet, including every Web site and every page of every Web site that the user visits, and in what order the pages were visited. The author proposes a new method to predict the end user behavior in a month-scale manner and discusses the advantages from some experiments with mobile clickstream data.

## 2 Purpose of Research and Backgrounds

### 2.1 Purpose of the Research

The mobile Internet depends on a thin channel with the volatile users. Users are easy-come and easy-go ones. The metrics to determine each user's regularity of

web visits are keys to promote mobile commerce sites. The aim of this research is to identify the visit regularity in monthly scale based user logs for the mobile Internet.

## 2.2 Related Works

As web-based commerce increases its coverage, clickstream analysis adds its importance. Extensive studies are done in the PC Internet clickstreams for e-commerce data mining for session analysis [1] [2], for page prediction [3], and path finding [4] and for personal recommendation [5]. It is also useful to identify the user behavior in the mobile Internet. However, the path-dependent analysis is not suitable for the mobile web because some of the web pages are short-lived due to the window space limitations in the mobile handsets. On the other hand, time-based analysis and regularity analysis are important to grasp mobile-specific user behaviors. The author explored how to use the user identifier provided by wireless carriers in mobile clickstream analysis [6]. Time zone analysis was done by Yamakami [7]. Helvey presented the significance of time of day in mobile clickstreams [8] to indicate the weekday/weekend user behavior differences. Hagen discussed that the mobile human interactions need to cope with methodological challenges [9]. There is no regularity analysis in the mobile Internet in the past literature.

## 3 System Design

### 3.1 Analysis System

The author implemented an analysis system for the mobile clickstreams. The analysis system configuration is illustrated in Fig. 3.1. The data sources are monthly mobile clickstream logs with user identifiers. The UID clustering part and the stat pre-processing part were implemented by PHP ver 4.12 [10]. The stat analysis part and the prediction analysis part were implemented by R ver 2.1.1 [11].

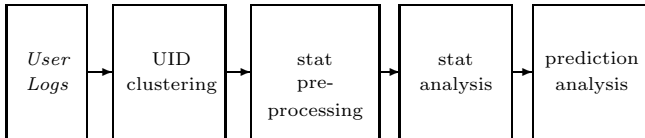


Fig. 1. An Analysis System Configuration

## 4 Method

In order to measure user loyalty in a mobile network service, a user log for monthly subscription services is used to identify the usefulness of observed measure for loyalty. The observed service is a paid premium service that requires monthly subscription fee (approximately 3 US dollars). In order to identify a threshold value for identifying next-month behavior, the author checks the following measures from the mobile clickstream.

- average access interval
- standard deviation of access intervals
- median of access intervals

In order to identify the stability and fits for prediction, the author performed the analysis for 2-month observation on clickstream logs shown in Fig. 2.

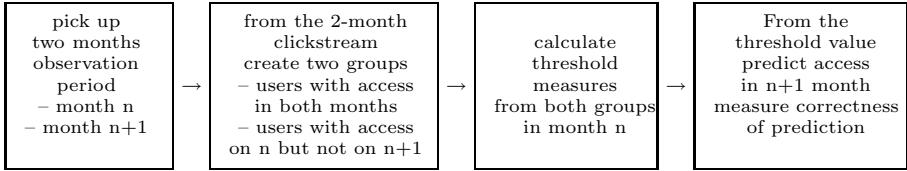


Fig. 2. Analysis flow from mobile clickstreams

### 4.1 Data Set

The observation target is a commercial news service in the mobile Internet. The service is available on the three different mobile carriers, with slightly different content menu. Each mobile carrier has different underlying network characteristics and different charging policies. The service is charged for monthly subscription fee. The log stores the unique user identifier, time stamped, command name and content shorthand name. The services were launched from 2000 to 2001, and continue up to today. The target service provides 40 to 50 news articles per week on weekdays. the commercial mobile service charges the monthly subscription fee to the users, approximately 3 US dollars per month. UID is usually 16 or more unique alphanumeric character long, e.g. “310SzyZjaaerYlb2”. The service uses Compact HTML [12], HDML [13], and MML that is a dialect of HTML. In this research, Compact HTML version service is used for analysis.

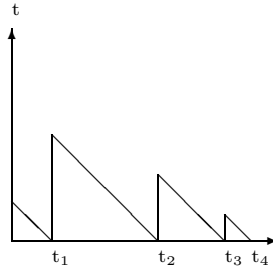
### 4.2 Interval Expectation Method

The author proposes an interval expectation method to estimate the user regularity in the mobile clickstream. The mobile clickstream contains user identifiers to enable accurate capturing each user’s command sequences. The interval to the next access in time-series patterns is shown in Fig. 3. The x-axis denotes time. The y-axis denotes the interval to the next access. It comes to a maximum value just after each access and monotonously decreases until the next access.

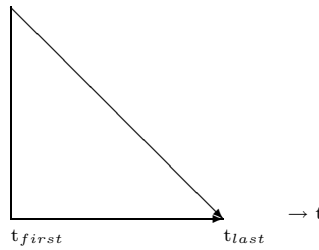
When there is no intermediate access during the observation, the sum of the triangle areas becomes the maximum, depicted in Fig. 4. The sum of triangle areas is called as *interval triangle area sum*.

When the number of access during the observation is same, the even interval case gives the minimum sum of triangle areas, as depicted in Fig. 5. Therefore this sum of triangle areas by time duration for next access gives a measure to compare regularity. The sum of triangle areas for user m,  $S_m$ , is given as

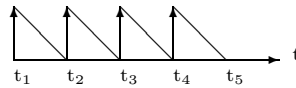
$$S_m = \Sigma \frac{1}{2}(t_n - t_{n-1})^2$$



**Fig. 3.** Interval Expectation



**Fig. 4.** No intermediate access case



**Fig. 5.** An even interval case

where  $t_n$  denotes time of  $n$ -th access for each user. When the two groups  $\{x_1, x_2, \dots\}$  and  $\{y_1, y_2, \dots\}$  and clickstreams in  $n$ -th month are given, the analysis system calculates  $\{S_{x_1}, S_{x_2}, \dots\}$  and  $\{S_{y_1}, S_{y_2}, \dots\}$ . Then use a threshold value  $p$  from  $\{S_{x_1}, S_{x_2}, \dots\}$  and  $\{S_{y_1}, S_{y_2}, \dots\}$  to determine  $(n+1)$ -th month's prediction of two groups. In this experiment, the author uses the middle value of the average of  $\{S_{x_1}, S_{x_2}, \dots\}$  and that of  $\{S_{y_1}, S_{y_2}, \dots\}$ .

This method has one drawback. When there are a small number of accesses in the middle of the month, the sum of interval triangle areas could be small even though access is very sparse during the month. This is common in the easy-come easy-go mobile Internet users. The mobile Internet analysis needs to deal with this sparseness in the monthly clickstream. In order to remove this anomaly, the analysis system has an option to automatically insert virtual access at the beginning of the month and the end of the month. This gives more accurate estimation of the sum of the interval triangle areas when the system uses monthly clickstreams.

The author used this sum of triangle areas as the measure to identify user regularity. The author picked one carrier's mobile clickstream dataset to verify



the effect of the proposed interval expectation. The February 2002 to July 2002 data were used. The preceding two-month clickstream is used to test the measure. Therefore, the predictions from April to July were obtained. The author extracted information to calculate a threshold value for  $(n+2)$ th month prediction from  $n$ -th and  $(n+1)$ th month clickstream.

In the experiment, the author uses average value from two groups to predict the third month behavior. First, the system extracts two groups, one from the users with access both in  $n$ -th and  $(n+1)$ th month and the other from users with access in  $n$ -th month and without access in  $(n+1)$ -th month. The sum of interval triangle areas for each user in both groups is calculated. Then, the average of the sums from the two groups are calculated to give a threshold value. This threshold value is used for  $(n+1)$ -th month to predict whether each user will access in the next month ( $(n+2)$ -th month). The prediction accuracy is determined by the real  $(n+2)$ -th month clickstream.

## 5 Result

The author performs the month-scale prediction in 2002 mobile clickstream from February to July. The prediction needs 2 months to get the threshold value, therefore, comparison from April to July is done. In order to compare the proposed method with other method without regularity analysis, the author also uses the access count-based threshold value to make prediction. The result is depicted in Fig. 5. The accuracy derived from the proposed method are ranged from 72.3 % to 83.1 %. It indicates improvement over a primitive method like account-count based prediction. It shows the validness of the proposed method in a month-scale manner. It should be noted that the primitive access count method is biased by the easy-go users with heavy access only within a narrow span of time. The proposed method shows effectiveness to filter this volatile heavy access users in the clickstream. The proposed method depends only on the two previous month behaviors, which means that it can cope with the mobile Internet dynamism. The proposed method can be flexibly applied to a wide range of mobile clickstream with user identifiers.

The mobile Internet users are easy-comers and easy-goers. The usage is very volatile. The month-scale regularity largely depends on appear/disappear behaviors in the observed case. The proposed method itself is simple and general, therefore, it can be applied to a wide range of applications.

In order to fully explore the mobile Internet user behaviors, it is important to investigate user behavior models. Markov models and variations are used in the past research [3]. The result obtained in this research does not provide information for the motivations and causes of behavior changes. The author considers that the reliable measure for regularity is a first step towards understanding mobile Internet behaviors. It is especially important in the long-term behavior transition studies to identify user clusters with different regularity. For the further user modeling, the clustering based on the observed regularity will help to capture the accurate user behavior estimation.

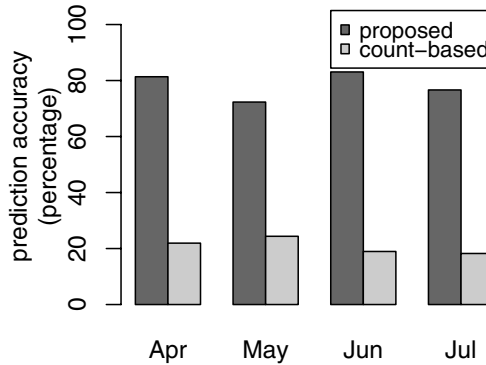


Fig. 6. Prediction accuracy using the interval expectation method

There are three different level of regularity in the mobile clickstreams: Sub-day-scale, day-scale, and month-scale. Sub-day-scale regularity is affected by each user's sub-day rhythm. Day-scale regularity is affected by each user's time of day rhythm. Month-scale regularity is affected by seasons or charging granularity(monthly subscription et al). This study is limited to the monthly-based prediction over a primitive method like access count-based one. The detailed analysis on these different scale levels is for further studies.

The mobile Internet poses a new challenge for knowledge discovery. Compared to the PC Internet, the analysis needs to focus access regularity not access heaviness. The time dimension of the mobile Internet behavior analysis needs further exploration. This research is a first step towards further exploration of this mobile-Internet specific characteristics which are critical for future mobile commerce and other application design.

## 6 Conclusion

The emergence of the mobile Internet requires mobile Internet-specific analysis of the mobile clickstream. Using the thin-channel with always-on characteristics, it is important to identify the user regularity as well as the traditional heaviness of usage. The author proposes an interval expectation method to predict the month-scale user regularity. The experiment shows prediction accuracy ranging from 72.3 % to 83.1 % and proves the advantage of the proposed method. It shows the mobile-specific regularity-based analysis has a fit to the easy-come and easy-go characteristics of the mobile Internet.

## References

1. Lee, J., Podlaseck, M., Schonberg, E., Hoch, R.: Visualization and analysis of clickstream data of online stores for understanding web merchandising. *Data Mining and Knowledge Discovery* **5**(1-2) (2005) 59–84

2. Andersen, J., Giversen, A., Jensen, A., Larsen, R., Pedersen, T., Skyt, J.: Analyzing clickstreams using subsessions. In: Proceedings of the third ACM international workshop on Data warehousing and OLAP. (2000) 25–32
3. Guenduez, S., Oezsu, M.: A web page prediction model based on click-stream tree representation of user behavior. In: ACM KDD2003. (2003) 535–540
4. Ali, K., Ketchpel, S.: Golden path analyzer: using divide-and-conquer to cluster web clickstreams. In: ACM KDD2003. (2003) 257–276
5. Kim, D.H., Atluri, V., Bieber, M., Adam, N., Yesha, Y.: Web personalization: A clickstream-based collaborative filtering personalization model: towards a better performance. In: ACM WIDM2004. (2004) 88–95
6. Yamakami, T.: Unique identifier tracking analysis: A methodology to capture wireless internet user behaviors. In: ICOIN-15, Beppu, Japan, IEEE Computer Society (2001) 743–748
7. Yamakami, T.: A mobile clickstream time zone analysis: implications for real-time mobile collaboration. In: Proceedings of KES2004 (Volume II). Volume LNCS 3214 of Lecture Notes in Computer Science., Springer Verlag (2004) 855–861
8. Halvey, M., Keane, M., Smyth, B.: Predicting navigation patterns on the mobile-internet using time of the week. In: WWW2005, ACM Press (2005) 958–959
9. Hagen, P., Robertson, T., Kan, M., Sadler, K.: Emerging research methods for understanding mobile technology use. In: Proc. of 19th conf. of SIGCHI of Australia (OZCHI 2005). (2005) 1–10
10. Group, T.P.: Php hypertext processor. available at <http://www.php.net/> (2003)
11. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2005) ISBN 3-900051-07-0.
12. Kamada, T.: Compact HTML for small information appliances. W3C Note, 09-Feb-1998, Available at: <http://www.w3.org/TR/1998/NOTE-compactHTML-19980209> (1998)
13. King, P., Hyland, T.: Handheld device markup language specification (1997)

# Reliable Communication Methods for Mutual Complementary Networks

Takashi Kaneyama<sup>1</sup>, Hiroshi Mineno<sup>1</sup>, Takashi Furumura<sup>2</sup>,  
Kunihiro Yamada<sup>3</sup>, and Tadanori Mizuno<sup>1</sup>

<sup>1</sup> Shizuoka University, 3-5-1 Johoku, Hamamatsu-shi, Shizuoka 432-8011, Japan  
kaneyama@mizulab.net

<sup>2</sup> Renesas Solutions Corporation, 4-1-6 Miyahara, Yodogawa-ku, Osaka 532-0003, Japan

<sup>3</sup> Tokai University, Kitakaname 1117, Hiratsuka-shi, Kanagawa 259-1292, Japan

**Abstract.** In the future home network, information appliances will be controlled and managed by the use of PLC (Power Line Communication) that enables the No New Wire and ZigBee concepts to be implemented in a ubiquitous sensor network. However, the arrival rate of PLC and the communication quality deterioration of ZigBee are significant problems because the control information for the appliances has to be transmitted reliably. To solve these problems, we examined communication methods that increase the arrival rate in a mutual complementary network environment. These methods improve reliability by mutually complementing, through PLC and ZigBee, the places where nodes can't communicate through only one interface. A comparison of these methods through ns-2 simulations shows that the Table Creation method is more reliable than the other methods we examined.

## 1 Introduction

The energy consumption of individual homes keeps on increasing. The unified management of home electric appliances would save energy and enhance home security. For these reasons, integrated management of home appliances will become important in the future, and a network that connects home electric appliances will thus be needed. In addition, the ownership ratio of computers in households has increased rapidly and Internet connections have become common. Many people hope to be able to control home appliances from inside and outside their homes. The establishment of a home network, including not only computers and peripherals but also home appliances, can enable such integrated management and remote control of home appliances; therefore, it will become important in the near future.

There are two problems of related th the costs of generalizing the home network. One is the cost of replacing household appliances with information appliances; the other is the cost of establishing the home network. In the latter case, while Ethernet is commonly used as the LAN in an office, it requires new wire to be laid in a home. In fact, a home network environment with No New Wire is desirable.

The use of the electric power lines that have already been laid in a home for communication, and ZigBee [1], which is the wireless standard is suitable for a low-cost home network, have problems with their reliability. The goal of this study is the achievement

**Table 1.** Wired and wireless standards

Wired	HomePNA 3.0	c.LINK	HomePlug 1.0
Band	5.5~9.5 MHz	770~1030 MHz	4.3~20.9 MHz
Speed	128 Mbps	270 Mbps	13.75 Mbps
Medium	Telephone line Coaxial cable	Coaxial cable	Power line

Wireless	IEEE802.11b/a/g	Bluetooth	ZigBee	UWB
Band	2.4 / 5 / 5 GHz	2.4 GHz	2.4 GHz	3.1~10.6 GHz
Speed	22 / 54 / 54 Mbps	1 Mbps	250 kbps	110 / 480 Mbps
Distance	100 / 50 / 100 m	10 m	70 m	10 / 3 m
Price	\$8	\$3	\$3	—
Power	1000 mW	100 mW	30 mW	200 mW

of reliable communications in home networks, and for this, we propose communication methods that improve reachability through a combination of low-reliability communication media.

The rest of the paper is organized as follows. In section II, we describe the necessity of the mutual complementary network which uses media with different features. Section III outlines the communication methods which improve reliability. Section IV describes the results of simulations. Section V ends the paper with a brief summary and concluding remarks.

## 2 Necessity of Mutual Complementary Network

### 2.1 Wired Network

While a home network needs to maintain the communications infrastructure wiring new cables costs a lot of money. Therefore, the best way to make a home network is to use the existing in a home. The cables that likely exist in a home are phone wires, coaxial cables for TV, and electric power lines. HomePNA, c.Link, and HomePlug are wired communication standards that use these communication media, and Table 1 lists their features.

Since these cables are easy to introduce, they are not troublesome for users. However when many appliances are to be connected to a home network, the interfaces of the phone wire and coaxial cable are not necessarily near enough to the appliances. In addition, because such interfaces may already be in use for telephone and TVs, branching filters will be needed. On the other hand, electric power lines are wired throughout houses, and most appliances are connected to outlets; thus power lines are a suitable home network medium. Under the radio law in Japan, PLC (Power Line Communications) is permitted in a bandwidth from 10 to 450 kHz, and it is a low-speed form (less than 10 kbps). However, the low bit rate is still enough for sending control commands, state information, etc.; hence, the PLC is effective for a home network.

### 2.2 Wireless Network

The wireless communication standards, 802.11b/a/g/n, Bluetooth, ZigBee, and UWB (Ultra Wide Band), are applicable to a home network. Table 1 lists their features.

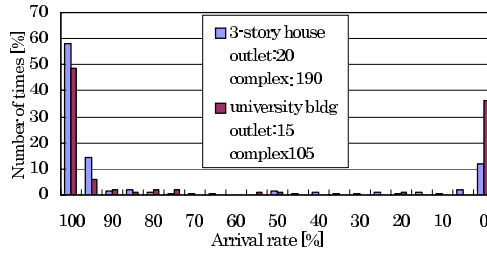


Fig. 1. Arrival rates of PLC in a home and university environment

It is likely that some sensors in the network will be placed where they cannot get power; therefore it is preferable that the power consumption of the communications be low. The range of wireless communications must also span the house. In light of these considerations, ZigBee is suitable for controlling home network appliance. The transmission range of ZigBee is about 30 m, and this is far enough for indoor communications. Its low bit rate of 250 kbps is of no matter, if appliances only exchange control information. Although the price of the ZigBee chip is currently about three dollars, it is likely to become cheaper especially if it becomes popular.

### 2.3 Mutual Complementary Network

A the home network must be reliable. PLC signals, however, are attenuated by different phases and appliances that exist between communicating equipment may decrease the arrival rate. Figure 1 shows the result of an investigation on the arrival rates for all combinations of outlets that can be used for PLC in a three-story house and a university building in a real environment. The figure shows that the arrival rate of PLC is characteristically either 100% or 0%. In ZigBee, communication quality is degraded by obstacles, transmission distance, and noise.

To solve these problems, we assume a mutual complementary network environment that improves arrival rate by using PLC and ZigBee, and propose communication methods that improve reachability efficiently. The mutual complementary network complements these places where nodes can't communicate through only one interface.

Figure 2 shows a model of the mutual complementary network. Two phases exist in the electric power line, and the arrival rate of PLC decreases significantly in communications between different phases. We assume that communication between different phases is impossible, and express the power line as two lines in the figure. There is a place where nodes cannot communicate because of the influence of the equipment connected to the power line (on the power line of phase A in the figure). Nodes are numbered in the figure, and each node has PLC and ZigBee interfaces. In PLC, only nodes 1 and 2 or 3 and 4 communicate. In ZigBee, it is possible for nodes in radio wave range to communicate.

### 2.4 Related Works

Referance [2] proposes a dual communication system that uses wired and wireless communication. The system is suitable for indoor use and is easy to install. The author

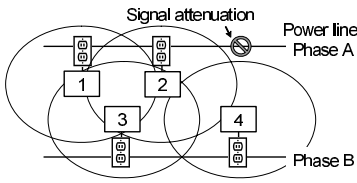


Fig. 2. Network model

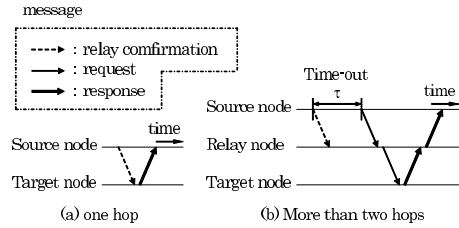


Fig. 3. Forwarding sequence of RC method

proposes a method of ensuring adequate communication quality by combining wired and wireless media. This is statistically shown to be able to improve reliability, and a simple routing [3] that operates under a given set of conditions is described. Our work differs from the literature noted above; We study efficient methods that can accommodate without special conditions.

Although there are several research intended to determine efficient path among routes based on parameters [4] [5], these assume same media as the communication media. Our research is differ from these in point of using media with different characteristics.

### 3 Communication Methods on a Mutual Complementary Network

#### 3.1 Prerequisites

We are studying communication methods for a mutual complementary network that efficiently increase the arrival rate by using media with different characteristics, in order to solve the cost problem of building a home network environment by using PLC and ZigBee. Our assumptions are as follows. The communication nodes have two network interfaces so that they can communicate with PLC and ZigBee. The node's interfaces are configured with different IDs, and the nodes can identify the IDs uniquely. Each node knows the interfaces IDs of the other node, and can assign the ID of a target node. We assume that data such as control information can be transmitted as one message.

#### 3.2 Broadcast (BC) Method

This is a simple method in which nodes repeatedly broadcasts and propagate messages to the target node. At the start, a transmission source node broadcasts a message in PLC and ZigBee at the same time. If the node receiving the message is the target node, it returns a response. If it isn't a target node and hasn't already received messages of the same ID, it operates as a relay node and it, as well as the source node, broadcasts the message. The message is delivered to the target node by repeating this procedure.

The advantage is that nodes can reliably deliver a message to the destination within a short time. The disadvantage is that this method has the possibility of causing broadcast storm and interfering with transmissions generated by another node.

### 3.3 Relay Confirmation (RC) Method

In this method, a source node determines whether there is a target node within one hop and confirms the necessity of relaying. If the target node is not within one hop, the message is relayed to the target node. Figure 3 shows the message flow of the RC method. The target node responds after the source node transmits the relay confirmation message when it is within one hop of the source node (Figure 3 (a)). If the target node is farther than two hops from the source node, the source node cannot receive a response to the relay confirmation, and thus it transmits a request message after waiting for the time-out period  $\tau$  (Figure 3 (b)).

The processing procedure of each node is as follows. First, the transmission source node sends RC messages to a target node with PLC and ZigBee, and it waits for the reply from the target node. If the source node receives the reply before  $\tau$  has elapsed, the target node exists within the range where the source node can communicate. Thus, the source node realizes that the message arrived at the target node and completes the communication. If time out occurs without the reply arriving, the source node broadcasts a request message with PLC and has another node relay it. The reason the source node broadcasts firstly by PLC is to have all nodes, which are connected with a same-phase line, relay messages; the messages are delivered early to nodes that are connected with different-phase lines. The relay node broadcasts a request message to the target node without sending relay confirmation. Note that a relay node transmits alternately on different media when it transmits to its neighbors. For example, a relay node must broadcast a message with ZigBee after receiving it with PLC. This enables nodes to prevent the broadcast storm while complementing the places in which they cannot communicate through only one interface.

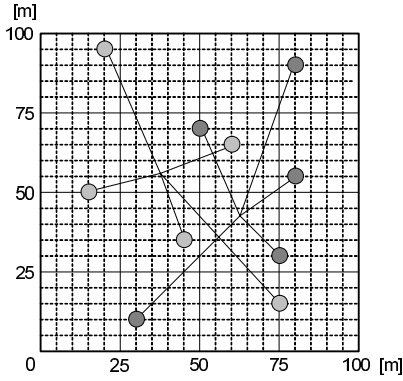
In the RC method, extra traffic can be suppressed because the source node relays only when it cannot transmit directly to a target node. However, the arrival time to the destination increases when there are more than two hops because the source node decides whether a transmission failure has occurred from the time-out.

### 3.4 Table Creation (TC) Method

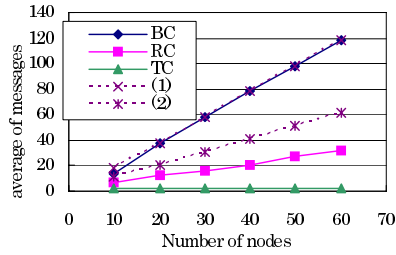
OLSR (Optimized Link State Routing) [6] is a routing algorithm that has been discussed by the MANET WG, and we have a plan for proposing a method that uses OLSR. However, we evaluated a TC method in which nodes create only tables of neighbor nodes, instead of the method using OLSR. In this method, all nodes periodically exchange Hello messages and make tables of neighbor nodes that communicate using PLC and ZigBee. Regarding the transmission of a request message, a source node first confirms the table, and then sends the request to the target node if one is found. If there is no target node in the table, the source node tries to detect a node to which it can send the message by either PLC or ZigBee. If a relay node cannot be chosen on these conditions, one is chosen at random from the table. A relay node that receives message chooses the next forwarding node in the same way that source node does.

Broadcasting is not used and the message is sent only to the selected relay node; thus traffic is kept to a minimum. However, when an inappropriate relay node is chosen, there is a possibility that message will not arrive by the shortest path or does not reach the target node.

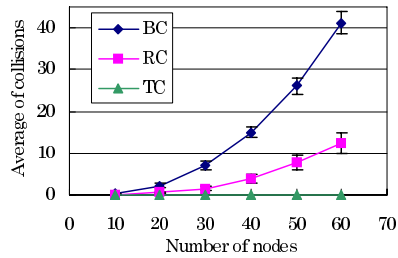




**Fig. 4.** Topology for evaluation on random arrangement of nodes



**Fig. 5.** Average number of messages (Simulation I)



**Fig. 6.** Average number of collisions (Simulation I)

## 4 Evaluation

### 4.1 Simulation Environment

We investigated the properties of the communication methods outlined in the previous section by using the network simulator ns-2 [7]. We assumed that each communication node had both PLC and ZigBee interfaces. As the wired interface, we set the bit rate to 7.5 [kbps] to simulate PLC, and set MAC protocol to Ethernet (IEEE802.3) in order to make a LAN, although SCP (Simple Control Protocol) adopts CSMA/CA. SCP is the communication protocol for home networks that we are planning to use in an actual PLC environment. As the wireless interface, we set the bit rate to 250 [kbps] to simulate ZigBee and set the transmission range to 12 [m], considering indoor communications. We used 802.11 as the MAC protocol instead of CSMA/CA used by ZigBee. Moreover, we set the size of the control information exchanged between each node to 30 [bytes] (referring to SCP) and the time-out  $\tau$  of the RC method to 200 [ms].

### 4.2 General Properties (Simulation I)

We investigated the general properties of each method by arranging nodes at random on the square lattice shown in Figure 4. The nodes were arranged at random in two groups, and each group formed a LAN. These two LANs imitated the power line network on which communication between different phases was impossible. Regarding the transmissions, the source and target nodes were chosen at random. We transmitted messages 100 times on networks with 10 to 60 (in increments of 10 nodes).

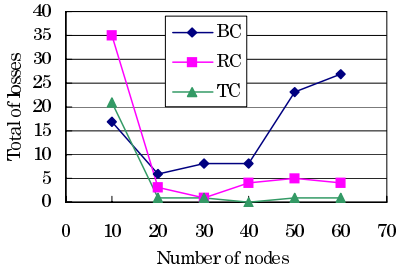


Fig. 7. Total number of losses (Simulation I)

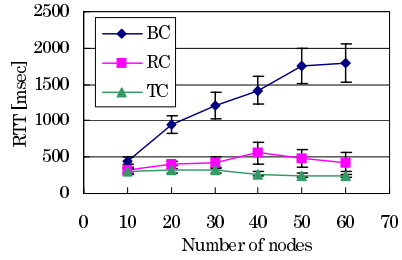


Fig. 8. Mean value of RTT (Simulation I)

To be able to compare each method numerically, we used the total number of messages. If the number of nodes is  $N$ , and the number of hops to target node is  $H$ , the total number of messages that are generated in one transmission request is calculated as follows.

$$T_{BC} = 2(N - 1) + H \tag{1}$$

$$T_{RC} = N + H + 1 \tag{2}$$

(if  $H=1$  then  $T_{RC} = 3$ )

$$T_{TC} = 2H \tag{3}$$

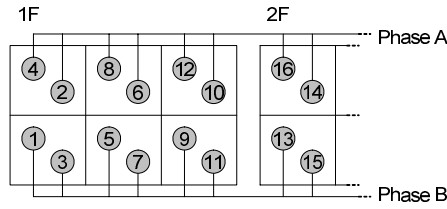
### 4.3 Results of Simulation I

Figure 5 shows the average number of messages forwarded by PLC and ZigBee in one transmission. The value calculated with expressions (1) and (2) are indicated in the figure as a reference. Note that the number of messages includes neither the Hello message nor the message of the MAC layer. Regarding the BC method, the number of messages increases in proportion to the number of nodes as shown in expression (1). Regarding the RC method, when relay confirmation message reaches a target node, the relay broadcasting is skipped. Therefore, the number of messages is smaller than the value calculated with expression (2).

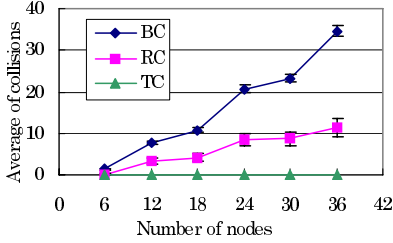
Figure 6 shows the average number of collisions in one transmission. It can be seen that the number of collisions increases with the number of nodes.

Figure 7 shows the total of loss after 100 transmissions. Messages were considered lost if the reply message did not reach the source node because of collisions and there was no reachable path to the target node. All methods had many lost messages on the network with only ten nodes. The reason is that a node not within wireless communication range could not forward to a node connected to power line of the opposite phase. Moreover, the BC method’s losses increased because of collisions in the networks with more than 20 nodes.

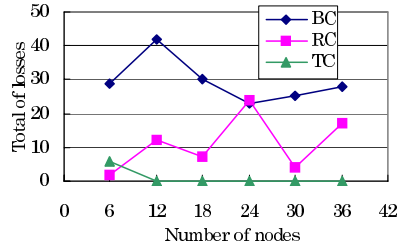
Figure 8 shows the mean value of the round trip time (RTT). In spite of its time-out of relay confirmation, the RC method had a much lower than the BC method had. Many messages are transmitted in the BC method; therefore, collision avoidance using CSMA/CA happened frequently. Consequently, the BC method had a large RTT.



**Fig. 9.** Topology for evaluation in a home environment



**Fig. 10.** Average number of collisions (Simulation II)



**Fig. 11.** Total losses (Simulation II)

**4.4 Evaluation in a Home Environment (Simulation II)**

We confirmed the properties of the proposed methods by using a topology for a house (Figure 9). The environment had six rooms on each floor and two nodes on the left side and right side of each room. The wireless range of the nodes is shown in the figure. For example, node 5 on the left side of a room was able to communicate with nodes 3, 6, and 7. On the other hand, node 6 on the right side of a room was able to communicate with nodes 5, 8, and 12. In addition, all nodes in the room of the upper side were connected with a power line of phase A, and those on the lower side were connected with a phase B line. We chose the source node and target node at random, and transmitted messages 100 times for networks of 6 to 36 nodes (increment of 6).

**4.5 Results of Simulation II**

Figure 10 shows the average number of collisions in one transmission, and Figure 11 shows the total losses in 100 transmissions. Compared with Figure 6 and Figure 10, there are more collisions in Figure 10. Regarding Figure 11, many losses occurred, especially in the BC method. It seems that the reason for the BC and RC methods having more losses is the influence of hidden terminals. The node topology was orderly (Figure 9); therefore, there is a possibility that interference could be caused by nodes that cannot communicate with each other.

**4.6 Considerations**

The results of the two simulations show there are many collisions and losses in networks using the BC and RC method. The TC method had better results. Therefore, we

found that the table method was the most effective. As for the BC and RC methods, it seems that the arrival rate was high without collisions. Thus, if a technique for reducing collision can be used, a better result might be obtained.

## 5 Conclusion

We discussed the necessity of a mutual complementary network which possessing transmission media having different features. We focused on the reliability of the network and proposed potential communication methods. The results of simulations, indicated that the availability of the TC method is the best.

We have to strengthen the reliability; thus, we will investigate ways to improve the methods discussed in this paper. Furthermore, we have a plan for deploying the methods in an actual environment and evaluating the reliability of their transmissions under realistic circumstances.

## References

1. ZigBee Alliance Website, <http://www.zigbee.org/>.
2. K. Yamada, et. al., "Dual Communication System Using Wired and Wireless Correspondence in Home Network," *KES2005*, LNAI 3681, pp.438-444, 2005.
3. K. Yamada, et. al., "Dual Communication System Using Wired and Wireless with the Routing Consideration," *KES2005*, LNAI 3681, pp.1051-1056, 2005.
4. H. Liu, et. al., "An Adaptive Genetic Fuzzy Multi-path Routing Protocol for Wireless Ad Hoc Networks," *SNPD2005*, pp.468-475, 2005.
5. S. Mueller, et. al., "Analysis of a Distributed Algorithm to Determine Multiple Routes with Path Diversity in Ad Hoc Networks," *WiOpt2005*, pp.277-285, 2005.
6. T. Clausen, P. Jacquet, et. al., "Optimized link state routing protocol," RFC 3626, <http://hipercom.inria.fr/olsr/rfc3626.txt>.
7. The Network Simulator ns-2, <http://www.isi.edu/nsnam/ns/>.

# Remote Plug and Play USB Devices for Mobile Terminals

Jungo Kuwahara<sup>1</sup>, Hiroshi Mineno<sup>1</sup>, Kiyoko Tanaka<sup>2</sup>, Hideharu Suzuki<sup>2</sup>,  
Norihiko Ishikawa<sup>2</sup>, and Tadanori Mizuno<sup>1</sup>

<sup>1</sup> Shizuoka University, 3-5-1 Johoku, Hamamatsu-shi, Shizuoka 432-8011, Japan  
jungo@mi.zulab.net

<sup>2</sup> Network Management Development Department, NTT DoCoMo, Inc.

**Abstract.** Advances in high performance mobile terminals and short distance wireless communication technology have been achieved. There has been renewed of interest in controlling peripheral devices with mobile terminals. However, in general peripheral devices are used in the form directly connected to a computer. We propose an intelligent USB (iUSB) where mobile terminals can remotely use USB devices for network-transparency. We achieved Plug and Play on the network (remote Plug and Play) with a USB devices. We confirmed that remote Plug and Play was effective if it had an environment with enough band width and low latency.

## 1 Introduction

Rapid advances in mobile-terminal performance have been achieved with mobile terminals equipped with Windows Mobile operating systems and Symbian operating systems. Short distance wireless communication technology such as IrDA, Bluetooth, and ZigBee has also been improved. There has been renewed interest in controlling peripheral devices with mobile terminals. For example, users may want to project an image of the mobile terminal to a television with a large surrounding screen. Users may also wish to hear audio from the portable terminal with high-quality speakers in the surroundings. The key technology in these scenarios is that the mobile terminals communicate with peripheral devices directly.

Our research group propose a mobile Personal Area Network (mPAN) as a new form of PAN [1]. mPAN is a network service that involves PAN through the use of mobile terminals and peripheral devices, and achieves service mobility where the mobile terminals which is a control points can start services even when moving. Mobile terminals with mPAN can use various high-quality I/O devices such as cameras and speakers throughout the network. However, in general these peripheral devices are used in the form directly connected to a computer. It is difficult to provide all interfaces (e.g., USB, SCSI, and IEEE1394) to a small mobile terminal although it is necessary to have all these interfaces when the mobile terminal uses these peripheral devices. We are therefore paying attention to technology that controls peripheral devices through network, e.g., UPnP or iSCSI [3][6][7][8].

In this paper, we focus attention on USB devices and propose intelligent USB (iUSB) where mobile terminals can remotely use USB devices in mPAN for network-transparency (Fig.1) . We achieved remote Plug and Play of a USB device on a networkD.

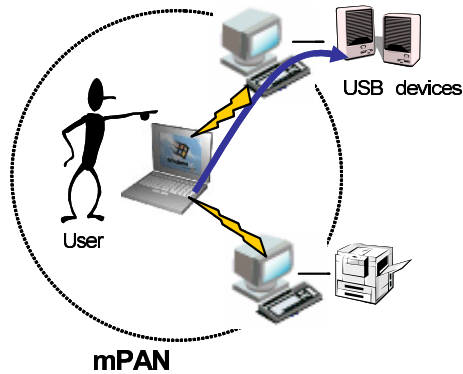


Fig. 1. iUSB in mPAN

## 2 Related Work

The Internet Small Computer Systems Interface (iSCSI) has been standardized by the IP Storage Working Group of the Internet Engineering Task Force [4]. By encapsulating the SCSI command and data in TCP/IP packets, the SCSI command system can not be seen from the outside. Connection is directly enabled to the IP network of the SCSI storage product. This has already been put into practical use, and it is used by network storage. The advantage of iSCSI is that connection between different kinds of operating systems is possible because it does not depend on the operating system or the architecture of the computer. However, only storage devices with SCSI command characteristics can be handled. It is difficult to use a variety of devices remotely in iSCSI.

USB/IP [2] has achieved the use of USB devices connected to other computers on a network in a Linux kernel. The Virtual Host Controller Interface (VHCI), the client side, detects and manages transactions for USB devices. The server installs a stub driver [5] as a kernel thread to access for the connected USB device. With USB/IP, the client encapsulates USB requests in IP packets, and they are transmitted to the server. Thus, almost all USB devices can be used throughout the network. However, these are now difficult to only use with Linux terminals. In addition, it is not possible to perform remote Plug and Play.

## 3 iUSB

An outline of iUSB is given here. iUSB assumes Windows operating system is the execution environment, and achieves network-transparency in remote USB devices. Terminals that use remote USB devices are defined as control points. Also, terminals USB devices are connected to are defined as server. A control point is centered in the iUSB, and a concentrated processing system controls a remote USB device connected to a server in the surrounding area. iUSB has main three main features.

**Full Functionality.** We introduced a kernel driver into iUSB and controlled a remote USB device. A remote USB device can be operated at not the application level but the

kernel level. The kernel driver is implemented in the lowest layer, so a remote USB device does not specialize in the application. That is, the user can operate all functions in the USB device.

**Network-Transparency.** Network-Transparency is defined as "Devices on the network that can be controlled by existing applications without any modification". The iUSB was designed to intercept the USB requests among kernel drivers. The control point encapsulates the USB request in the IP packet, and transmits to the server. Applications do not notice the USB request transmitted to the server. That is, the user can operate it as well as the USB device connected locally without considering it when there is a remote USB device.

**Zero-Configuration.** We achieved remote Plug and Play of a USB device on the network between the control point and server. It was possible to automatically set the USB device because it involved remote Plug and Play. Therefore, the control point can only use it by connecting the USB device to the server immediately.

## 4 Design of iUSB Architecture

### 4.1 Virtual Host Controller Interface

The device-access mechanism for a traditional operating system is not taken into consideration for devices on IP networks. Therefore, USB devices can only be used directly connected to a computer right now. We discuss a virtual-connection mechanism for USB devices on an IP network within the limits of existing traditional operating systems. This mechanism enables modifications minimizing operating system device access. By virtually connecting remote USB devices within the limits of existing traditional operation systems, file access and device access using applicable, existing software resources is maximized. In other words, users can operate remote USB devices while using existing applications that they are accustomed to and familiar with.

We propose an iUSB Host Controller Interface (iHCI) to implement a virtual-connection mechanism. The iUSB extends the USB driver stack over network using iHCI (Fig.2). The USB request is passed from the USB driver to the USB device in USB communication, so it can be intercepted with a bus driver layer. The iHCI receives the USB request from the USB Functional Device Driver (FDD), encapsulate it in IP packets, and transmits the request to the USB device over the network. All transaction processing of the remote USB device is done with iHCI. The USB FDD does not need to modify it because the iHCI in the lowest layer of USB driver stack receives the USB request from the USB FDD. Thus, an existing device driver can be used, and it is possible to deal with many USB devices.

Since iHCI is a virtual host controller interface composed of software, the USB interface does not need to take on the user side. That is to say, it is possible for a mobile terminal with a limited number of interfaces to control a remote USB device because it uses iHCI. As a result, a mobile terminals which is a control point can operate USB devices in mPAN.

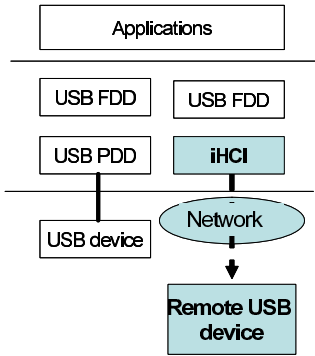


Fig. 2. Extended USB Driver Stack

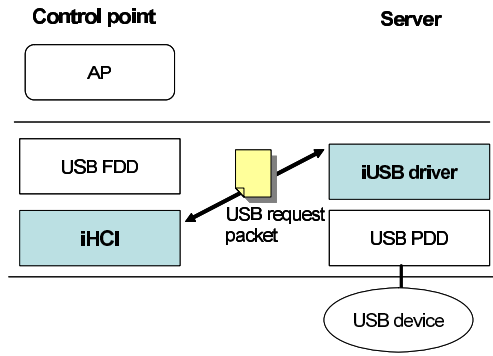


Fig. 3. iUSB Architecture

### 4.2 iUSB Architecture

Fig.3 outlines the iUSB architecture. iHCI is introduced into the control point, and it is operated as a virtual host controller interface. There is a USB FDD above the iHCI, which corresponds to the USB FDD that controls the USB device that is connected to the server. The AP that exists in the highest layer of the control point discovers the USB device connected to the server. However, the iUSB driver is introduced into the server. The iUSB driver is a FDD, and the server controls the USB device connected to it by using the iUSB driver. The USB request in iUSB generated by the control point is a request by the USB device connected to the server. Thus, the iUSB driver’s function is to receive the USB request transmitted from the control point, and to pass it to the connected USB device. iUSB sets the iUSB driver to an individual USB device that is connected to the server, and the iHCI of the control point manages and controls collectively.

The control point has two functions. First, it discovers the USB device connected to the server, and processes Plug and Play between the control point and server. Details on these operations will be described in the next section. Second, it processes USB requests by iHCI. Basically, the USB Physical Device Drive (PDD) transmits a USB request to a USB device connected to a local computer. However, because the USB device used is connected to the server in iUSB, it is necessary to transmit the USB request to the server’s USB device. With Windows operating system, the USB request is used as a USB Request Block (URB) by communication between USB drivers. Receiving URB from a high-layer driver, iHCI encapsulate URB in a UDP/IP packet. URB by UDP/IP packet encapsulation (USB request packet) is transmitted to the server. The USB request packet is processed by the server, and the processing result is returned to the control point. iHCI receives the processing result, and passes it to a high-layer driver. As previously mentioned, processing of the USB request generated by the control point is completed.

The server also has two functions. First, the server similarly processes Plug and Play between itself and the control point. Second, it behaves as a remote device based on requests received from the control point, and returns the processing result. To put it more precisely, the iUSB driver receives the USB request packet, and passes URB to



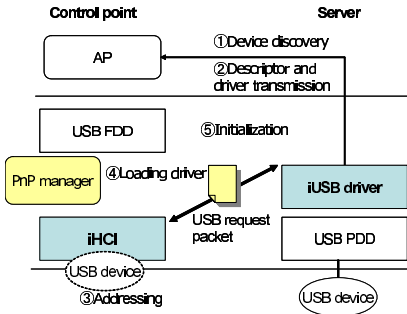


Fig. 4. Remote Plug and Play process flow

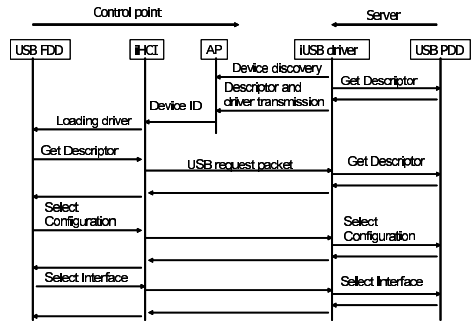


Fig. 5. Remote Plug and Play process sequence

USB PDD in the server. After this, the processing result is returned to the control point. Accordingly, the server’s USB device is controlled as a remote device by the control point.

## 5 Remote Plug and Play

We describe remote Plug and Play in this section. When the USB device connects to a computer in the USB protocol, an enumeration process is performed between the host and device. This process discovers the connected USB device, enabling the host to communicate with the USB device. That is, the enumeration process and Plug and Play are the same. This process in the USB protocol is as follows.

1. The host discovers the device,
2. The host acquires the descriptor of the device,
3. The host gives a unique address to the device,
4. The host loads the device driver, and
5. The host initializes the device.

Fig.4 outlines the remote Plug and Play process flow in iUSB. The remote Plug and Play process sequence is in Fig.5. Each processes is as follows.

### 1. Device discovery

The server transmits the device discovery message to the control point when the USB device is connected to it. The server’s iUSB driver generates the device discovery message. The USB device is recognized as being connected to the server when AP receives the device discovery message.

### 2. Descriptor and driver transmission

It is necessary for the control point to know what information there is on the connected USB device. The descriptors define the characteristics and the attributes of the USB device. To achieve this, the iUSB driver in the server acquires the descriptors and transmits them to AP in the control point. By recognizing the descriptors, the control point confirms what kind of device is connected to the server.

The driver for the connected USB device also transmits to the control point because it assumes mobile terminals are the control point in iUSB. As a result, the control point does not need to have a USB device driver.

### 3. Addressing

The iHCI writes an object for each discovered unit of USB devices. When the iHCI writes an object, one serial number is allotted to it. As a result, the USB device connected to the server can be uniquely managed.

### 4. Loading driver

The iHCI loads an appropriate device driver into the discovered USB device on the basis of the descriptors. When the driver is loaded in Windows operating system, the device is identified from the vendor, product, and revision codes of the USB device described in the descriptors. A device driver corresponding to the device identifier is automatically loaded by passing the device identification to the PnP manager.

### 5. Initialization

The control point processes USB requests necessary for initializing the USB device of the server when the driver for the USB device is loaded. The necessary USB requests to initialize the device are three requests. Get Descriptor, Select Configuration, and Select Interface. When these requests are intercepted, the iHCI transmits them to the server as USB request packets. After receiving the USB request packets, the iUSB driver of the server initializes the USB device by passing these requests to a low-layer driver. When the USB device has initialized, the control point recognizes it connected normally to the server. Finally, the remote Plug and Play for iUSB ends.

## 6 Experiment

We implemented an experimental environment for iUSB, as outlined in Fig.6. The control point and server were connected via a hub to provide an ideal mPAN environment without interference. Both the control point and server run Windows XP with iUSB modules. The remote Plug and Play processing time measured USB mass storage class devices (CD-ROM Drive, Flash Memory, Floppy Disk Drive, and the Digital Camera). The local Plug and Play processing time was also measured to compare them with the remote Plug and Play processing time.

Fig.7 plots average Plug and Play processing times. We evaluated Plug and Play in three kinds of experiments. First, the Plug and Play processing time of the control point was measured when a USB device was connected to the control point (Fig.7 (a)). That is, this was the time required to do Plug and Play processing when the USB device was directly connected to a local computer. Second, the remote Plug and Play processing time was measured when the control point had a driver for the USB device (Fig.7 (b)). Third, the remote Plug and Play processing time was measured when the USB driver was transmitted from the server to the control point (Fig.7 (c)). Fig.7 (c) shows the time a USB storage system device driver (26 kbytes) was transmitted in. In each experiment, Plug and Play was attempted ten times, and the average Plug and Play processing time was calculated.

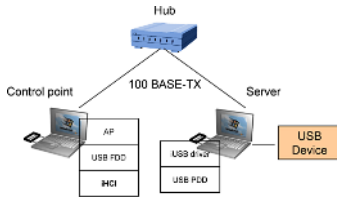


Fig. 6. Experimental environment

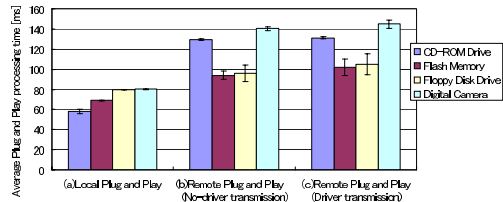


Fig. 7. Plug and Play processing time

We see from Fig.7 that no-driver transmission (b) and driver transmission (c) of average Plug and Play processing time are twice that of local Plug and Play (a). This is because a USB request packet was sent and received between the control point and server. However, the average remote Plug and Play processing time was approximately 100 ms , so we can see that remote Plug and Play was completed in a short period of time. Users will be able to use USB devices without feeling any difference to those connected locally. On the other hand, the average Plug and Play processing time for no-driver transmission (b) and driver transmission (c) are almost the same. We can see that the control point does not need to have a USB device driver. Therefore, A mobile terminal without the USB driver can receive and use USB device on the server.

## 7 Conclusion

We proposed an iUSB architecture to use remote USB devices with network transparency. The iUSB achieved remote Plug and Play for USB devices by implementing iHCI. We confirmed in experiments that remote Plug and Play was completed immediately.

We plan for iUSB to use a variety of USB devices. We also examine the packet loss problem with USB requests and the QoS problem with iUSB.

## References

1. K. Tanaka, et al. gProposal of Mobile Personal Area NetworksCh WiNFCppD241-245C2004D
2. T. Hirofuchi, E. Kawai, K. Fujikawa, and H. Sunahara, gUSB/IP - a Peripheral Bus Extension for Device Sharing over IP Network, h USENIX, 2005D
3. UPnPhttp://www.upnp.orgD
4. J. Satran, et al. Internet Small Computer Systems Interface (iSCSI). RFC3720, Apr, 2004.
5. T. Sato, K. Nakayama, Y. Kobayashi, M. Maekawa, gA Kernel-Level Framework for Network-Transparent Peripheral Control, h SIGOS, No.92-16, 2001.
6. C.-Y. Huang, et al. gA Cyclic-Executive-Based QoS Guarantee over USB, h RTAS, 2003.
7. Inside Out NetworksDAnywhereUSB. http://www.ionetworks.com/
8. J. Sato, E. Kawai, Y. Nakamura, K. Fujikawa, and H. Sunahara, gThe Design and Implementation of a Generic Framework for Remote Device Control, hSIGOS, 2003.

# SIP-Based Streaming Control Architecture for Mobile Personal Area Networks

Hiroshi Mineno<sup>1</sup>, Naoto Adachi<sup>1</sup>, Tsutomu Eto<sup>1</sup>, Kiyoko Tanaka<sup>2</sup>,  
Hideharu Suzuki<sup>2</sup>, Norihiro Ishikawa<sup>2</sup>, and Tadanori Mizuno<sup>1</sup>

<sup>1</sup> Faculty of Informatics, Shizuoka University, 3-5-1 Johoku, Hamamatsu, 432-8011, Japan  
mineno@inf.shizuoka.ac.jp, mizuno@inf.shizuoka.ac.jp

<sup>2</sup> NTT DoCoMo Inc., 3-5 Hikarino-oka, Yokosuka, 239-8536, Japan  
{tanakakiy, suzukihid, ishikawanor}@nttdocomo.co.jp

**Abstract.** This paper addresses seamless streaming control architecture among networked media devices at home network based on the Session Initiation Protocol (SIP). We assume the mobile phone acts as a control point of the virtual device, and call the network a mobile Personal Area Network (mPAN). The architecture is based on combining REFER method and 3PCC mechanisms of SIP. From the experimental results, the additional Handover supported SIP UA dramatically reduced retrieval time compared with non-supported one. The indirect approach with B2BUA cut down transfer time when renegotiation between parties was unnecessary.

## 1 Introduction

The appearance of dual-mode Cellular/Wireless LAN (WLAN) mobile phones provides a potential for new IP-based multimedia communications in collaboration with long-range and short-range wireless networks. While mobile devices are improving with more enhanced capabilities for IP-based multimedia communications, they remain limited in terms of bandwidth, display size, audio quality and computational power. At the same time, recent consumer electronics are media devices networked through the broadband Internet access service, and they have much more great capabilities than mobile phones. Combining the advantages of both into a single virtual device complements the capabilities of mobile phones and provides a seamless interaction among consumer electronics, mobile phones and PC devices.

This paper addresses seamless streaming control architecture among networked media devices at home network based on the Session Initiation Protocol (SIP) [1]. Since the SIP has been chosen by the Third Generation Partnership Project (3GPP) as its standard for session establishment in the Internet Multimedia Subsystem (IMS) and it is being deployed in hardware and software IP multimedia clients, the architecture complementally based on the SIP and its extensions provide a great affinity for future ubiquitous multimedia communications. We assume the mobile phone acts as a control point of the virtual device, and call the network a mobile Personal Area Network (mPAN). Users can control seamlessly input/output devices among the networked media devices within mPAN.

## 2 Related Works

Several existing approaches address the seamless transfer of IP-based multimedia sessions between devices to enable a user to switch terminals in the middle of a session. Generally session mobility architecture are divided into two areas: Indirection approaches [2] [3] and SIP-based approaches [4] [5]. There are both pros and cons respectively.

The indirection approaches use a special-purpose proxy to handle all session migration and media adaptation to different terminals. Therefore the advantages of indirection approach are that the ability of trans-coding at the proxy and hiding the session transfer from corresponding node. In contrast, the disadvantages are that the software is limited to the supported one and the user data flow between two calling parties must always traverse the proxy, regardless of whether session transfer is desired, introducing triangular routing.

The SIP-based approaches uses third-Party Call Control (3PCC) [6] and the REFER method [7] those are supported by SIP. The advantages of this approach are that the SIP is a signaling protocol and it uses its message bodies to describe the media so that it can be transported using RTP or another protocol. Using a SIP for signaling promotes the interoperability with multimedia communication software. In contrast, it requires the end-to-end signaling exchange, even though the device for forwarding has same capabilities and is nearby. It results in long update delays and high signaling overhead in the backbone network.

## 3 Streaming Control Within mPAN

### 3.1 mPAN Architecture

Figure 1 shows the architecture of our system. It consists of a Correspondent Node (CN), a Control Point (CP) like mobile phone, SIP-enable devices (Dev) on home network, and a Gateway (GW) that connects home network with the Internet. These components are containing a SIP User Agent (SIP UA) for standard SIP call setup, as well as a specialized SIP-handling capability for Session Mobility (SIP SM), and a Service Discovery Protocol (SDP) like Service Location Protocol (SLP) or Simple Service Discovery Protocol (SSDP) of UPnP. CP is a dual-mode mobile phone and has two network interfaces, such as cellular for long-range mobile communications and WLAN for short-range wireless communications. Each interface is configured with a global IP address and a private IP address, respectively. Each Dev on home network has a private IP address that was assigned by the address configuration protocol, such as DHCP or AutoIP. CP can communicate with Devs on home network through its private IP address, and acts as a controller of mPAN.

On the gateway, a Back-to-Back User Agent (B2BUA) handles the internetworking between home network and the Internet. The Internet's original uniform address architecture has been replaced with new de facto Internet address architecture, consisting of a global IP address realm and many private IP address realms interconnected by Network Address Translators (NAT). SIP negotiates various parameters including port numbers, IP addresses, and details of the media stream such as type of encoding based

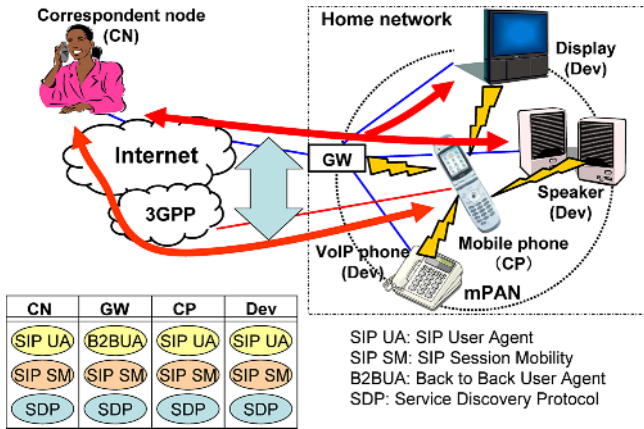


Fig. 1. System architecture

on a Session Description Protocol (SDP). Therefore, more recently the use of SIP has been extended to include information required to establish direct UDP, and indirect TCP connectivity between hosts behind NAT boxes and firewalls [8].

The B2BUA is an application-level gateway (ALG) that has the ability to modify SIP message headers and body contents in order to switch between public and private addresses. If the common codec does not exist between CN and Dev, different codec can be used between the B2BUA and each of them with the additional function for transcoding between the streams. Moreover, if the forwarding Dev supports same codec with the previous Dev, our architecture does not require the end-to-end signaling exchange. B2BUA hides the session transfer from the CN. It results in short update delays and low signaling overhead in the backbone network. This hybrid approach has the both of advantages of indirection and SIP-based approach.

These components and mPAN capabilities provide the CP to control seamlessly input/output devices among the neighboring devices like a virtual device. SIP is designed for establishing media streams (audio or video), it can also operate each media stream through the different input/output devices independently.

### 3.2 SIP REFER Method and 3PCC

A standard session transfer of SIP is done by sending a REFER request. After bidirectional media session is established between User-A and User-B, User-B sends INVETE request to User-C for transfer the session to User-C. If User-B receives “200 OK” from User-C and sends the ACK, User-B sends User-A REFER request that contains a SDP of User-C in Refer-To header and a SDP of User-B in Referred-By header. If User-A accepts the session transfer, User-A replies User-B “202 Accepted” response and sends the NOTIFY that reports its status. Then User-A sends User-C INVITE request, and bidirectional media session is established between User-A and User-C. After User-C confirms the session establishment with User-A, User-C sends User-B BYE request and User-B sends User-A BYE request to disconnect the previous session between User-A

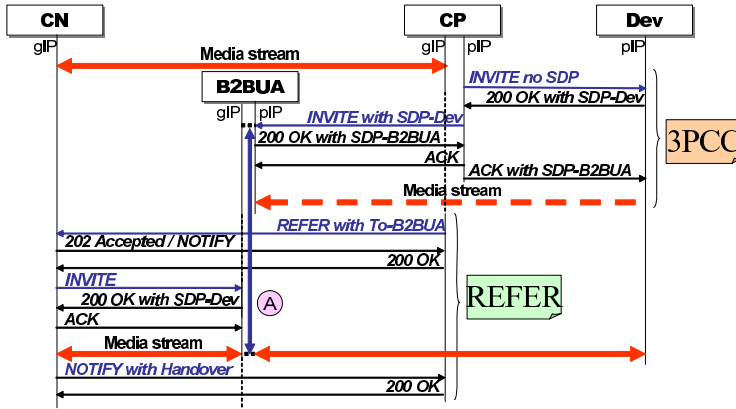


Fig. 2. Call flow of mPAN basic streaming transfer

and User-B. The basic session transfer with REFER method is done by these message exchange.

In third-party call control (3PCC), the controller establishes a SIP session between User-A and User-B. Controller sends User-A INVITE request without its SDP in the body. User-A replies its SDP to the controller so that controller sends User-B the SDP of User-A in the INVITE request. This results that User-B knows the User-A's SDP and User-B can send media data to User-A. Then, User-B replies controller a "200 OK" response with a SDP of User-B, and controller sends User-A the SDP of User-B in the ACK. This results that User-A can send media data to User-B and the bidirectional media session is established between User-A and User-B.

### 3.3 mPAN Basic Streaming Transfer

The basic streaming transfer within mPAN is done by combining REFER method and 3PCC mechanisms described previously. We assume CP has two network interfaces, one is configured with a global IP address (gIP) for communicating with CN in the Internet directly, another is configured with a private IP address (pIP) for communicating with Devs in home network. The private IP address is assigned by such as DHCP or AutoIP. If CP has already formed mPAN with nearby electronics devices those are not connected with home network and these private IP addresses are assigned, a dynamic address configuration occurs to solve the address conflicts when the mPAN merge with home network [9]. Moreover, all nodes can make out each other with service discovery protocol as mentioned above.

The call flow for mPAN basic streaming transfer is shown in Fig. 2. The first, CP makes a media session between Dev and B2BUA with 3PCC mechanism. The session is established with pIPs in home network. Then, CP uses a REFER method to make a media session between CN and B2BUA. The session is established with gIPs in the Internet. The data loss does not occur in the streaming transfer from CP to Dev because of the media session between Dev and B2BUA is established before executing REFER method.

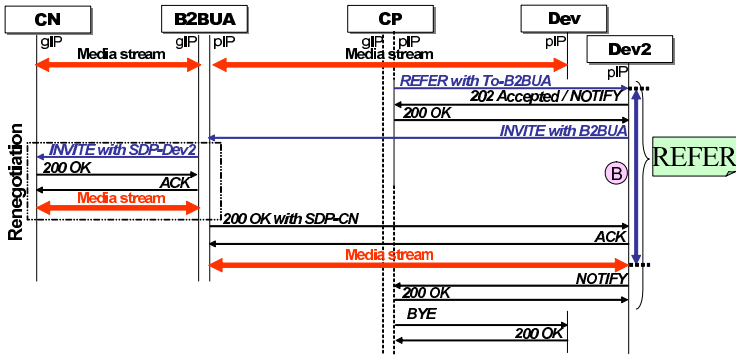


Fig. 3. Call flow of mPAN streaming transfer between devices

The mPAN basic streaming transfer can be done by normal SIP mechanism. However we add an extended header “Handover” to the NOTIFY method for making a CP manages the session between CN and CP though the media streaming is transferred to Dev. The Handover header has some statuses: sleep, sleep\_audio, sleep\_video, wakeup, wakeup\_audio and wakeup\_video. The statuses related to sleep are used to suspend the existing media streaming, and the statuses related to wakeup are used to resume the existing media streaming. If there is an existing media streaming when the extended NOTIFY is received, the receiver does not send BYE request to the previous party and suspends the existing media streaming. Because renegotiation of media streaming between parties requires time, we extend the NOTIFY method to be able to achieve fast retrieval from Dev to CP.

### 3.4 mPAN Streaming Transfer Between Devices

The streaming transfer between devices within mPAN is implemented in REFER method like 3PCC. Fig. 3 shows the call flow after Fig. 2. The media streaming between CN and CP was transferred to indirect two media streaming, one is established between CN and B2BUA and the other is established between B2BUA and Dev. CP sends Dev2 a REFER request that contains a SDP of B2BUA in Refer-To header and a SDP of CP in Referred-By header. Dev2 replies CP a “202 Accepted” response and sends the extended NOTIFY message. Although CP receives the extended NOTIFY, CP does not suspend because there is not an existing media streaming between CP and Dev2. Then, Dev2 sends B2BUA INVITE request.

If necessary for B2BUA to renegotiate a media streaming with CN, B2BUA sends CN INVITE request that contains SDP of Dev2. Such situation occurs when Dev2 has different codec or resolution with Dev. If not necessary, the renegotiation time is cut off and CN does not notice the transfer between devices in other party. After Dev2 confirms the establishment with B2BUA, Dev2 sends CP an extended NOTIFY. CP does not suspend because not existing media streaming between CP and Dev2. Then, CP sends Dev a BYE request to disconnect the previous media streaming between B2BUA and Dev. This call flow is the same as an original SIP REFER method although CP acts as a controller.



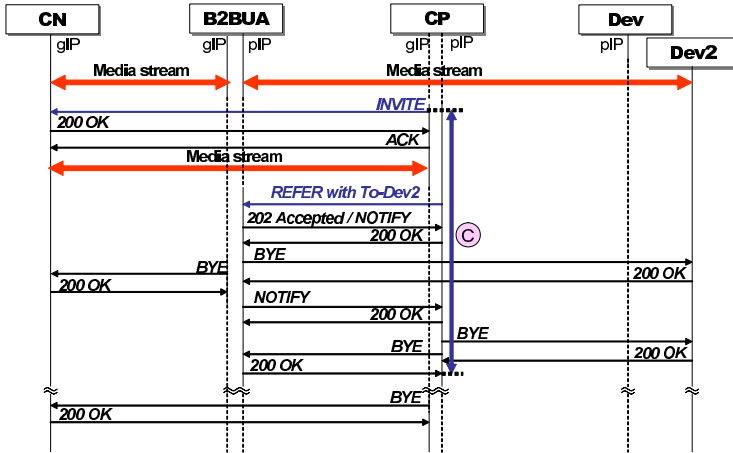


Fig. 4. Call flow of mPAN streaming retrieval and session disconnect

### 3.5 mPAN Streaming Retrieval and Session Disconnect

The streaming retrieval to CP from Dev is done by sending INVETE once again. Fig. 4 shows the call flow after Fig. 3. As CP sends CN INVETE request though gIP of CP, transferred media streaming between B2BUA and Dev is retrieved to CP. After retrieval of media streaming, CP sends B2BUA a REFER request that contains Subscription-State header with “expire=0” and Refer-To header with Dev2. This executes the node sends a BYE request to the node indicated by Refer-To header. In this case, B2BUA sends Dev2 a BYE request. Then, B2BUA sends CN a BYE request to disconnect the session between CN and B2BUA and notifies CP the confirmation. By sending BYE requests to Dev2 and B2BUA from CP respectively, the streaming retrieval to CP from Dev is completely finished.

The session disconnection between CN and CP is done by sending a BYE request from either party. When the BYE request from CN to CP occurs during media streaming is transferred to device, CP sends BYE requests to B2BUA and device.

## 4 Performance Evaluation

### 4.1 Effect of Handover Header

We developed an mPAN-supported SIP UA and a B2BUA based on the SIP Communicator that is a Java based SIP UA with audio/video capabilities built on top of the JAIN SIP Reference Implementation (nist-sip-1.2) and the Java Media Framework. We run the SIP UA on CN, CP, Dev and Dev2 those are Windows XP laptop-type machines. We run the B2BUA on GW that is a Windows XP desktop-type machine as a prototype test-bed environment. Moreover, SIP Express Router (SER) was applied as a SIP server on Linux machine. All nodes run on same LAN so as to minimize the effect of communication delay.

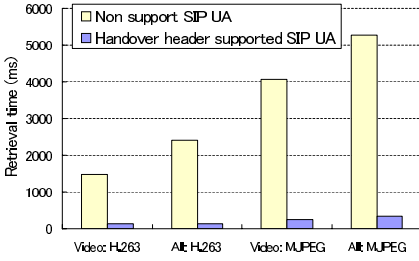


Fig. 5. Streaming retrieval time

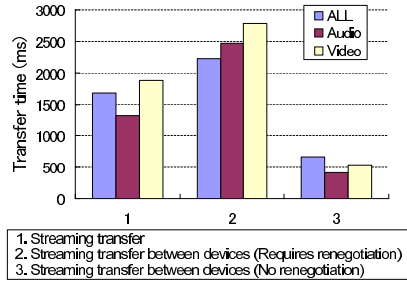


Fig. 6. Streaming transfer time

First of all, we evaluated the effect of additional Handover header for reducing renegotiation time between CN and CP. As shown in We measured time required to retrieve a media streaming from Dev, as shown phase C in Fig. 4, compared with handover-supported SIP UA and non-supported one. Two codec for video streaming, H.263 and Motion-JPEG (MJPEG), were compared with only a video streaming retrieval and both video/audio streaming retrieval, respectively. H.263 is a codec suitable for narrowband environment and MJPEG is a codec suitable for broadband environment. In addition, the resolution was QVGA (320 × 240) that assumes recent mobile phones.

The results are shown in Fig. 5. The Handover supported SIP UA dramatically reduced retrieval time compared with non-supported SIP UA that took few seconds. As the codec, the results of using H.263 were approximately 130 ms both only a video streaming retrieval and video/audio streaming retrieval. The results of using MJPEG were approximately 245 ms for only a video streaming retrieval, and 340 ms for video/audio streaming retrieval. This is because MJPEG requires more data and processing than H.263. However, the minor performance differences do not have a big influence on the streaming retrieval.

### 4.2 Effect of Indirect Approach with B2BUA

We evaluated the effect of indirect approach with B2BUA. When a media streaming between CN and CP is transferred to device within mPAN, CP makes an indirect media streaming between B2BUA and device with 3PCC and REFER mechanisms We measured time required to make the indirect media streaming between CN and Dev from a direct media streaming between CN and CP, as shown phase A in Fig. 3. Then, we measured time required to transfer a media streaming from Dev to Dev2, as shown phase B in Fig. 3, compared with requiring renegotiation between CN and B2BUA and no renegotiation. In the case of streaming transfer between devices within mPAN, if the media information of forwarding device is same as previous device, renegotiation with CN is unnecessary.

The results are shown in Fig. 6. The time required to make the indirect media streaming between CN and Dev was approximately 1680 ms for video/audio streaming, 1325 ms for audio (uLaw) streaming, and 1880 ms for video (MJPEG) streaming. As mentioned in 3.3, the data loss did not occur in the mPAN basic streaming transfer. Because

the majority of transfer time depends on the processing at parties, it will be able to cut down with improving implementation of SIP UA and changing codec.

As the time required for transfer a media streaming from Dev to Dev2, The time required to transfer the media streaming that require renegotiation between CN and B2BUA was approximately 2230 ms for video/audio streaming, 2475 ms for audio streaming and 2780 ms for video streaming. In contrast, the time that did not require renegotiation between CN and B2BUA was approximately 667 ms for video/audio streaming, 414 ms for audio streaming and 530 ms for video streaming. The indirect approach with B2BUA dramatically reduced transfer time when renegotiation between parties was unnecessary. This transfer time will also be able to cut down with improving implementation and changing codec.

## 5 Conclusion

We presented a discussion on seamless streaming control among networked media devices with a mobile phone. The architecture is based on combining REFER method and 3PCC mechanisms of SIP. The architecture completely based on the SIP and its extensions provide a great affinity for future ubiquitous multimedia communications. The effects of additional Handover header and indirect approach with B2BUA were evaluated through prototype test-bed. From the experimental results, the additional Handover supported SIP UA dramatically reduced retrieval time compared with non-supported one. The indirect approach with B2BUA cut down transfer time when renegotiation between parties was unnecessary.

We have plans for improving implementation to cut down transfer time, implementing a service discovery protocol and creating the system on the Microsoft Windows Mobile platform. Furthermore, we also have to consider security and privacy concerns.

## References

1. J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session Initiation Protocol," *IETF RFC 3261*, 2002.
2. H. Wang, et al., "ICEBERG: An Internet-core Network Architecture for Integrated Communications," In *IEEE Personal Communications*, 2000.
3. M. Hasegawa, et al., "Cross-Device Handover Using the Service Mobility Proxy," In Proceedings of the *WPMC'03*, Vol. 2, pp.1033–1037, 2003.
4. K. El-Khatib, et al., "Personal and Service Mobility in Ubiquitous Computing Environments," In *Journal of Wireless communications and Mobile Computing*, Vol.4, pp.595-607, 2004.
5. R. Shacham, et al., "The Virtual Device: Expanding Wireless Communication Services through Service Discovery and Session Mobility," In Proc. of *WMCNC*, pp.73-81, 2005.
6. J. Rosenberg, et al., "Best Current Practices for Third Party Call Control (3pcc) in the Session Initiation Protocol (SIP)," *IETF RFC 3725*, 2004.
7. R. Sparks, "The Session Initiation Protocol (SIP) Refer Method," *IETF RFC 3515*, 2003.
8. S. Guha, Y. Takeda, and P. Francis, "NUTSS: A SIP based approach to UDP and TCP connectivity," In Proceedings of the *SIGCOMM'04 Workshops*, pp.43-48, 2004.
9. K. Weniger, "PACMAN: Passive Autoconfiguration for Mobile Ad hoc Networks," *IEEE Journal on Selected Areas in Communications (JSAC)*, Vol. 23, No. 3, pp.507-519, 2005.

# Development of Middleware GLIA for Connective Wide Collaborative Space with Networked I/O Device

Takaya Yuizono<sup>1</sup>, Shinichi Nishimura<sup>2</sup>, and Jun Munemori<sup>3</sup>

<sup>1</sup> Department of Knowledge System Science, School of Knowledge Science,  
Japan Advanced Institute of Science and Technology,  
1-1, Asahidai, Nomi, Ishikawa 923-1292, Japan

`yuizono@jaist.ac.jp`

<sup>2</sup> Department of Mathematics and Computer Science, Interdisciplinary Faculty of  
Science and Engineering, Shimane Univ.,  
1060, Nishikawatsu, Matsue, Shimane 690-8504, Japan

<sup>3</sup> Department of Design and Information Science, Faculty of Systems Engineering,  
Wakayama Univ.,

930, Sakaedani, Wakayama-shi, Wakayama 640-8510, Japan

`munemori@sys.wakayama-u.ac.jp`

**Abstract.** Middleware named GLIA has been developed to realize wide collaborative space with combining many desktop computers and shared work over all desktop space with a networked mouse. The middleware has been implemented with Java/Swing API and a networked mouse has been realized with mobile objects both to require no server and to guarantee scalability. The GLIA was applied to experiments to evaluate the networked mouse performance by the task that a user clicks at a target image, which was randomly appeared on the wide workspace. The results showed as follows; the mouse performance of GLIA was not inferior to an usual single mouse and supported collaboration environment of three persons with wide screens. The performance of UDP/IP implementation using GLIA was superior to both TCP/IP and Bluetooth implementation. GLIA seems to be effective for the space of idea generation method which need wide space of screen.

## 1 Introduction

Collaboration technology moves toward large workspace and emergent collaboration over ubiquitous computing environment [1], [2]. The movement is originally derived from knowledge media [3] and a concept “seamlessness” of creative support environment [4]. On the other hand, we have developed the groupware for a new idea generation that allows us to collect more data with mobile devices such as PDA [5]. However, more the number of data, more difficult we handle the data. To handle more data, two approaches have been taken: first is that many data has been analyzed by artificial intelligence technology [6], second approach is that human interface technology which assists human thinking [7], [8].

In this research, to choose second approach, a middleware has been developed to realize large shared workspace. The second approach is also a tool approach for supporting human thinking via interactive media such as groupware technology. The groupware technology for multi-user interface has been studied such as Group-Kit [9], Pebbles Draw [10] and GDA [11]. However, they have some difficulty to develop wide collaborative workspace because they do not support building collaborative space both with multi output for combining desktop monitor and with multi input for multiuser cursors.

This paper describe about middleware named GLIA (Groupware-kit for Linking Interactive Action), which has been developed for combining many PC displays and support multiuser cursors via network. Next, the performance of a networked mouse is examined with a target-clicking task.

## 2 Middleware GLIA

### 2.1 System Configuration

The goal of middleware named GLIA is to make shared space for supporting a large workspace and multiuser interface for group interactive action for catching emergent intelligence. The middleware allows us to combine GUI displays with a mouse cursor via network. A programming language for the development of GLIA is a Java programming language that has good code transportation, multi thread for better implementation of multiuser interaction, and object serialization for code mobility to realize virtual networked devices. Fig. 1 shows a concept structure of middleware GLIA.

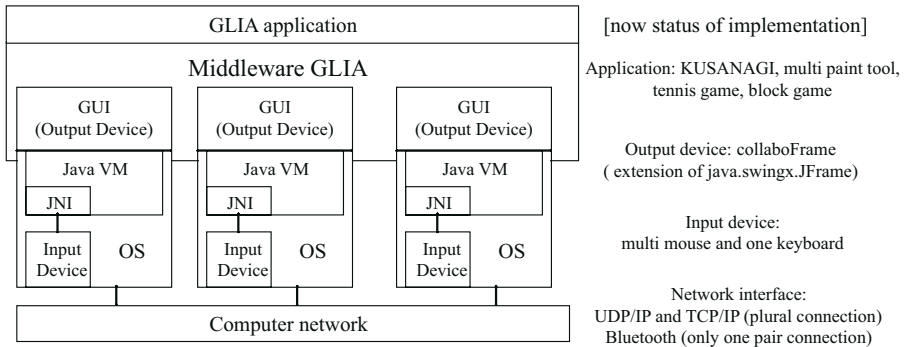


Fig. 1. Concept of middleware named GLIA and the present implementation

The middleware aims to combine output devices for scalable workspace with multi input devices via computer network. The right side of Fig. 1 indicates the present implementation status of GLIA. The implementation is described below;

the GUI screen as output device is extension of JFrame class that is a core class of JAVA/SWING API, the input device is multi mouse and one keyboard per a computer. All input devices and output devices are combined by GLIA's network with UDP socket, TCP socket and Bluetooth API. The Bluetooth API has been developed with Java Native Interface and the present implementation supports only one pair connection. The number of code lines of GLIA is approximately 4,000.

A mouse cursor via network is realized by a mechanism shown in Fig. 2. A mouse monitor observes a mouse device. The monitor inspects a state of the mouse device every 20 milliseconds. The state reflects the position of a mouse cursor on workspace via a mouse agent. The mouse agent can move one computer to other computer via network. When the mouse cursor crosses to the other display, the mouse agent moves to the other PC with the display.

This networked mouse has one cursor realized by a pair of software objects those are a mouse monitor and a mouse agent. No server for managing all cursors guarantees a scalability of networked mouse.

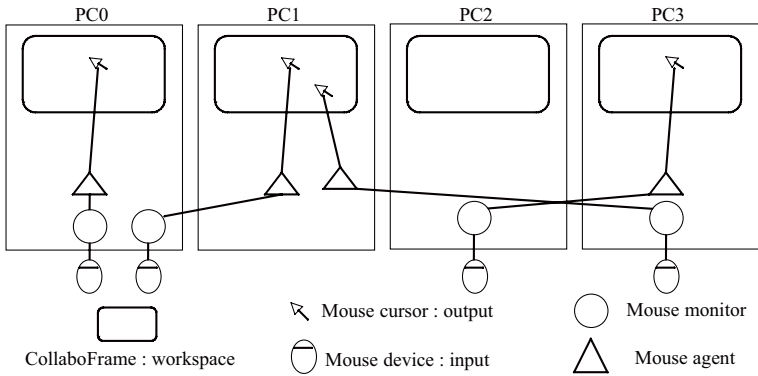


Fig. 2. GLIA mechanism of a networked mouse

## 2.2 GLIA Application

A groupware named KUSANAGI [12] for a new idea generation has been developed. The groupware supports an execution of KJ method with multi PCs. The KJ method is well known in Japan as an idea generation method from collected data. In Fig. 3, five computers are horizontally arranged in a line. The GLIA allows us to access all five personal computers row with only one mouse. The middleware are applied to the execution of the KJ method of 211 data labels on width 5,760 and height 814 pixels workspace as shown in Fig. 3.

On the other hand, GLIA supports Bluetooth wireless communication. Therefore, the middleware can make an ad hoc connected workspace shown in Fig. 4. In the figure, a multipaint application is executed on twice screen size workspace on two notebook computers.



**Fig. 3.** Overview of GLIA extended workspace (width 5760 pixel x height 814 pixel) that carried out groupware named KUSANAGI for a knowledge creation



**Fig. 4.** Two notebook computers combined with GLIA via Bluetooth and an example application is a multiuser paint tool

### 3 Performance Experiments of Networked Mouse

In order to evaluate a performance of networked mouse with GLIA, some experiments had been done. The task of experiments was mouse clicking at a target image (32-pixel width and 32-pixel height) randomly appeared on a workspace. Fig. 5 shows a case of combined 5 PC displays. When a user clicked at the target image, next target was randomly appeared on the workspace. Then, the time was measured since the next target was clicked and the distance between the location of previous task and the next target was measured.

### 4 Results and Discussion

Table 1 shows the networked mouse performance of varying the number of connected PCs. The result shows that more number of PCs, which realize a larger



**Fig. 5.** An example screen of networked mouse experiments with five computers combined with GLIA that realized a 5760-pixel x 814-pixel size workspace

workspace, makes mouse speed faster. Because, faster you move a mouse, the faster the cursor corresponding to the mouse moves. This result shows that the performance of networked mouse does not go down though enlarging the workspace. In other hand, the load test of the number of mouse devices was investigated. The result showed that about ten mouse devices had the same user performance as one connected mouse.

**Table 1.** Networked mouse performance of varying the number of connected PCs

N=300	Number of PCs	Workspace size(pixel)	Distance (pixel)	Time (ms)	Speed (pixel/s)	Error rates
	7	8064 x 814	2618	2614	1001.5	0.19
Networked	6	6912 x 814	2415	2331	1035.9	0.16
	5	5760 x 814	1994	1936	1029.8	0.15
mouse	4	4608 x 814	1542	1600	963.6	0.14
	3	3456 x 814	1208	1460	827.6	0.11
with GLIA	2	2304 x 814	866	1300	666.5	0.15
	1	1152 x 814	523	1045	500.3	0.10
Usual cursor	1	1152x814	525	966	543.1	0.17

Table 2 shows the networked mouse performance of varying the number of participants. This results shows that networked mouse by multi user had good performance for the target-clicking task. The time in the cases of three persons were shorten than the cases of two persons. And, the speed in the case of the usage of 7 PCs by three persons was the fastest in all cases.



**Table 2.** Networked mouse performance of varying the number of participants

N=100	Number of PCs	Distance (pixel)	Time (ms)	Speed (pixel/s)	Error rates
Three participants with GLIA	7	3137	1573	1993.7	0.13
	5	2024	1224	1653.6	0.20
	3	1182	1161	1017.5	0.23
	1	488	888	549.4	0.15
Two participants with GLIA	7	2801	1810	1547.0	0.11
	5	1851	1445	1281.3	0.13
	3	1248	1189	1049.8	0.17
	1	517	999	517.4	0.15
Usual cursor	1	525	966	543.1	0.17

**Table 3.** Effect on networked mouse performance of varying the method of communication in the case of using two PCs

	Media	Distance (pixel)	Time (ms)	Speed (pixel/s)
UDP socket	Ether cable	862.3	1498.9	554.9
TCP socket	Ether cable	852.6	2285.7	367.5
Bluetooth	Wireless	773.6	1947.6	389.5

Table 3 shows the networked mouse performance of varying the method of communication. The result shows that the performance of UDP socket is superior to both TCP socket and Bluetooth. In a context of the workspace shown as Fig. 5, unreliable communication UDP socket is not a matter, because the workspace locates in reliable local area network.

In other hand, the moving time of a cursor from one PC to another PC was measured. The cursor was controlled with a mouse agent in the case that those PCs were not connected with a mouse device directly. The moving time was about 40 ms and almost stable between two PCs. Therefore, the performance of networked mouse was not fallen down by the number of connected PCs and the cursor moving time was affected by the number of PCs that the cursor walked across.

## 5 Conclusion

Middleware named GLIA has been developed to support a large workspace with the combination of GUI displays with networked mouse. The performance of networked mouse was experimented by the task of mouse clicking at a target image. The results showed that the networked mouse of the middleware kept good performance with the increasing of the combined number of personal computers. The UDP/IP implementation was not superior to usual one mouse performance and better than TCP/IP implementation and Bluetooth implementation. SO,

GLIA seems to be effective for the space of idea generation method which need wide space of screen.

In near future, GLIA will be applied to the evaluation of a knowledge creation on large workspace and evolve to large collaboration space between distance sites.

## References

1. Streitz, N. et al. : i-LAND: an interactive landscape for creativity and innovation, Proc. of CHI'99 (1999) 120-127
2. Russel, D., Streitz, N. and Winograd, T. : Building Disappearing Computers, Com. of the ACM, Vol. 48, No. 3 (2005) 42-48
3. Stefik, M. : The Next Knowledge Medium. The AI Magazine, Vol. 7, No. 1 (1986) 34-46
4. Stefik, M. and Brown, J. S. : Toward Portable Ideas. in Technological Support for Work Group Collaboration, Edit. by Olson, M.H., Lawrence Erlbaum Associates (1989) 147-165
5. Yuizono, T. , Munemori, J. and Nagasawa, Y. : Application of Groupware for a New Idea Generation Consistent Support System using PDA for Input Device. Proc. of ICPP Workshops on INDAP, IEEE Press (1999) pp.394-399
6. Chen, H. et al. : Automatic Concept Classification of Text from Electronic Meetings. Com. of the ACM, Vol. 37, No.10 (1994) 56-73
7. Misue, K. et al.: Enhancing D-ABDUCTOR towards a diagrammatic user interface platform. Proc. of KES'98 (1998) 359-368
8. Shigenobu, T., Yoshino, T. and Munemori, J. : Idea Generation Support System GUNGEN DX II Beyond Papers. Proc. of KES'03, LNAI 2774, Springer-Verlag (2003) 741-747
9. Roseman, M. and Grenberg, S. : Building real-time groupware with GroupKit, a groupware toolkit. TOCHI, Vol. 3, ACM Press (1995) 66-106
10. Myers, B. A. , Stiel, H. and Gargiulo, R. : Collaboration using multiple PDAs connected to a PC. Proc. of CSCW'98, ACM Press (1998) 285-294
11. Munemori, J., Noda, T. and Yoshino, T. : Group Digital Assistant: Combined or Shared PDA Screen Proc. of ICDCS'04, IEEE Press (2004) 682-689
12. Yuizono, T. et al. : A Proposal of Knowledge Creative Groupware for Seamless Knowledge. Proc. of KES 2004, LNAI 3214, Springer-Verlag (2004) 876-882

# Development and Evaluation of No-Record Chat System Against Screen Capture

Takashi Yoshino and Takuma Matsuno

Systems Engineering, Wakayama University, Sakaedani 930,  
Wakayama, Japan

yoshino@sys.wakayama-u.ac.jp

<http://www.wakayama-u.ac.jp/~yoshino/>

**Abstract.** To prevent leakage of private information becomes serious issues in the field of information technology. The security of the telecommunication line is almost enough. However, the environment that does a secret conversation is insufficient through the network. We found that it can be embarrassed that the record of the conversation remains though a person wants to communicate privately with another. In this paper, we proposes the chat system that can take communications to solve the above-mentioned problem. We have developed a chat system named 'you-me Chat.' They cannot capture the screen of the conversation though users can read the conversation sentences in you-me Chat while chatting. In you-me Chat, the character is resolved to imperfect parts as image data. They are displayed continuously. They are imperfect characters even if a person preserved them on a PC as data. A person can read them for image lag of eyes. We present the effectiveness of the system from the experiments.

## 1 Introduction

A rapid spread of the Internet increases the chance of communications in the future through the network. To prevent leakage of private information becomes serious issues in the field of information technology. The security of the telecommunication line is almost enough. However, the environment that does a secret conversation is insufficient through the network. Ministry of Internal Affairs and Communications in Japan said that 63.3% of the Internet users feel dissatisfied and insecurity in the use of the Internet [1]. There are some researches about image-based user authentication systems [2,3]. Our research uses the image-based as well as the above-mentioned researches. In our system, however, the different point is to use the lag of images of human's eyes. We found that it can be embarrassed that the record of the conversation remains though a person wants to communicate privately with another. In this paper, we proposes the chat system that can take communications to solve the above-mentioned problem. We have developed a chat system named you-me Chat. They cannot capture the screen of the conversation though users can read the conversation sentences in

you-me Chat while chatting. In you-me Chat, the character is resolved to imperfect parts as image data. They are displayed continuously. They are imperfect characters even if a person preserved them on a PC as data. A person can read them for image lag of eyes. We present the effectiveness of the system from the experiments.

## 2 You-Me Chat

### 2.1 Realization Method

The features of you-me Chat are the followings.

- The conversation sentences can be only read while chatting.
- The conversation sentences cannot be preserved by the screen capture.

We devised the following realization method. The imperfect image data of conversation sentence is continuously displayed. The proposal method draws continuously in parts as shown in Figure 1. The image of the original text of Figure 1 shows an image before it is resolved. Three images of bottom of Figure 1 show the images that the image of the original text is resolved. The user sees the resolved images. The method can be expected in the following effects.

- A person can read them for image lag of eyes while chatting.
- They are imperfect characters even if a person preserved by screen capture them on a PC as data.

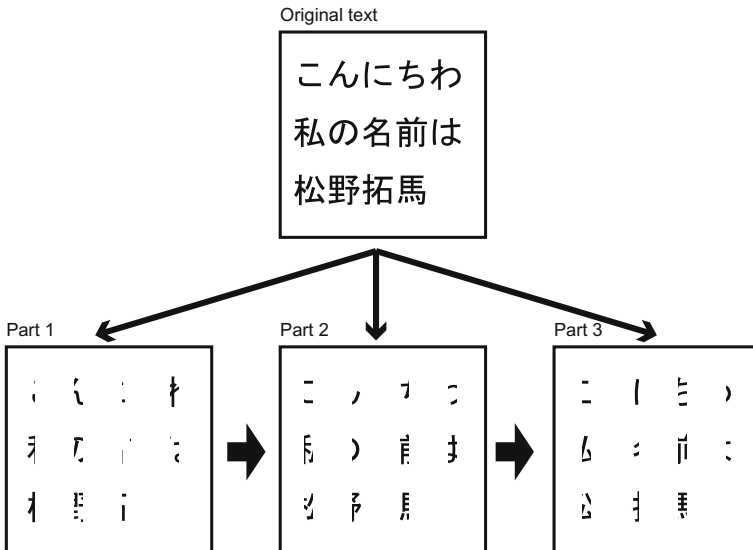


Fig. 1. Realization method of you-me Chat

## 2.2 Drawing Pattern

There are a lot of making methods of imperfect characters. We should look for an appropriate drawing method. In you-me Chat, we selected the following three kinds of parameters of drawing pattern. Figure 2 shows the procedures of making drawing patterns. The mask pattern is overprinted at the original text. In other words, the image operation ‘OR’ is carried out to the original text and the mask pattern.

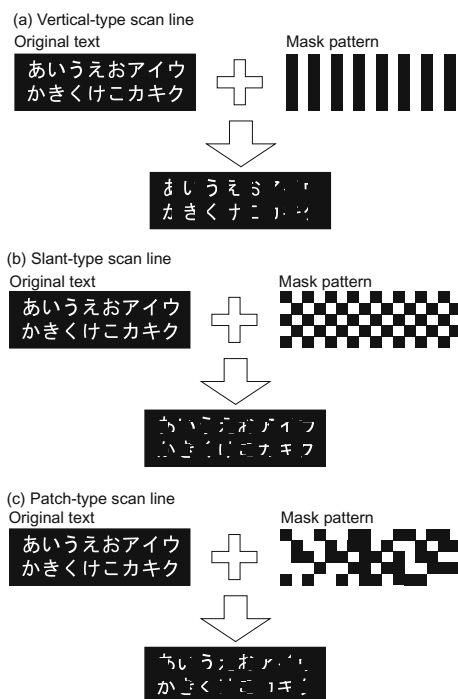


Fig. 2. Procedures of making drawing pattern

- Mask area ratio (5 kinds)

The mask area ratio is a ratio where the character is hidden. When the mask area ratio is large, the ratio that conceals the screen grows. At that time, the character comes not to read easily. This time, we use five kinds (47%, 52%, 58%, 65%, and 71%).

- Rectangular size (3 kinds)

A rectangular size is a minimum size rectangular on a pattern. This becomes a resolution when the character is concealed. We use 3 kinds (2pixelC4pixel, and 8pixel).

- Drawing methods (scan line) (3 kinds)

We adopted the following three kinds of methods for drawing methods. We call this ‘scan line’.

- Vertical-type scan line
- Slant-type scan line
- Patch-type scan line

Then, we have 45 drawing patterns shown in Table 1.

**Table 1.** 45 kinds of drawing patterns

Mask area ratio	(a) Vertical-type scan line			(b) Slant-type scan line			(c) Patch-type scan line		
	Rectangular size			Rectangular size			Rectangular size		
	2pixel	4pixel	8pixel	2pixel	4pixel	8pixel	2pixel	4pixel	8pixel
52%									
58%									
65%									
71%									
47%									

### 3 Un-Readability Patterns on Screen Capture Image

We have to find un-readability patterns on screen capture image out of 45 patterns. We experimented to find the appropriate patterns. Subjects are 29 students of Wakayama University. Table 2 shows the appropriate patterns, the combination of scan line, mask area ratio and rectangular size, from the experiments. No one of the subjects was able to guess the sentence meaning. Moreover, no one of the subjects was able to read two words or more.

**Table 2.** Combination of Scan line, mask area ratio and rectangular size

(Mask area ratio, Rectangular size)	(a) Vertical-type scan line	(b) Slant-type scan line	(c) Patch-type scan line
(65%, 2 pixel)			
(71%, 2 pixel)			
(65%, 4 pixel)			
(71%, 4 pixel)			

### 4 Experiments

We have developed a chat system you-me Chat. Figure 3 shows a screenshot of you-me Chat. A user can change the drawing patterns on you-me Chat. We experimented to evaluate the effectiveness of you-me Chat.

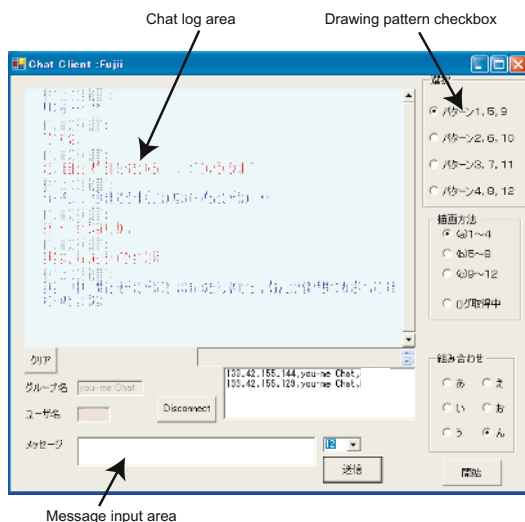


Fig. 3. Screenshot of you-me Chat

#### 4.1 Experiment 1: Readability on the Screen of You-Me Chat

In experiment 1, we use the patterns of Table 2. In the patterns of Table 2, we confirmed that no one of the subjects was able to guess the sentence meaning. The experiment method is shown as follows. At first, a subject reads a different sentence in 12 patterns on you-me Chat. Moreover, a subject selects one pattern of the most legible in 12 patterns. Subjects are 36 students of Wakayama University.

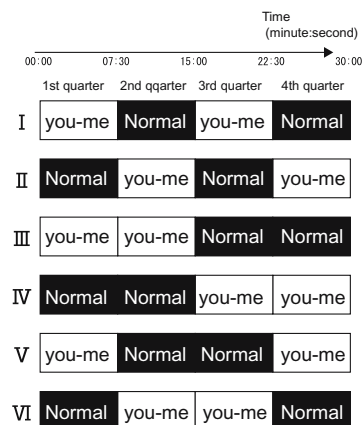


Fig. 4. Experiment Combination between you-me Chat and normal chat

## 4.2 Experiment 2: Chat Experiments

The purpose of experiment 2 is a comparison with you-me Chat and a normal chat. Subjects are 36 students of Wakayama University. The time of the chat is 30 minutes, and the 30 minutes were divided into four parts: Two quarter for you-me Chat, the rest two quarter for a normal chat. Figure 4 shows the combination of quarters. We asked the subjects to talk about something secret in advance. All subjects chose own romance as the topic. A subject counts the his/her secret messages while chatting.

## 5 Results and Discussion

### 5.1 Readability on the Screen of You-Me Chat

Table 3 shows the results of readability on the screen of you-me Chat. Circle mark shows the pattern evaluated that all subjects can read. Table 4 shows the number of people answered as the most seeing easily. It is natural that 34 people of 36 people chose the patterns of 65% mask ratio. 21 people chose slant-type, 2 pixel rectangular size, and 65% mask area ratio.

**Table 3.** Results of readability on the screen of you-me Chat

Scan line	Rectangular size	Mask area ratio	
		65%	71%
Vertical-type	2		
	4		
Slant-type	2		
	4		
Patch-type	2		
	4		

### 5.2 Number of Secret Messages

Table 5 shows the comparison with the number of secret messages and the number of messages on the experiments. There is no difference in the number of

**Table 4.** Number of people answered as the most seeing easily

Scan line	Rectangular size	Mask area ratio	
		65%	71%
Vertical-type	2	3	0
	4	9	0
Slant-type	2	21	2
	4	1	0
Patch-type	2	0	0
	4	0	0
Total		34	2



messages between you-me Chat and a normal Chat. However, we found the significant difference in the number of secret messages between you-me Chat and a normal Chat. We found that the subject can chat about secret easily using you-me Chat.

**Table 5.** Comparison with number of secret messages and number of messages on the experiments

	you-me Chat	Normal chat
Number of secret messages	1.64**	0.75**
Standard deviation	2.01	1.04
Number of messages	28.17	27.10
Standard deviation	13.52	13.77

\*\* : p-value < 0.01

### 5.3 Questionnaire Result

Table 6 shows the result of the questionnaire. We had a 5-point scale (with 1: Strongly disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly disagree) for subjects to evaluate each item on the questionnaire. Table 7 shows the impression of the experiments. From Table 6 (1), most of students use the chat software in daily life. From Table 6 (1) and (2), the subjects communicated with the partner on you-me Chat. From Table 6 (4), the score of the question of ‘My eyes were tired’ is 2.8. We cannot disregard the value. However, 16 of 36 people do not feel the unpleasantness. From Table 6 (5), most subjects talked about the secrets under experiment. This result shows that most experiments were executed according to our design. From Table 6 (6), a subject was able to speak the quarter of you-me Chat more comfortably than the quarter of a normal chat. From Table 7 many people answered that I felt comfortably using you-me Chat. This result is corresponding to increasing of the number of secret messages in Table 5 .

**Table 6.** Result of questionnaire survey

Questionnaire items	Evaluated number of people					Average Value
	1	2	3	4	5	
1. I usually use chatting.	3	1	0	8	24	4.4
2. I smoothly communicated with the partner on you-me Chat.	0	3	2	17	14	4.2
3. I was able to read the partner's message on you-me Chat.	0	1	0	15	20	4.5
4. My eyes were tired.	5	11	7	11	2	2.8
5. I told some secrets each other on this chatting.	0	2	8	12	14	4.1
6. I was able to speak the quarter of you-me Chat more comfortably than the quarter of normal chat.	0	1	7	8	20	4.3

**Table 7.** Impression of the experiments

- I heard that the content of the normal chat was recorded. So I felt comfortably using you-me Chat.
- I felt that it was not easy to chat at the quarter of a normal chat. At the quarter of you-me Chat I felt that the talk was rapidly developed.
- At the quarter of a normal chat, I became shameful feelings clearly seeing the messages.
- I felt atmosphere of sending the eyes-only document.
- I am not concerned about the remainder of data.

## 6 Conclusion

We have developed a chat system ‘you-me Chat.’ The conversation sentences can be only read while chatting using the you-me Chat. However, the conversation sentence cannot be preserved by the screen capture. From the experiments, we found the people tend to communicate of a secret message easily on you-me Chat.

## References

1. Ministry of Internal Affairs and Communications, Result of Communication use trend investigation in 2004, [http://www.soumu.go.jp/s-news/2005/050510\\_1.html](http://www.soumu.go.jp/s-news/2005/050510_1.html).
2. Jermyn, I., Mayer, A., Monrose, F., Reiter, M. K. and Rubin, A. D.: The design and analysis of graphical passwords, Proc. 8th USENIX Security Symposium, Aug. 1999.
3. Harada, A., Isarida, T., Mizuno, T., Nishigaki M.: A User Authentication System Using Schema of Visual Memory, Journal of Information Processing Society of Japan, Vol. 46, No. 8, pp. 1997-2013, 2005.

# Home-Network of a Mutual Complement Communication System by Wired and Wireless

Kunihiro Yamada<sup>1</sup>, Takashi Furumura<sup>2</sup>, Yasuhiro Seno<sup>2</sup>, Yukihiisa Naoe<sup>2</sup>,  
Kenichi Kitazawa<sup>2</sup>, Toru Shimizu<sup>3</sup>, Koji Yoshida<sup>4</sup>, Masanori Kojima<sup>5</sup>,  
Hiroshi Mineno<sup>6</sup>, and Tadanori Mizuno<sup>6</sup>

<sup>1</sup>Tokai University

yamadaku@tokai.ac.jp

<sup>2</sup>Renesas Solutions Corporation

<sup>3</sup>Renesas Technology Corporation

<sup>4</sup>Syounan Institute of Technology

<sup>5</sup>Osaka Institute of Technology

<sup>6</sup>Shizuoka University

**Abstract.** Although the system of the individual purpose exists in a home, a network for home use does not exist. There are three important problems to home network realization. It is cost's being cheap, and there being no necessity for construction, and being able to communicate anywhere in a network. In order to realize this, the network system which works simultaneously two communication systems, wireless (IEEE 802.15.4Zigbee) and wired (Power line carrier communication ; PLC ) communication, was examined .

## 1 Introduction

In public institutions, such as an office, a factory, a store, an airplane, the Shinkansen, a hospital, a school, and a public office, a network is put in practical use. Furthermore, we are going to build the new interface between environment and human being by adding a sensor and a tag tip to a network[1]. As natural environment here, they are the weather, and the atmosphere and an earthquake. Moreover, as artificial environment, it is a highway and a car, and they are a city and a building[2]. It is tried to carry out to the past data or prediction as an example using mobile apparatus in addition to the present position of course recommendation service, a man, or a move object[3][4].

Although the system with individual PC Internet, hot-water supply system, AV system, home telephone, interphone, security system, etc. exists in a home[5], the home network which cooperates each other does not exist. What a domestic network aims at does not stop at one domestic problem, but has a spread to the area, a country, or a whole-world level. It consists of three items greatly.

One is reservation of domestic safety. The 2nd is energy control which supports cleaning of an earth scale, and both sides of household economy. Energy consumption control and working control of the power generation system by sunlight or wind force are performed. The 3rd is support of the convenience of a home life.

Also from the thing of a domestic network to aim at, this domestic network will be installed in all the homes of one area and one country, it will begin, and the purpose will be attained. Considering installation of the network in a home with various conditions, a quite high level will be required of the communication performance and communication quality of this network. Then, the conditions to such realization of a home network are expense's being cheap, and there being no necessity for construction, and being able to communicate anywhere in a network, if it is domestic again. Here, it is considering as 300m of floor area with 3 stories of ferro- concrete as a size.

## 2 Communication Characteristic

In order to guess the communication performance of the mutual complement network by two communication systems, wired and wireless, an individual communication performance is measured first [6]. Power line communication ( PLC ) is used as wired communications. This has the very important feature in this research of not needing installation of a new wired. Moreover, as wireless communications, permission of related administration uses unnecessary IEEE802.15.4(ZigBee) [7] [8]

### 2.1 Wired-Communications Performance

By the communication system using the electric light line of ordinary homes, it is called PLC; power line communication. By home use, it thinks as one of the leading communication systems [9].

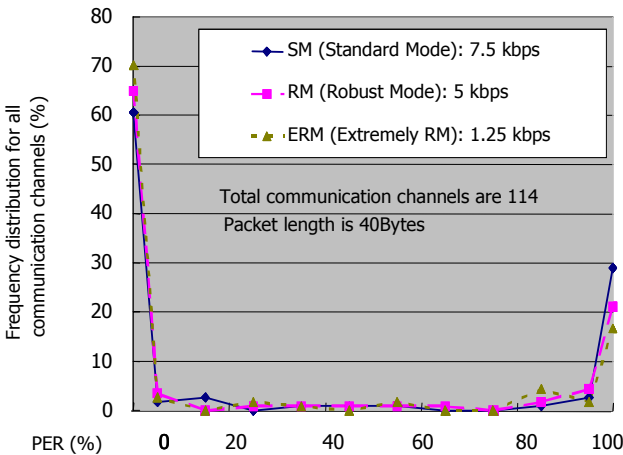


Fig. 1. Testing result of PLC (100kHz–400kHz) in typical Japanese home

Moreover, it has obstacles by home apparatus, such as crosstalk of a noise, low impedance, and a signal. Therefore, reservation of communication quality was conventionally difficult. In recent years, the communication performance is improving by

adoption of spectrum diffusion technology. [10]. 40 or more home apparatus was operated and the communication performance of PLC was measured. Fig.2.is as a result of measurement. The X-axis is that express Packet Error Rate (PER), then 0% will be the best and error calls it 0%.

The Y-axis expresses the frequency distribution of all communication courses with %. 0% of PER is 70% of frequency distribution from Fig. 2, and 100% of PER is 20% of frequency distribution. From 0% to 100% of PER in the meantime of PER is 10% of frequency distribution.

## 2.2 Wireless-Communications Performance

As wireless, IEEE802.15.4 is used by one of the WPAN (Wireless personal area network) plans. Following ZigBee is called. This has the performance of a low rate (20Kbps or 250Kvps)by low power consumption. The communication performance of this ZigBee was measured. In prospect distance open air, electric field intensity is decreased, as it separates from near an output antenna so that naturally. Especially, as the second floor shows to Fig. 5 from the first floor, PER shows 40% and a very large value in the house of ferro-concrete. Moreover, on the third floor, PER shows 70% and a still larger value from the first floor, and a communication performance becomes still worse.

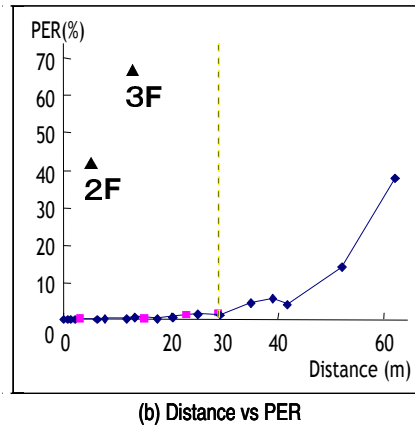


Fig. 2. IEEE802.15.4 test result data ( b )

## 3 Mutual Complement Network of Wired and Wireless

The preceding clause showed the characteristic of PLC and ZigBee. Both do not need installation of a wired. That is, although it is the communication system of construction needlessness, it is difficult to perform an home network by one of communication systems.

PLC has the problem of single phase 3 line electric supply, and its communication performance is it bad between the wall sockets between differnt phase.

Although it is satisfactory in the space which can keep seeing ZigBee, it is weak with the obstacle which interrupts a field of view. Then, the communication performance over a floor is bad like the first floor and the second floor Since it has generated according to respectively independent conditions, each other is suppliable with the characteristic of these two communication systems. The communication performance when working two communication systems simultaneously below is guessed.

### 3.1 Examination of the Communication Quality

We examined the communication quality for a detached three-storied house with a 300m<sup>2</sup> floor area using the data in Sections 2. Although in the case of the communication quality is greatly affected between different phases, we decided not to use a phase coupler assuming that no physical changes should be made in a simply communication system.

The "Truth" category made up 70%, the "Truth?" category 10%, and the "False" category 20% as shown in Fig.2. In wireless communication, the percentage of the "Truth" was 82%, "Truth?" 14%, and "False" 4% after they were weight-averaged, as shown in Fig.3.

Shown in Fig.3. are the communication reliabilities when the wired and wireless media are used simultaneously. The communication quality is improved from 70% to 94.6% for the independent wired communication and from 82% to 94.6% for the independent wireless communication, which can be interpreted as an improvement to 96% if the "Truth" and "Truth?" categories are considered as the "Truth" category.

		Wireless		
		Truth. 82	Truth? 14	False. 4
Wired	Truth. 70	57.4	9.8	2.8
	Truth? 10	8.2	1.4	0.4
	False. 20	16.4	2.8	0.8

Fig. 3. The Communication quality in % when the wired and wireless communication media are used simultaneously

### 3.2 Verification Experiment

Next, a verification experiment of the mutually complementing wired and wireless network system was conducted in a three-story detached house (Fig.4). In the demonstration test, a success rate of 100% was obtained in data communication for all combinations from the 1st floor to the 3rd floor, as shown in Fig.4. The table also shows the success rate between single unit cells for wired and wireless communication.

The average success rate was 80% for wired single, and 81.7% for wireless single. In wired single, a combination showed a success rate of about 98% or lower even on the same floor, which was supposed to be coming from the different phase connection in the power wired, and this agreed with the result of an actual check of the power distribution on each floor (Fig.5.).

		Overall success rate			Wired single success rate			Wireless single success rate		
		Floor of reception								
		1F	2F	3F	1F	2F	3F	1F	2F	3F
Floor of transmission	1F	100	100	100	98	80	74	100	83	55
	2F	100	100	100	82	81	75	81	87	84
	3F	100	100	100	75	76	76	60	83	93

Fig. 4. The house of an evaluation experiment. This unit is % display

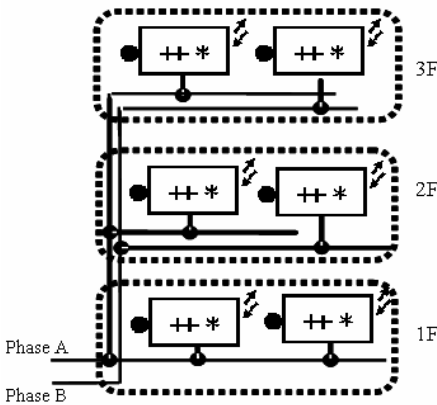


Fig. 5. An actual check of the power distribution condition

		Floor of reception		
		1F	2F	3F
Floor of transmission	1F	100%	97%	88%
	2F	97%	97%	96%
	3F	90%	96%	98%

Fig. 6. Assumed success rate between unit cells

For wireless communication, the success rate was 93% between two unit cells on the same floor, about 83% between adjacent floors, and about 58% between two floors separated by two or more floors in between, and thus the range of transmission decreased as the difference in floors increased.

If backup for a communication error arising during wired single communication or wireless single communication is done by selecting wired and wireless in the unit cells, the error rate is:  $\text{Wired single communication error rate} \times \text{Wireless single communication error rate}$ . According to the rate of a single medium, the success rate is as shown in Fig.14.

The average success rate in Fig.14 is 95.7%, which almost agrees with the simulation value, 96%.

### 3.3 Consideration for Generation

The communication performance between the same phase of the general adaptation PLC is 98% from fig 12; 1F & 1F of mutual complement network of PLC and ZigBee.

This is based on the fact that 1F are wired only by the same phase. Since all the communication performance containing the same phase and different phase are 70%, so the rate of different phase becomes 42%.

Next, the communication performance in the case of different stories should be considered. Of course, from fig 5, the characteristic of the performance becomes worse in proportion about distance in two points of experimental points. From fig 12, the communication performance of the same floors would be 93.3%. Moreover, the average value of the data for the first floor of 1F-2F and 2F-3F should be 82.8%, and similarly the data of the second floor of 1F-3F is 57.5%.

From these values, the communication performance based on the mutual complement between each floor is calculated and is shown in Fig A1. The communication performance of the PLC is 70% as mentioned above.

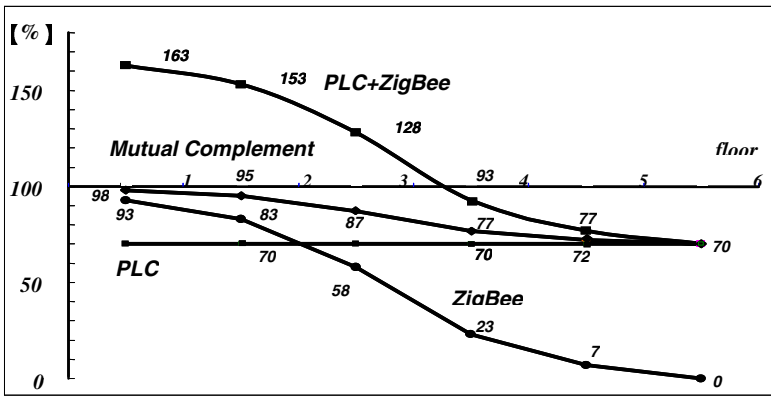


Fig. 7. PLC/ZigBee Mutual Complement

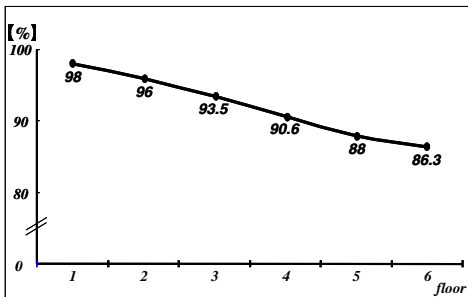


Fig. 8. PLC / ZigBee Transmission rate (Expected)

		Receiving node No.					
		1	2	3	4	5	6
Sending node No.	1	—	Z	P	x	P	x
	2	Z	—	x	P	x	P
	3	P	x	—	Z	P	x
	4	x	P	Z	—	x	P
	5	P	x	P	x	—	Z
	6	x	P	x	P	Z	—

P : Success by PLC communication  
 Z : Success by 802.15.4 communication  
 X : Communication impossible

Fig. 9. Communication between two nodes



ZigBee performance without experimental is considered with guess value about the 4<sup>th</sup>, 5<sup>th</sup> and 6<sup>th</sup> floor. And the communication characteristic of the home network with 1-6 floors is shown in Fig A8. Increasing floor number, ZigBee communication performance ,i.e., the radio performance becomes worse shown in Fig A7. The plc communication would be OK within 6<sup>th</sup> floors distance level.

#### 4 Improvement of Communication Quality with Node Relay

As shown on previous discussion on Fig A2; such as radio and wired mutual complement method, the communication performance becomes worse according to increasing floor number like 98% to 86% at 6<sup>th</sup> floor building. In order to improve this, in the different floor level, PLC; wired communication should be used and in the same floor level, phase free communication such as ZigBee communication is useful. This means that routing method between PLC and ZigBee should be considered as a perfect double complement communication.

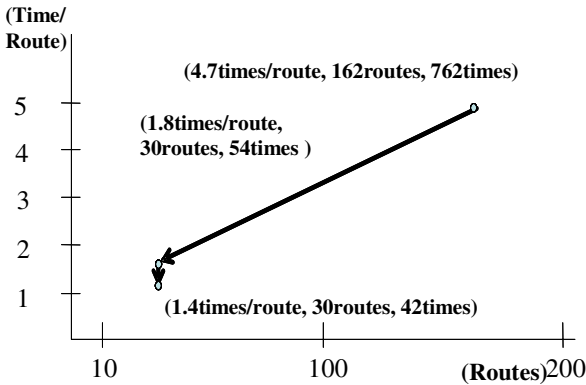


Fig. 10. Verification result of algorithm



Fig. 11. Picture transmission by PLC

#### 5 Still Image Transmission by Using PLC

As an actual data transmission, JPEG still image transmission between 2 nodes has been tried. Original data was 320x280 dots; 4Kbytes data by JPEG compression. It took about 4 seconds to transmit the data. Actual data is shown on Fig.11.

#### 6 Conclusions

We studied the possibility of developing networks for relatively small spaces the size of family homes which are easy to install and operate using two different communication media, and found it possible to improve the communication quality with this technology. Each communication performance of wired communications and wireless

communications was 70% and 82%, and in the mutual complement system, although the simulation which becomes 96.0% had been obtained.

But after actual experimental evaluation, communication performance was different according to building floor numbers. This is caused by the ZigBee's characteristic such as worse communication performance because of ceiling and floor barrier.

In order to improve this, the relay method such as PLC and ZigBee routing method should be considered.

## References

1. Y.Shiraishi,M.Arikawa"A Personal Spatial Information System with Function for sensor Data Mapping "IPSJ Symposium Series Vol.2004.no7.pp33-36.
2. Runesas Technology "Solution Seminar Text book" Nov. 2003.
3. J.Sawamoto, et al, "Multi-Agent System Development Framework for Location Based Services" IPSJ Symposium Series Vol.2004.no7.pp41-44.
4. T.Nakamura, et al "Implementation of Green Town", IPSJ Symposium Series Vol.2004.no7.pp293-296
5. T.Okuda, et al, "Networking Gas Central Heating System", Matushita Technical Journal Vol.50 No.3 Jun.2004
6. K. Yamada,etc"Dual Communication Systemu using Wired and Wireless Correspondence in aSmall Space ",kes2004,P-II,pp898-904
7. URL <http://www.zigbee.org/>
8. M16C/6S <http://www.renesas.com>
9. URL <http://www.yitran.com>
10. M.Kubo,A.Kurobe,S.Yoshida,M.Watanabe"Power Line Communication and Development of Power Line Communication Modules", Matushita Technical Journal Vol.49 No.1

# Interaction Between Small Size Device and Large Screen in Public Space

Chunming Jin, Shin Takahashi, and Jiro Tanaka

Department of Computer Science, University of Tsukuba  
{kin, shin, jiro}@iplab.cs.tsukuba.ac.jp

**Abstract.** We propose a simple way to interact with a large screen which is situated in a public space by using a small size device. In order to get the connection information easily, we use QR-Code to provide connection information such as IP address, position information and password to users. When users are connected to the large screen, the system will provide a part of the large screen to a user as his personal space at right in front of him, so that the user can display the information in his PDA and manage the information on the large screen. The user can exchange the data with the large screen and the other users through the personal space on the large screen. We keep developing the prototype system according to this approach.

**Keywords:** Ubiquitous, Interaction, Large screen display, Public space, PDA, QR-Code.

## 1 Introduction

Small size devices such as cell-phones and PDAs (Personal Digital Assistant) are handy to carry. By utilizing wireless network such as WiFi and Bluetooth, users should have the capability to connect to the internet and handle information anywhere and anytime. For example, users can manage their schedule or access the internet with their cell-phone or PDA. Users can also send a mail or even can shop by accessing the internet. However, a small size device has a small size display. It cannot display much information on one screen, which forces users to switch the pages all the times. On the other hand, a large screen can be useful to provide information to users in public space. Currently, large screens are becoming ubiquitous and being used in public spaces such as subway stations, shopping malls, book stores, etc. It will be useful if we could use the large screen in public space with a small size device. In this paper, we will describe how to interact with a large screen in public space by using a small size device.

Our goal is to build a system where a PDA can be used as an input device to operate data on a large screen display in public space. We suppose that we have a wall-sized screen and a small size device which can connect to the internet. Our system provides a limited space to a user as his personal screen area on the large screen, thus the user can display the information in his PDA on the large screen. Users can exchange the information in a simple way with a drag & drop interface on the large screen. Our system also supports a multi-user environment. Therefore, users can

operate information with other users, such as exchanging data with each other through the large screen.

## 2 System Overview

This section describes the overview of our system. To start using the system, users have to connect with the large screen server. We use the QR-Code [9] to provide connection information to users. When a user is connected to the large screen, the user is given a personal screen area by the system. Thus the user has the capability to interact with the large screen through the personal screen area. The user can also operate the data freely with the drag & drop interface on the personal and the public screen area.

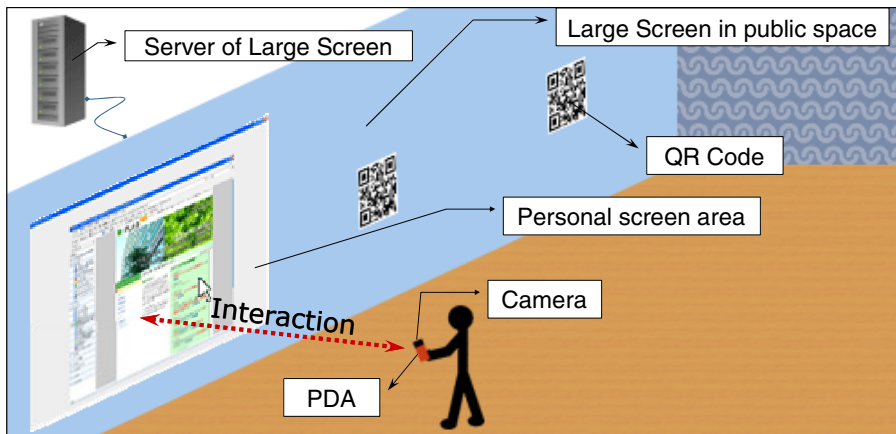


Fig. 1. System overview

### 2.1 Connection

We assume a large screen is situated in a public space, and users near the large screen can access the internet through the wireless network environment. Many users who are physically near the large screen will be in same network through the wireless network and may use the large screen at the same time. In order for a user to reserve a physical part of this screen and open a personal area adapted to his needs, all he has to do is to read a QR-code positioned at a particular screen space. A QR-Code is a kind of two-dimensional bar code that is easy to read by analyzing the image. Each QR-Code contains three pieces of information allowing users to connect to reserve the personal screen area. The first piece of information is the server's IP address, the second piece is the coordinates on the screen of the part that the user wants to use; the last one is the password required for the connection. This protocol follows the ShownPass technique [6], since users who are not physically near the large screen cannot obtain the password for the connection.

A user could take a picture of the QR-Code with the camera embedded in his PDA and analyze the picture to get the connection information, then send the position information and the password back to the server with the IP address of his PDA in the same format to the server in charge of distributing a space on the large screen. A personal screen area will be ready for the user right in front of him, and user can start using his PDA as an input device to operate on the large screen.

## 2.2 Personal Screen Area and Public Screen Area

When the large screen server is displaying the personal screen area to the user, there are two kinds of different areas on the large screen. One is the personal screen area, an area which is given to a user as his personal area that is used by only one user to display personal information of the user. The other one is public screen area, an area outside of personal screen areas used by all of the users to display the common information. As shown in Fig.2 (upper) there are several QR-Codes displayed on the public screen area. In this situation this area is called the public screen area. However, as shown in Fig.2 (lower), since another user opened up a personal screen area by using a QR-Code in the public screen area, the area will become a personal space, and the information which displayed on that area will slip out of this area. So the public screen area is not an absolute public screen area, yet some parts of the public screen area may switch to a personal screen area.

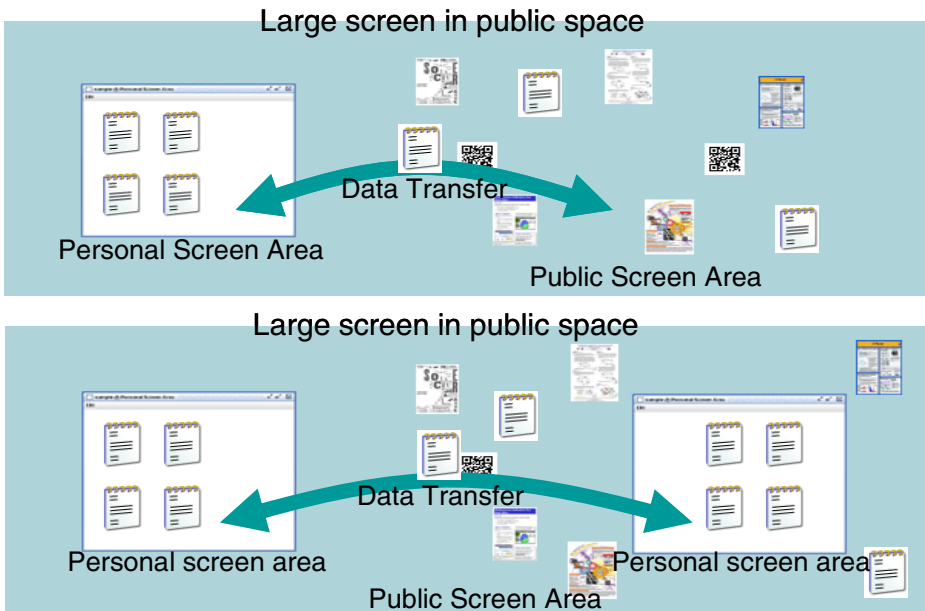


Fig. 2. Personal Screen Area and Public Screen Area

The personal screen area can be used in two ways. First, it can be used as a PDA's extended display. The user can start an applications installed in his PDA. In this case, it will not cause any security problem in general, because all the information is displayed in the personal screen area, and it does not transmit any private information to the large screen server.

Secondly, it can be used as a personal area for using the application on the large screen server. A user can exchange data between the large screen and his PDA with a drag & drop interface. For instance, the user can drag a file on the large screen from the public screen area to his personal screen area to save it on his PDA. Conversely, the user can drag data from his personal screen area to the public screen area. The user can also move his personal screen area to another available position on the screen. We should note that it may cause the security problem in this case.

When the user drags a file from his personal screen area to the public screen area, the system will copy this file from his PDA to the server of the large screen. Everyone who uses this large display has permission to copy this file into his PDA and modify this file. But nobody has the right to delete the file on the public screen area of the large screen directly except the user who published the file. When the user drags the file from the public screen area to his personal screen area again, the system will copy the file to his PDA, and nobody else has permission to copy or modify this file. In fact, a user does not have the permission to operate the data in other user's personal screen area.

This system could be used in many public spaces such as subway stations, shopping malls, book stores, etc. Users can download information such as part-time job information, message from the other users, posters to his PDA or cell-phone. Users can also upload such information from his PDA to the large screen.

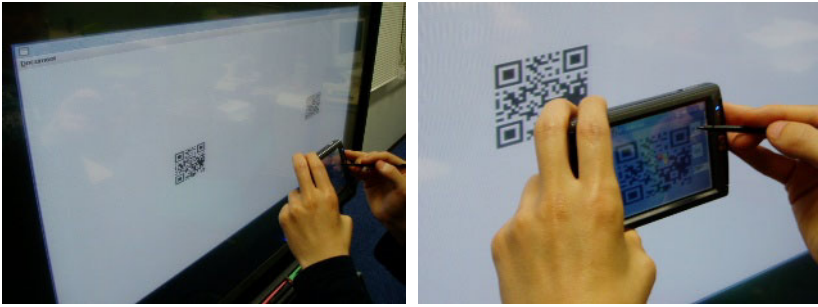
### 3 Prototype System

We have implemented a prototype system that can display the data of the PDA and can browse web pages on the large screen. As the personal screen area, we implemented it as the second way described above that use it as a personal area for using application on the large screen. Our prototype consists of two parts. One is the PDA side, and the other one is the large screen side. These two parts communicate through the wireless network environment.

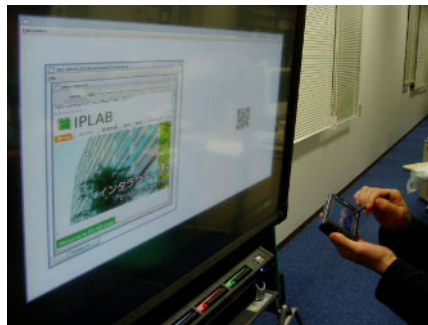
Normally, users handle the information of the PDA within a normal PDA interface. When they find a large screen (Fig.3 left), users can start a program installed in the PDA to capture the QR-Code (Fig.3 right) and analyze the QR-Code image to get the connection information. Then the program will try to connect with the large screen server by using the connection information.

When the user is connected to the large screen server, the server will display a personal screen area to the user. Currently, we implemented a web browser to access the internet web-site (Fig.4), a file browser is available in personal screen area to display files in the user's PDA (Fig.5).

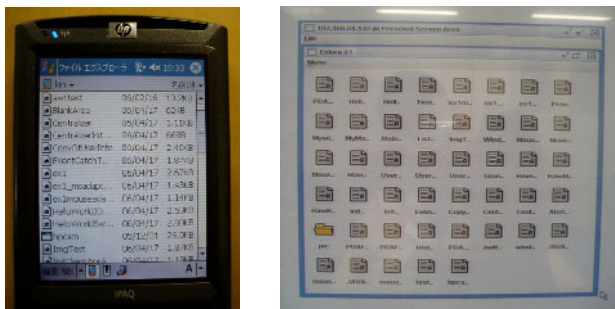
Users can operate on the large screen by using their PDA as input devices. For instance, users can move their personal cursor on the large screen by touching their PDA's screen. The system can generate particular events on the large screen based on the PDA's events. Users therefore can start some applications to manage information on the large screen.



**Fig. 3.** Display a QR-Code in the usable section to provide the connection information to users (left) and the user capture the QR-Code and analyzing the image to get the connection information (right)



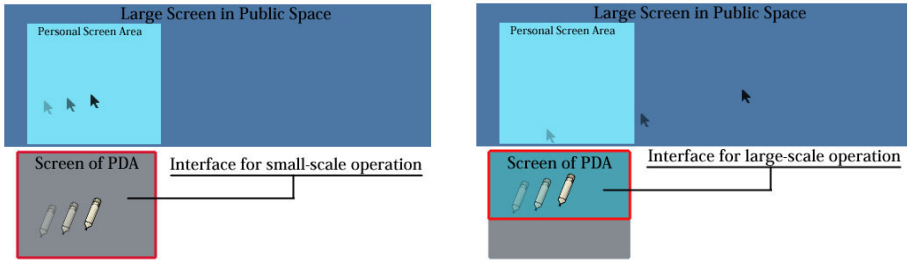
**Fig. 4.** The large screen will display the personal screen area to the user. The user can open up some applications installed on the large screen.



**Fig. 5.** Files in the PDA (left) and the large screen (right)

When a user operates on a large screen there are two different scales of operation. The first one is related to small-scale operations within a user's personal screen area. The second one is operations at large distances outside a user's personal screen area. To handle small-scale operations, we developed a pointing interface where the user

moves cursor relatively, a user thus can operate on the PDA's surface using a pen the same way as touch-pad on the notebook computers (Fig.6 left). To handle large-scale operations, we developed a pointing interface where the user can point directly on the miniature of the large screen on the PDA, the user therefore can operate on the PDA's surface the same way as touch-panel on the tablet PC (Fig.6. right). Fig.7 shows the two different interfaces for the small-scale operation and the large-scale operation.



**Fig. 6.** Interface for small-scale operation (left) & Interface for large-scale operation (right). When the user moves the pen on the interface for small-scale operation the cursor will moves like touch-pad on the notebook. When the user moves the pen on the interface of large-scale operation, the cursor will moves like touch-panel on the tablet PC.



**Fig. 7.** Interface for small-scale operation (left) & Interface for large-scale operation (right)

## 4 Related Work

There are a number of studies on the interaction between a large screen and a small size device. For example the WebWall [5] is a system which enables multi-user communication through a large screen. WebWall allows user to access the web-based applications such as simple sticky notes within small size devices such as mobile phones or PDAs. The ContentCascade [4] is a system which enables the user to download some contents from the public exhibition or displays situated in the public space such as shopping malls, book stores, etc, to the personal devices such as mobile phones. The C-Blink [2] system allows users to control the cursor on a large screen by using a cell-phone with a colored screen. This system is based on a program that



rapidly changes the hue of the phone screen and the user's waving the phone screen in front of the camera mounted above the large screen. The camera tracks the relative position of this signal to control the cursor on the large screen. The "Sweep and Point & Shoot" system [1] allows users to control the cursor on a large screen directly by using a cell-phone. This system keeps taking pictures with the camera embedded on the cell-phone and compares the pictures to determine the relative motion, thus allowing a user to control the cursor on the large screen. The paper [3] describes the interaction with situated displays using mobile phone. They situated displays out side of the office, and visitors are able to use the system in order to download information from the office door display through Bluetooth such as the owner's contact details. The paper [7] allows users to get information from the real world by using a PDA and RFID chips. The paper [8] is a research about how people move from individual to group work through the use of both PDAs and a shared public display. These systems allow users to handle directly the information. However, they do not use any concept of personal screen area. The novelty of our system is the concept of a personal screen area within a multi-users environment.

## 5 Summary

We have proposed a system to support public space communication by using a large screen in public space and a PDA. In this system, we use QR-Code to provide the connection information to the user. The user runs a program installed in the PDA to capture the QR-Code with the camera embedded in the PDA and to analyze the image to get the connection information. When the user connected to the large screen server, the server displays a personal screen area to the user that the user can manage the information through the personal screen area. We have also described our prototype system. Currently, we have implemented a system that we can control the cursor so that the user can browse the data in his PDA and browse web pages on the large screen. We will focus on the interface and keep on developing the prototype system.

## References

1. Rafael Ballagas, Michael Rohs, Jennifer G.Sheridan. Sweep and point and shoot: phonecam-based interaction for large public displays, ACM CHI'05, 2005, pp.1200-1203.
2. Kento Miyaoku, Suguru Higashino, Yoshinobu Tonomura. C-Blink: A hue-Difference-Based Light Signal Marker for large Screen Interaction via Any Mobile Terminal. ACM UIST'04, 2004,pp.147-156.
3. Keith Cheverst, Alan Dix, Daniel Fitton, Chris Kray, Mark Rouncefield, George Saslis-Lagoudakis, Jennifer G. Sheridan. Exploring Mobile Phone Interaction with Situated Displays. PERMID workshop at Pervasive 2005, Munich, PERVASIVE 2005,pp.43-37.
4. Himanshu Raj, R. Gossweiler, D. Milojevic, "ContentCascade Incremental Content Exchange between Public Displays and Personal Devices", in Proc. of the first Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous'04), Boston, Massachusetts, USA, pp.374-381, 2004.
5. Ferscha, A. and S. Vogl, Pervasive Web Access via Public Communication Walls: Pervasive Computing, Springer LNCS 2414, Zurich, Switzerland, pp.84-97, 2002.

6. Yuji Ayatsuka, Michimune Kohno, Jun Rekimoto, "Real-World Oriented Access Control Method with a Displayed Password" in Computer Human Interaction (APCHI 2004), LNCS 3101, Jun. 2004, pp.9-29
7. Pasi Valkkynen, Timo Tuomisto: Physical Browsing Research. PERMID workshop at Pervasive 2005, Munich. PERVASIVE 2005,pp.35-38.
8. Greenberg, S., Boyle, M., and Laberg, J., "PDAs and Shared Public Displays: Making Personal Information Public, and Public Information Personal." Personal Technologies, 1999. March. pp.54-64.
9. QR-Code: <http://qrcode.com>

# Development of a Medical Information Gathering System Using JMA Standard Health Insurance Claim System ORCA on IPv6 Network

Takashi Yoshino<sup>1</sup>, Yoko Yukimura<sup>1</sup>, Kunikazu Fujii<sup>1</sup>,  
Yoshiki Kusumoto<sup>2</sup>, and Masayuki Irie<sup>3</sup>

<sup>1</sup> Faculty of Systems Engineering, Wakayama University, 930 Sakaedani,  
Wakayama, Japan

yoshino@sys.wakayama-u.ac.jp

<http://www.wakayama-u.ac.jp/~yoshino/>

<sup>2</sup> CYBER LINKS Co., Ltd., Japan

<sup>3</sup> Wakayama Medical University, Japan

**Abstract.** Recently, regional medical information sharing becomes hot topic. It is difficult for a doctor to obtain the medicine name that a certain patient got in other medical institutions. This is one of the causes of the medical malpractice. An electronic medical record is one of the useful solutions. However, the initial and operational costs are huge. The diffusion rate in Japan is only about 5%. Then, we propose the alternate method to share regional medical information. In this proposal method, we connect a health insurance claim system ORCA. This system has a lot of basic medical information, such as name of medicines, disease names. In addition, we developed a medical information gathering system on the IPv6 network.

## 1 Introduction

Recently, an electronic medical record system and the telemedicine system are in progress in medical field. However, only one-hospital computerization is insufficient. Regional medical information sharing becomes hot topic. It is difficult for a doctor to obtain the medicine name that a certain patient got in other medical institutions. This is one of the causes of the medical malpractice. An electronic medical record is one of the useful solutions. However, the initial and operational costs are huge. The diffusion rate in Japan is only about 5%. Then, we propose the alternate method to share regional medical information. In this proposal method, we connect a health insurance claim system ORCA. This system has a lot of basic medical information, such as name of medicines, disease names. In addition, we developed a medical information gathering system on the IPv6 network.

## 2 Related Work

1. Medical information gathering systems There are some medical information sharing systems. However, most systems are running only in the same hospitals or in the same medical institution group. The system targets the hospitals that use a different system.
2. IPv6 applications  
IPv6 hardware and software products are developed in the world. However, there are almost fundamental software [1]. Recently, we think that these fundamental software has been almost completed. Then, application software on the fundamental software is necessary.

## 3 Medical Information Gathering System

### 3.1 Main Issues and Solutions

It is necessary to exchange data to use a medical information sharing system. All institutions should use the same software usually. However, this is not realistic, especially for wide-area, regional medical information sharing. Then, the paper adopts the following solutions.

1. XML format  
As for XML format, both the interchangeability and readability of data is high.
2. Standard protocol  
We use http for network protocol because http can pass over various networks.

### 3.2 Why Do We Use IPv6?

Why do we use IPv6? Because IPv6 is better than IPv4 IPv6 is sometimes also called the Next Generation Internet Protocol. Many countries tries to find the killer application for the next Internet generation [2].

### 3.3 What Is ORCA?

Our system connects a health insurance claim system ORCA [3]. ORCA has been developed by JMA (Japan Medical Association Research Institute). ORCA is open-source based health insurance claim software, and collects vast amount of medical data to making information source for evidence-based medicine. ORCA is built on Debian (one of distribution for Linux) and IPv6 stack. That is, the initial and operational costs are very low. The database of ORCA has a lot of basic medical information, such as name of medicines, disease names. For the above-mentioned features, we think that ORCA can become a sustainable integrating medical information infrastructure for medical institutions.

### 3.4 System Configuration

Figure 1 shows the system configuration. The central servers are connected with ORCA servers in each hospital. ORCA server is a system that operates ORCA and our system. The central servers consists of the application server and a database server. The central servers work about the data gathering. The application server uses Apache 2.2.0 and PHP 5.1.1. The database server uses PostgreSQL 8.1.1. In this system, the patient data is stored in the ORCA server of each medical institution. In other words, patient's information is not preserved on the server at all. In client PC, a web browser is only used. We use a Zaurus SL-C3100 (Sharp Co., Ltd) for a PDA. The OS of Zaurus is Linux. We succeeded to use IPv6 with the Zaurus.

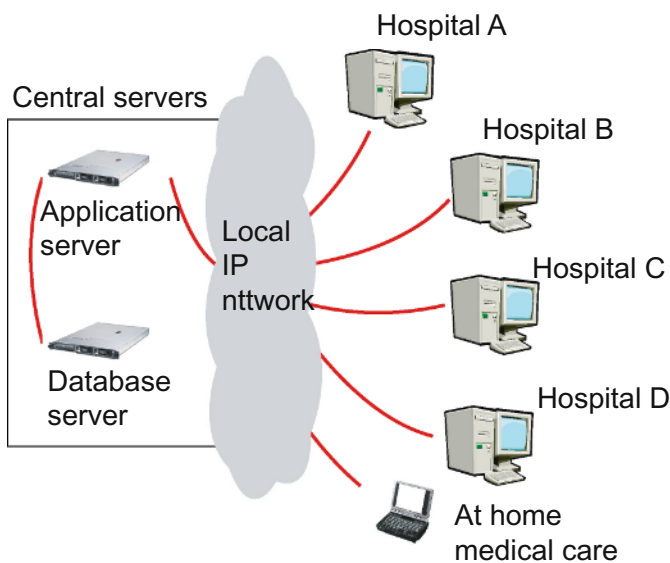


Fig. 1. System configuration of medical information gathering system

## 4 Functions of Medical Information Gathering System

A medical information gathering system has the following functions.

1. Patient information gathering function
2. Realtime medical information statistics function

We assume the doctor and the nurse as a user of this system. This system can communicate on both IPv4 and IPv6. IPv6 has a lot of advantages compared with IPv4. However, IPv6 is not so widespread. Our system is a case to show the utility of IPv6. Especially, our system treats an individual medical information. Therefore, we expect the high security of IPv6. Naturally, only by using IPv6, the

system does not become secure. IPv6 has the high potential of security because IPv6 supports IPsec as one of standard functions.

### 4.1 Patient Information Gathering Function

Figure 2 shows the example result of the medicine names that a certain patient got in other medical institutions. Figure 3 shows the screen of a PDA. The system can use two type of Network configuration, a client-server (CS) type and a peer to peer (P2P) type.

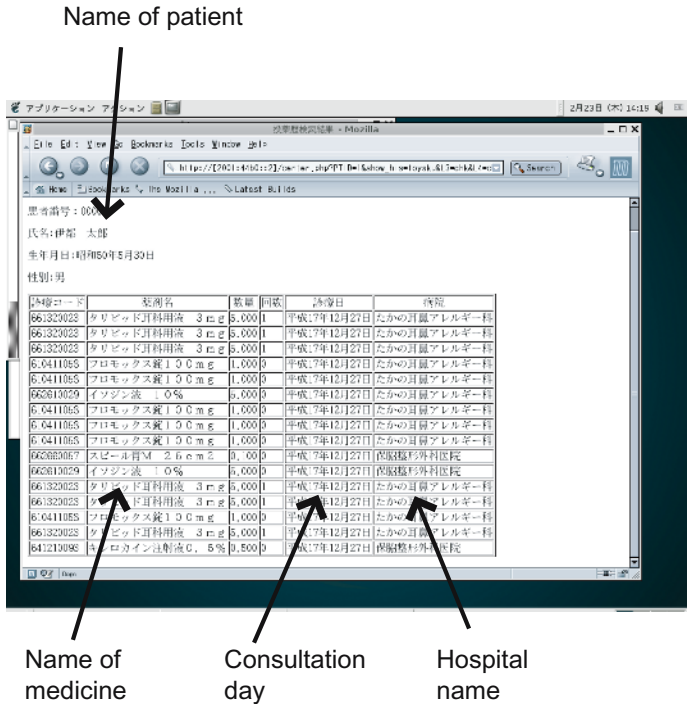


Fig. 2. Example result of the medicine name that a certain patient got in other medical institutions

Figure 4 shows the network configuration of peer-to-peer (P2P) type. In the P2P type, a certain ORCA server directly accesses the other ORCA servers. As for each client, it is necessary to know IP addresses of ORCA servers beforehand. Figure 5 shows the network configuration of client-server (CS) type. In the CS type, after a central server collects the each hospital patient data, data is passed to the client. This type assumes the use of a large amount of client. P2P type has the advantage of not needing central servers. Moreover, P2P type has the advantage that a patient data can be quickly gathered. CS type can connect of a large amount of ORCA server. Especially, CS type has a dramatic effect on to the collection of the statistical data.

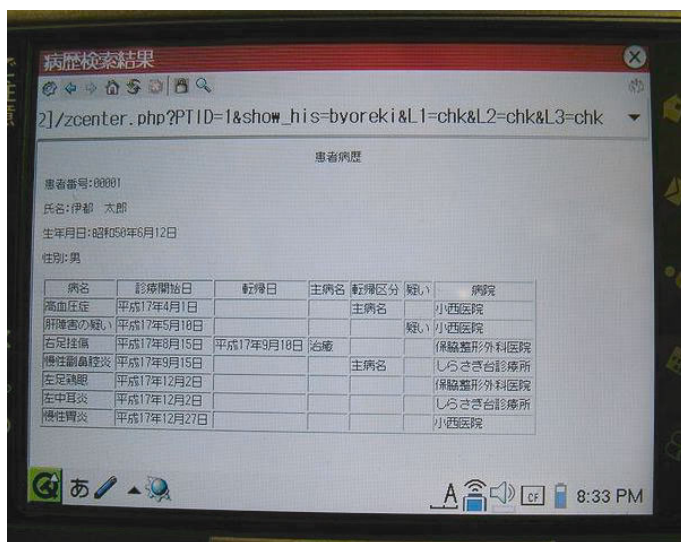


Fig. 3. Screen of a PDA

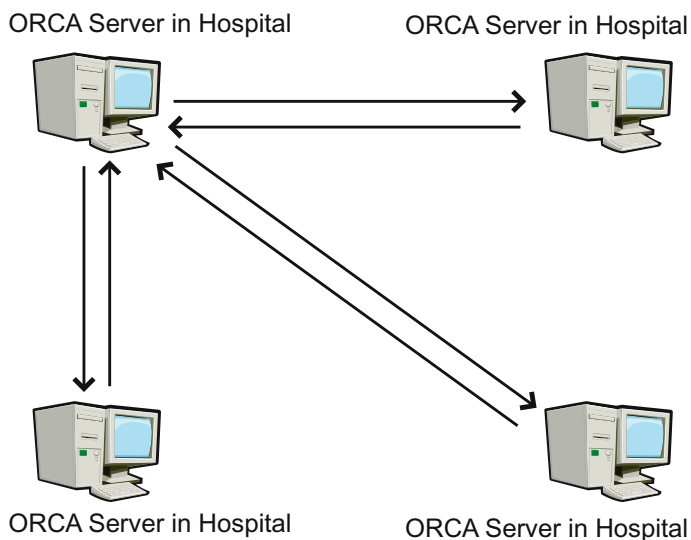


Fig. 4. Network configuration of a peer to peer (P2P) type

#### 4.2 Realtime Medical Information Statistics Function

This system can show the statistic information monthly, seasonal , by age , by area from ORCA server. Figure 6 shows the example result of statistic information. We can extract a lot of effective information by using the network in a short time.

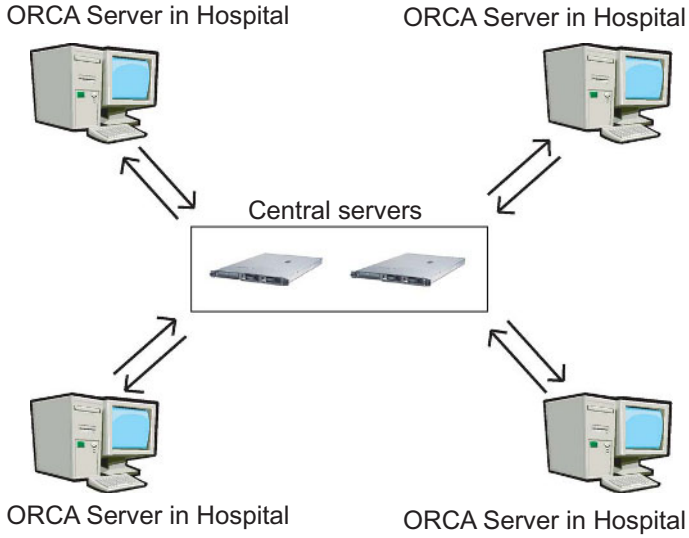


Fig. 5. Network configuration of a client-server (CS) type

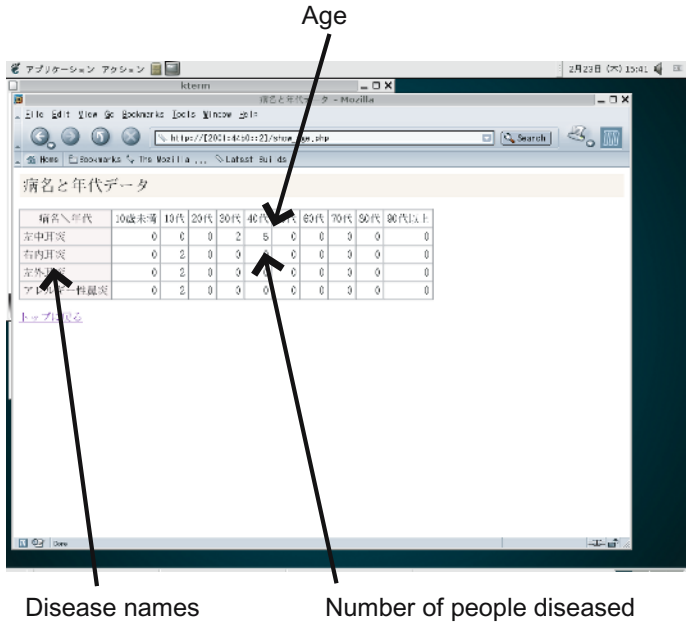


Fig. 6. Example result of statistic information

### 4.3 Technical Problems

We had some technical problems for development of the system. First, a PDA is imperfect for IPv6. We had a plan to use PocketPC2003. This is because the OS



of PocketPC2003 supports IPv6. However, a web browser and a driver of wireless LAN do not support IPv6. Then, we used a Zaurus SL-C3100. The commercial SL-C3100 does not support IPv6. We can use IPv6 because the OS of SL-C3100 is Linux.

Second, we have to be careful to the version of the software. There are a lot of IPv6-unsupported software in the default software. Table 1 shows IPv6-support version of each software.

**Table 1.** IPv6-support version

Category	Software	Version
OS	Windows 2000	Patch is necessary.
	Windows Xp	Supported
	WindowsCE	Supported from CE.Net 4.1iPocketPC2003)
	WindowsServer2003	Supported
	Linux	Supported from Kernel 2.1
	FreeBSD	Supported from Release-4
Database	PostgreSQL	Supported from version 7.4
Web server	Apache	Supported from version 2.0
Web browser	Firefox on Windows	Supported from version 1.6a
	InternetExplorer on Windows	Supported from version 6.0
	InternetExplorer on Windows CE	Supported from version 5.5
	lynx	Supported from version 2.8.4

## 5 Conclusion

We have developed a medical information gathering system using JMA standard health insurance claim system ORCA on IPv6 Network. Technical problems are almost solved now. However, the field of the medical is very sensitive to security. It is necessary to obtain the reliance to the medics in the future.

## References

1. Yukio Hiranaka, Masato Ohnuma, Akihisa Yoshida, Toshihiro Taketa, Tatsumi Hosokawa, Takashi Yamagata, Seiichi Okoma, Yuuji Hirose, Tsuyoshi Yoneda, Shigehiro Takeda, Teruaki Arashida, Hiroki Nakagawa, Takahiro Kudaira, Kouji Tanaka: Multimedia and Routing Specific Applications on IPv6 Networks, Proceedings of the 2004 International Symposium on Applications and the Internet Workshops (SAINTWf04), 136-139, 2004
2. Shu-Fen Tseng, Hsi-Chieh Lee, Te-Ching Kung, Shou-Lien Chou, Jing-Yi Chen: The Development of Global IPv6 Products: An Exploration, Proc. 19th IEEE International Conference on Advanced Information Networking and Applications (AINA 2005), pp. 845-850, March 2005.
3. Eizen Kimura, Tateishi Norihiko, Ken Ishihara: The Design of Virtual Private Network Topology for Migrating to IPv6 World, Proceedings of the 2004 International Symposium on Applications and the Internet (SAINTf04), pp. 313-316, 2004.

# A Prototype of a Chat System Using Message Driven and Interactive Actions Character

Junko Itou<sup>1</sup>, Kenji Hoshio<sup>2</sup>, and Jun Munemori<sup>1</sup>

<sup>1</sup> Faculty of Systems Engineering, Wakayama University,  
930, Sakaedani, Wakayama 640-8510, Japan

<sup>2</sup> cavia inc.

1-4-30-25, Mori Bldg. 18, Roppongi, Minatoku, Tokyo 106-0032, Japan  
{itou, munemori}@sys.wakayama-u.ac.jp

**Abstract.** In this article, we present a chat system in which embodied characters behave as agents of users and automatically act on messages of the users and the other character's action. We display and exchange nonverbal expressions including gestures, eye-gazes, noddings, and facial expressions in daily conversation. Nonverbal expressions convey various kinds of information that is essential to make our face-to-face communication successful. In the previous work on social psychology, it is known that there are interdependences among nonverbal expressions between those from different persons in conversation with each other. We apply this knowledge to the chat between embodied characters, so that 3D characters interactively act by user's messages. The system evaluation results demonstrated higher validity than the system that the user explicitly indicates the character's actions.

## 1 Introduction

Online communication is widely spreading and tools of online communication become diverse, for example e-mail, chat system, remote meeting system, distance learning, and so on. There are also varieties of proposed tools on chat system from conventional text-based one to graphical one that agents in place of users talks in virtual 3D space. By using embodied character chat system, the users obtain messages by watching embodied character's actions as well as by reading plain texts so that chat systems like this are expected to serve as a human interface easier for the users to acquire information than conventional text-based systems. A character in the chat systems plays a role as a agent of a user not only to express the user's emotional state which cannot be displayed by a chat message, but also to make the chat alive. As the result, it becomes clear what meaning the user implies for the chat messages.

In order to employ embodied characters for chatting users, we need to control their nonverbal expressions which include gestures, eye-gazes, noddings, and facial expressions, and so on. As far as characters have their own faces and bodies in order to be "embodied", their users read various meanings in the nonverbal expressions displayed by the faces and the bodies of the characters

even if the characters are not actually designed to send nonverbal expressions to their users but only to speak. Thus, for any embodied characters, we need to control their nonverbal expressions properly so that they convey appropriate meaning to the users. For example, it would be strange if a character does not smile at all although the user tells a funny story. As another example, when one character talks or smiles to the partner character, if the partner character freeze with no response to the action, it looks also strange.

Previous work on controlling nonverbal expressions of embodied characters mainly discusses the consistency of nonverbal expressions with the speech utterances or the goal of conversation for each agent [1][2]. However, when we consider dialogues between a pair of embodied characters, we need to consider interdependences between nonverbal expressions displayed by those two characters. As explained in more detail in the next section, it is reported in the field of social psychology that nonverbal expressions given by humans during their talks are not independent with each other[3]-[6].

In the remainder of this article, we will illustrate a design approach of our Message Driven and Interactive Action Character (MEDIAC) chat system which is applied this knowledge of those interdependences between nonverbal expressions from the embodied characters.

This paper is organized as follows: in section 2, we will describe the knowledge about the interdependences between nonverbal expressions in human communication reported in the previous work on social psychology. In section 3, we will propose a chat system to maintain the interdependences between nonverbal expressions of embodied characters in their chat. Validation test of our system will be given in section 4. Finally, we discuss some conclusions and future work in section 5.

## 2 Interdependence Between Conversation Partners

It has been investigated in social psychology what features are found in nonverbal expressions given by humans during their conversation. Through those investigations, it is known that nonverbal expressions given by humans in real conversation have some interdependences and synchronicity.

For example, it was reported that the test subjects maintained eye contact with their partners in lively animated conversation, whereas they avoided eye contact when they are not interested in talking with their partners [3][4]. In the experiments by Matarazzo, noddings by the listeners in conversation encouraged utterances of the speakers, and as the result, animated conversation between the speakers and the listeners is realized [5]. In another experiments by Dimberg, facial expressions of the test subjects were affected by those of their partners[6]. The subjects smiled when their partners gave smiles to them, whereas they gave expressions of tension when their partners had angry faces.

These results implies positive correlations or synchronicity between the nonverbal expressions given by the conversation partners for eye gazes, noddings and facial expressions.

### 3 MEDIAC Messenger Chat System

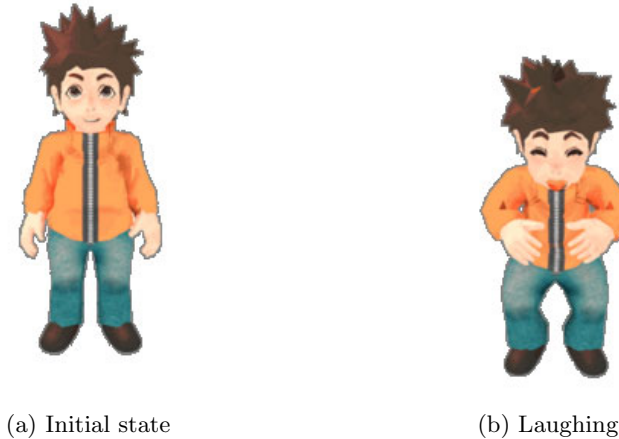
#### 3.1 Goal

In the most current graphical chat systems, the user should choose and specify the character's action explicitly. By employing the knowledge described in the previous section, we aim to realize automatic actions and reactions of embodied characters in order to make users' chat alive. It is preferable that character's actions and reactions can be produced by the user's message because the user attends to only his/her chat and it is very troublesome for the user to set manually all actions and reactions at each message during his/her chat. We apply the interdependences between nonverbal expressions described in section 2 to the relationship between actions and reactions.

In the remainder of this section, we will propose a system to realize this adjustment process especially for eye gazes, gestures, and facial expressions of the characters for the first step towards our goal.

#### 3.2 Actions of Character

Let us consider a pair of embodied characters  $A$  and  $B$  to explain our chat system. We denote user  $A$ 's agent by character  $A$  and user  $B$ 's agent by character  $B$ . The overview of the character is shown in Fig. 1.



**Fig. 1.** 3D Character in MEDIAC Messenger

This character has 34 types of actions, which are related to about 250 words. The example of the mapping is given in Table.1.

As discussed in section 2, the actions of character  $A$  are not only determined by the messages of user  $A$  but also the actions of character  $B$ . For example, when character  $B$  laughs by the message of user  $B$ , character  $A$  should smile even if user  $A$ 's message does not contain the word of happy or smile.

**Table 1.** Character actions and keywords

Action	keyword *
bow	Hi, Hello, Good morning, Thank you, Nice to meet you
laugh	laugh, laughter, boff
smile	smile, happy, :-)
cry	sob, cry, weep, X-(
panic	dismay, uh-oh, nix, oops, darn
surprise	surprise, great, aghast, consternation
think	think, hm, hmm, um

\* : In MEDIAC Messenger, all keywords are written in Japanese.

Our system has 21 types of reaction. By Dimberg's experiments in section 2, the action of "smile" which is displayed on one character, responds to the other character's reaction "smile". On the other hand, when the action of "angry" is displayed, the reaction of "surprise" is expressed. When a message includes the keywords corresponding to "think", the partner shows the reaction "nodding" to stimulate their dialogue according to the research by Matarazzo. Furthermore, Watanabe[7] pointed we do not only exchange words, but also we share gestures and physical rhythm such as breath, so that we can feel an identification with the conversation partner. From this indication, kinds of synchronicity should be reflected to the relationships between an action and a reaction. Table.2 shows a part of the reaction rules which are defined as noted above.

**Table 2.** Character reaction rules

Action	Reaction	Action	Reaction	Action	Reaction
bow	bow	point at you	point at me	angry	surprise
laugh	laugh	point at me	point at you	achcha	nod
smile	smile	quake	quake	think	nod
handwave	handwave	no no	no no	cheer	smile
bummage	bummage	frustrate	no no	deny	strand

### 3.3 System Structure

Our chat system MEDIAC Messenger is constructed of a sever and multiple clients. Users start this client system then some windows come up as shown in Fig. 2

Main Window shows information about login users. In the Chat Window, users chat and User's Agent and Others Agent which is a character of chat partner, variously act according to the users' input chat text. These characters directly render on a user's desktop and the user can move the characters to the various points on the screen.

After user A inputs a message in the field of Chat Window and sends to server, the sever extracts keyword from the sent message by morphologic analysis so that

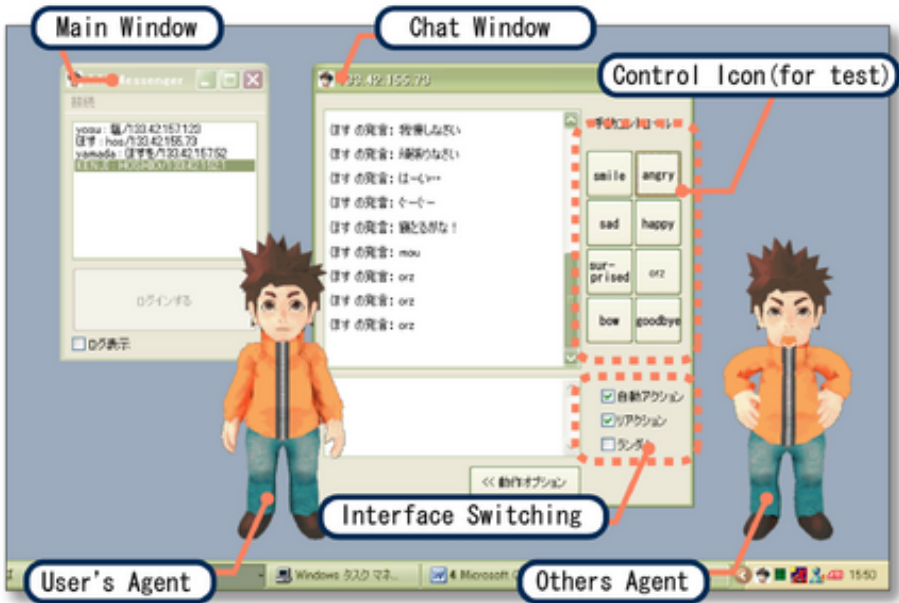


Fig. 2. Overview of MEDIAC Messenger

the server determines which action character  $A$  should behave. The action and user  $A$ 's message are sent from the server to each client.

The client of user  $B$  which receives the action and the message, determines the reaction for the character  $A$ 's action according to the reaction table, and send to the server. Again the server sends the reaction to each clients. In this way, each character's action and reaction are performed.

Each action and reaction takes one or two seconds. A user can input during an action or a reaction but the action or the reaction is not displayed until the last action ends.

The chat messages are not displayed around the characters as a dialogue balloon, but are stored in the Chat Window. We employ this style through the result of the exploratory experiment in which some comments were obtained: (1) It hard to read the balloon message when the long message were sent. (2) The user tended to write a message reading the chat log. (3) Users felt the balloon message was more impactful than the actions of the character, that is, users watched the message in spite of the acting character.

## 4 Experimental Results

We implemented the MEDIAC Messenger described in the previous section. To validate our system, we conducted a experiment using systems: one of the system that a user inputs actions of him/her character by icon, and another one is our system.

Experimental subjects were 8 college students. They were divided into 4 groups, and we instructed each pair to chat for 30 minutes using the systems. The 30 minutes are composed of 3 sections: the first 10 minutes is section A in chatting with action control icon, second 10 minutes is section B in chatting with message driven action character, and the final 10 minutes is section C in chatting with message driven and interactive action character. In each section, the subjects were imposed no restrictions on the chatting topic.

The result of a questionnaire is described in Table.3, and the number of messages and performed actions is described in Table.4.

**Table 3.** Questionnaire specifics

questionnaire item	Sec. A	Sec. B	Sec. C
felt stressed (1:yes - 5:no)	2.6	3.5	3.3
felt relaxed (1:no - 5:yes)	2.6	4.6	3.9
could chat casually(1:no - 5:yes)	3.0	4.6	4.5
chat became more joyful (1:no - 5:yes)	3.5	4.1	4.1
felt joyful for existence of character (1:no - 5:yes)	4.0	4.3	3.9
want this character for chat (1:no - 5:yes)	3.0	3.9	2.9

**Table 4.** Number of messages and performed actions

(a) Number of messages                      (b) Number of performed actions

	Sec. A	Sec. B	Sec. C
pair A	65	70	58
pair B	27	63	75
pair C	52	54	46
pair D	90	130	122
average	58.5	79.3	75.3

	Sec. A	Sec. B	Sec. C
pair A	42	122	98
pair B	10	67	142
pair C	26	79	149
pair D	127	172	244
average	51.3	110.0	158.3

We totally obtained fine rating for our system. The number of messages increased in Section B and Section C. It reveals the users could communicate smoothly and actively by using our system. The number of performed actions went up in proportion to the number of messages. You can tell the participants felt the relaxed and more joyful by the Table. 3 although the character showed their actions after another (one action occurred every 5.5 seconds in section B and every 3.8 seconds in section C).

However, there is not significant difference in section B from section C. According to the comment of the questionnaire survey on the items, some users felt that the character is too disturbing to chat. Our proposed system dose not yet treat the timing or the duration of actions and reactions displayed by characters. The evaluation will come up and show the significant difference by the modification of the point.

## 5 Conclusion

In this article, we proposed a chat system to display automatically action and reaction in response to the user's message aiming to make a chat alive. Considering the knowledge obtained by the previous work on social psychology, we defined the reaction as a response of actions, and realized adjustment of nonverbal expressions.

In the experiments, we confirm the validity of our method based on the comparison and the evaluation by many subjects. On the other hand, there are some comments from test subjects that the character make the users feel gloomy because the character moves too busy. We should plan to adopt the timing between a action and a reaction.

This system has more two imperfections. We refer to the investigations in the field of social psychology, which focuses on the dialogue: face-to-face communication. Therefore, this system only deal with the person-to-person chat. This problem will be solved by treating the many-to-many communications as a set of one-to-one communication.

Finally, we also should plan to improve the reaction table is proper and to add different types of reaction to the reaction table reflected the previous investigations.

## References

1. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H. and Yan, H.: "Embodiment in Conversational Interfaces: Rea" CHI-99, pp.520-527, 1999
2. De Carolis, B., Pelachaud, C., Poggi, I., De Rosis, F.: "Behavior Planning for a Reflexive Agent" in: Proc. of International Joint Conference on Artificial Intelligence (IJCAI 2001), pp. 1059-1066, 2001.
3. Beattie, G. W. : "Sequential patterns of speech and gaze in dialogue" *Semiotica*, Vol.23, pp. 29-52, 1978.
4. Kendon, A.: "Some functions of gaze direction in social interaction" *Acta Psychologica*, Vol.26, pp. 22-63, 1967.
5. Matarazzo J.D., Saslow, G., Wiens, A.N., Weitman, M., and Allen, B.V.: "Interviewer Head Nodding and Interviewee Speech Durations" *Psychotherapy:Theory, Research and Practice*, Vol.1, pp.54-63, 1964.
6. Dimberg U.: "Facial Reactions to Facial Expressions" *Psychophysiology*, No.6, Vol.19, pp.643-647, 1982.
7. Watanabe T.: "E-COSMIC: Embodied Communication System for Mind Connection" in: Proc. of the 9th International Conference on Human-Computer Interaction (HCI International 2001), Vol.1, pp.253-257, 2001.



# Genetically Optimized Fuzzy Set-Based Polynomial Neural Networks Based on Information Granules with Aids of Symbolic Genetic Algorithms

Tae-Chon Ahn, Kyung-Won Jang, and Seok-Beom Roh

Department of Electrical Electronic and Information Engineering, Wonkwang University,  
344-2, Shinyong-Dong, Iksan, Chon-Buk, 570-749, South Korea  
{tcahn, nado}@wonkwang.ac.kr

**Abstract.** In this paper, we propose a new architecture of Fuzzy Set-based Polynomial Neural Networks (FSPNN) with a new fuzzy set-based polynomial neuron (FSPN) whose fuzzy rules include the information granules (about the real system) obtained through Information Granulation. Although the conventional FPNN with Fuzzy Relation-based Polynomial Neurons has good approximation ability and generalization capability, there is an important drawback that FPNN is very complicated. If we adopt fuzzy set-based fuzzy rules as substitute for fuzzy relation-based fuzzy rules, we can get an advantage of the rule reduction. We use FSPN as a node of Fuzzy Polynomial Neural Networks to reduce the complexity of the FPNN. The proposed FPNN with Fuzzy Set-based Polynomial Neuron can achieve compactness. Information Granulation can extract good information from numerical data without expert's knowledge which is important for building Fuzzy Inference System. We put Information Granulation to the proposed FSPN. The structure of the proposed FPNN with FSPN is determined with aids of symbolic gene type genetic algorithms.

## 1 Introduction

To build models with substantial approximation capabilities, there should be a need for advanced tools. As one of the representative advanced design approaches, comes a family of self-organizing networks, (called "FPNN" as a new category of neuro-fuzzy networks) [1], [4]. The conventional FPNN with Fuzzy Relation-based Neuron shows good approximation ability. FPNN with FRPN is implemented based on GMDH structure. Generally, the structure of fuzzy model based on GMDH is very complicate.

To reduce the complexity of FRPN, we propose Fuzzy Set-based Polynomial Neuron(FSPN) based on Yamakawa's fuzzy model. Moreover, we put the FSPN to a basic node of FPNN. In result, the proposed FPNN with FSPN can be more compact than FPNN with FRPN.

In this paper, in considering the above problems coming with the conventional FPNN [1], [4], we introduce a new structure of fuzzy rules named Fuzzy Set based Polynomial Neuron as well as a new symbolic gene type genetic design approach. In other hand, from a point of view of a new understanding of fuzzy rules, information

granules seem to melt into the fuzzy rules respectively. The determination of the optimal values of the parameters available within an individual FSPN leads to a structurally and parametrically optimized network through the symbolic gene type genetic approach.

## 2 The Architecture and Development of Fuzzy Set-Based Polynomial Neural Networks (FSPNN)

The FSPN encapsulates a family of nonlinear “if-then” rules. When put together, FSPNs results in a self-organizing Fuzzy Set-based Polynomial Neural Networks (FSPNN). The FSPN consists of two basic functional modules. The first one, labeled by **F**, is a collection of fuzzy sets (here denoted by  $\{A_k\}$  and  $\{B_k\}$ ) that form an interface between the input numeric variables and the processing part realized by the neuron. The second module (denoted here by **P**) refers to the function – based nonlinear (polynomial) processing that involves some input variables This nonlinear processing involves some input variables ( $x_i$  and  $x_j$ ), which are capable of being the input variables (Here,  $x_p$  and  $x_q$ ), or entire system input variables. Each rule reads in the form.

$$\begin{aligned} \text{if } x_p \text{ is } A_k \text{ then } z \text{ is } P_{pk}(x_i, x_j, \mathbf{a}_{pk}) \\ \text{if } x_q \text{ is } B_k \text{ then } z \text{ is } P_{qk}(x_i, x_j, \mathbf{a}_{qk}) \end{aligned} \tag{1}$$

Where,  $\mathbf{a}_{qk}$  is a vector of the parameters of the conclusion part of the rule while  $P(x_i, x_j, \mathbf{a})$  denoted the regression polynomial forming the consequence part of the fuzzy rule. The activation levels of the rules contribute to the output of the FSPN being computed as a weighted average of the individual condition parts (functional transformations)  $P_K$  (note that the index of the rule, namely “ $K$ ” is a shorthand notation for the two indexes of fuzzy sets used in the rule (1), that is  $K = (l, k)$ ).

$$\begin{aligned} z &= \sum_{l=1}^{\text{total inputs}} \left( \frac{\sum_{k=1}^{\text{total\_rules related to input } l} \mu_{(l,k)} P_{(l,k)}(x_i, x_j, \mathbf{a}_{(l,k)})}{\sum_{k=1}^{\text{total\_rules related to input } l} \mu_{(l,k)}} \right) \\ &= \sum_{l=1}^{\text{total inputs}} \left( \sum_{k=1}^{\text{rules related to input } l} \tilde{\mu}_{(l,k)} P_{(l,k)}(x_i, x_j, \mathbf{a}_{(l,k)}) \right) \end{aligned} \tag{2}$$

In the above expression, we use an abbreviated notation to describe an activation level of the “ $K$ ” th rule to be in the form.

$$\tilde{\mu}_{(l,k)} = \frac{\mu_{(l,k)}}{\sum_{k=1}^{\text{total\_rules related to input } l} \mu_{(l,k)}} \tag{3}$$

When developing an FSPN, we use genetic algorithms to produce the optimized network. This is realized by selecting such parameters as the number of input variables, the order of polynomial, and choosing a specific subset of input variables. Based on the genetically optimized number of the nodes (input variables) and the polynomial order, we construct the optimized self-organizing network architectures of the FSPNNs.

### 3 Information Granulation Through Hard C-Means Clustering Algorithm

Information granules are defined informally as linked collections of objects (data points, in particular) drawn together by the criteria of indistinguishability, similarity or functionality [11]. Granulation of information is a procedure to extract meaningful concepts from numeric data and an inherent activity of human being carried out with intend of better understanding of the problem.

#### 3.1 Definition of the Premise and Consequent Part of Fuzzy Rules Using Information Granulation

We assume that given a set of data  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  related to a certain application, there are some clusters which are capable of being found through HCM. The center point and the membership elements represent each cluster. The center point means the apex of the membership function of the fuzzy set. Let us consider building the consequent part of fuzzy rule. The fuzzy rules of Information Granulation-based FSPN are as followings.

$$\begin{aligned} \text{if } x_p \text{ is } \mathbf{A}_k^* \text{ then } z\text{-}m_{pk} &= P_{pk}((x_i - v_{pk}^i), (x_j - v_{pk}^j), a_{pk}) \\ \text{if } x_q \text{ is } \mathbf{B}_k^* \text{ then } z\text{-}m_{qk} &= P_{qk}((x_i - v_{qk}^i), (x_j - v_{qk}^j), a_{qk}) \end{aligned} \quad (4)$$

Where,  $\mathbf{A}_k^*$  and  $\mathbf{B}_k^*$  mean the fuzzy set, the apex of which is defined as the center point of information granule (cluster) and  $m_{pk}$  is the center point related to the output variable on cluster  $_{pk}$ ,  $v_{pk}^i$  is the center point related to the  $i$ -th input variable on cluster  $_{pk}$  and  $\mathbf{a}_{qk}$  is a vector of the parameters of the conclusion part of the rule while  $P((x_i - v^i), (x_j - v^j), a)$  denoted the regression polynomial forming the consequence part of the fuzzy rule which uses several types of high-order polynomials (linear, quadratic, and modified quadratic) besides the constant function forming the simplest version of the consequence. If we are given  $m$  inputs and one output system and the consequent part of fuzzy rules is linear, the overall procedure of modification of the generic fuzzy rules is as followings.

**Step 1)** Build the universe set.

**Step 2)** Build  $m$  reference data pairs composed of  $[\mathbf{x}_1; Y]$ ,  $[\mathbf{x}_2; Y]$ , and  $[\mathbf{x}_m; Y]$ .

**Step 3)** Classify the universe set  $U$  into  $l$  clusters such as  $c_{i1}, c_{i2}, \dots, c_{il}$  (subsets) by using HCM according to the reference data pair  $[\mathbf{x}_i; Y]$ .

**Step 4)** Construct the premise part of the fuzzy rules related to the  $i$ -th input variable ( $x_i$ ) using the directly obtained center points from HCM.

**Step 5)** Construct the consequent part of the fuzzy rules related to the  $i$ -th input variable ( $x_i$ ).

**Sub-step 1)** Make a matrix as equation (5) according to the clustered subsets.

$$A_j^i = \left[ \begin{array}{cccc|c} x_{21} & x_{22} & \cdots & x_{2m} & y_2 \\ x_{s1} & x_{s2} & \cdots & x_{sm} & y_s \\ x_{k1} & x_{k2} & \cdots & x_{km} & y_k \\ \vdots & \vdots & \cdots & \vdots & \vdots \end{array} \right] \quad (5)$$

Where,  $\{x_{k1}, x_{k2}, \dots, x_{km}, y_k\} \in c_{ij}$  and  $A_j^i$  means the membership matrix of  $j$ -th subset related to the  $i$ -th input variable.

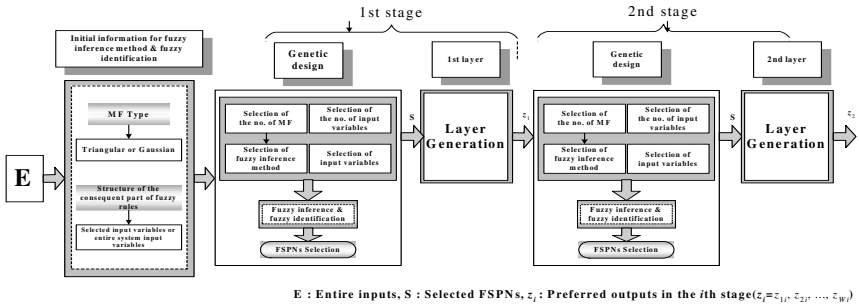
**Sub-step 2)** Take an arithmetic mean of each column on  $A_j^i$ . The mean of each column is the additional center point of subset  $c_{ij}$ .

$$center\ points = \left[ v_{ij}^1 \quad v_{ij}^2 \quad \dots \quad v_{ij}^m \quad \left| \quad m_{ij} \right. \right] \tag{6}$$

**Step 6)** If  $i$  is  $m$  then terminate, otherwise, set  $i=i+1$  and return step 3.

### 4 Genetic Optimization of FSPNN

GAs is aimed at the global exploration of a solution space. Symbolic GAs use serial method of symbolic type, roulette-wheel as the selection operator, one-point cross-over, and an invert operation in the mutation operator [2]. To retain the best individual and carry it over to the next generation, we use elitist strategy [3]. Symbolic GAs is different in the point of the type of chromosome. Symbolic GAs uses some symbol( $x_1, x_2$ , etc) as chromosome, while simple GAs uses the binary type chromosome. The overall genetically-driven structural optimization process of FSPNN is shown in Fig. 1.



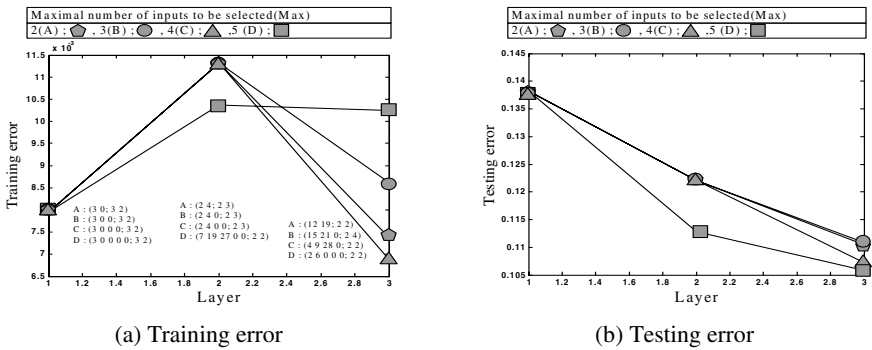
**Fig. 1.** Overall genetically-driven structural optimization process of FSPNN

The framework of the design procedure of the genetically optimized FSPNN comprises the following steps.

- [Step 1] Determine system’s input variables.
- [Step 2] Form training and testing data.
- [Step 3] Specify initial design parameters.
- [Step 4] Decide FSPN structure using genetic design.
- [Step 5] Carry out fuzzy-set based fuzzy inference and coefficient parameters estimation for fuzzy identification in the selected node (FSPN).
- [Step 7] Check the termination criterion.
- [Step 8] Determine new input variables for the next layer.

## 5 Experimental Studies

We illustrate the performance of the network and elaborate on its development by experimenting with data coming from the gas furnace process. The delayed terms of methane gas flow rate,  $u(t)$  and carbon dioxide density,  $y(t)$  are used as system input variables such as  $u(t-3)$ ,  $u(t-2)$ ,  $u(t-1)$ ,  $y(t-3)$ ,  $y(t-2)$ , and  $y(t-1)$ . Fig. 2 depicts the performance index of each layer of IG-gFSPNN with Type T\* (which means that all input variables are used for building the first layer of IG-gFSPNN) according to the increase of maximal number of inputs to be selected.



**Fig. 2.** Performance index of IG-gFSPNN (with Type T\*) with respect to the increase of number of layers

Table 1 summarizes a comparative analysis of the performance of the network with other models.

**Table 1.** Comparative analysis of the performance of the network; considered are models reported in the literature

Model				Performance index		
				PI	PI <sub>s</sub>	EPI <sub>s</sub>
Box and Jenkin's model[7]				0.710		
Tong's model[8]				0.469		
Sugeno and Yasukawa's model[9]				0.190		
Pedrycz's model[5]				0.320		
Oh and Pedrycz's model[6]				0.123	0.020	0.271
Kim et al.'s model[10]					0.034	0.244
Proposed IG-gFSPNN	Type III (SI=6)	Triangular	3 <sup>rd</sup> layer(Max=3)		0.008	0.110
		Gaussian- like	3 <sup>rd</sup> layer(Max=3)		0.008	0.099

PI - performance index over the entire data set,

PI<sub>s</sub> - performance index on the training data, EPI<sub>s</sub> - performance index on the testing data.

## 6 Concluding Remarks

In this study, we have surveyed the new structure and meaning of fuzzy rules and investigated the GA-based design procedure of Fuzzy Set-based Polynomial Neural Networks (IG-FSPNN) with information granules along with its architectural considerations. The whole system is divided into some sub-systems that are classified according to the characteristics named information granules. Each information granule seems to be a representative of the related sub-systems. A new fuzzy rule with information granule describes a sub-system as a stand-alone system. A fuzzy system with some new fuzzy rules depicts the whole system as a combination of some stand-alone sub-system. The GA-based design procedure applied at each stage (layer) of the FSPNN leads to the selection of the preferred nodes (or FSPNs) with optimal local characteristics (such as the number of input variables, the order of the consequent polynomial of fuzzy rules, and input variables) available within FSPNN. The comprehensive experimental studies involving well-known datasets quantify a superb performance of the network in comparison to the existing fuzzy and neuro-fuzzy models.

**Acknowledgement.** This work was supported by Research Institute of Engineering Technology Development, Wonkwang University.

## References

1. Oh, S.-K., Pedrycz, W.: Self-organizing Polynomial Neural Networks Based on PNs or FPNs : Analysis and Design. *Fuzzy Sets and Systems* **142**(2) (2004) 163-198.
2. Michalewicz, Z.: *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, Berlin Heidelberg New York (1996).
3. Jong, D.K.A.: Are Genetic Algorithms Function Optimizers? *Parallel Problem Solving from Nature 2*, Manner, R. and Manderick, B. (eds.) North-Holland, Amsterdam (1992).
4. Oh, S.K., Pedrycz, W.: Fuzzy Polynomial Neuron-Based Self-Organizing Neural Networks. *Int. J. of General Systems* **32** (2003) 237-250.
5. Pedrycz, W.: An identification algorithm in fuzzy relational system. *Fuzzy Sets and Systems* **13** (1984) 153-167
6. Oh, S. K., Pedrycz, W.: Identification of Fuzzy Systems by means of an Auto-Tuning Algorithm and Its Application to Nonlinear Systems. *Fuzzy sets and Systems* **115**(2) (2000) 205-230
7. Box, D. E., Jenkins, G. M.: *Time Series Analysis, Forecasting and Control*. Holden Day, California (1976)
8. Tong, R. M.: The evaluation of fuzzy models derived from experimental data. *Fuzzy Sets and Systems* **13** (1980) 1-12
9. Sugeno, M., Yasukawa, T.: A Fuzzy-Logic-Based Approach to Qualitative Modeling. *IEEE Trans. Fuzzy Systems* **1**(1) (1993) 7-31
10. Kim, E. T., et al.: A simple identified Sugeno-type fuzzy model via double clustering. *Information Science* **110** (1998) 25-39
11. Zadeh L. A.: Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems* **90** (1997) 111-117

# Evaluation of the Distributed Fuzzy Contention Control for IEEE 802.11 Wireless LANs

Young-Joong Kim, Jeong-On Lee, and Myo-Taeg Lim

Department of Electrical Engineering, Korea University,  
1, 5-ka, Anam-dong, Sungbuk-ku, Seoul 136-701, Korea  
{kyjoong, apsd, mlim}@korea.ac.kr  
<http://cml.korea.ac.kr>

**Abstract.** In our previous works, we focused on run-time optimization of the IEEE 802.11 protocol to improve its performance using a well-known fuzzy logic approach. Specifically, we derived the simple, and more accurate, approximation of the network contention level and the average size of contention window to maximize the theoretical throughput limit. In addition, we proposed the distributed fuzzy contention control (DFCC) mechanism using a fuzzy logic approach. In this paper, we propose the extension of the DFCC mechanism with a priority mechanism. To verify efficiency and robustness of our mechanism, the performance of the IEEE 802.11 standard protocol with the extension of DFCC mechanism are investigated through more realistic scenarios.

## 1 Introduction

This paper focuses on the IEEE 802.11 wireless local area network (WLAN) presented in [1], [2]. Some researchers have investigated the enhancement of the IEEE 802.11 media access control (MAC) protocol to increase its performance when it is used in WLANs. Through a performance analysis, it has been studied the tuning of the standard's parameters [3], [4], [5]. In [6], solutions have been proposed for achieving a more uniform distribution of the accesses. Recently, other researchers have pointed out that the standard protocol can be very inefficient [7], [8], [9]. Specially, the average size of the contention window that maximizes the theoretical throughput limit was derived analytically, and it was shown that depending on the network configuration, the standard can operate very far from the theoretical throughput limit. In addition, an appropriate tuning of the backoff algorithm that can drive the IEEE 802.11 protocol close to the theoretical throughput limit was proposed by [9]. That is a distributed algorithm that enables each station to tune its backoff algorithm at run-time. To perform this tuning, a station must have exact knowledge of the network contention level; unfortunately, in a real case, a station cannot have exact knowledge of the network contention level (i.e., number of active stations and length of the message transmitted on the channel), but it, at most, can estimate it [10], [11].

Hence, a distributed mechanism for contention control in IEEE 802.11 WLANs was proposed and evaluated by [10], [11]. This mechanism, named

*Asymptotically Optimal Backoff* (AOB), dynamically adapts the backoff window size to the current network contention level and guarantees that an IEEE 802.11 WLAN asymptotically achieves its optimal channel utilization for a large number of stations. The AOB mechanism measures the network contention level by using two simple estimates: the *slot utilization*  $S_U$  and the average size of transmitted frames. These estimates are simple and can be obtained by exploiting information that is already available in the standard protocol. AOB can be used to extend the standard IEEE 802.11 access mechanism without requiring any additional hardware. According to AOB mechanism, its control is based on the parameter, named *Probability of Transmission*,  $P_T$ , whose value depends on the  $S_U$ . However, this control is not effective for a small number of stations. Moreover, since the  $P_T$  depends on the ratio,  $S_U/S_{U_{optimal}}$ , so  $P_T$  is always small. Therefore, it always offers a little opportunity.

In the previous work, we derived the simple, and more accurate, approximation of the network contention level and the average size of contention window to maximize the theoretical throughput limit, and we proposed a new  $P_T$  formula using a well-known fuzzy logic approach. Moreover, we proposed and a new *distributed fuzzy contention control* (DFCC) mechanism using the proposed  $P_T$  formula [12].

In this paper, we extensively investigate the performance of the IEEE 802.11 protocol enhanced with the proposed DFCC mechanism in more realistic scenarios. Specifically, we analyze the behavior of DFCC mechanism when the network operates under steady-state conditions with more real mixed traffics. In addition, we analyze the robustness of proposed mechanism in transient conditions. It is interesting to note that it is possible to extend the basic DFCC mechanism with a priority mechanism. The extension of DFCC mechanism with priorities is proposed.

The contents of this paper are as follows. In the section 2, we sketch the portions of the AOB mechanism and the approximated theoretical throughput limit which are relevant for this paper, and the extension DFCC mechanism with a priority mechanism is proposed. To verify our proposed mechanism, we make the steady-state and transient analysis, and performance evaluations of the priority mechanism in the section 3. Finally, the section 4 gives our conclusions.

## 2 Run-Time Fuzzy Optimization of IEEE 802.11

The drawbacks of the IEEE 802.11 backoff algorithm, explained in the previous works, indicate a direction for improving the performance of a random access scheme by exploiting the information on the current network congestion level that is already available at the MAC level. Specifically, the utilization rate of the slots called  $S_U$  presented in [10], [11] observed on the channel by each station is used as a simple and effective estimate of the channel congestion level. A simple and intuitive definition of the  $S_U$  is then given by:

$$S_U = \frac{Num\_Busy\_Slot}{Num\_Available\_Slot} \quad (1)$$



where  $Num\_Busy\_Slot$  is the number of slots in the *Backoff Interval* (BI) where one or more stations start a transmission attempt, and  $Num\_Available\_Slot$  is the total number of slots available for transmission in the BI, i.e., the sum of idle and busy slots. In the IEEE 802.11 standard mechanism, every station performs a *Carrier Sensing* activity and thus, the  $S\_U$  estimate is simple to obtain and no additional hardware is required [10], [11].

The current  $S\_U$  estimate can be used by each station to evaluate the opportunity to either perform or defer the scheduled transmission attempt. When the probability of a successful transmission is low, it should defer its transmission attempt. This can be achieved in an IEEE 802.11 network by exploiting the DFCC mechanism proposed in [12]. The DFCC mechanism can be to dynamically tune the backoff window size to achieve the theoretical capacity limit of the IEEE 802.11 protocol. This mechanism is based on the  $P\_T$  parameter which depends on the current contention level of the channel, i.e.,  $S\_U$  and the function of  $q$  value, named *Asymptotic Contention Limit*  $M \cdot p_{min}(q)$ . Here,  $M$  is the number of current stations. A detailed description can be found in [12]. This mechanism guarantees that the optimal channel utilization is asymptotically achieved for  $M > 10$  values.

### 2.1 Approximated Theoretical Throughput Limit

By the help of [9], we approximate  $p_{min}$  with the  $p$  value satisfying the following relationship:

$$E[Coll] \cdot E[N_c] = (E[N_c] + 1) \cdot E[Idle\_p] \tag{2}$$

where  $E[Coll]$  is the average time the channel is busy due to a collision,  $E[Idle\_p]$  is the average number of consecutive idle slots, and  $E[N_c]$  is the average number of collisions in a virtual transmission time. The expressions in (2) are defined in [9]. However, since this  $p_{min}$  derivation is too complex for our purpose, we use the Taylor series expansion. By applying Taylor series expansion, we derived the following new relationship in our previous work [12].

$$\{2l(q) - 1\}(Mp)^3 - 2\{l(q) - 1\}(Mp)^2 - 6Mp + 6 = 0, \tag{3}$$

where

$$l(q) = \frac{1 + 2q}{1 - q^2}. \tag{4}$$

The solution of (3),  $Mp$  is an input of the fuzzy system defined in [12]. In addition, through the comparative results, we showed that the proposed estimate obtained by (3) is very closer to analytical estimate than the asymptotic estimate. A detailed description can be found in [9], [12].

### 2.2 Distributed Fuzzy Contention Control Mechanism

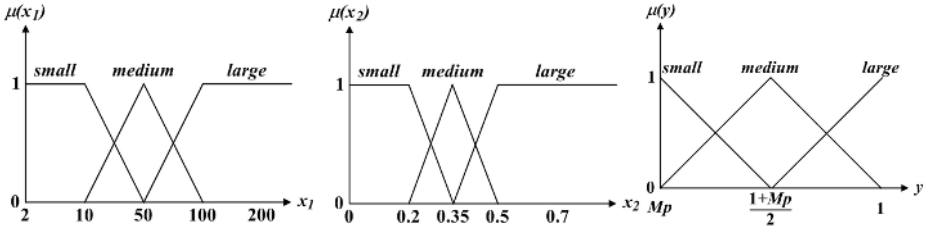
In this paper, we propose a new  $P\_T$  parameter using fuzzy logic approach and the heuristic formula is as follows:

$$P\_T(S\_U, N\_A, P\_L, y) = (1 - S\_U^{N\_A \cdot P\_L}) \cdot y \tag{5}$$

where  $P\_L$  is the *Priority Level* parameter and  $y$  is the output variable of the following fuzzy IF-THEN rules:

$$\begin{aligned}
 \text{Rule}^{(1)} &: \text{IF } x_1 \text{ is } S \text{ and } x_2 \text{ is } L, \text{ THEN } y \text{ is } L \\
 \text{Rule}^{(2)} &: \text{IF } x_1 \text{ is } S \text{ and } x_2 \text{ is } M, \text{ THEN } y \text{ is } L \\
 \text{Rule}^{(3)} &: \text{IF } x_1 \text{ is } S \text{ and } x_2 \text{ is } S, \text{ THEN } y \text{ is } M \\
 \text{Rule}^{(4)} &: \text{IF } x_1 \text{ is } M \text{ and } x_2 \text{ is } L, \text{ THEN } y \text{ is } L \\
 \text{Rule}^{(5)} &: \text{IF } x_1 \text{ is } M \text{ and } x_2 \text{ is } M, \text{ THEN } y \text{ is } M \\
 \text{Rule}^{(6)} &: \text{IF } x_1 \text{ is } M \text{ and } x_2 \text{ is } S, \text{ THEN } y \text{ is } S \\
 \text{Rule}^{(7)} &: \text{IF } x_1 \text{ is } L \text{ and } x_2 \text{ is } L, \text{ THEN } y \text{ is } M \\
 \text{Rule}^{(8)} &: \text{IF } x_1 \text{ is } L \text{ and } x_2 \text{ is } M, \text{ THEN } y \text{ is } S \\
 \text{Rule}^{(9)} &: \text{IF } x_1 \text{ is } L \text{ and } x_2 \text{ is } S, \text{ THEN } y \text{ is } S
 \end{aligned} \tag{6}$$

where  $x_1$  is an input variable as the number of current stations  $M$ ,  $x_2$  is an input variable as the proposed estimate of  $p_{min}$ , the linguistic variables  $S$ ,  $M$ , and  $L$  mean “small,” “medium,” and “large,” respectively. Moreover, each proposed membership function is presented in Fig. 1.



**Fig. 1.** (i) The number of current stations,  $x_1$ , as a linguistic variable that can take fuzzy sets “small”, “medium,” and “large” as  $M$  values in the left plot. (ii) The proposed estimate,  $x_2$ , as a linguistic variable that can take fuzzy sets as  $p_{min}$  values in the center plot. (iii) The output of fuzzy rules,  $y$ , as a linguistic variable that can take fuzzy sets as  $[Mp_{min}, 1]$  values in the right plot.

For each given frame to transmit, the *Priority Level* parameter could be mapped on the Type of Service values defined by the application level. When the  $P\_L$  is greater than 1, it introduces a fast increment in the priority of the station, with respect to the number of transmission attempts performed. The proposed  $P\_T$  parameter can be used to evaluate the opportunity to perform a transmission on the shared channel. When the station decides to defer the transmission, it reschedules a new attempt, as in the case of a collision occurred. Specifically, the proposed algorithm adopted by each station is sketched in Algorithm 1.

**Algorithm 1:** DFCC mechanism with a priority mechanism.

```

...
if (Backoff_Counter == 0) /* A slot for transmission is reached */
then
  calculate the S_U;
  calculate the M_p;
  obtain the y; /* y is the output of the fuzzy IF-THEN rules */
  if (Rand() < P_T(S_U, N_A, P_-L, y))
  then
    BYPASS the transmission indication to the HW;
  else
    DEFER the transmission;
  if ((transmission deferred) or (collision occurred))
  then
    NOTIFY the collision occurred;
    /* schedule a new retransmission */
...

```

The proposed mechanism with a priority mechanism can be used to extend the standard 802.11 access mechanism without requiring any additional hardware.

### 3 Performance Evaluation of DFCC Mechanism

In this section, to verify efficiency of our mechanism, the performance of the IEEE 802.11 standard protocol with AOB and proposed mechanism is investigated through simulations. The physical characteristics and parameter values of the investigated system are reported in Table 1.

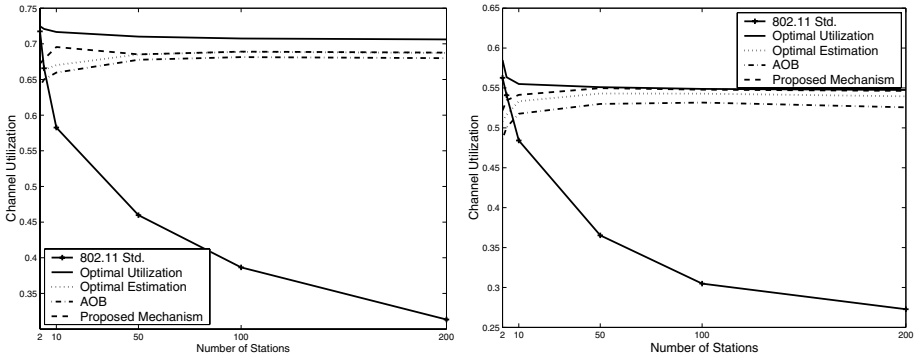
**Table 1.** Physical parameters for simulations

Parameters	Values
Number of current stations ( $M$ )	2 to 200
$CW_{min}$	16
$CW_{max}$	1024
Channel transmission rate	2Mb/s
Payload size	Geometric distribution with $q$
Acknowledgement size	200 $\mu$ sec
Header size	136 $\mu$ sec
Slot Time ( $t_{slot}$ )	50 $\mu$ sec
SIFS	28 $\mu$ sec
DIFS	128 $\mu$ sec
Propagation time	< 1 $\mu$ sec

### 3.1 Steady-State Analysis of the DFCC Mechanism

To analyze the DFCC behavior in a more realistic scenario, we assume that the message length distribution is bimodal. Specifically, we assume that “long messages” have an average length of 100 slots while “short messages” have an average length of 2.5 slots, and a slot corresponds to 100 bits.

Fig. 2 show the protocol capacity of the IEEE 802.11 protocol with and without the additional mechanisms. Simulation results indicate that the channel utilization with the DFCC mechanism is near-optimal and the DFCC mechanism is more effective than the AOB mechanism.



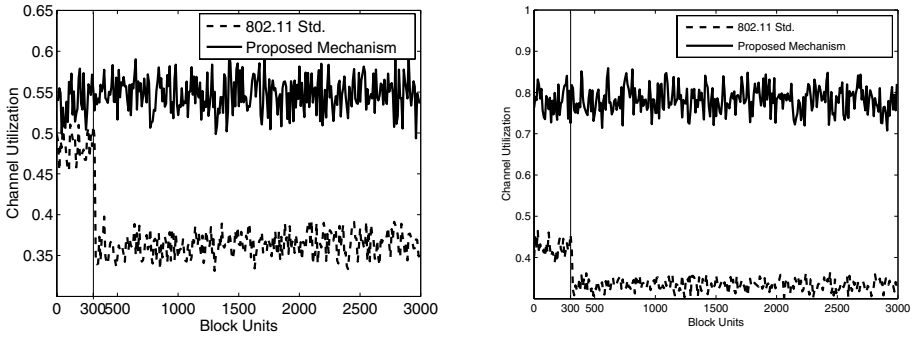
**Fig. 2.** (i) The left plot is the channel utilization of the IEEE 802.11 protocol with and without the additional mechanisms versus mixed traffic condition with  $q = 0.5$ . (ii) The right plot is the channel utilization of the IEEE 802.11 protocol with and without the additional mechanisms versus mixed traffic condition with  $q = 0.2$ .

### 3.2 Analysis of the DFCC Mechanism in Transient Situations

In this section, we investigate robustness of the DFCC Mechanism to react to rapid change in the number of active stations. Specifically, we start from a system operating in steady-state conditions with  $M = 10$  or  $M = 100$ . All stations transmit mixed messages. Message length is sampled from geometric distribution with an average of 2.5 and 100 slots for short and long messages, respectively. After 300 block units an additional 10 or 100 stations become active. Fig. 3 show the effectiveness and robustness of the DFCC Mechanism. In the Fig. 3, the vertical bar indicates burst arrival time. In the case of DFCC mechanism, the rapid increase in the number of active stations produces a negligible effect in channel utilization. On the contrary, standard protocol is negatively affected by these changes.

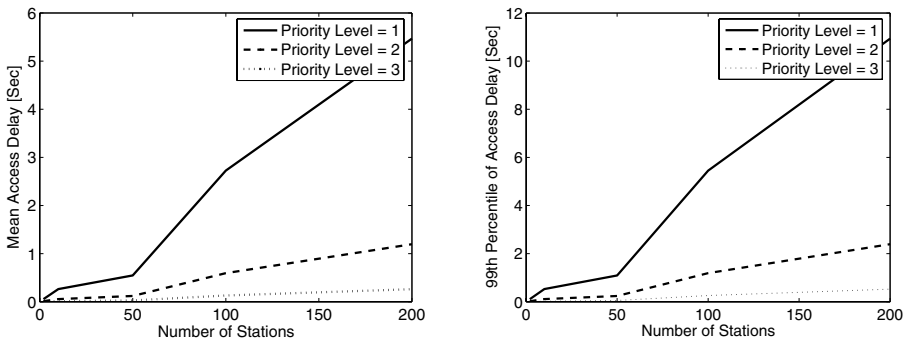
### 3.3 Analysis of the DFCC Mechanism with a Priority Mechanism

In this section, we investigate the effectiveness of the DFCC Mechanism with a priority mechanism. Fig. 4 (i) shows the effectiveness of the priority mechanism,



**Fig. 3.** (i) The left: A burst of 10 new stations is activated when the network is operating in steady-state conditions with 10 active stations. (ii) The right: A burst of 100 new stations is activated when the network is operating in steady-state conditions with 100 active stations.

in terms of the mean access delay obtained by three sets of stations belonging to three different priority levels. Fig. 4 (ii) confirms the effectiveness of the DFCC Mechanism with a priority mechanism, in terms of 99-th percentile of the access delay. Fig. 4 indicate that the priority level influences the access delay of all stations. Therefore, the proposed the DFCC Mechanism with a priority mechanism can be used to achieve a better access delay for a high priority station.



**Fig. 4.** (i) The left plot is a mean access delay with three different priority levels. (ii) The right plot is 99-th percentile of access delay with three different priority levels.

### 4 Conclusion

In this paper, we extensively investigate the performance of the IEEE 802.11 protocol enhanced with the proposed DFCC mechanism in more realistic scenarios. To increase the portion of bandwidth available for the high-priority traffics, the extension of DFCC mechanism with priorities is proposed. The effectiveness and

robustness of the proposed mechanism are investigated. The extension of DFCC mechanism with priorities becomes effective allowing high-priority frames to be transmitted with an enhanced Quality of Service.

## References

1. IEEE Standard for Wireless LAN - Medium Access Control and Physical Layer Specification, P802.11. Nov. (1997)
2. Stallings, W.: *Local & Metropolitan Area Networks*. Englewood Cliffs, Prentice Hall (1996)
3. Crow, B. P., Widjaja, I., Kim, J. G., Sakai, P. T.: IEEE 802.11 Wireless Local Area Networks. *IEEE Commun. Mag.*, Sept. (1997) 116-126
4. Chhaya, H. S., Gupta, S.: Performance Modeling of Asynchronous Data Transfer Methods in The IEEE 802.11 MAC Protocol. *ACM/Balzer Wireless Netw.*, Vol. 3. (1997) 217-234
5. Bianchi, G., Fratta, L., Oliveri, M.: Performance Evaluation And Enhancement of The CSMA/CA MAC Protocol for 802.11 Wireless LANs. *Proc. PIMRC*. Taiwan, Oct. (1996) 392-396
6. Weinmiller, J., Woesner, H., Ebert, P., Wolisz, A.: Analyzing and Tuning the Distributed Coordination Function in the IEEE 802.11 DFWMAC Draft Standard. *Proc. Int. Workshop on Modelling, MASCOT* (1996)
7. Cali, F., Conti, M., Gregori, E.: IEEE 802.11 Wireless LAN: Capacity Analysis and Protocol Enhancement. *Proc. INFOCOM Conf.*, Mar./Apr. (1998) 142-149
8. Cali, F., Conti, M., Gregori, E.: Dynamic IEEE 802.11: Design, Modeling and Performance Evaluation. *IEEE J. Selected Areas in Comm.*, Vol. 18(9). Sept. (2000) 1774-1786
9. Cali, F., Conti, M., Gregori, E.: Dynamic Tuning of The IEEE 802.11 Protocol to Achieve A Theoretical Throughput Limit. *IEEE/ ACM Trans. Networking*, Vol. 8(6). Dec. (2000) 785-799
10. Bononi, L., Conti, M., Donatiello, L.: Design And Performance Evaluation of A Distributed Contention Control (DCC) Mechanism for IEEE 802.11 Wireless Local Area Networks. *J. Parallel And Distributed Computing*, Vol. 60(4). Apr. (2000).
11. Bononi, L., Conti, M., Gregori, E.: Runtime Optimization of IEEE 802.11 Wireless LANs Performance. *IEEE Trans. on Parallel and distributed Systems*, Vol. 15(1). Jan. (2004) 66-80
12. Kim, Y., J., Lim, M., T.: Run-Time Fuzzy Optimization of IEEE 802.11 Wireless LANs Performance. *ICNC 2005, LNCS 3612*, Sep. (2005) 1079-1088

# Improved Genetic Algorithm-Based FSMC Design for Multi-nonlinear System

Jung-Shik Kong<sup>1</sup> and Jin-Geol Kim<sup>2</sup>

<sup>1</sup>Dept. of Automation Eng., Inha University, Yonghyun-Dong, Nam-Gu, Incheon, Korea  
selkirk@paran.com

<sup>2</sup>School of Electrical Eng., Inha University, Yonghyun-Dong, Nam-Gu, Incheon, Korea  
john@inha.ac.kr

**Abstract.** A new motion controller for the multi-nonlinear system is proposed by using a fuzzy sliding mode controller (FSMC) based on improved genetic algorithm (GA). In controlling the nonlinear element of the system, there are some critical problems such as the limit cycle. As the system has nonlinearities, a robust controller is one of the optimal solutions. The FSMC is a kind of the robust methods to control nonlinearities effectively in a system. Prior to applying a FSMC, genetic algorithm is used for identifying system without manual tuning and obtaining optimal fuzzy set of FSMC. The suggested GA is an improved type to find optimal solution. It uses new type of crossover and mutation with a sigmoid function that is applied to improve the searching ability. Also, an additional compensator and motion controller are suggested in order to improve position tracking. All the processes are investigated through simulations and experimentally verified in a real motor system.

## 1 Introduction

The motor system is one of the general actuators that require a control algorithm. The system includes many kinds of nonlinearities such as saturation, Coulomb friction, backlash and so on. These nonlinearities result in generating limit cycle and system instability. Many studies have been performed to reduce these problems [1-3]. E. J. Davison [4] introduced a describing function as a method of frequency analysis and Takamasa Hori [5, 6] suggested a disturbance observer to reduce the limit cycle by a nonlinear parameter. K.T. Woo and C. W. Tao [7, 8] proposed a method using an AI algorithm. In addition, the control method to track the trajectory is one of the most important issues in the motor system, since most motor systems are employed for generating various motions according to given trajectories. Thus, the proper control algorithm must be included in the controller to improve the performance of tracking control.

In this paper, the FSMC is applied to the nonlinear systems as a robust controller. Even if the FSMC can efficiently represent an arbitrary nonlinear control law, the lack of parameter information for configuration remains one of the main problems for deployment of practical applications. Although optimal solutions can be obtained through simulation, inaccurate system information causes the special tuning through

the real experiment. Therefore, identification of motor systems is required for applying optimal results to the real system. Here the GA is one of the effective methods used for calculating the optimal gain of the FSMC and obtaining precious mathematical model by identification. At that time, improved GA is applied to gain more optimal solutions. Following this sequence, nonlinearities can be removed from the FSMC. However, tracking performance may not be satisfactory, even though the limit cycle is removed by the FSMC. Therefore, the tracking algorithm must be included in the motor system. In this paper, a motion controller with a compensator is applied to track the given trajectory. All the processes are performed using simulation programs and are verified with an actual motor control system.

## 2 Improved GA

GA is considered as one of the searching algorithms in a global area. Finding the optimal solution without solving a differential equation is the main advantage of the GA.

In this paper, an improved type of GA is suggested. The previous method was separated into two sections. One was a binary-coded genetic algorithm (BGA) and the other was a decimal convex genetic algorithm (DGA). Fig. 1 shows the crossover and mutation processes of the BGA and DGA, respectively.

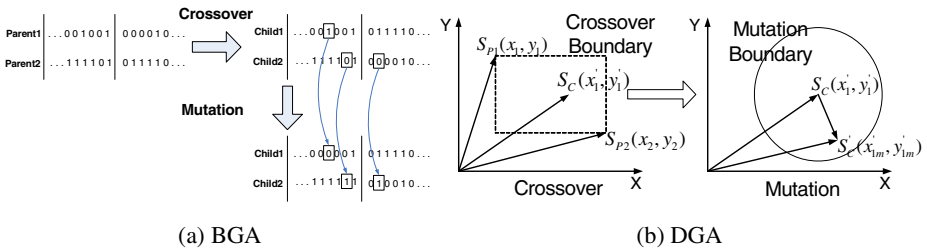


Fig. 1. Crossover and mutation of the BGA and DGA

In Fig. 1(b),  $S_{p1}(x_1, y_1)$  and  $S_{p2}(x_2, y_2)$  are a set of parents,  $S_c(x'_1, y'_1)$  is the result of crossover and  $S'_c(x'_{1m}, y'_{1m})$  is the output of the mutation. Here  $x$  and  $y$  represent the parameters to be determined using GA. BGA has the advantage of being able to apply genetic concepts easily. However, accuracy of BGA limits the searching precision for an optimal solution. In addition, DGA can search for a more precise optimal solution than the BGA because it uses real values. However, it settles down with increasing time, because this algorithm is dictated by time with the number of generations.

In this study, a more advanced algorithm was proposed to improve the problems associated with the previous algorithms. This method basically includes a DGA. As shown in Fig. 1(b), the DGA has the maximum crossover boundary between two parents during the crossover. In a conventional decimal-coded GA, the fitness function is settled after passing the generation count and the optimal data is searched for at



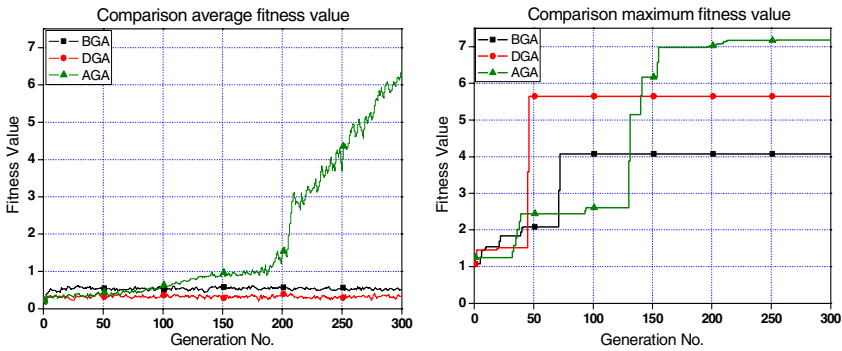
this time. The data, which remains close to the optimal value, may not be the desired data. If crossover and mutation boundary is reduced while the generation count has passed, a more optimal solution can be searched. In order to solve this problem, a sigmoid function was used in crossover and mutation. Equations (1-2) show the proposed crossover and mutation.

$$\begin{aligned} x_1' &= x_1 + \delta(t) \times (1 - \lambda_1) x_2, & x_2' &= x_2 + \delta(t) \times (1 - \lambda_2) x_1 \\ y_1' &= y_1 + \delta(t) \times (1 - \lambda_3) y_2, & x_2' &= y_2 + \delta(t) \times (1 - \lambda_4) y_1 \end{aligned} \tag{1}$$

$$\begin{aligned} x_{1M}' &= x_1' + \delta(t) \times \text{sgn}(\alpha) \times d \times \lambda_m \\ y_{1M}' &= y_1' + \delta(t) \times \text{sgn}(\alpha) \times d \times \lambda_m \end{aligned} \tag{2}$$

$$\delta(t) = \left( 1 - \frac{1}{1 + e^{-(10t+5)}} \right), \quad t : \text{generation count}$$

where  $\lambda$  is the random number ( $\lambda \in (0,1)$ ). Since the parents influence the child's crossover, the child value should be within the crossover boundary. In addition, mutation  $d$  denotes the maximum boundary altered by the mutation, and  $\alpha$  is a random number. By using  $\delta(t)$ , relative boundaries of the crossover and the mutation are reduced according to the time. It will be selected the proper value to search more optimal solution. Fig. 2 shows the comparison of average and maximum of the fitness values.



**Fig. 2.** Comparison of average and maximum fitness values

AGA is an acronym for an advanced decimal-coded genetic algorithm. Sigmoid function is used in AGA when the GA is performed. It can improve the possibility for crossover result that can be closer than the DGA. From Fig. 2, the average fitness from the proposed GA has a relatively high fitness value. As the identification results show, a larger fitness value indicates a larger similarity to the real motor models.

### 3 Controller Design

#### 3.1 Motor Identification

Motor identification is needed to search for accurate information on the motor. An accurate motor model can assist in determining the optimal controller gains. However, most motor models have nonlinear parameters, which make it difficult to determine a mathematical model. To solve this problem, many researches are proposed [9-11]. GA is especially good to approach model that has no accurate parameters. Fig. 3 shows the identification system using the GA.

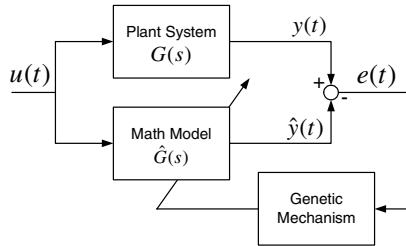


Fig. 3. Block diagram of the motor system identification

In Fig. 3, a plant system block  $G(s)$  represents a real model, while  $\hat{G}(s)$  denotes a mathematical model controlled by the GA. Table 1 shows the parameters of the GA for a motor identification. Here  $P_R$  is the position value of a real system and  $P_M$  is the position value of a mathematical model.

Table 1. Parameters of the GA for a motor identification

Parameters	value	Parameters	value
Generation No.	300	Crossover Rate.	0.6
Population No.	100	Mutation Rate	0.1
Fitness function	$fit_i = 1 / \sum  P_R - P_M $		

#### 3.2 Design of FSMC

A new controller is proposed to solve the combined nonlinearity efficiently, as shown in Fig. 4. Here three types of control units are inserted in order to control nonlinearities and motion. One is FSMC that can reduce effect by nonlinearities. Another is a compensator that is also proposed to improve position tracking in a motion controller. In Fig. 4, the compensator module has a FSMC applied to real hardware, and only has saturation with nonlinearity. The other is the motion controller that is based on a PID controller.

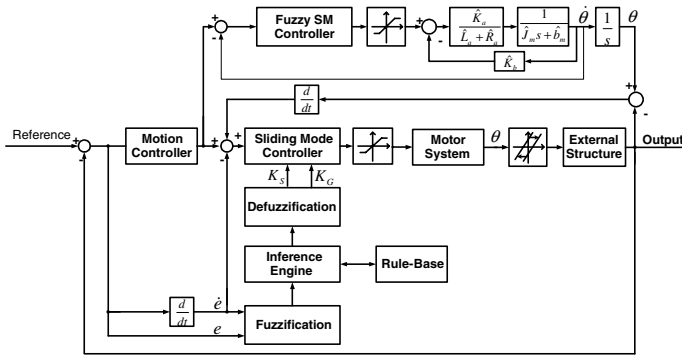


Fig. 4. Block diagram of proposed controller

To apply the fuzzy mechanism in the sliding mode controller (SMC), the following equation is defined.

$$u = K_G |e| \operatorname{sgn}(e + K_s \dot{e}) \tag{3}$$

where  $K_s$  and  $K_G$  are surface and proportional gains of SMC, respectively. A fuzzy algorithm is used for finding optimal  $K_s$  and  $K_G$ . Here  $e$  and  $\dot{e}$  are velocity error and differential velocity error, represented by a 7-triangle membership function. In addition, 9-triangle membership functions are employed for the fuzzy output sets, and the centre-average method is used in the process of defuzzification.

GA is used for searching for optimal vertexes of output membership function of FSMC. Fig 5 shows the output membership function. At this stage, all membership functions are symmetrically positioned with respect to the centre of the universes of discourse. Therefore, the vertexes of the universes of discourse are classified into 4 levels such as NB(PB), NL(PL), NM(PM), and NS(PS). Table 2 shows the genetic parameters applied to search for vertexes of output membership function of FSMC.

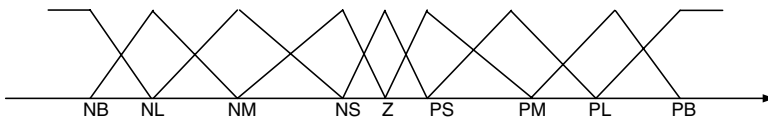


Fig. 5. Output membership function

Table 2. GA parameters of applying at FSMC

Parameters	Value	Parameters	Value
Generation No.	300	Crossover Rate.	0.7
Population No.	100	Mutation Rate	0.2
Fitness function	$fit_s = (l_{\max} - l_{\min} + 1) / \sum (R_t - Y_t)$		

where  $l_{max}$ ,  $l_{min}$ ,  $R_t$  and  $\gamma_t$  are the maximum level of the limit cycle, the minimum level of the limit cycle, the reference value and the output value at each time  $t$ , respectively. The limit cycle is considered in the fitness function of FSMC.

### 4 Experiment

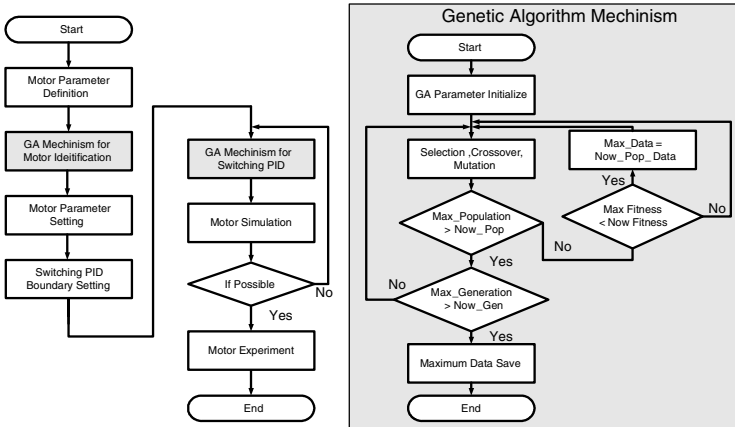
In order to verify this suggestion, some simulations are performed using the motor control system, which was developed for our humanoid robot system. Table 3 details the parameters used in the motor model and Table 4 shows the tuning results of the FSMC gains by the GA. All sequences used for finding the GA are shown in Fig. 6.

**Table 3.** Detailed parameters of a motor

Sym.	Value	Sym.	Value	Sym.	Value	Sym.	Value
$L_a$	0.000342	$J_m$	$8.09 \times 10^{-7}$	$K_a$	0.0193	$f_s$	0.00194
$R_a$	6.09146	$f_m$	$1.51 \times 10^{-5}$	$K_b$	0.0169	$f_k$	0.00657

**Table 4.** FSMC gains

	NB	NL	NM	NS
$K_G$	129.75	119.57	98.48	84.23
$K_S$	0.00416	0.00626	0.00418	0.00273



**Fig. 6.** Block diagram of all the process

The experiments are carried out using the proposed controller with a TI TMS320F2810 DSP controller. In the experiment, the backlash gap of the motor system was 4.3 degree and the reduction ratio of the gearbox was 231:1. And

proportional and integral gains of the motion controller are selected at 1.7 and 0.1, respectively.

In order to confirm the performance of the proposed controller, a real trajectory in a robot system is applied. Fig. 7 presents the result and error by the PID controller and the SMC. In Fig. 7, using the PID controller and SMC, the optimal gains are searched using the GA. And Fig. 8 shows the result and error by the proposed controller.

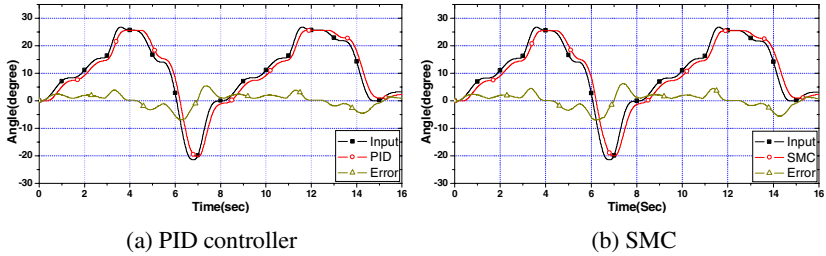


Fig. 7. Tracking position and error by PID controller and SMC

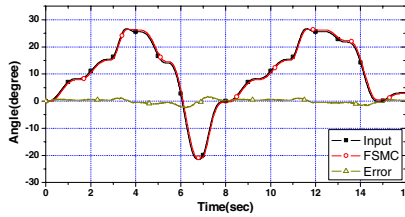


Fig. 8. Tracking position and error by the proposed controller

Table 5 shows the means and the standard deviation for error of each controller. From table 5, it is verified that the proposed controller can control motor system with better performance.

Table 5. Means and standard deviations for error of each controller

	PID	SMC	Proposed Controller
Mean	1.979736	2.127173	0.681042
Standard deviation	1.459002	1.641408	0.421955

## 5 Conclusion

In this paper, two types of GA were applied to the motor system with nonlinearities. The first was used for obtaining the motor parameters in order to identify the motor system. The second was employed to obtain the optimal output membership function of FSMC. This proposed controller based on FSMC could efficiently remove the internal and external nonlinearities such as the saturation, backlash, and some friction.

In order to reduce these nonlinearities, an accurate motor model was derived using the improved GA, and its output membership function determined by GA without any tedious transformations. In addition, a compensator was suggested to reduce the effect of external nonlinear elements. Also, motion controller helps motor system to follow given trajectory. Through these processes, the limit cycle caused by the nonlinearities could be efficiently reduced and motor system could have more superior performance in the tracking system. As a result, the suggested controller based on FSMC was very efficient to control the motor system with nonlinearities.

In the future, more advanced genetic algorithms will be considered. The current algorithm has a time limitation caused by computational load when used in a real time. In order to apply a GA to a real time system, the computational mechanism has to be changed into a simpler form. Furthermore, more advanced algorithms for multi-nonlinear processes will also be needed to provide a more accurate and faster control operation.

## Acknowledgement

This work was supported by INHA UNIVERSITY Research Grant.

## References

1. Nizar J. Ahmad and Farshad Khorrani: Adaptive Control of Systems with Backlash Hysteresis at the Input, Proceedings of the American Control Conference (1999), 3018-3022
2. Serkan T., Impram and Neil Munro: Limit Cycle Analysis of Uncertain Control Systems with Multiple Nonlinearities, Conf. on Decision and Control (2001), 3423-3428
3. Weiqing Huang and Lilong Cai: New Hybrid Controller for systems with Deterministic Uncertainties, IEEE Trans. on Mechatronics, Vol. 5, No. 4 (2000), 342-348
4. E.J. Davison: Application of the Describing Function Technique in a Single-Loop System with Two Nonlinearities, IEEE Trans. on Automatic Control (1968), 168-170
5. Satoshi Komada, Noriyoshi Machii and Takamasa Hori: Control of Redundant Manipulators Considering Order of Disturbance Observer, IEEE Trans. on Industrial Electronics, Vol. 47, No. 2 (2000), 413-420
6. Kohji Yamada, Satoshi Komada, Muneaki Ishida, and Takamasa Hori: Characteristics of Servo System Using High Order Disturbance Observer, Conf. on Decision and Control (1996), 3252-3257
7. K.T. Woo, Li-Xin Wang, F.L. Lewis, and Z.X. Li: A Fuzzy System Compensator for Backlash, IEEE Int. Conf. on Robotics and Automation (1998), 181-186
8. C. W. Tao: Fuzzy Control for Linear Plants with Uncertain Output Backlashes, IEEE Trans. on systems, Man and Cybernetics, Vol. 32, No. 3 (2002), 373-380
9. Lu, S. and Basar, T.: Robust Nonlinear System Identification Using Neural-network Models, IEEE Transactions on Neural networks, Vol. 9 (1998), 407-429
10. Yamada, T. and Yabuta, T.: Dynamic System Identification Using Neural Networks, IEEE Transactions on Systems Man and Cybernetics Vol. 23 (1993), 204-211
11. Vachkov, G. and Fukuda, T.: Identification and Control of Dynamical Systems Based on Cause-effect Fuzzy Models, Int. Conf. IFSA Vol. 4 (2001), 2072-2077

# The Interactive Feature Selection Method Development for an ANN Based Emotion Recognition System

Chang-Hyun Park and Kwee-Bo Sim

School of Electrical and Electronic Engineering, Chung-Ang University, 221,  
Heukseok-Dong, Dongjak-Gu, Seoul, 156-756, Korea  
kbsim@cau.ac.kr

**Abstract.** This paper presents an original feature selection method for Emotion Recognition which includes many original elements. Feature selection has some merit regarding pattern recognition performance. Thus, we implemented a simulator called an 'IFS system' and the results of the IFS were applied to an emotion recognition system(ERS). Our innovative feature selection method was based on a Reinforcement Learning Algorithm and since it required responses from human users, it was denoted an 'Interactive Feature Selection'. By performing an IFS, we were able to obtain three top features and apply them to the ERS.

## 1 Introduction

Emotion recognition research has been typically attempted using four kinds of medium. They are speech, image, physiological signal, and gesture. In addition, our IEEE survey papers published from 1990 to 2005 show that papers using the speech medium have been published more often than others have. The reason for this result is probably due to feature set extraction from speech and image being easier than physiological signal or gesture and the possibility of classification is higher. In particular, EEG, ECG, and SC sensors are used to obtain a physiological signal but the signal from those sensors may be obstructed by electrical signals from fluorescent lamps or electric home appliances. This problem is the one obstacle in emotion recognition using a physiological signal. For an image, this means facial expression recognition and the main problem in this case is usually lighting conditions, which often change, or personal accessories like glasses which affect recognition performance. A problem of gesture recognition is similar to that of image recognition and the bigger problem is that it may not include much information regarding emotion. Apart from above the problems which these three media present, speech signal can send much more information regarding emotion. For example, talking over the telephone, one can recognize emotions and this shows the validity of speech signal for emotion recognition. Even a cheap microphone can be used sufficiently as a sensor for collecting speech signals and noise will not affect the extraction of the feature set unless it is too loud to be classified as a coming from the source of the signal. These reasons are why most researchers have focused on speech signal and

why we have selected this medium for our paper. The commonly used feature set for emotion recognition from speech consists of pitch, energy, formant, and speech rate. Some researchers select all four of the feature sets, others select only one, and the features are generally extracted statistically from the four feature sets. In [1], 87 features were extracted from pitch, energy, and formant and they were classified into five emotions. In [2], 17 features were extracted from pitch, energy, speech rate and so on with sex also being classified. In addition, In [3], 11, 40, and 13 features were extracted. The fact that feature set selection is not fixed suggests that features may or may not be relevant to emotion recognition. This problem will plague researchers in this field until exceptional results are obtained. For this case, there is a GA based selection method, Floating search method and so on which can somewhat reduce difficulties for researchers [4]. Especially, the Sequential Forward Selection and Sequential Backward Selection methods of a Floating search method have been frequently used. In [2], a Forward Selection (FS) method was used and in [1], the 10 best features were selected out of 87 features by using a Sequential Floating Forward Selection algorithm (The extended version of SFS). In [5], a Sequential Forward Selection algorithm was also used and the best feature subset was selected out of 39 candidate feature sets and In [6], a good feature set was found using genetic programming for the music genre classification related problem. These feature selection methods provided a good solution for "The curse of dimensionality" and contributed to the performance of pattern recognitions. In addition, feature selection methods included supervised and unsupervised cases. Generally, a supervised case is employed more often than an unsupervised case. This is due to unsupervised feature selection methods having a high probability of incorrect results for corresponding patterns regarding perceived speech [7]. Although, there are many cases that cannot obtain an explicit supervised value, the unsupervised method has advantages. We propose a method using reinforcement learning taking advantage of both the supervised and unsupervised method, which can alleviate the shortcomings of both methods. Researches of the reinforcement learning have been proceeded using many methods, i. e. Dynamic programming, Monte Carlo method, TD method, Q learning etc. proposed by Sutton and Barto. Since there is such a variety of methods and the main elements such as "state", "action" and "reward" may be freely defined and implemented by a developer, this method is thought to be a very important one for machine learning techniques [8]. In this study, rather than using a specific reinforcement learning method, we propose a method which selects feature sets by calculating rewards received when an action is performed in a state. In particular, this method does not only calculate the frequency of emotion transit but also the sum of the rewards for the evaluation of a feature selection. Therefore, this method has the advantage that the more frequently it contacts a user, the better its performance becomes. The outline of the paper is as follows, In Section II, it explains the emotion recognition method and Section III explains the proposed algorithm. The Section IV shows a simulation and result of using the proposed algorithm. Section V conclude and shows future works.



## 2 Emotion Recognition Method

This paper addresses emotion recognition by extracting features of speech. The emotion recognition with speech is largely divided into cases using acoustic information and language or discourse information. The former is a method that uses some feature sets such as pitch, formant, speech rate, timbre, etc. and the latter uses the meaning of a word. That is, whether the word is positive or negative to whether it represents a happy or sad state. The process of emotion recognition consists of collecting emotional speech, the acoustic analysis, implementing DB, feature set extraction and such features are trained and classified with emotions using a pattern classification method.

### 2.1 Database and Preparation

Emotional speeches were collected from 10 male graduate students. Their ages ranged from 24 to 31 years old and they were asked to say with 10 short sentences emotionally. The choice of the 10 sentences (scripts) was decided upon from the result of another survey or experiment. In the first stage, 30 sentences had been prepared and all the 30 sentences were asked to say by the subjects. After the recording, the speeches were listened to by other people and they were asked the question "What emotion do you feel when listening to the given recording?". The emotions conveyed in the 10 sentences that were read and the answers given by the subjects in this experiment were in agreement 90% of the time. In addition, the length of the prepared sentences was limited from 6 to 10 syllables. The recording format was 11Khz, 16bit, mono and the subjects were asked to keep a distance of 10 cm between themselves and the microphone. Since the distance of the microphone affected loudness or intensity, maintaining the required distance was very important. Recorded files were preprocessed and stored in DB (MS-ACCESS). In the preprocessing stage, there were several processes to signals such as FFT (extracting spectrum), Pitch extraction (by an autocorrelation method), IR (Increasing Rate) of pitch, CR (Crossing Rate), VR (Variance), and statistical values etc [9].

### 2.2 Pattern Classification Method

We used an artificial neural network for pattern classification, which commonly performs well and is robust to a signal with noise. It has been the most popular method to use in the pattern recognition field. This method commonly uses a Back Propagation Algorithm for tuning network parameters. In this study, we fixed the setting to ANN as follows, The number of Input Units and Hidden Units and Output Units and Learning rate and Tolerance and Sigmoid function are 3~5 and 11 and 2 and 0.003 and 0.25 and  $\frac{1}{1+e^{-3x}}$ , respectively.

## 3 The Interactive Feature Selection Algorithm

Typically, researchers in the emotion recognition use various feature sets. Some researchers looked into the relation between acoustic analysis and emotion and

used the feature sets based on that relation. However, because this method is subjective, it may easily lead to local minima. For this reason, recent studies consider a feature selection method for finding small superior features (4~10) out of as many as 30 to 90 features. Most researchers do not use all features because they cannot confirm whether they are valid or not and noises with every features may deteriorate. Therefore, feature selection methods are popular in the pattern classification field[7].

### 3.1 Sequential Forward Selection(SFS) Algorithm

Sequential Forward Selection is the simplest greedy search algorithm. In this paper, we will briefly explain this algorithm. The following figure shows the algorithm. Starting from the empty set, sequentially add the feature  $x^+$  that results in the highest objective function  $J(Y_k + x^+)$  when combined with the feature  $Y_k$  that has already been selected.

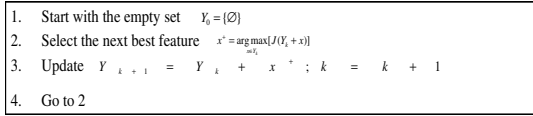


Fig. 1. SFS algorithm

### 3.2 Interactive Feature Selection Algorithm(IFS)

The Interactive Feature Selection algorithm we are proposing is an algorithm based on reinforcement learning. Specially, popular algorithms such as SFS, SBS and so on, are deductive algorithms but our proposed algorithm is inductive. Also, these feature selection algorithms are based on the rationale of correlation and information-theoretic measures. Correlation is based on the rationale that good feature subsets contain features highly correlated with a class, yet are uncorrelated with each other. The IFS is also based on the correlation concept. Moreover, the feature selection algorithms consist of a search strategy and an evaluation by objective function but the conventional methods are incompetent in the search strategy part. Therefore, an IFS focuses on both the search strategy and evaluation by objective function. Fig 2(a) shows an IFS process. We assume that an emotion recognition system that includes this algorithm will be applied to a home robot or appliance. If such a situation is considered, users may be comfortable inputting emotional speech and a user’s emotional state at that time(as a supervisor value). Due to this characteristic, this algorithm is a user adaptive system that can efficiently solve a problem and the more a user is in contact with this system, the better it will perform. The fig 2(b) shows an example of the IFS algorithm and is based on the Fig 2(a). First, this algorithm starts with a full feature set and when a new feature set and an emotion identifier is inputted, it assigns a +1 or -1 to the “return sign”(if an old emotion ID equals

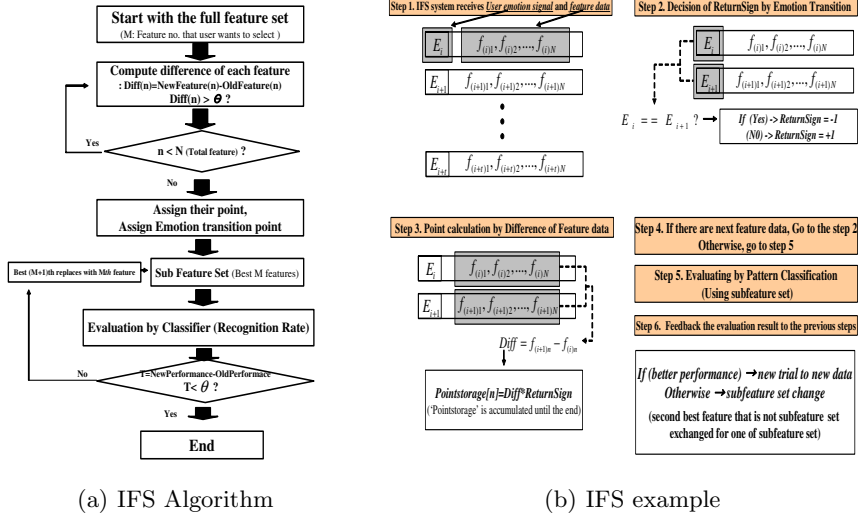


Fig. 2. Emotion Recognition System and Feature Selection DB implemented by VC++

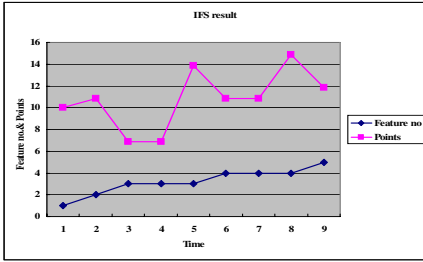
a new emotion ID then +1, Otherwise -1). Thereafter, the product of “return sign” and the difference of each feature is stored in an array “Pointstorage”. This iteration is repeated for one episode(user can arbitrarily define an episode). After the episode, the feature set that was selected first is applied to an objective function(Pattern Classification System) and the evaluation result is stored. If the next evaluation result is worse than the previous, the worst feature of the selected feature set will be replaced with the best feature among those that were not selected(step 6 in Fig 2(b)).

## 4 Experimental Results

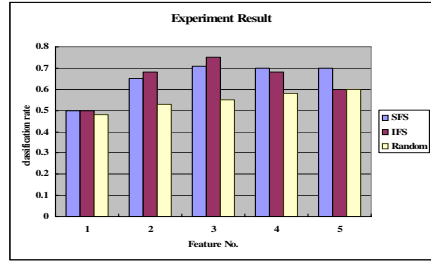
### 4.1 Simulation and Results

We applied 11 original features to the IFS simulator;Pitch features( max, min, mean, median, crossing rate, increasing rate), Loudness ;Intensity(max, min, mean),Sect.No and Speech rate. This program was made only for an IFS and the results from this program were applied to the Emotion Recognition System using ANN. That is, the feature set applied to ERS was the features previously outputted by IFS and then the emotion recognition experiment was performed. Classification was attempted using three methods. Results are shown in the fig 3(a) and 3(b).

From the Fig 3(a),we can see that the IFS system searched for better results and improved gradually. Also, the lower graph represents the feature no. at each instant. In the algorithm,because the searching work was performed again when the new evaluation set result was worse than the previous one, there was



(a) IFS Result on time axis



(b) Emotion Classification Rate Comparison

Fig. 3. Experiment Result

Feature No.	Features
1	Pitch Mean
2	Pitch Mean,Speech rate
3	Pitch Mean,Speech rate,Loudness
4	Pitch Mean,Speech rate,Loudness, Sect.No.

Fig. 4. Selected Feature Sets

some range in the steady state(The range of 3~5 and 6~8 in the Time axis). In addition, from Fig 3(b), this graph compares three methods(IFS, SFS and random selection) with the changed feature no. As expected, random selection performed poorly but IFS and SFS similarly performed better. In the IFS and SFS case, the results show a subtle distinction but the IFS with features 1,2 and 3 was better. However, with features 4 and 5, SFS showed better results.

Fig 4.1 shows the selected feature set from IFS system. As we can see, popular feature sets were selected.

## 5 Conclusions

This paper presents a solution to feature selection when there is an emotion recognition problem. The solution called an IFS performed as well as an SFS. In particular, it is reinforcement based learning and supplements the role of search strategy in the feature selection process. Using the IFS simulator, we found some of the best features and used them in the emotion recognition experiment and results were compared to those of SFS and Random selection. Performance was slightly better than SFS. However, IFS has some disadvantages. If the amount of training data is too small, selection results may be not good. SFS does not require much training data. It is also sufficient that training data be only one set. If an objective function is clear, SFS will be adequate. However, in the case of emotion recognition, SFS may not perform as well as it had. In this case, the correlation-based method like the IFS will be better.

## Acknowledgments

This research was supported by the Brain Neuroinformatics Research Program by Ministry of Commerce, Industry and Energy

## References

1. D.Ververidis and C.Kotropoulos :Emotional speech classification using Gaussian mixture models, Proceedings of ISCAS, vol. 3, May (2005) 2871-2874
2. C.M.Lee and S.S Narayanan :Toward detecting emotions in spoken dialogs, IEEE Transactions on Speech and Audio Processing, vol.13, March (2005) 293-303
3. J.Wagner, J.H.Kim and E.Andre :From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification, Proceedings of ICME, July (2005) 940-943
4. P.Pudil and J.Novovicova :Novel Methods for Subset Selection with Respect to Problem knowledge, IEEE Intelligent Systems, March (1998) 66-74
5. Y.L.Lin and W.Gang :Speech Emotion Recognition based on HMM and SVM, Proceedings of Machine Learning and Cybernetics, Vol.8, Aug (2005) 4898-4901
6. F.Morchen, A.Ultsch, M.Thies and I.Lohken :Modeling Timbre Distance With Temporal Statistics From Polyphonic Music, IEEE transaction on Audio, Speech and Language Processing, Vol.14, Issue 1, Jan. (2006) 81-90
7. E.F.Combarro, E.Montanes, I.Diaz, J.Ranilla, and R.Mones :Introducing a Family of Linear Measures for Feature Selection in Text Categorization, IEEE transactions on Knowledge and Data Engineering, Vol.17, No.9, Sept.(2005) 1223-1232
8. R.S.Sutton and A.G.Barto, Reinforcement Learning :An Introduction, A bradford book, London, (1998)
9. C.H.Park and K.B Sim :The Implementation of the Emotion Recognition from Speech and Facial Expression System, Proc. of ICNC'05-FSKD'05, Aug. (2005) 85-88

# Supervised IAFC Neural Network Based on the Fuzzification of Learning Vector Quantization

Yong Soo Kim<sup>1</sup>, Sang Wan Lee<sup>2</sup>, Sukhoon Kang<sup>1</sup>, Yong Sun Baek<sup>3</sup>,  
Suntae Hwang<sup>4</sup>, and Zeungnam Bien<sup>2</sup>

<sup>1</sup>Department of Computer Engineering, Daejeon University,  
Daejeon, 300-716, Korea

kystj@dju.ac.kr, shkang@dju.ac.kr

<sup>2</sup>Department of Electrical Engineering and Computer Science, KAIST,  
Daejeon, 305-701, Korea

bigbean@ctrsys.kaist.ac.kr, zbien@ee.kaist.ac.kr  
<http://ctrgate.kaist.ac.kr/~bigbean>

<sup>3</sup>Department of Computer Web Information, Daeduk College,  
Daejeon, 305-715, Korea

dodo029@ddc.ac.kr

<sup>4</sup>Department of Information and Communications Engineering, Daejeon University,  
Daejeon, 300-716, Korea

hwang@dju.ac.kr

**Abstract.** In this paper, a fuzzy LVQ(Learning Vector Quantization) is proposed which is based on the fuzzification of LVQ. The proposed FLVQ(Fuzzy Learning Vector Quantization) uses the different learning rate depending on the correctness of classification. When the classification is correct, the amount of update is determined by consideration of location of the input vector relative to the decision boundary. When the classification is not correct, the amount of update is determined by the degree of belongingness of the input vector to the winning class. The supervised IAFC(Integrated Adaptive Fuzzy Clustering) neural network 3, which uses FLVQ, is introduced in this paper. The supervised IAFC neural network 3 is both stable and plastic because it uses the control structure which is similar to that of Adaptive Resonance Theory(ART)-1 neural network. We used iris data set to compare the performance of the supervised IAFC neural network 3 with those of LVQ algorithm and backpropagation neural network. The supervised IAFC neural network 3 yielded fewer misclassifications than LVQ algorithm and backpropagation neural network.

## 1 Introduction

A neural network is a network of interconnected neurons. These neurons are interconnected via weights. These weights are adjusted to improve the performance of neural network. Therefore, a learning rule which controls the adjustment of weights plays an important role on the performance of neural network.

LVQ is one of supervised learning rules. LVQ moves the weight of a winner toward the input vector if the classification is correct[1,2]. On the other hand, LVQ

moves the weight of a winner away from the input vector if the classification is incorrect. Chung and Lee proposed FLVQ which incorporates fuzzy membership value with LVQ[3,4]. They derived FLVQ by optimizing an appropriate fuzzy objective function which takes into accounts of two goals. The first goal is minimizing the network output error which is the class membership differences between target and actual membership values. The second goal is minimizing the distances between the input patterns and the prototypes of classes. They solved the underutilization problems of LVQ and got better result than that of LVQ. This FLVQ updates the prototypes of classes regardless of winning or losing. The amount of update depends on the difference between the target membership value and the actual membership value in addition to the difference between the input pattern and the prototype of class. The problem of this FLVQ is that it requires target membership value. But, it is not easy to get target membership value in the real situations. Karayiannis also fuzzified LVQ[5,6]. He derived FLVQ by minimizing the average generalized mean between the input vectors and the prototypes using gradient descent. The prototypes are updated through an unsupervised learning process. All prototypes are updated and the amount of update depends on the difference between the input vector and the prototype, the fuzzy membership value, and a learning rate for each prototype. Because it uses batch learning, each prototype is updated with respect to all input vectors. But, it uses a large amount of memory. Karayiannis proposed weighted FLVQ[5]. It is derived by minimizing the weighted generalized mean of the squared distance between the input vector and the prototypes. It is similar to FLVQ by Karayiannis. Tsao et al. also proposed FLVQ[7]. It is similar to FLVQ by Karayiannis.

This paper proposes a fuzzy LVQ which fuzzified LVQ. The proposed FLVQ uses a function of iterations,  $\Pi$  membership function, and the fuzzy membership value instead of the learning rate of LVQ. The  $\Pi$  membership function reduces the effect of outliers to the prototype of class. LVQ uses the same learning rate regardless of the classification is correct or not. However, the proposed FLVQ uses the different learning rates depending on the correctness of classification. The proposed FLVQ uses the difference between one and the fuzzy membership function if the classification is correct. But it uses the fuzzy membership value when the classification is not correct. When the classification is correct, the weighting factor of the data point, which locates near the decision boundary, for updating amount of the weight is larger than the weighting factor of the data point, which locates far from the decision boundary, for updating amount of the weight. This reduces the effect of outliers to the decision boundary. The outliers deteriorate the decision boundary, because the outliers tend to move away the prototypes of the classes from the proper locations for the decision boundary. The proposed FLVQ prevents the outliers from deteriorating the proper decision boundary, because it uses the difference between one and the fuzzy membership value. The fuzzy membership value of an outlier in the class, where it belongs to, is larger than the fuzzy membership value of the data point, which locates near the decision boundary, in the class where it belong to. The larger the fuzzy membership value, the smaller the difference between one and the fuzzy membership value. Because the FLVQ uses the difference between one and the fuzzy membership value, it considers the data point, which locates near the decision boundary, more important when it updates the prototype of class. When the

classification is not correct, the proposed FLVQ uses the fuzzy membership value to update the prototype of the selected class. The updating amount of the prototype of the selected class is proportional to the amount of belongingness of the misclassified data point in the selected class.

The proposed FLVQ is integrated into the supervised IAFC neural network 3. The supervised IAFC neural network 3 has both the stability and the plasticity as the ART-1 neural network because it uses the control structure which is similar to that of the ART-1 neural network[8](Fig.1). It is stable to preserve significant past learning but plastic to incorporate new input point whenever it might appear. It controls the number of clusters and the size of clusters by the vigilance parameter. In the supervised IAFC neural network 3, the vigilance parameter is related to a distance threshold or cluster diameter[9,10]. Even though the ART-1 neural network processes binary data, the supervised IAFC neural network 3 processes continuous-valued data. The supervised IAFC neural network 3 uses the Euclidean distance to choose the nearest prototype and calculate thresholds[9,10].

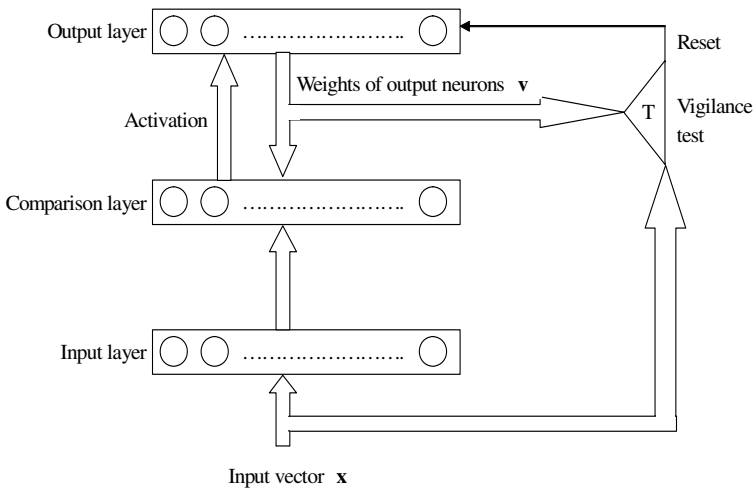


Fig. 1. The structure of the supervised IAFC neural network 3

We compared the performance of the supervised IAFC neural network 3 with those of LVQ algorithm and backpropagation neural network using the iris data set, which is a benchmark data set for comparing the performance of clustering algorithms.

## 2 Supervised IAFC Neural Network 3

After the input vector is applied to the supervised IAFC neural network 3, competition among output neurons occurs in a winner-take-all fashion. The output neuron, of which weight has the minimum Euclidean distance to the input vector, wins the competition. The  $I$ -th output neuron,



$$I = \min_i \|\mathbf{x} - \mathbf{v}_i(t)\|, \tag{1}$$

where  $\mathbf{x}$  is the input vector and  $\mathbf{v}_i(t)$  is the weight of the  $i$ -th output neuron, wins the competition.

After selecting a winning output neuron, the supervised IAFC neural network 3 performs the vigilance test according to the following vigilance criterion :

$$e^{-\mu_i} \|\mathbf{x} - \mathbf{v}_i(t)\| \leq T, \tag{2}$$

where  $T$  is the vigilance parameter. The fuzzy membership value  $\mu_i$  is defined as follows :

$$\mu_i = \frac{\left[ \frac{1}{\|\mathbf{x} - \mathbf{v}_i(t)\|^2} \right]^{\frac{1}{m-1}}}{\sum_{j=1}^n \left[ \frac{1}{\|\mathbf{x} - \mathbf{v}_j(t)\|^2} \right]^{\frac{1}{m-1}}} \tag{3}$$

where  $n$  is the number of committed output neurons, and  $m \in [1, \infty]$  is a weight exponent.  $m$  is experimentally set to 2. However, when the number of committed output neurons is one, the vigilance criterion is  $\|\mathbf{x} - \mathbf{v}_1(t)\| \leq T$ .

The dissimilarity measure in Eq. (2) is the relative distance which considers both the Euclidean distance and the relative location of the input vector to the prototypes of the existing classes[9]. The weighting factor for the input vector, which locates far from the decision boundary, is smaller than the weighting factor for the input vector, which locates near the decision boundary. This weighting factor is multiplied by the Euclidean distance between the input vector and the prototype of the winning class. We can compare this relative distance with Mahalanobis distance which considers statistical properties of data[11]. In the case of the Mahalanobis distance, weighting factor is large when the covariance is small. On the other hand, weighting factor is small when the covariance is large.

If the winning output neuron satisfies the vigilance test, the supervised IAFC neural network 3 updates the weight of the winning output neuron as follows:

$$\mathbf{v}_i(t+1) = \mathbf{v}_i(t) + f(t) \cdot \pi[\mathbf{x}, \mathbf{v}_i(t), T] \cdot (1 - \mu_i) \cdot [\mathbf{x} - \mathbf{v}_i(t)] \tag{4}$$

if  $\mathbf{x}$  is classified correctly,

$$\mathbf{v}_i(t+1) = \mathbf{v}_i(t) - f(t) \cdot \pi[\mathbf{x}, \mathbf{v}_i(t), T] \cdot \mu_i \cdot [\mathbf{x} - \mathbf{v}_i(t)] \tag{5}$$

if  $\mathbf{x}$  is classified incorrectly,

$$\mathbf{v}_i(t+1) = \mathbf{v}_i(t) \quad \text{for } i \neq I, \tag{6}$$

where  $f(t)$  is the function of iterations.  $f(t)$  is defined as  $\frac{1}{k(t-1)+1}$ , where  $k$  is the constant which controls convergent speed.  $\pi(\mathbf{x}, \mathbf{v}_i(t), T)$  is defined as

$$\pi(\mathbf{x}, \mathbf{v}_i(t), T) = \begin{cases} 1 - 2 \left( \frac{\|x - \mathbf{v}_i(t)\|}{T} \right)^2, & \text{when } 0 \leq \|x - \mathbf{v}_i(t)\| \leq \frac{T}{2} \\ 2 \left( 1 - \frac{\|x - \mathbf{v}_i(t)\|}{T} \right)^2, & \text{when } \frac{T}{2} \leq \|x - \mathbf{v}_i(t)\| \leq T \\ 0, & \text{when } \|x - \mathbf{v}_i(t)\| \geq T. \end{cases} \tag{7}$$

When the classification is correct, the proposed FLVQ moves the weight of the winning class toward the input vector. In Eq. (4), the proposed FLVQ considers the location of the input vector for updating the weight of the winning output neuron using  $1 - \mu_i$ . In Fig. 2, the fuzzy membership value of the input vector B in the class 1 is larger than that of the input vector A in the class 1. The input vector near the decision boundary has more information about the proper decision boundary. The input vector, which locates far from the decision boundary like the input vector B, moves the decision boundary away from the proper position. It deteriorates the decision boundary. Using  $1 - \mu_i$  can prevent the input vector, which locates far from the decision boundary like the input vector B, from deteriorating the decision boundary. By using  $1 - \mu_i$  in Eq. (4), the weighting factor for the input vector A is larger than that for the input vector B. On the other hand, when the classification is not correct, the proposed FLVQ moves the weight of the winning class away from the input vector. The proposed FLVQ uses the fuzzy membership value to update the weight of the winning class as in Eq. (5). The updating amount of the weight of the winning output neuron is proportional to the fuzzy membership value.

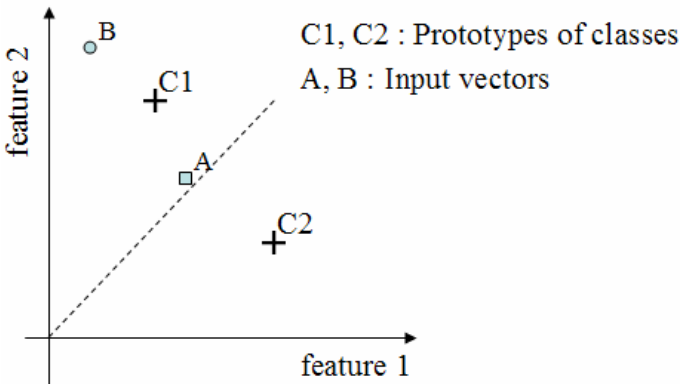
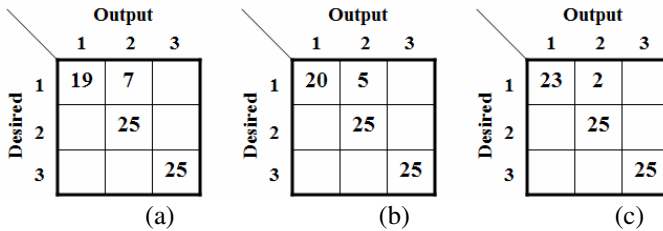


Fig. 2. Consideration of location of the input vector with respect to the decision boundary

### 3 Test and Result

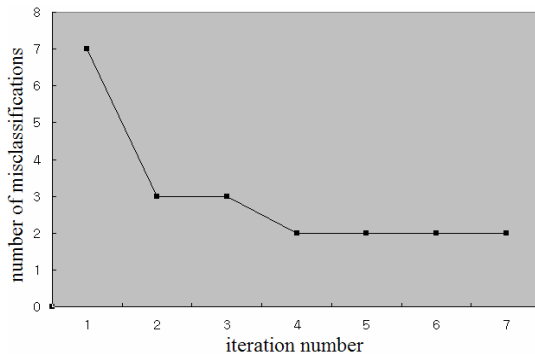
We used iris data set, which is a benchmark data set for comparing the performance of clustering algorithms, to compare the performance of the supervised IAFC 3 with those of LVQ algorithm and backpropagation neural network. Iris data set consists of 150 four-dimensional data, and 3 subspecies[12]. Each subspecies has 50 data. We chose 75 data arbitrarily from 150 data, and used as a training data set. We chose 25 data from each subspecies. And we used the other 75 data as a testing data set.

During the training, if  $\|v(t) - v(t-1)\|$  is less than 0.01, we considered the weights to be converged experimentally and stopped the iterative operations of training for the supervised IAFC neural network 3. After that, we tested the supervised IAFC neural network 3. Fig. 3 shows the comparison between the result of the supervised IAFC neural network 3 and the results of LVQ algorithm and backpropagation neural network. The supervised IAFC neural network 3 iterated 7 iterations to train, and yielded 2 misclassifications when  $T$  is 1.7 and  $K$  is 0.5. We tested LVQ algorithm using MATLAB Toolbox. We tested LVQ algorithm under the condition that the learning rate is 0.01. LVQ algorithm yielded 7 misclassifications. And backpropagation neural network yielded 5 misclassifications. Fig. 4 shows the number of misclassifications versus iteration number when we trained the supervised IAFC neural network 3.



(a) LVQ algorithm (b) Backpropagation neural network (c) Supervised IAFC neural network 3

**Fig. 3.** Comparison of results using iris data set



**Fig. 4.** The number of misclassifications versus the iteration number when the supervised IAFC neural network 3 was trained

## 4 Conclusion

We proposed FLVQ which is based on the fuzzification of LVQ. The proposed FLVQ uses the different learning rate depending on the correctness of classification. The proposed FLVQ considers the location of input vector relative to the decision boundary. This prevents the outlier from deteriorating the decision boundary.

We used iris data set to compare the performance of the supervised IAFC neural network 3 with those of LVQ algorithm and backpropagation neural network. The supervised IAFC neural network yielded fewer numbers of misclassifications than LVQ algorithm and backpropagation neural network. It required a few iterations to converge experimentally.

## References

1. C-T Lin and C. S. G. Lee: Neural Fuzzy Systems – A Neuro-Fuzzy Synergism to Intelligent Systems. New Jersey : Prentice-Hall. (1996)
2. J.C. Bezdek, E. C. Tsao, and N.R. Pal.: Fuzzy Kohonen Clustering Networks. Proceeding of the First IEEE conference on Fuzzy Systems, San Diego. pp.1035-1043. (1992)
3. F. -L. Chung and T. Lee: A fuzzy Learning Model for Membership Function Estimation and Pattern Classification. Proceedings of the third IEEE Conference on Fuzzy Systems. vol. 1. pp. 426-431. (1994)
4. F. -L. Chung and T. Lee: Fuzzy Learning Vector Quantization, Proceedings of 1993 International Joint Conference on Neural Networks, Nagoya, Vol. 3, pp. 2739-2743. (1993)
5. N. B. Karayiannis: Weighted Fuzzy Learning Vector Quantization and Weighted Fuzzy C-Means Algorithms, IEEE International Conference on Neural Networks, Vol. 2, pp. 1044-1049. (1996)
6. N. B. Karayiannis, and Bezdek, J.C.: An Integrated Approach to Fuzzy Learning Vector Quantization and Fuzzy C-Means Clustering, IEEE Transactions on Fuzzy Systems, Vol. 5, pp. 662-629. (1997)
7. E. C. -K. Tsao, J. C. Bezdek, and N. R. Pal: Fuzzy Kohonen Clustering Networks, Pattern Recognition, Vol 27., No. 5, pp. 757-764. (1994)
8. G. A. Carpenter and S. Grossberg: A Massively Parallel Architecture for A Self-Organizing Neural Pattern Recognition Machine. Computer Vision, Graphics, and Image Processing. vol. 37. pp. 54-115. (1987)
9. Y. S. Kim and S. Mitra: An adaptive integrated fuzzy clustering model for pattern recognition. Fuzzy Sets and Systems. vol. 65. pp. 297-310. (1994)
10. B. Moore: ART1 and Pattern Clustering. Proceedings of the 1988 Connectionist Models Summer School. San Mateo. pp.174-185. (1989)
11. J. T. Tou and R. C. Gonzalez: Pattern Recognition Principles. Massachusetts: Addison-Wesley. (1974)
12. E. Anderson : The IRISes of the Gaspe Penninsula. Bulletin American IRIS Society. Vol. 59. pp. 2-5. (1935)

# Walking Pattern Analysis of Humanoid Robot Using Support Vector Regression

Dongwon Kim and Gwi-Tae Park\*

Department of Electrical Engineering, Korea University, 1, 5-ka, Anam-dong, Seongbuk-ku,  
Seoul 136-701, Korea  
{upground, gtpark}@korea.ac.kr

**Abstract.** This work presents walking pattern analysis of a humanoid robot using support vector regression. The humanoid robot is highly suitable to work in human environments but the dynamics involved are highly nonlinear and unstable. So we are establishing empirical relationships based on the walking pattern analysis as dynamic stability of motion. Zero moment point is usually used as a basic component for dynamically stable motion. Kernel method and support vector machines (SVM) have become very popular as methods for learning from examples. We apply SVM to analyze humanoid robot walking. The experimental results show that the SVM based on the kernel substitution provides a promising alternative to model robot movements but also to control actual humanoid robots.

## 1 Introduction

Biped locomotion is a popular research area in robotics due to the high adaptability of a walking robot in an unstructured environment. When attempting to automate the motion planning process for a biped walking robot, one of the main issues is assurance of dynamic stability of motion. This can be categorized into three general groups [1,2]: body stability, body path stability, and gait stability. A zero moment point (ZMP), a point where the total forces and moments acting on the robot are zero, is usually employed as a basic component for dynamically stable motion. These days, studies on the ZMP modeling for feasible walking motion of humanoid robot are increased. In [3,4] fuzzy system and neuro-fuzzy systems have been developed and applied to model ZMP trajectory of a biped walking robot. Constructed models have been used to qualitatively examine walking behaviors. However, there still exist other soft computing techniques not yet evaluated. Investigating their applicability to humanoid robot modeling is highly demanded since some of them may exhibit better ability than previous methods, thereby providing more improved insight into physical walking mechanisms.

In this study, a support vector regression (SVR) [5,6,10] is first applied to model a biped walking robot. Support vector regression or machines and kernel methods (KMs) have become in the last few years one of the most popular approaches to learning from examples with many potential applications in science and engineering. The

---

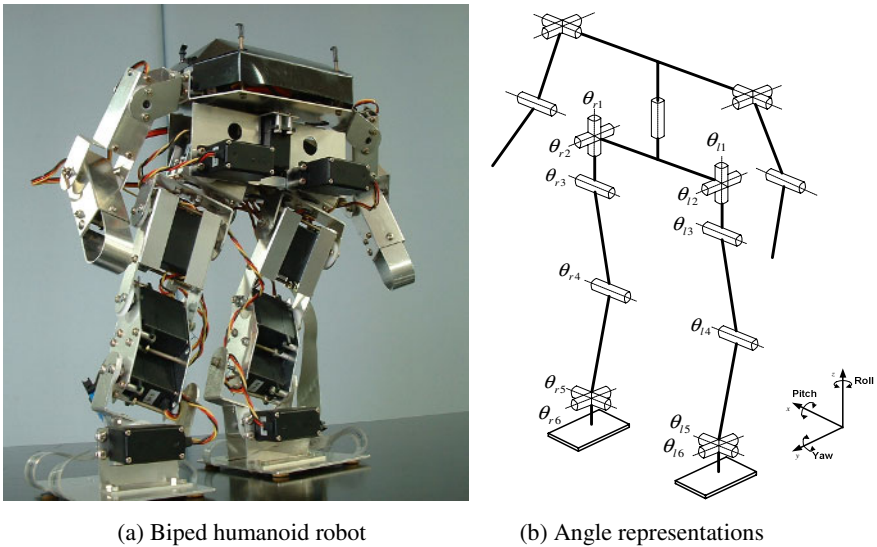
\* Corresponding author.

basic theory is well understood and applications work successfully in practice. In many applications, SVM has been shown to provide higher performance than traditional learning machines [7] and has been introduced as powerful tools for solving classification problems.

The rest of the paper is organized as follows. Actual ZMP trajectory of humanoid robot throughout the whole walking phase will be described in Section 2. Application research covering the use of SVM in ZMP trajectory is discussed in Section 3. Experimental results are given in Section 4. Section 5 presents concluding remarks.

## 2 ZMP Trajectories of Humanoid Robot

In practice, we have design and implement a biped humanoid robot as shown in Fig. 1(a). The robot has 19 joints and the locations of the joints during motion are shown in Fig. 1(b).

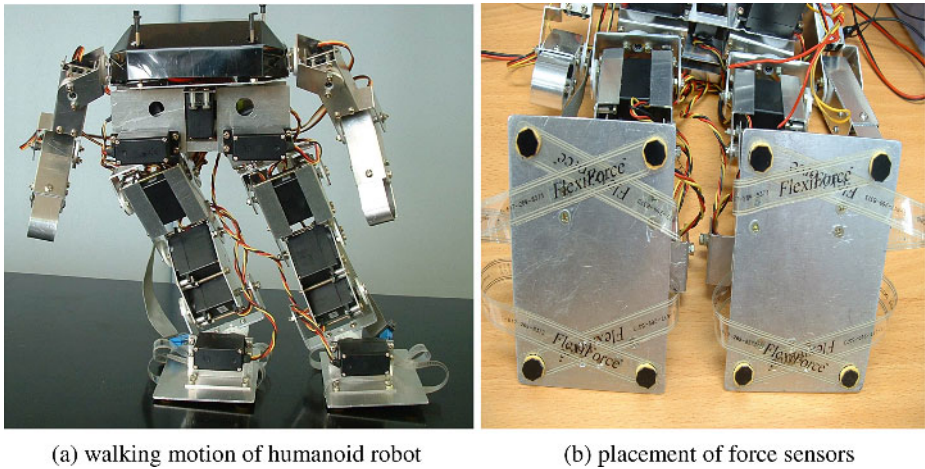


(a) Biped humanoid robot

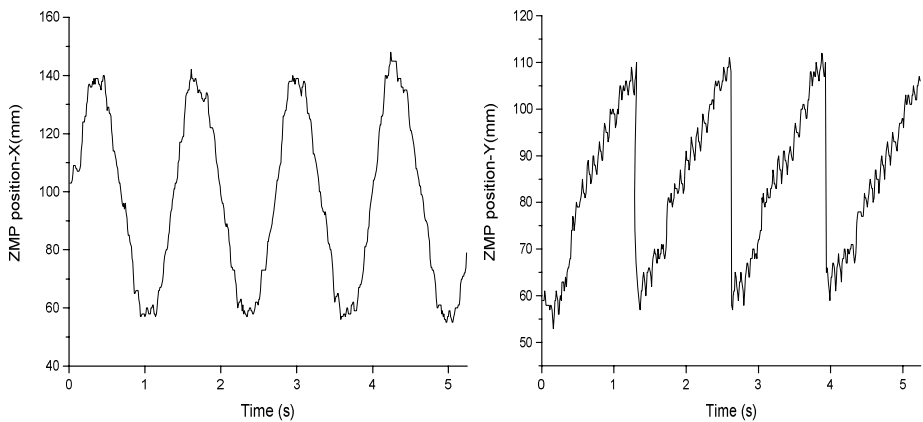
(b) Angle representations

**Fig. 1.** Actual humanoid robot and its joint angle representation

The height and the total weight are about 308mm and 1700g including batteries. Each joint is driven by the RC servomotor that consists of a DC motor, gear, and simple controller. Each of the RC servomotors is mounted in the link structure. Our biped walking robot is able to walk under the condition which one step is 48 mm per 1.4 s on the flat floor. More detailed information about specification and block diagram of the robotic system are in [3,4]. The walking motion of the robot is shown in Fig. 2 (a) which shows front view of the robot when the robot is on a flat surface. The measured ZMP trajectory is obtained from 10 degree of freedom (DOF) data. Two DOF are assigned to the hips and ankles, and one DOF to each knee. Using these joint angles, a cyclic walking pattern has been realized.



**Fig. 2.** Front view of the humanoid robot and force sensors at the four corners



**Fig. 3.** ZMP positions of our humanoid robot

Regarding to the ZMP measurement system, we employed a direct approach as reported in [3,4] which is to use data measured using sensors mounted on the robot's feet. Fig. 2 (b) depicts the used sensors and their placement on the sole of the robot's feet. The type of force sensor used in our experiments is FlexiForce A201 sensor [8]. They are attached to the four corners of the sole plate. Sensor signals are digitized by an ADC board, with a sampling time of 10 ms. Measurements are carried out in real time. The foot pressure is obtained by summing the force signals. Using the sensor data it is easy to calculate the actual ZMP values. The ZMPs in the local foot coordinate frame are computed using (1), where  $f_i$  represents a force applied to the right and left foot sensors and  $r_i$  is a sensor position.

$$P = \frac{\sum_{i=1}^8 f_i r_i}{\sum_{i=1}^8 f_i} \tag{1}$$

Experimental results are shown in Figs. 3-4 which show the actual ZMP positions,  $x$ -coordinate and  $y$ -coordinate, of the four-step motion of the biped walking robot and its corresponding ZMP trajectories.

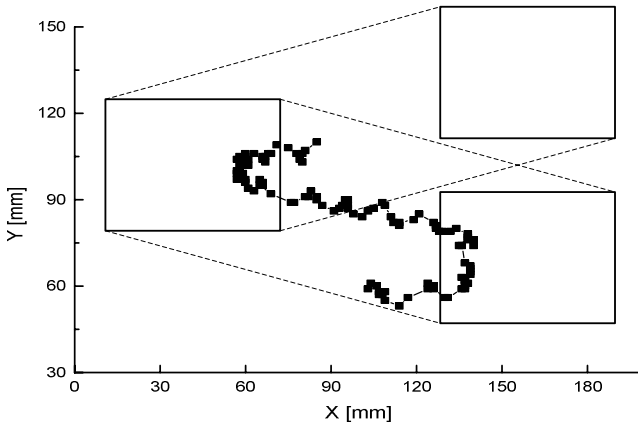


Fig. 4. ZMP trajectories of the humanoid robot corresponding to Fig. 3

### 3 Support Vector Regression

The SVMs can be applied to regression problems by the introduction of an alternative loss function [9]. The basis idea in SVR is to map the input data  $x$  into a higher dimensional feature space via a nonlinear mapping  $\Phi$  and then a linear regression problem is obtained and solved in this feature space. The following presents some basic concepts of SVR as described by prior research. A detailed explanation may be found in [5]. In SVM method, the regression function is approximated by the following function:

$$y = \sum_{i=1}^l w_i \Phi_i(x) + b \tag{2}$$

where  $\{\Phi_i(x)\}_{i=1}^l$  are the feature of inputs,  $\{w_i\}_{i=1}^l$  and  $b$  are coefficients. The coefficients are estimated by minimizing the regularized risk function.

$$R(C) = C \frac{1}{l} \sum_{i=1}^l L_\epsilon(d_i, y_i) + \frac{1}{2} \|w\|^2 \tag{3}$$



where

$$L_\varepsilon(d_i, y_i) = \begin{cases} 0 & \text{for } |d - y| < \varepsilon, \\ |d - y| - \varepsilon & \text{otherwise} \end{cases} \quad (4)$$

and  $\varepsilon$  is a prescribed parameter.

In Eq. (3),  $L_\varepsilon(d_i, y_i)$  is  $\varepsilon$ -insensitive loss function, which indicates that it does not penalize errors below  $\varepsilon$ .  $\frac{1}{2}\|w\|^2$  is used as a flatness measurement of Eq. (2) and  $C$  is a regularized constant determining the tradeoff between the training error and the model flatness. Introduction of slack variables  $\zeta, \zeta^*$  leads Eq. (3) to the following constrained function

$$\text{Minimize } R(w, \zeta^*) = \frac{1}{2}\|w\|^2 + C^* \sum_{i=1}^l (\zeta_i + \zeta_i^*) \quad (5)$$

$$\begin{aligned} \text{s.t. } w\Phi(x_i) + b - d_i &\leq \varepsilon + \zeta_i, \\ d_i - w\Phi(x_i) - b &\leq \varepsilon + \zeta_i^*, \quad \zeta_i, \zeta_i^* \geq 0. \end{aligned} \quad (6)$$

Thus, function (2) becomes the explicit form

$$\begin{aligned} f(x, \alpha_i, \alpha_i^*) &= \sum_{i=1}^l w_i \Phi_i(x) + b = \sum_{i=1}^l (\alpha_i - \alpha_i^*) \Phi(x_i)^T \Phi(x) + b \\ &= \sum_{i=1}^l (\alpha_i - \alpha_i^*) K(x_i, x) + b \end{aligned} \quad (7)$$

In formula (7), Lagrange multipliers  $\alpha_i$  and  $\alpha_i^*$  satisfy the constraints  $\alpha_i^* \alpha_i^* = 0, \alpha_i \geq 0, \alpha_i^* \geq 0$  and they can be obtained by maximizing the dual form of function (5)

$$\begin{aligned} \Phi(\alpha_i, \alpha_i^*) &= \sum_{i=1}^l d_i (\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) \\ &\quad - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) K(\alpha_i, \alpha_j) \end{aligned} \quad (8)$$

with constraints

$$\sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0, \quad 0 \leq \alpha_i \leq C, \quad 0 \leq \alpha_i^* \leq C \quad (9)$$

Based on the nature of quadratic programming, only a number of coefficients among  $\alpha_i$  and  $\alpha_i^*$  will be nonzero, and the data points associated with them refer to support vectors. The form  $\Phi(x_i)^T \Phi(x)$  in Eq. (7) is replaced by kernel function with the form

$$K(x, y) = \Phi(x)^T \Phi(y) \quad (10)$$

There are some different kernels for generating the inner products to construct machines with different types of nonlinear decision surfaces in the input space. We employed three kind of kernel functions as follows

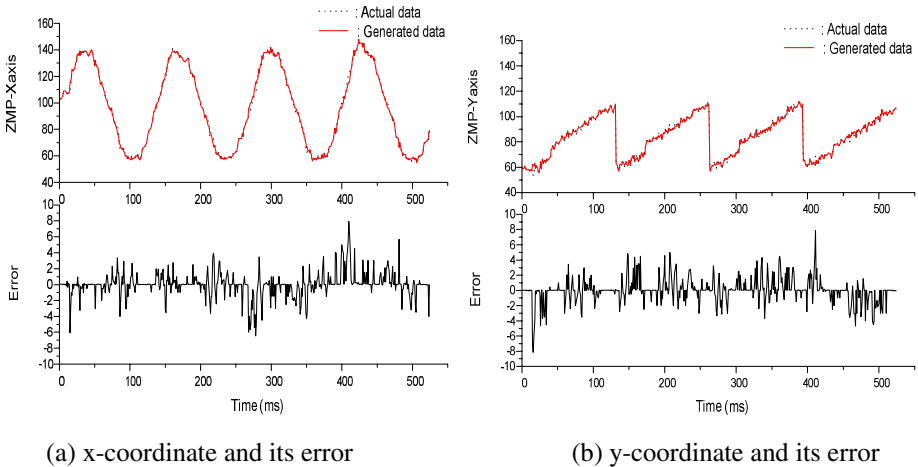
$$\begin{aligned}
 \text{linear} : & \quad K(x, y) = x^T y \\
 \text{polynomial} : & \quad K(x, y) = (xy + 1)^d \\
 \text{RBF} & \quad : \quad K(x, y) = \exp(-\frac{1}{\sigma^2} \|x - y\|^2)
 \end{aligned}
 \tag{11}$$

### 4 Experimental Results

Using the three types of kernel functions such as linear, polynomial, and radial basis function for SVR, approximated models are constructed and their results are compared. The accuracy was quantified in terms of mean squared error (MSE) values. The SVR was applied to model the ZMP trajectory of the humanoid robot depicted in previous section using measured data. The measured data are employed as the process parameters. In Table 1, MSE values corresponding three types of kernel functions are listed and we can compare them with respect to various kernel functions.

**Table 1.** Kernel functions and corresponding accuracy of humanoid robot

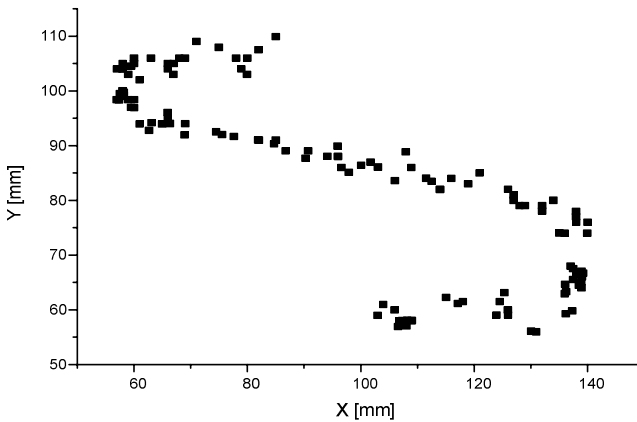
kernel type	x-coordinate	y-coordinate
linear	47.184	60.704
polynomial	9.695	18.568
RBF	<b>2.524</b>	<b>2.721</b>



**Fig. 5.** Generated ZMP xy-coordinate and error

One of the advantages of linear kernel is that there is no parameter to tune except the constant  $C$ . For the nonlinear case there is an additional parameter, the kernel parameter, to tune. First, this study uses three kernel functions such as linear function, polynomial function, and radial basis function. As constant  $C$ , we set the value as 1000. Moreover, the degree of polynomial and width of RBF are set to 2.

From the Table1, the polynomial kernel provides worse results than the RBF kernel. In addition, it takes a longer time in the procedure. The generated ZMP positions from the RBF kernel, and its errors between actual data and generated data are shown in Fig 5. In Fig. 6, we can see the corresponding ZMP trajectories that are generated from the RBF kernel. From the figure, the generated ZMP is very similar to actual ZMP trajectory of the biped humanoid robot.



**Fig. 6.** Generated ZMP trajectories of the humanoid robot corresponding to Fig. 5

## 5 Conclusions

Support vector regression modeling at the ZMP trajectory of a practical humanoid robot has been presented. The trajectory of the ZMP poses an important criterion as dynamic stability of motion, but the complex dynamics involved make the robot control difficult. To establish empirical relationships between walking robot and ground, we employed the support vector machines with respect to various kernel functions to walking robot. As a result, SVM based on kernel method have been found to work well. Especially SVM with RBF kernel function provides the best results.

## Acknowledgment

The authors would like to thank the financial support of the Korea Science & Engineering Foundation. This work was supported by grant No. R01-2005-000-11044-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

## References

1. Yagi, M., Lumelsky, V.: Biped Robot Locomotion in Scenes with Unknown Obstacles. Proc. IEEE Intl. Conf. Robotic & Autom. (1999) 375-380.
2. Vulobratovic, M., Frank, A.A., Juricic, D.: On the Stability of Biped Locomotion. IEEE Trans. Biomed. Eng. **17** (1970) 25-36.
3. Kim, D., Kim, N.H., Seo, S.J., Park, G.T.: Fuzzy Modeling of Zero Moment Point Trajectory for a Biped Walking Robot. Lect. Notes Artif. Int. **3214** (2004) 716-722.
4. Kim, D., Seo, S.J., Park, G.T.: Zero-moment point trajectory modeling of a biped walking robot using an adaptive neuro-fuzzy system. IEE Proc.-Control Theory Appl. **152** (2005) 411-426.
5. Wang, W., Xu, Z.: A heuristic training for support vector regression. Neurocomputing. **61** (2004) 259-275.
6. Lin, C.F., Wang, S.D.: Fuzzy Support Vector Machines. IEEE Trans. Neural Networ. **13** (2002) 464-471
7. Burges, C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2** (1998)
8. FlexiForce A201 Sensor Model: <http://www.tekscan.com/flexiforce/flexiforce.html>
9. Vapnik, V.: The Nature of Statistical Learning Theory. New York, John Wiley (1995).
10. Gunn, S.: Support vector machines for classification and regression. ISIS technical report, Image Speech & Intelligent Systems Group University of Southampton (1998).

# Low-Cost Stabilized Platform for Airborne Sensor Positioning

F.J. González-Castaño, F. Gil-Castiñeira, J.M. Pousada-Carballo,  
P.S. Rodríguez-Hernández, J.C. Burguillo-Rial, and I. Dosil-Outes

Departamento de Ingeniería Telemática, Universidad de Vigo, Spain  
javier@det.uvigo.es

**Abstract.** We present a low-cost design of a stabilized platform to position sensors in aerial vehicles. The stabilization system has been designed to provide enough precision for vehicles flying at an altitude of <2000 m and a speed of <400 km/h. The design comprises a 3-axis accelerometer/gyroscope kit, an embedded controller (executing a Kalman filter and a postprocessing smoothing algorithm) and a kit of 3-axis orthogonally mounted high-torque servos.

**Keywords:** Stabilization, servo, gyroscope, accelerometer, sensor.

## 1 Introduction

The work in this paper is motivated by the participation of the authors in Spanish MCyT project VEM2003-20088-C04-02, to develop a hyperspectral system for oil spill detection. It comprises an optics block, a processor and set of location sensors. The optical components are installed in an electronically stabilized platform, as described in this paper.

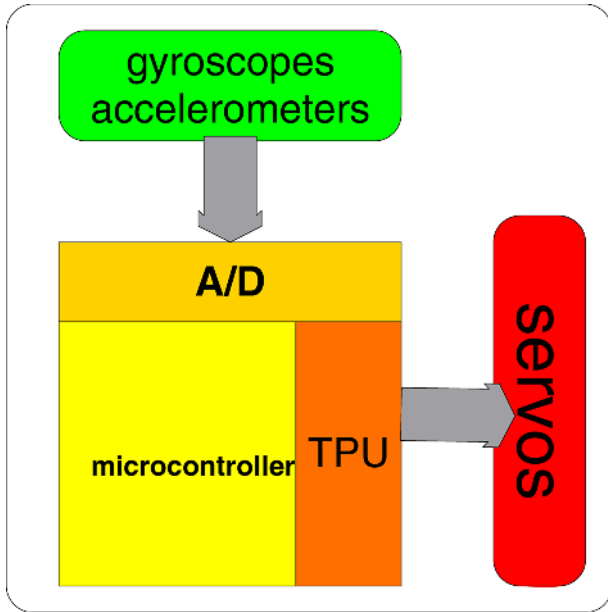
A review of the many recent hardware and software positioning systems is beyond the length of this paper. However, it is evident that airborne stabilization systems are nowadays of paramount importance [6]. We refer the reader to [8] for the theory behind. Previous authors [7] have exploited movement measurement units for geolocation purposes. In this research, we consider that such units (accelerometers and gyroscopes that feed the stabilization platform) are indeed useful sensors, to take into account in the spill detection process. In other words, we propose to fusion geolocation data with hyperspectral images to improve detection.

## 2 System Details

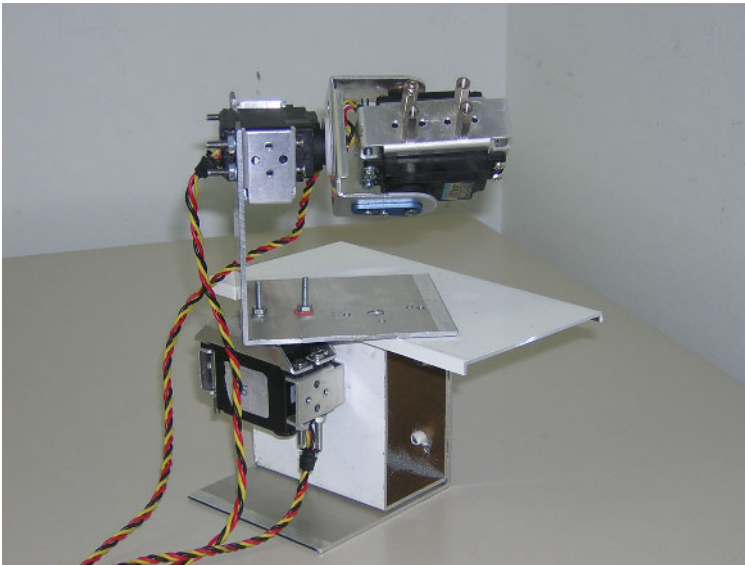
The stabilization system aims an airborne hyperspectral camera. The movement sensors consist of a 3-axis kit of accelerometers and gyroscopes, attached to the vehicle. An embedded microcontroller board reads their output and actuates on the platform: a 3-axis set of orthogonally mounted servos. Figure 1 shows a general view of the system.

### *Hardware*

The platform is based on three advanced HSR-5995TG Hitec coreless amateur servos (figure 2) [2]. They have a considerably high performance: a torque of 30 kg×cm and a speed of 60°/0.12 seconds at 7.4V.



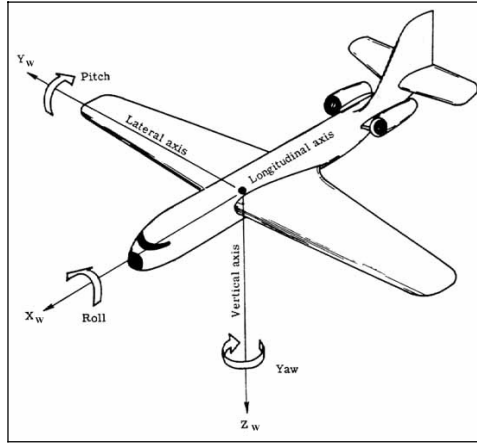
**Fig. 1.** General view of the system



**Fig. 2.** Stabilization platform

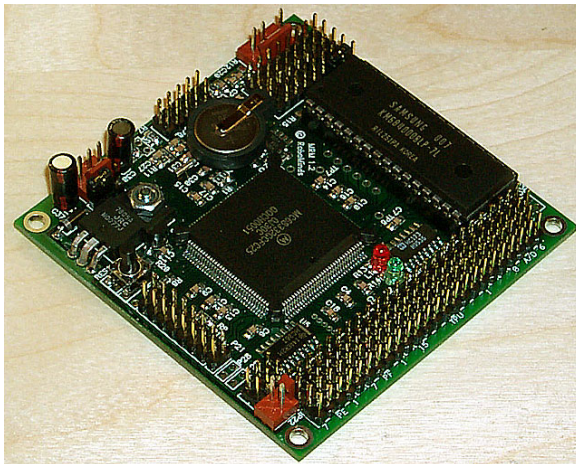
The gyroscopes and accelerometers are miniature angular-rate sensors of very simple construction, made of single piezoelectric columns printed with electrodes [3]. They track aircraft movements as shown in figure 3.

These devices are intended for applications such as shake detector for hand-held cameras, computer mouse control and RC helicopters. There exist units with a higher precision [4], but they are much more expensive.



**Fig. 3.** Aircraft movements to be tracked

The processor that controls the servos (from movement sensor data) is a Motorola 68332 operating at 25 MHz. It is housed in a board with 32K $\times$ 8 RAM and 512K $\times$ 8 Flash memory (figure 4) [1]. This board has a 7-channel 8-bit 1MPS+ A/D converter, which we have found really useful to acquire the output of the accelerometers and the gyroscopes. Other interesting features are the two servo ports (isolated power) and the 16-channel TPU. With them we have implemented a standard hobby servo PWM output port.



**Fig. 4.** Microcontroller board (microcontroller at the centre, ports on the right side, flash memory at the top)

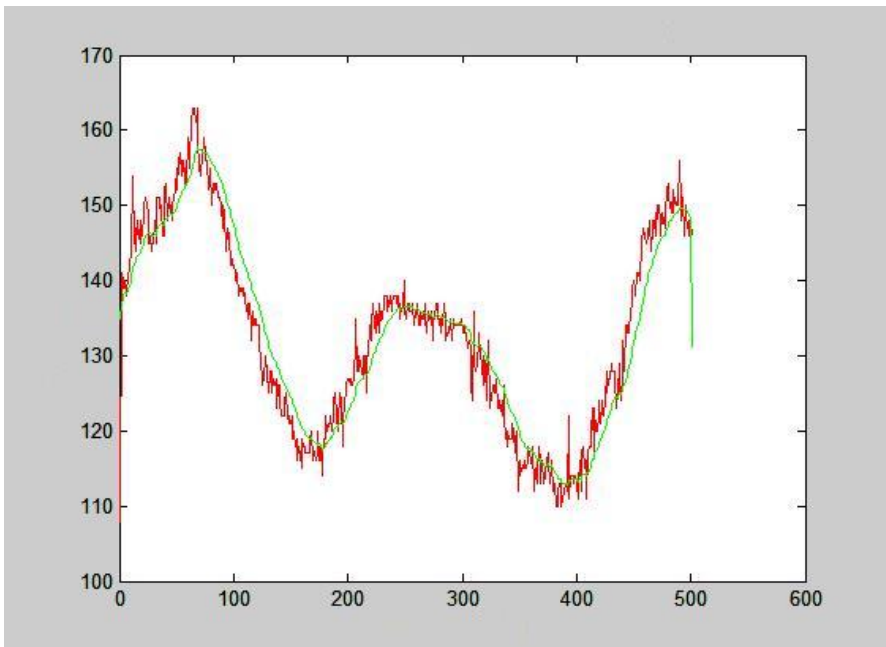
### Software

The main challenges we had to overcome were real-time processing and gyroscopes/accelerometers noise filtering. For the second goal we implemented a Kalman filter [5]. This kind of filter is widely used in robotics, aerospace, underwater vehicles, automotive industry, virtual reality and others.

After applying the Kalman filter to the output of the accelerometers, we run a simple postprocessing algorithm to further reduce noise in the discrete-time domain. Figure 5 shows the combined effect of the Kalman filter and the postprocessor on one of the axis. The noisy output of the movement sensors is stored, to be fused with hyperspectral images prior to the detection stage. Let  $x[n]$  be the output of the Kalman filter. For each axis, the postprocessing algorithm proceeds as follows:

- a) IF  $x[n]-x[n-1] < \alpha$  THEN  $n \leftarrow n+1$ ; GOTO a)
- b)  $y[n]=(x[n]+x[n-1])/2$
- c) IF  $y[n]-x[n-1] < \alpha$  THEN update Kalman filter;  $x[n]=y[n]$ ;  $n \leftarrow n+1$ ; GOTO a)
- d) IF the accelerometers register activity GOTO f)
- e) IF  $y[n] < \beta$  THEN  $n \leftarrow n+1$ ; GOTO a)
- f) Activate servo to correct  $y[n]$ ;  $x[n]=y[n]$ ;  $n \leftarrow n+1$ ; GOTO a)

The microcontroller actuates on the stabilization platform by tracking the resulting smoothed signals (one per axis).



**Fig. 5.** Accelerometers/gyroscopes output (noisy signal) and joint effect of the Kalman filter and the postprocessing stage (smooth signal)



### 3 Performance

Currently, the system is able to command all three servos in the platform each 0.1 s, with a resolution of  $0.2^\circ$ . In figure 6 we see that the highest possible operating altitude is desirable for a given aircraft speed (less angular displacement for the same response time). Obviously, the aircraft must fly at some optimum altitude range for the hyperspectral camera to take useful images.

For example, if the aircraft flies at 400 Km/h, then  $d = 22.22$  m each 0.2 s. At an altitude of 400 m, to track a given point on earth surface there must be a fixed correction of  $3.2^\circ$  along the flight path. Since the servos may sweep  $50^\circ$  each 0.1 s, leaving a processing margin of 0.1 s it is possible to compensate a pitch of  $46.8^\circ$ , which should be enough in steady flights. Roll and yaw are easier to handle, as far as the system does not take images during aircraft turns.

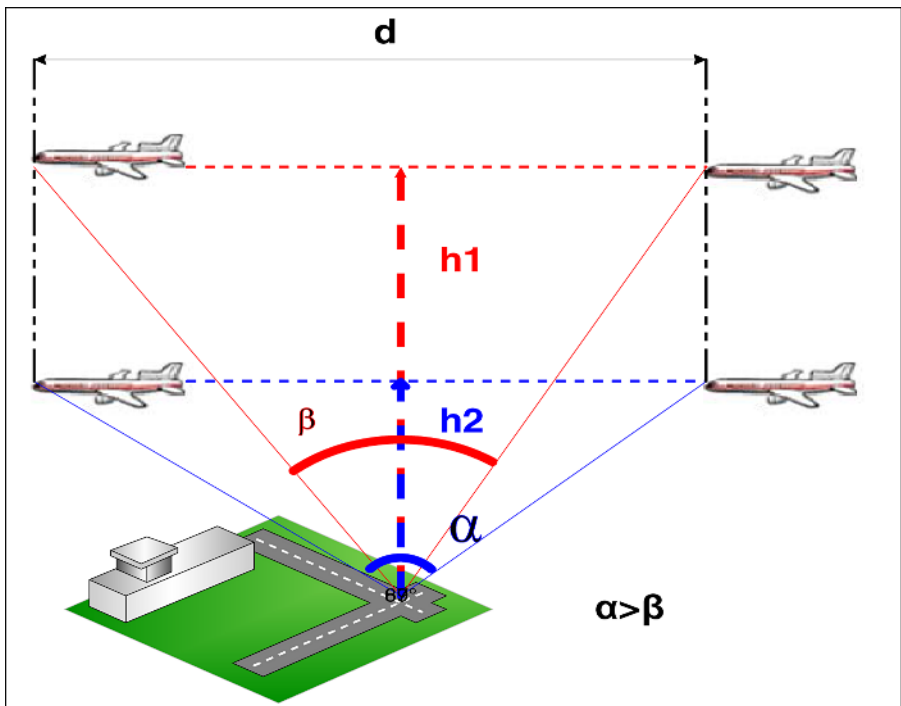


Fig. 6. Operating the system in flight

### 4 Conclusions

In this paper, we gave described a low cost stabilization platform for airborne sensor positioning. We have described its elements and a simple algorithm to smooth movement sensor data. We plan to enrich the input of the oil spill detector with unprocessed accelerometers/gyroscopes signals to improve its performance.

## Acknowledgments

This work has been supported by Spanish MCyT grant VEM2003-20088-C04-02.

## References

1. Robominds: <http://www.robominds.com>.
2. Hitec: <http://www.hitecrd.com>.
3. Tokin: <http://www.tokin.com>
4. MTI: Miniature Attitude and Heading Reference System. Xsens Technologies: <http://www.xsens.com>
5. The Kalman Filter: <http://www.cs.unc.edu/~welch/kalman/>
6. K. J. Held, B. H. Robinson, "Tier II Plus Airborne EO Sensor LOS Control and Image Geolocation", Hughes Aircraft Company, 1997.
7. M. Bäumker, F.J. Heimes, "New Calibration and Computing Method for Direct Georeferencing of Image and Scanner Data Using the Position and Angular Data of a Hybrid Inertial Navigation System", <http://www.fh-bochum.de/fb5/baeumker/baheimesoepe.pdf>.
8. M. Bäumker, F.-J. Heimes, H. Hahn, W. Klier, R. Brechtken, T. Richter. "Mathematical modelling, computer simulation, control and applications of a stabilized platform of an airborne sensor", International Archives of Photogrammetry and Remote Sensing. Vol XXXIII, Part B2, 2000.

# Comparison of Squall Line Positioning Methods Using Radar Data<sup>\*</sup>

Ka Yan Wong and Chi Lap Yip

Dept. of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong.  
{kywong, clyip}@cs.hku.hk

**Abstract.** Squall lines are strong indicators of potential severe weather. Yet, automated positioning and tracking algorithms are not common. We propose three different ways to model and identify squall lines using radar images. The three methods are ellipse fitting, Hough transform, and the use of a genetic algorithm-based framework. They model a squall line as an ellipse, a straight line, and adjoining segments of arc respectively. We compare the advantages and limitations of each method in terms of speed, flexibility, stability and sensitivity to parameter settings. It is found that ellipse fitting is the most efficient, followed by Hough transform. Both methods lack flexibility and stability. The genetic algorithm-based framework is stable, has flexibility in modelling and analysis, but comes with a cost of efficiency. The proposed methods provide independent and objective information sources to assist weather forecast.

## 1 Introduction

Squall lines indicate areas of extreme instability and severe turbulence [1]. They are narrow bands or lines of squally thunderstorms that bring heavy rain, strong winds and lightning. They are strong indicators of potential severe weather, which can cause injuries, loss of lives and property damage.

Weather centers are equipped with weather radars which provide reliable platforms for monitoring the development and movement of rainstorms and squall lines. However, automatic squall line positioning and tracking algorithms are few and far between. Forecasters have to keep a close watch on the radar monitors during the event. Design of automated algorithms for locating, tracking and forecasting of squall lines is thus important. Yet, computer scientists and meteorologists often focus on the problem of automatic rainstorm detection and tracking. Literature on automatic squall lines identification are not common. It is only addressed by [2] in which a Machine Intelligent Gust Front Algorithm is designed to detect convergence lines in radar images. The algorithm utilizes a set of independent feature detectors, which look for a variety of radar signatures such as reflectivity thin lines to produce “interest images”. Convergence lines are then detected by combining individual interest images. However, the method requires the design of feature detector kernels which may not be intuitive.

---

<sup>\*</sup> The authors are thankful to the Hong Kong Observatory for the provision of data and expert advices.

In this paper, we model the problem of squall line positioning as the location of a narrow band or lines of high reflectivity echoes. We propose three methods for identifying squall lines from radar data. These methods can provide independent and objective information sources to assist forecasters in identifying and positioning squall lines. The first method is based on ellipse fitting of reflectivity echoes, which is the most common approach in rainstorm detection [3]. The second one relies on detecting a straight line of high reflectivity echoes using Hough transform [4]. Finally, we use the arc-shaped weather system model and a genetic algorithm-based framework proposed in our recent work [5] for positioning. We compare the underlying advantages and the limits of each method, and give suggestions on the choice of algorithms and parameters.

In Sections 2 and 3, we first discuss the use of ellipse fitting and Hough transform in squall line positioning. Then we present the adjoining arc model and the genetic algorithm-based framework in Section 4. The methods are compared and evaluated in Section 5. A summary in Section 6 concludes the paper.

## 2 Positioning by Ellipse Fitting

Ellipse fitting is one of the most common approaches in rainstorm positioning. Rainstorms are identified by grouping pixels over some predefined intensity threshold and described by ellipses [3][6]. Each ellipse is then labeled and tracked by centroid tracking. The ellipse representation provides information such as orientation, area of precipitation and dimension. To apply ellipse fitting to squall line positioning, thresholding of radar reflectivity data is needed, since squall lines typically have high reflectivity values. The procedure is as follows. A radar reflectivity image is first thresholded with reflectivity threshold at  $t_d$  dBZ to make the squall line stand out. A  $5 \times 5$  median filter is then applied to reduce noise in image. Fig. 1(a) shows the original and preprocessed radar images. An ellipse is then fitted to the cloud of points of the preprocessed image by solving an eigensystem. The locations  $p_i = (x_i, y_i)$  of the points are first translated so that the centroid is at the origin. They then form the population set  $\mathbf{P} = \{p_1, p_2, \dots, p_n\}$ . The covariance matrix  $\mathbf{C}$  is found using data from  $\mathbf{P}$ , and the eigenvalues  $\lambda_i$  and their corresponding unit eigenvectors  $\mathbf{e}_i$  are estimated by solving  $\mathbf{C}\mathbf{e}_i = \lambda_i\mathbf{e}_i$  for  $i = 0$  and  $1$ . The directions and magnitudes of the major and minor axes of the ellipse are then given by  $\mathbf{e}_i$  and  $4\sqrt{\lambda_i}$  respectively. The centroid of the cloud of points gives the ellipse centroid.

## 3 Positioning by Hough Transform

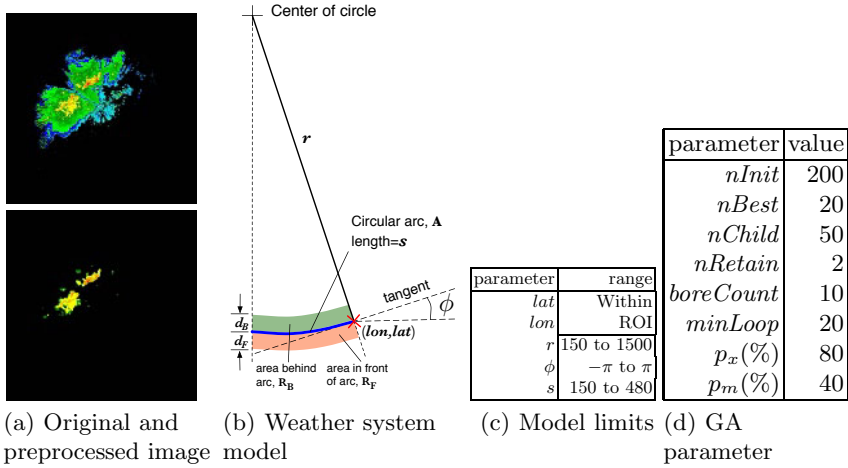
A squall line can be modeled as a line of high reflectivity echoes. Thus, Hough transformation, a method for straight line location, can be applied on radar data over some predefined reflectivity threshold to detect squall lines. The idea of Hough transform is as follows. Given a point on the Cartesian plane, there are an infinite number of straight lines passing through it. Each of these lines passing through  $(x, y)$  in the Cartesian plane can be represented by the polar

equation  $\rho = x \cos \theta + y \sin \theta$  for some combinations of  $\rho$  and  $\theta$ . Here,  $\rho$  is the perpendicular distance of the line from the origin, and  $\theta$  is its slanting angle. Note that since fixing  $x$  and  $y$  gives a relation between  $\rho$  and  $\theta$ , a single point in the Cartesian plane maps to a trajectory in the  $(\rho, \theta)$  plane. When a number of points in the  $(x, y)$  plane are collinear and pass through a particular straight line  $\rho_0 = x \cos \theta_0 + y \sin \theta_0$ , the point  $(\rho_0, \theta_0)$  in the Hough space can be traced by all the trajectories that correspond to those collinear points. This means that the straight line passing through the most number of points can be found by identifying the  $(\rho, \theta)$  value intersected by the largest number of trajectories in the Hough space. This allows us to design the squall line positioning algorithm as follows. A radar reflectivity image is first thresholded with a reflectivity threshold of  $t_d$  dBZ and smoothed using a  $5 \times 5$  median filter. Hough transformation is then applied on the preprocessed image and the straight line that passes through the largest number of reflectivity points can then be found. Finally, the segment identifying the squall line can be extracted by removing the parts covering zero reflectivity areas at their ends. This line segment becomes the proposed squall line of the algorithm.

#### 4 Positioning Using a Genetic Algorithm-Based Framework

In our recent work [5], a generic model of weather systems, along with a genetic algorithm-based framework for finding weather systems from numerical weather prediction data, is proposed. We have adapted the model and the framework to process radar data for squall line positioning. The model approximates a weather system as adjoining segments of arcs. Each arc is described by a vector with five elements  $\langle lat, lon, r, \phi, s \rangle$  (Fig. 1(b)). The values  $lat$  and  $lon$  give the latitude and longitude of the beginning point on the arc. The value  $r$  is the radius of the arc,  $\phi$  the angle of the tangent of the arc at  $(lon, lat)$ , and  $s$  the length of the arc drawn clockwise from  $(lon, lat)$ . The arc at radius  $r$  can be extended to an area using two auxiliary arcs with radii  $r + d_F$  and  $r - d_B$  with the same angular span, where  $d_F$  and  $d_B$  are both positive constants specified in the experiments.

The model and framework can be used to position a squall line by processing radar images thresholded at  $t_d$  dBZ and  $5 \times 5$  median-filtered. The fitness function of spatial average of reflectivity values along the arc is then used in the genetic algorithm (GA) matching. The algorithm is as follows. Initially,  $n_{Init}$  candidate arc parameters  $\langle lat, lon, r, \phi, s \rangle$  are generated randomly in a Region Of Interest (ROI). An ROI is the bounding rectangle of the areas with radar echoes, or a user-defined rectangular area within the radar range. The values of  $lat$  and  $lon$  generated by the algorithm are limited by the ROI and the limits of the other three parameters are determined by values of typical squall lines (Fig. 1(c)). After the initial set of candidate arcs is generated, the algorithm enters an iterative phase. The fitness value for each candidate arc is calculated and the fittest  $n_{Best}$  arcs are retained as parents.  $n_{Child}$  children arcs are then generated using the parents with crossover and mutation probabilities of  $p_x$  and  $p_m$  respectively, with



**Fig. 1.** Preprocessed image, weather system model and parameter limits

at least *nRetain* of them verbatim copies of the parents. Crossover splices and exchanges parameter values of the parent arcs after a randomly selected crossover point, whereas mutation alters only one of the genes randomly. The iterative phase ends when the best score does not improve for *boreCount* iterations after the algorithm runs for at least *minLoop* iterations, or when the score is over a user-defined threshold *minScore*. Fig. 1(d) summarizes the GA parameters used.

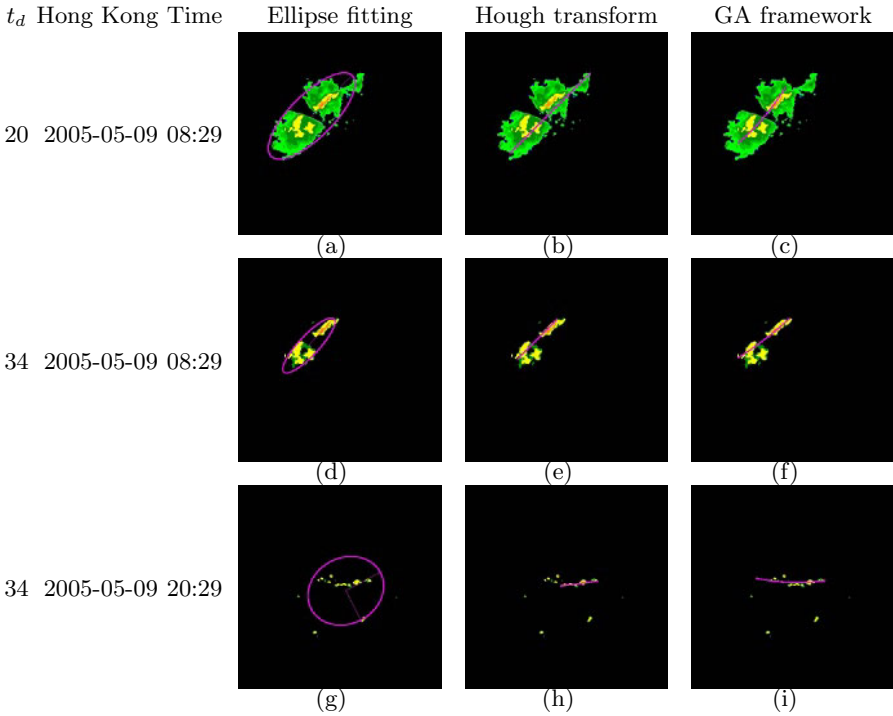
The algorithm has an option of using temporal information, which is suggested in our previous work [7]. To mandate the use of temporal information, the fittest *nBest* candidates of the last iteration of the current frame are stored and used as part of the initial set of candidates in the next frame. The effect of using temporal information is also investigated in this paper.

## 5 Results and Comparison

To evaluate the performance of the three proposed methods, a Java-based system prototype is built. Sequences of radar reflectivity images at zero degree Plan Position Indicator (PPI) with radar range of 512 km, captured every 12 minutes, were used for testing. These include 80 images of a squall line case, HKT 2005-05-09 08:05 to 23:53, from its mature stage until dissipation. We compare the underlying advantages and limits of each method, and give suggestions on the choice of algorithms and parameters.

Fig. 2(d) to 2(f) show some of the results of squall line positioning using the proposed methods. In general, all the methods can identify the squall line provided that areas with low reflectivity values are removed.

The three methods model a squall line differently. Ellipse fitting encircles an area of reflectivity echoes, which represents a squall line using its position, orientation and dimension. Hough transform identifies a straight line of reflectivity



**Fig. 2.** Results comparing the three methods

echoes, which describes a squall line using its position, orientation and length. These two methods are simple but inflexible as the results depend only on the spatial distribution of reflectivity values.

In contrast, the GA-based identification method models a squall line as adjoining arcs of high reflectivity echoes using their position, orientation, length, and curvature. The arcs can be modified to represent narrow curved areas by setting the parameter  $d_F$  and  $d_B$ . To give a fairer comparison with the other two methods which use a single geometric shape to describe a squall line, we limit the number of arcs to one in our experiments. Still, the model can be tuned by limiting the ranges of the parameters  $\langle lat, lon, r, \phi, s \rangle$ . This flexibility allows change in shape and curvature of the arc, giving a more detailed description of a squall line among the three methods.

Ellipse fitting and Hough transform work by analyzing individual remote sensing images as if they are independent. Temporal information cannot be used in both approaches. The GA-based identification, in contrast, has an option of using temporal information. Fig. 3 compares temporal stability of the methods, using graphs of the changes of area (ellipse fitting) or length (Hough transform or GA-based identification) across frames. From the graphs and statistics, it can be seen that results using ellipse fitting is the least stable among the three methods, while the temporal GA-based identification method is the most stable.

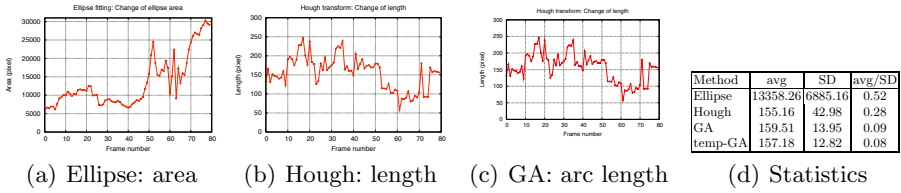


Fig. 3. Temporal stability

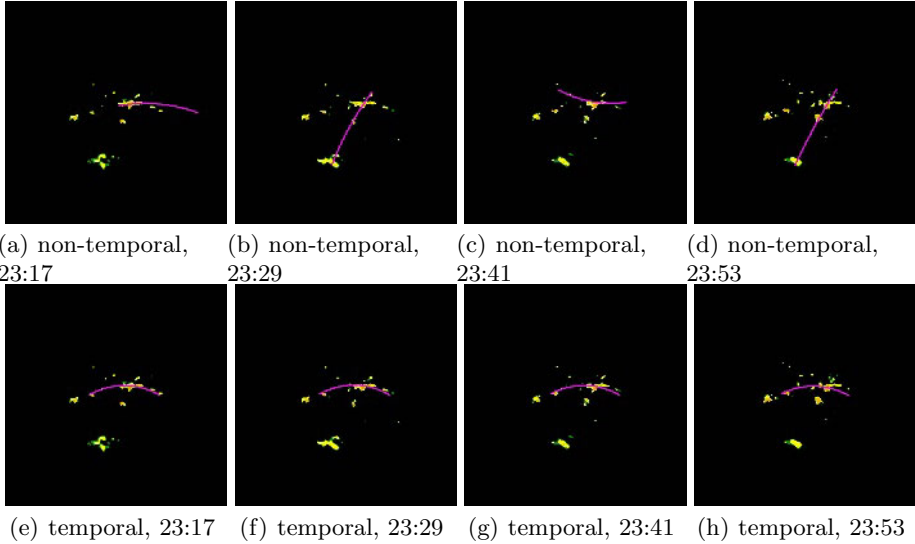


Fig. 4.  $t_d = 34$ , non-temporal vs. temporal GA, HKT 2005-05-09 23:17 to 23:53

Although squall lines are relatively rapid-changing weather systems, its structure would not change much between two frames that is 12 minutes apart. Among the presented methods, only the temporal GA-based one can take advantage of this property by using the fittest candidates of the previous frames as candidates of the current frame. This allows information such as the length and curvature of a squall line to be maintained, giving more stable results.

The advantage of temporal stability can also be illustrated using the results when the squall line is dissipating (Fig. 4). With the information from the previous frame, the temporal GA-based method discovers arcs that are stable in shape and positionally continuous with respect to time, an important requirement in weather system identification. This property of temporal continuity is not always given by the other two approaches or GA-based method without the use of temporal information.

Both ellipse fitting and Hough transform are standard techniques, with no flexibility in analysis. On the contrary, the GA-based framework provides a platform where different fitness functions can be used. The framework also allows tuning



**Table 1.** Comparison of squall line positioning methods

	Line	Area	Model flexibility	Temporal	Temporal stability	Analysis flexibility	Parameter tuning	Threshold sensitivity	Speed	Answer any time
Ellipse	×	✓	×	×	Moderate	×	×	High	Fastest	×
Hough	✓	×	×	×	Fair	×	×	High	Good	×
GA	✓	✓	✓	×	Good	✓	✓	Low	Good	✓
GA (temporal)	✓	✓	✓	✓	Best	✓	✓	Low	Good	✓

of parameters to improve its performance, providing flexibility in analysis. The downside is that experimentation or educated estimation of initial parameters is needed for GA-based algorithms.

The sensitivity of the algorithms to the threshold  $t_d$  is also investigated. Fig. 2(a) to 2(c) and 2(d) to 2(f) show the results using  $t_d = 20$  and  $t_d = 34$  respectively. Ellipse fitting encircles the entire area with radar echoes including the low reflectivity ones; while Hough transform just identifies a straight line of radar echoes. Both methods only take care of the spatial pixel distribution, without taking actual reflectivity values into consideration. The detected ellipse or line may not always be the one with the highest reflectivity values, thus not always on a squall line. In contrast, the GA framework identifies the squall line with high reflectivity values even if the low reflectivity regions are retained. However, the weakness of the fitness function is that it tends to find a line of maximum reflectivity values, without considering the pixel distribution. This may lead to the problem that only part of the squall line is identified instead of the whole.

The problem of Hough transform and ellipse fitting methods in not considering the reflectivity values can also be illustrated using the identification results during the squall line dissipation phase (Fig. 2(g) to 2(i)). It can be seen that only the GA-based identification can locate the weakening squall line, while ellipse fitting just encircles the entire area with radar echoes and Hough transform just locates a line having the most radar echoes.

In terms of speed, using a PowerMac G5 with 1.6 GHz PowerPC processor and 1.5 GB RAM running Mac OS 10.4.5, ellipse fitting processes around 13.3 images a minute on average, Hough transform processes around 5.7, GA-based identification 3.8, and temporal GA-based 4. All methods are acceptable because they can complete the identification before the next radar scan completes in 12 minutes. Ellipse fitting, which requires scanning of pixel distribution only once, runs the fastest. The efficiency of the other three methods are comparable. The efficiency of the GA-based method is controlled by its parameters. The parameter  $nInit$  and  $nChild$  can be changed to a lower value to reduce the time of candidate generation. The termination control values  $minScore$ ,  $boreCount$ , and  $minLoop$  can also be adjusted to allow early return of results. By the iterative nature of GA, the algorithm can also be queried at any time for the best answer found so far to meet the necessary time constraints.

Table 1 gives a comparison of the proposed methods. In short, ellipse fitting is the most efficient, while the speed of the other three methods are comparable. Ellipse fitting and Hough transform analyze pixel distribution and provide rough identification results. However, they lack flexibility and temporal stability, and

their accuracies depend on the predefined threshold  $t_d$ . The GA-based identification method provides a flexible model and platform for analysis. It maintains the best temporal stability, which comes at a cost of efficiency and a need to specify parameters, due to a more complex iterative nature of the algorithm. Nevertheless, tuning of parameters for efficiency is possible. Also, the algorithm can be queried at any time for an answer. The algorithm performance would further improve if temporal information is incorporated.

## 6 Summary

We have introduced three methods for squall line positioning: ellipse fitting, Hough transform and the use of a genetic algorithm-based framework. The three methods model a squall line differently, as an ellipse, a straight line, and adjoining segments of arcs respectively. Experiments show that ellipse fitting is the most efficient, while the speed of the other methods are comparable. Ellipse fitting and Hough transform lack flexibility and temporal stability, and are sensitive to the preprocessing parameters. To use them effectively, low reflectivity regions of the radar data have to be removed. In contrast, the GA-based identification allows parameter tuning, and provides a flexible model and framework for analysis. It is practically insensitive to the preprocessing parameters, and maintains the best temporal stability. The GA-based method is relatively less efficient, but can be sped up by tuning the GA parameters and incorporating temporal information. Our study contribute to the field of meteorological computing by providing independent and objective information sources to assist forecasters in identifying and positioning squall lines.

## References

1. Ahrens, C.D.: *Meteorology Today: An Introduction to Weather, Climate, and the Environment*. West Pub. Company (2000)
2. Troxel, S., Pughe, W.: *Machine Intelligent Gust Front Algorithm (MIGFA) for the WSP*. Lincoln Laboratory, Massachusetts Institue of Technology. (2002)
3. Li, P.W., Lai, S.T.: Short range quantitative precipitation forecasting in Hong Kong. *J. Hydrology* **288** (2004) 189–209
4. Duda, R.O., Hart, P.E.: Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM* **15** (1972) 11–15
5. Wong, K.Y., Yip, C.L., Li, P.W.: Identifying weather systems from numerical weather prediction data. In: *Proc. 18th ICPR, Hong Kong (2006)* 841–844
6. De Jongh, I., De Lannoy, G., Verhoest, N., De Troch, F.: Rainstorm characteristics derived from weather radar images. In: *Proc. of Euro. Conf. on Radar Meteor. (2002)* 222–226
7. Wong, K.Y., Yip, C.L.: Tropical cyclone eye fix using genetic algorithm with temporal information. In: *Proc. 9th KES. LNAI-3681, Australia (2005)* 854–860

# On Clustering Performance Indices for Multispectral Images

C. Hernández, J. Gallego, M.T. Garcia-Sebastian, and M. Graña\*

Computational Intelligence Group, Dept. CCIA  
University of the Basque Country  
San Sebastian 20018 Spain

manuel.grana@ehu.es, josune.gallego@ehu.es, mamen.hernandez@ehu.es

**Abstract.** Clustering of multispectral image pixels can be a exploratory tool to analyze the contents of the image in the absence of ground truth information. The validity of the clustering algorithms can be quantified computing several performance indices. Each performance index enhances some statistical property of the obtained data partitions. Performance indices are not equivalent, and they can even lead to quite different conclusions from the same data partitions. To show this, we have applied two well known clustering algorithms (K-means, Fuzzy c-means) and some supervised classification algorithms to a well known multispectral image. We compare the ground truth partition with the ones found by the clustering and supervised algorithms. The values of the diverse performance indices over the same partitions vary and can lead to quite different conclusions.

## 1 Introduction

A key problem in the selection of an appropriate clustering algorithm for the unsupervised processing of multispectral images forced by the lack of ground truth information is the quantification of the comparison of algorithm results. Recent works [1,2] address the methodological problem of the validation of classification results when no ground truth data exists from a general methodological perspective assuming the neutral effect of the selection of the computed performance index. We focus here on the influence that the selected performance index may have in the formulation of conclusions regarding the validity of clustering algorithm results. In the absence of ground truth information, the performance indices usually compare the different partitions of the data generated by the different algorithms or repetitions of the same algorithm with random initial conditions. When some ground truth is available, this information is considered as the gold standard of the partitions to compute the performance indices.

To factor out the unknown influence of the clustering algorithms we have used two well known clustering algorithms: K-means [5] and Fuzzy c-means [3], [8].

---

\* The Spanish Ministerio de Educacion y Ciencia supports this work through grants DPI2003-06972 and VIMS-2003-20088-c04-04.

In a similar reasoning we have used as the experimental data a well known multispectral image. We have computed a comprehensive set of performance indices found in the literature, which are summarized in [6]. Furthermore the results of supervised classification algorithms can be viewed as an alternative way to obtain data partitions which, per force, must be in better agreement with the ground truth because of the unavoidable class identification ambiguity in unsupervised algorithms. This also may allow to ascertain the extent to which the performance indices are able to discriminate between supervised and unsupervised approaches. Surprisingly, some indices do not distinguish between them. The supervised classification algorithms were: MLE, MDM and Fisher discriminant [9].

First we introduce some of the tested external performance indices and then we comment on the experiment results. Finally, we give some conclusions.

## 2 External Clustering Performance Indices

External performance indices compare the results obtained by different clustering instances obtained either by different algorithms or by different realizations of the same algorithm, through the computation of the relation between the data partitions produced by the clustering algorithm. Obviously, the ground truth may be used as an alternative partition of the data. Then the performance index is a validation measure. When both partitions come from repetitions of the same clustering algorithm the external performance indices measure the variance of the clustering results. Good clustering algorithms provide consistent results across repetitions. When the partitions come from different clustering algorithms the external performance index shows the equivalence between algorithms.

We denote  $P$  a partition of data set  $X$  of cardinality  $n$  into a set of  $p$  clusters:  $P = \{C_1, \dots, C_p\}$  where  $\bigcup_{i=1}^p C_i = X$  y  $C_i \cap C_j = \emptyset$  with  $i, j = 1, \dots, p$  and  $i \neq j$ . The comparison of the partitions of the data set  $X$  obtained by different procedures is based on the study of pairs of data elements. Suppose two partitions  $P$  and  $Q$  of  $X$  into  $p$  y  $q$  classes, resp.,  $p \leq q$  obtained by different algorithms. Let them be

$$P = \{C_1, \dots, C_p\} \quad Q = \{C'_1, \dots, C'_q\} \quad (1)$$

We take all the data item pairs  $X$ ,  $(x, y) \in X \times X$  and analyze the agreement of the relative classification of the items in the two partitions. The two partitions agree when the two items either fall in the same class or not in both partitions. The two partitions disagree about the data item pair if its elements fall in the same class for one partition and in different classes for the other partition. We denote  $r$  the positive agreements (data pairs falling in the same class in both partitions),  $s$  the negative agreements (data pairs falling in different classes in both partitions),  $v$  the number of rejections when partition  $P$  puts the items in the same class and partition  $Q$  in different classes, and finally  $u$  the number of rejections when partition  $Q$  puts the items in the same class and partition  $P$  in different classes.

Customarily, it is accepted that there are three categories of external indices : (1) similarity indices that measure the similarity between partitions, (2) dissimilarity indices that compute the dissimilarity between partitions, and (3) mixed indices that compute both similarities and dissimilarities.

The most famous similitude index is the Kappa index, which measures the proportional difference between positive acceptance and the expected random agreement between partitions:

$$IKA(P, Q) = \frac{n \sum_{i=1}^p n_{ii} - \sum_{i=1}^p n_i.n_i}{n^2 - \sum_{i=1}^p n_i.n_i}, \tag{2}$$

where the  $n_{ij}$  are the entries of the contingency matrix between partitions. The IKA index is maximum when the contingency matrix trace is maximum. The Kappa index requires that the two partitions have the same number of classes.

Other similitude indices are shown in table 1.

**Table 1.** External similitude indices

Indices	Expression	Range
Rusell/Rao	$\frac{r}{r+u+v+s}$	[0, 1]
Rand / Kendall Simple Matching Sokal y Michener	$\frac{r+s}{r+u+v+s}$	[0, 1]
Rand adjusted	$\frac{r-E(r)}{Max(r)-E(r)}$	[-1, 1]
Hubert	$\frac{(r+u+v+s).r-(r+u)(r+v)}{\sqrt{(r+u)(r+v)(v+s)(u+s)}}$	[-1, 1]
Rogers y Tanimoto	$\frac{r+s}{r+2(u+v)+s}$	[0, 1]
Hamman Gower y Legendre	$\frac{(r+s)-(u+v)}{r+u+v+s}$	[-1, 1]
Sokal y Sneath V / Gower / Ochiai II	$\frac{rs}{\sqrt{(r+u)(r+v)(s+u)(s+v)}}$	[0, 1]
Sokal y Sneath I	$\frac{2(r+s)}{2(r+s)+u+v}$	[0, 1]

The dissimilarity indices measure the differences between two partitions of the same data set, they provide the complementary view of the similarity indices. Table 2 presents some of the dissimilarity indices tested below.

### 3 Experimental Results

The experimental data come from the multispectral image *flc1* provided in the CD that accompanies Landgrebe’s book [9] . The image size is  $949 \times 220 \times 12$ . We have used the first 3 principal components, corresponding to the 77%of the bands

**Table 2.** External dissimilarity indices

Indices	Expression	Range
Mirkin	$2(u + v)$	$[0, +\infty]$
normalized Mirkin	$\frac{4(u+v)}{n^2-n}$	$[0, 1]$
van Dongen (IVD)	$2n - \sum_{i=1}^p \max_{j=1\dots q} n_{ij} - \sum_{j=1}^q \max_{i=1\dots p} n_{ij}$	$[0, +\infty]$
van Dongen normalized	$\frac{IVD(P,Q)}{2n}$	$[0, 1]$
Mac Nemar	$\frac{u-v}{\sqrt{u+v}}$	$[-1, 1]$
Mac Nemar corrected	$\frac{ u-v -1}{\sqrt{u+v}}$	$[0, 1]$
Sokal y Sneath III	$\frac{r+s}{u+v}$	$[0, +\infty]$
Q0	$\frac{u \cdot v}{r \cdot s}$	$[0, \infty]$
Average Squared	$\frac{u+v}{r+u+v+s}$	$[0, 1]$
Distancia Euclídea	$\sqrt{\frac{u+v}{r+u+v+s}}$	$[0, 1]$
Distancia de Soergel-Tanimoto	$\frac{u+v}{r+u+v}$	
Bray-Curtis	$\frac{u+v}{2r+u+v}$	$[0, 1]$
Shannon	$2(u + v) \cdot \log 2$	$[0, +\infty]$
Binary pattern difference	$\frac{uv}{(r+u+v+s)^2}$	$[0, 1]$
Pattern variance	$\frac{u+v}{4(r+u+v+s)}$	
$uv$	$\frac{4uv}{(r+u+v+s)^2}$	$[0, 1]$
Binary size difference	$\frac{(u+v)^2}{(r+u+v+s)^2}$	$[0, 1]$
Binary shape difference	$\frac{u+v}{r+u+v+s} - \left(\frac{u-v}{r+u+v+s}\right)^2$	$[0, 1]$

variance. There are 10 classes in the ground truth that must be discovered by the clustering algorithms. We use the 11.451 labeled pixels provided with the image for classification experiments. We have computed 100 repetitions of the K-means with different initialization strategies (denoted cluster, sample, uniform in the plots). We have also computed 100 repetitions of the Fuzzy c-means algorithm with values  $m = 2$  and  $m = 1.5$  of the exponent parameter that tunes the fuzziness of the approach. In all the cases the number of clusters looked for are 10. Besides we have computed the following supervised classification algorithms: MLE, MDM and Fisher discriminant.

The computed similarity indices and their codes in the plots are: Russel-Rao (IRR), Jaccard (IJ), Dice (ID), Ochiai I (IO1), Kulczynski I (IK1), Rogers-Tanimoto (IRT), Baroni-Urbani (IBU), Kulczynski II (IK2), Faith (IF), Braun-Blanquet (IBB), Simpson (IS), Michael (IMIC), Forbes (IFO), Sokal y Sneath I (ISS1), Sokal y Sneath II (ISS2), Sokal y Sneath III (ISS3), Sokal y Sneath IV (ISS4), Sokal y Sneath V (ISS5), Rand (IR), Rand ajusted (IRN), Hubert (IH), Kappa (IKA).

The computed dissimilarity indices and their codes in the plots are: Q0 (IQ0), Mirkin normalized (IMN), Van Dongen normalized (IVDN), Hamming distance (IDH), Averagesquared (IAS), Euclidean distance (IDE), Soergely Tanimoto (ISO), Bray-Curtis (IBC), Binary pattern difference (IBPD), Binary pattern variance

(IBPV), uv index (IUV), Binary size difference (IBSD), Binary shape difference (IBSHD). Similarity indices have a growing value when the agreement is high. We normalize them dividing them by the ground truth self similarity (Ideal). Dissimilarity indices have decreasing values with the agreement between partitions. Thus, when evaluating against the ground truth zero is the optimal value.

To compute the Kappa index and the classification accuracy it is need to perform an assignment of clusters to ground truth classes. We did it maximizing the trace of the confusion matrix.

Figure 1 shows the average similarity indices computed between the clustering results and the ground truth partition. The first thing that can be noticed is that it is almost impossible to distinguish between them for all the computed indices. This effect may be due to the data used, but it is very surprising. The interesting fact in the figure is that there is a large variance of the indices over exactly the same partitions. Some of the indices may lead to conclude that the clustering results are extremely good while others allow to conclude that they are very bad. In the extended paper we will try to reason these particular results.

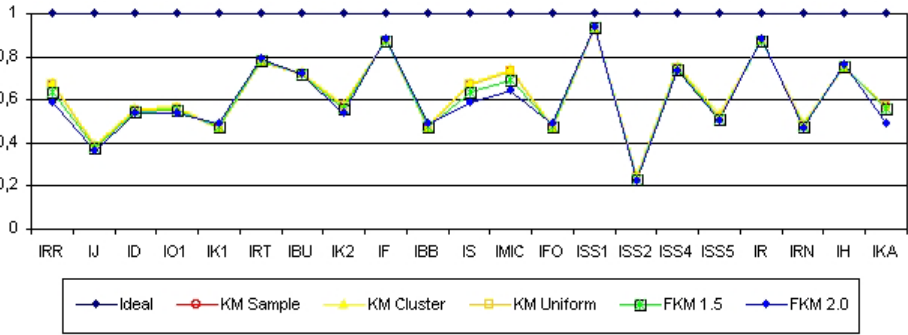


Fig. 1. Similarity indices computed for the unsupervised clustering approaches

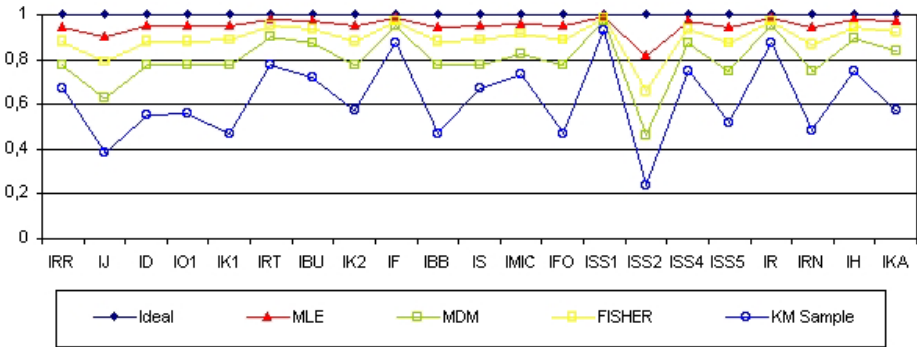


Fig. 2. Similarity indices comparing supervised and unsupervised approaches

Figure 2 shows the average similarity indices of one unsupervised approach (KM-sample) and the supervised approaches, against the ground truth. Naturally, the latter benefit from the knowledge of the ground truth information and their index values are above the unsupervised approach for all the indices. However it must be noted that the relative difference between approaches varies largely with the chose index. Even, one of the indices does not distinguish between supervised an unsupervised approaches.

Figure 3 shows the the average dissimilarity indices computed between the clustering results and the ground truth partition. Again, the diverse unsuper-vised approaches are nearly indistinguishable for all the computed performance indices. Again, the index value variance is relatively high, and few indices agree in their evaluation of the results. Finally, figure 4 shows the average dissimilarity indices of one unsupervised approach (KM-sample) and the supervised approaches against the ground truth. Again, the supervised approaches perform better than the unsupervised approach, however several indices are unable to distinguish between the supervised and unsupervised approaches.

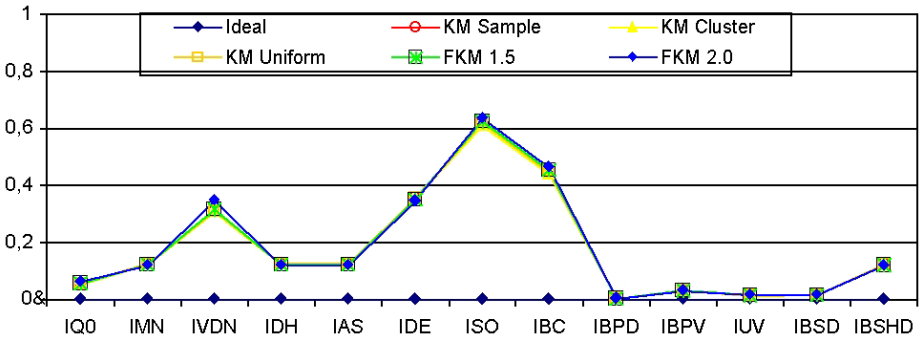


Fig. 3. Dissimilarity indices computed for the unsupervised clustering approaches

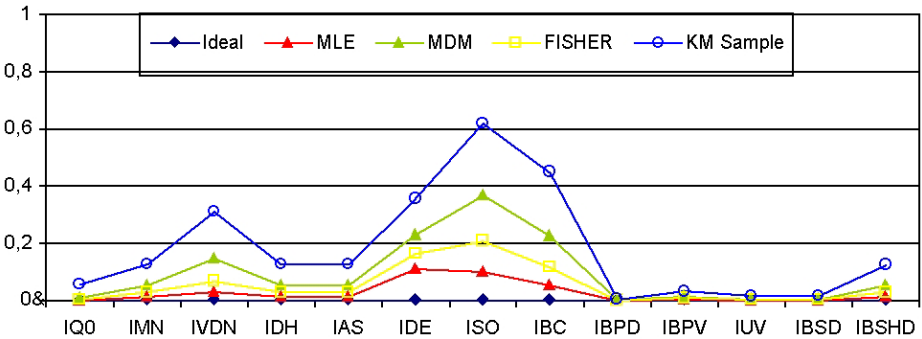


Fig. 4. Dissimilarity indices comparing supervised and unsupervised approaches



## 4 Conclusions

The question addressed in this paper is does the selected performance index influence the conclusions that can be extracted from the classification algorithms? We are mainly concerned with unsupervised algorithms, namely clustering algorithms. Nevertheless we have included in our battery of classification algorithms some supervised ones. We have used standard implementations of the clustering algorithms, not biasing them with personalized tricks. We have used a well known multispectral image that can be easily obtained, so our computational results must be easily reproducible. We did compare all the algorithms against the ground truth given with the image. This ground truth is not taken as a golden standard or the absolute truth but as a benchmark partition not biased by any algorithm.

The first observation from the results in the figures is that the values of the performance indices vary widely inside their own category. Meaning that different performance indices may produce quite different conclusions. The second observation is that most of the indices are unable to distinguish among the K-means and the Fuzzy c-means. Finally some of them are also unable to distinguish supervised from unsupervised approaches.

## References

1. Baraldi, A., Bruzzone, L., Blonda, P., Carlin, L. (2006) *Quality assessment of classification and cluster maps without ground truth knowledge* IEEE Trans. Geoscience and Remote Sensing 44(1):214-235
2. Baraldi, A., Bruzzone, L., Blonda, P. (2005) *Badly Posed Classification of Remotely Sensed Images; An Experimental Comparison of Existing Data Labeling Systems* IEEE Trans. Geoscience and Remote Sensing 43(4):857-873
3. Bezdek J.C., Ehrlich R., Full W. (1984) *FCM: Fuzzy C-Means Algorithm*. Computers and Geosciences 10(2-3):191-203
4. Da Silva Meyer A. (2002) *Comparação de coeficientes de similaridade usados em análises de agrupamento com dados de marcadores moleculares dominantes*. Dissertação (mestrado) Escola Superior de Agricultura "Luz de Queiroz". Piracicaba. Universidade de São Paulo. p.106.
5. Duda R.O., Hart P.E., Stork D.G. (2001) *Pattern Classification. Second Edition*. Wiley-Interscience Publication. Wiley & Sons.
6. Hernández C., Graña M., Gallego J. (2005) *Survey of clustering performance indices* Research Report, Dept. CCIA, UPV/EHU, Facultad Informática, San Sebastián.
7. Halkidi M., Vazirgiannis M., Batistakis I. (2000) *Quality scheme assessment in the clustering process*. Proceedings of Principles Knowledge Discovery and Data Mining, PKDD'2000, Lyon, France.
8. Keller J., Krisnapuram R., Pal N.R., Bezdek J.C. (2005) *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. New York: Springer Verlag
9. Landgrebe D. A. (2003) *Signal Theory Methods in Multispectral Remote Sensing*. Wiley & Sons.
10. Youness G. (2004) *Contributions à une méthodologie de comparaison de partitions*. Thèse CEDRIC, 2004. Thèse de doctorat, Université Pierre et Marie Curie.

# Aerial Photo Image Retrieval Using Adaptive Image Classification\*

Sung Wook Baik<sup>1</sup>, Moon Seok Jeong<sup>1</sup>, and Ran Baik<sup>2</sup>

<sup>1</sup> College of Electronics and Information Engineering, Sejong University,  
Seoul 143-747, Korea  
{sbaik, mjeong}@sejong.ac.kr

<sup>2</sup> Department of Computer Engineering, Honam University,  
Gwangju 506-090, Korea  
baik@honam.ac.kr

**Abstract.** The paper presents a method for content based image retrieval (CBIR) using an adaptive image classification with Radial Basis Function networks. It supports geographical image retrieval over digitized historical aerial photographs, in a digital library, which are gray-scaled and low-resolution images. CBIR is achieved on the basis of texture feature extraction and image classification. Feature extraction methods for geographical image analysis are Gabor spectral filtering and Laws' energy filtering, which are the most widely used in image classification and segmentation. Image classification supports effective CBIR through composite classifier models dealing with multi-modal feature distribution. The method is evaluated over a digital library that contains collections of thousands of small-sized texture tiles obtained from large-sized aerial photograph images with geographical features.

## 1 Introduction

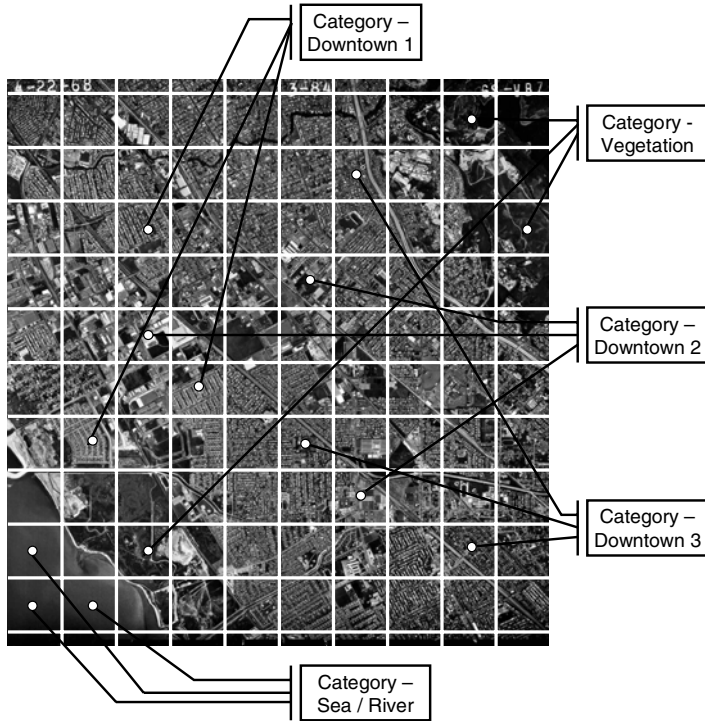
The content-based image retrieval [1-5] in digital libraries helps to relieve the tedious work of manually finding the geographical region of interest. Content-based image retrieval requires the integration of image processing and information retrieval technologies. This is the retrieval of images on the basis of features automatically derived from the images themselves. Texture, color and shape are the most widely used in most researches to describe features in the image. However, the retrieval of historical aerial image photographs is based only on texture features because they are gray-scaled and low-resolution images. Therefore, more robust feature extraction methods are required to allow effective retrieval results. The feature extraction is described in the next section.

In low resolution aerial images, we cannot apprehend the appearance of any objects in detail. Therefore, we need to use regions with a collection of tiny and complicated structures--such as man-made features including buildings, roads, parking lots, airports and bridges--in order to deal with them as texture motifs for image classification/segmentation. We can also regard the shapeless regions of natural resources such as forests, rivers and oceans as texture motifs.

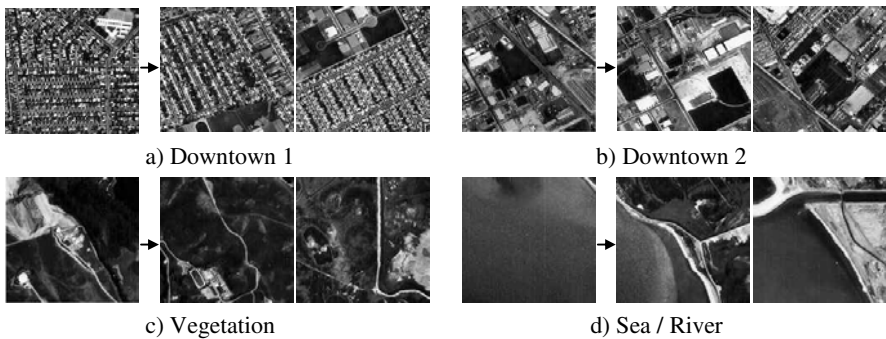
---

\* This study was supported by a grant of the Seoul R&BD Program.

This work presents a texture-based geographical image retrieval system with adaptive image classification. It provides geographical image retrieval over a digital library containing collections of thousands of small-sized blocks of pixels obtained from large-sized aerial photograph images with geographical features. Fig. 1 presents an aerial photograph image and 100 non-overlapping blocks of pixels with geographical features such as downtowns, water, vegetation, and so on. Fig. 2 shows texture



**Fig. 1.** An aerial photograph image and 100 non-overlapping sub-images



**Fig. 2.** Texture tiles with geographical features appearing in Fig. 1

tiles with geographical features appearing in Fig. 1. The left-side image in each graphical feature is a query image and the right-side images are result images corresponding to the query image after image retrieval.

## 2 Texture Based Feature Extraction for Analysis of Aerial Images

To obtain geographical features for aerial image retrieval, we have used two texture feature extraction methods 1) Gabor spectral filtering and 2) Laws' energy filtering, both of which have been widely used by researchers and perform very well for various classification and image segmentation tasks. These filters are useful to represent directionality, coarseness, and regularity of patterns appearing on aerial images when they are particularly applied to digitized historical aerial photographs that are gray-scaled and low-resolution images.

Two-dimensional Gabor functions were proposed by Daugman [6] to model the spatial summation properties of simple cells in the visual cortex. A bank of Gabor filters has been obtained through a systematic mathematical approach in which the parameters of these functions are changed to represent variable local frequency and orientation information [7-8]. A Gabor function consists of a sinusoidal plane of particular frequency and orientation modulated by a two-dimensional Gaussian envelope. A two-dimensional Gabor filter is given by:

$$G(x, y) = \exp\left[\frac{1}{2}\left(\frac{x}{\sigma_x^2} + \frac{y}{\sigma_y^2}\right)\right] \cos\left(\frac{2\pi x}{n_0} + \alpha\right) \tag{1}$$

- $n_0$  is the number of pixels per cycle (pixels/cycle),
- $\phi$  is the phase of a sinusoidal plane wave along an axis,
- $\sigma_x (=1.5 \times n_0)$  is a space constant (standard deviation) of the Gaussian envelope along the x axis, and
- $\sigma_y (=0.5 \times \sigma_x)$  is a space constant (standard deviations) of the Gaussian envelope along y axis.

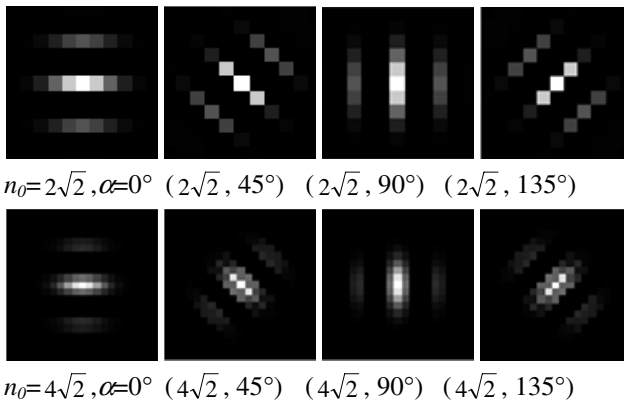


Fig. 3. An example of a set of Gabor filters

By orienting the sinusoid at an angle  $\alpha$  and changing the frequency  $n_0$ , many Gabor filtering sets can be obtained. An example of a set of eight Gabor filters (Fig. 3.) is decided with different parameter values ( $n_0 = 2\sqrt{2}$  and  $4\sqrt{2}$  pixels/cycle and orientations  $\alpha = 0^\circ, 45^\circ, 90^\circ, \text{ and } 135^\circ$ ).

Laws' convolution kernels based on five dimensional vectors [9] are used as an energy filter bank for image retrieval. It consists of 25 filters (Table 1), which can be derived from their weights. Convoluting and transposing each other produces various square masks of 25 filters. Each filter is 5x5 matrices and is designed as 5x5 windows. Each filter ( $M_i$ ) is used to convolve an original texture image ( $f$ ) to generate a filtered image ( $g$ ):

$$g_i(j, k) = \sum_{m=-a}^a \sum_{n=-a}^a M_i(m, n) f(j + m), (k + n) \text{ for } i = 1, \dots, 8 \quad (2)$$

- $j$  and  $k$  indicate a coordinate in the image, and
- $(2a + 1) \times (2a + 1)$  indicates the size of a filter for  $a = 2$

**Table 1.** Laws' Filter Weights and Specifications

Specification	Weight of filter
L5 (Level)	(1,4,6,4,1)
E5 (Edge)	(-1,-2,0,2,1)
S5 (Spot)	(-1,0,2,0,-1)
R5 (Ripple)	(1,-4,6,-4,1)
W5 (Wave)	(-1,2,0,-2,1)

For efficient extraction of texture characteristics, two additional filtering steps (i.e. averaging and non-linear filtering) are required. The averaging filter is applied to estimate the local energy response of the filter. This is achieved by replacing each value of the pixel in all filtered images by the average of the absolute values over a small overlapping window centered at the pixel.

$$Q_i(j, k) = 1/(\#S) \sum \sum_S |g_i(m, n)| \text{ for } i = 1, \dots, n \quad (3)$$

- $S$  corresponds to the local averaging window of pixels.

It is important to choose the size of the window is important. A small window produces an overly sensitive response whereas a large window may produce a meaningless response. A  $7 \times 7$  square window is used for averaging.

Non-linear filtering is applied to eliminate the smoothing effect at the borderline between distinctive homogeneous areas. The non-linear filter computes standard deviation over five small windows spread around a given pixel. The mean for the lowest deviation window is returned as the output.

### 3 Image Retrieval Using Adaptive Image Classification

Radial Basis Function classifiers have been used to model image feature distributions for a variety of research objectives such as image classification and segmentation. A

modified Radial Basis Function (RBF) classifier [10], with Gaussian distribution as a basis, was chosen for texture data modeling and classification. This is a well-known classifier, widely used in pattern recognition and suited for engineering applications. Its well-defined mathematical model allows for further modifications and on-line manipulation with its structure and parameters. It can be easily implemented as a neural network. The RBF classifier models a complex multi-modal data distribution through its decomposition into multiple independent Gaussians.

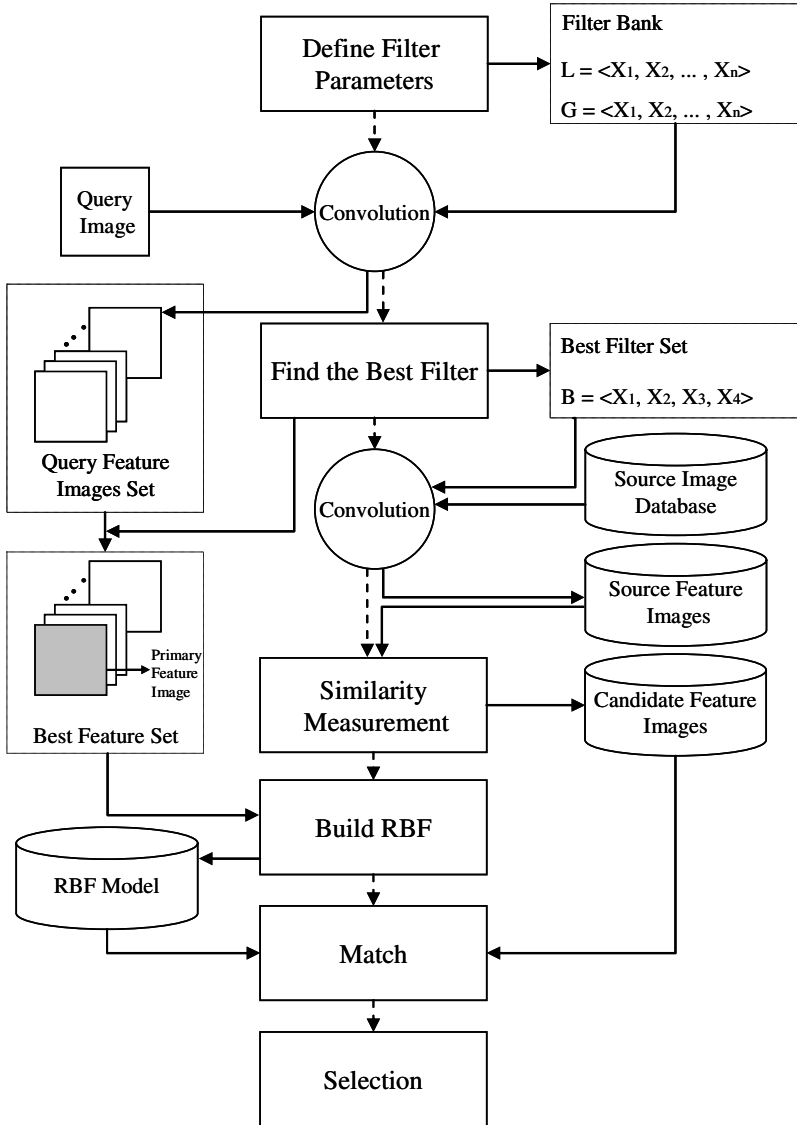


Fig. 4. Algorithm for image retrieval using the RBF classifier

A RBF function  $F_r$  consists of a set of basis functions that form localized decision regions. Overlapping local regions formed by simple basis functions can create a complex distribution. For Gaussian distribution, as a basis function, each region is represented by its center and width corresponding to a mean vector and a covariance matrix  $(\mu, \Sigma)$ . For a multi-modal distribution of a class  $r$ , a RBF can be formed through the following linear combination of these basis functions:

$$F_r(X) = w_0 + \sum_i w_i f_{ri}(X) \quad (4)$$

where:  $w_i$  is the trainable weight vector (for  $i = 0, \dots, N_r$ );  $r$  is the class membership number;  $N_r$  is the number of nodes (basis functions) in class  $r$ ; and

$$f(X) = \exp[-1/2(X - \mu)^T \Sigma^{-1}(X - \mu)] \quad (5)$$

Each group of nodes corresponds to a different class. The combination of nodes is weighted. Each node is a Gaussian function with a trainable mean vector and a covariance matrix. Classification decision yields a class  $r$  of the highest  $F_r(X)$  value for a sample vector  $X$ .

Fig. 4 describes how to search the whole image database to retrieve the best matches for any particular image query. The description of the proposed image retrieval using the RBF classifier is summarized as follows:

We defined the parameters of functions representing filters for feature extraction before image retrieval. The formation of filters is decided by changing parameter values gradually with the help of a human geographical analyst. These filters are located in the filter bank. The first task is to find primary geographical features appearing on the query image through these filters. Each geographical feature can be defined by a filter out of the filter bank. A set of filters representing salient features within the query image is called the best filter set. The best one in the set is called the primary filter. A homogeneous region can be segmented with the salient feature represented by the filter if the salient feature partially dominates the query image. The second task is to find images similar to the query image with a threshold for similarity measurement if a dominant region in each image is detected when the image is convolved by the primary filter of the query image. A collection of these candidate images is an intermediary retrieval result obtained by the primary filter. The third task is to build a classification model with the best filter set for the complete image retrieval. The classification model is defined by the radial basis functions described above. The final task is to match the candidate images in the intermediary result with the RBF classification model in order to select the final retrieval result.

## 4 Experiments

We have used Gabor and Laws' energy filtering set to extract graphical features from texture tiles obtained from large-sized aerial photograph images. The numbers of Gabor filters and Law's energy filters are 96 (24 orientations and 4 scales in the frequency domain) and 25, respectively. We evaluate the performance of our image retrieval method for each one of four geographical features such as downtowns including buildings, downtowns including roads, vegetation, and sea/river.

Experimental data are provided by the UC Berkeley Library Web [11]. They are 184 aerial photographs of the San Francisco Bay area, California, made in April, 1968 by the U.S. Geological Survey. The scale of the originals is 1:30,000. Each photograph image has the size of approximately 1300 X 1500 pixels with 256 grey-level (the size and resolution of each image are little different from each other). It is cut into about 195 (13 X 15) overlapped sub-images (texture tiles) of size 200 X 200. A test bed for image retrieval has about 35,880 (184 X 195) texture tiles. For evaluation purposes, 5 types of visually similar patterns for each class are provided by human observers. From 10 to 50 texture tiles are also selected for each type according to human visual experiences and indexed for retrieval performance evaluation, during which the rest of the images not selected are also used together with the indexed images. Table 2 summarizes the performance of the aerial image retrieval methods according to several experimental results.

**Table 2.** Performance of the aerial image retrieval methods  
(Step 1: Using only Filters, Step 2:Using RBF)

	Downtown with Buildings		Downtown with Roads		Vegetation		Water/Forest	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
<b>Step 1</b>	78%	15%	82%	21%	91%	71%	93%	76%
<b>Step 2</b>	85%	65%	88%	70%	95%	79%	96%	81%

## References

1. X. Qi and Y. Han, A novel fusion approach to content-based image retrieval, *Pattern Recognition*, 2005, vol. 38, pp. 2449-2465.
2. J. Vogel and B. Schiele, Performance evaluation and optimization for content-based image retrieval, *Pattern Recognition*, 2006, vol. 39, pp. 897-909.
3. N.V. Shirahatti and K. Bernard, Evaluation image retrieval, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 955-961.
4. A. Gasteratos, P. Zafeiridis and I. Andreadis, An intelligent system for aerial image retrieval and classification, *Lecture notes in computer science*, 2004, vol. 3025, pp. 63-71.
5. J.P. Eakins, Towards intelligent image retrieval, *Pattern Recognition Society*, 2001, vol.35, pp. 3-14.
6. J.G. Daugman: Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, *Journal of the Optical Society of America*, 1985, vol. 2, pp. 1160-1169.
7. S.E. Grigorescu, N. Petkov and P. Kruizinga: Comparison of texture features based on Gabor filters, *IEEE Transactions on Image Processing*, Vol. 11, Issue 10, 2002, 1160-1167.



8. L. Chen, G. Lu and D. Zhang, Effects of Different Gabor Filter Parameters on Image Retrieval by Texture, Proceedings of the 10<sup>th</sup> International Multimedia Modeling Conference, pp. 273-278, 2004.
9. A. Gasteratos, P. Zafeiridis, I. Andreadis, An Intelligent System for Aerial Image Retrieval and Classification, LNCS, Vol. 3025, pp. 63-71, 2004.
10. S. W. Baik and P. Pachowicz, On-Line Model Modification Methodology for Adaptive Texture Recognition, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 32, Issue. 7, 2002.
11. <http://sunsite.berkeley.edu/AerialPhotos/vbzj.html#index>

# Integration of Spatial Information in Hyperspectral Imaging for Real Time Quality Control in an Andalusite Processing Line

A. Prieto, F. Bellas, F. López-Peña, and R.J. Duro

Integrated Group for Engineering Research  
Universidade da Coruña  
abprieto@udc.es, fran@udc.es, flop@udc.es, richard@udc.es

**Abstract.** This paper presents an ANN hyperspectral classification system specifically developed to perform the quality control of an andalusite processing line. The main problem with these types of tasks is related with the way the ground truth is obtained, leading to labels that correspond to large areas with inhomogeneous contents. Thus, when any type of learning algorithm is used in order to train ANN based classifiers, one has to be sure that the samples presented to the networks really contain spectra that correspond to the labels. Therefore, a previous study on the size of the windows to be used by the ANNs as well as the way the information from the different pixels in these windows are combined must be carried out. The ANNs in the segmentation operator are based on Gaussian functions. The results obtained have shown that success rates, which were very poor when working with the spectral information of individual pixels, can be improved to better than 95%.

## 1 Introduction

Multispectral and hyperspectral imaging spectrometers for use in remote sensing applications from aerospace platforms have undergone swift development during the 1980s and 1990s [1] [2]. The subsequent operation of hyperspectrometers for remote sensing has proven the power of this technique for sorting and discrimination tasks, opening the door to their application in industrial environments. In the last decade the number of such applications has been steadily increasing encompassing many different tasks concerning inspection and process control, particularly those requiring color or spectral discrimination [3].

Hyperspectral images provide many spectral components at each spatial element. Today there are several commercially available hyperspectral imaging systems ranging various spectral regions from the near ultraviolet to visible, near-infrared and short-wave infrared wavebands covering the range from 0.4 microns to 2.5 microns depending on the application. Regardless of the different technologies used for the construction of these instruments, each one of them should perform the same set of basic functions, i.e., fast acquisition of large amounts of data from an imaging sensor, data processing to obtain an appropriate image segmentation classifying the different elements present, and the final post-processing and representation of the results.

This paper is concerned with the development of a hyperspectral system specifically designed to perform quality control in an andalusite processing plant. Andalusite is an anhydrous, aluminum silicate mineral used in refractories both in bricks and in monolithic applications. The two main characteristics in determining quality and uses for andalusite are grain size and purity [4]. The purpose of this work is to develop an ANN based hyperspectral system able to detect in real time in a given section of the processing line the andalusite quality based on these two parameters. The key premises for this application are that the spectra of the andalusite should change with the level of impurities and with its granulometry. This problem can be taken as an example of one of the main difficulties when using hyperspectral based sensor systems: the way ground truth is provided.

Whenever a hyperspectral based classification system has to be implemented, some expert usually labels ground truth images indicating what areas correspond to which compounds, whether in the form of endmembers or mixtures of them. Obviously, such expert does not label the ground truth image pixel by pixel, as this would be a tedious and error prone process, but through a coarser method indicating areas corresponding to a given product or mixture. The problem arises when a classifier must be designed or trained, especially in the case of ANN based classifiers.

For the classifier to be trained, a training set providing true input-output pairs must be given. Inputs are usually spectra and target outputs are the classification category provided by the expert. If the area that has been labeled as belonging to a particular category is not spectrally homogeneous, that is, its pixels contain different spectra, some of which may even belong to other categories, the training process degrades, sometimes becoming even impossible. One very good example of this can be found in the literature on the processing of the Indian Pines hyperspectral image [5][6] regarding the areas classified as corn and soybean. In these areas the plants were very young and thus only represented a small percentage of the pixels, being most of the ground area corresponding to a single pixel partly or totally earth. This leads to a very difficult training problem, as the labels do not correspond to a given spectrum or a range of similar spectra and some pixels in both areas contain spectra that are the same but that are labeled differently.

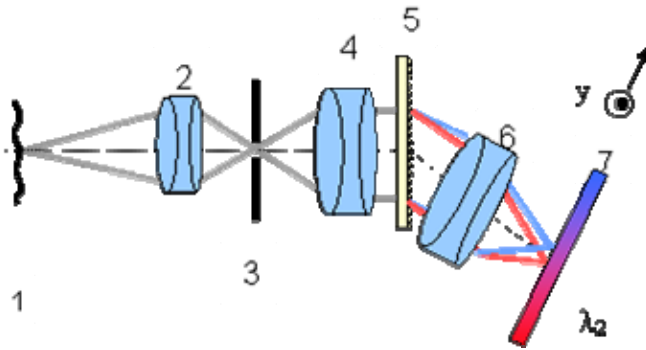
An approach based on a previous preprocessing of the image taking into account spatial information is proposed here. The idea is to determine a size for the processing window and a combination of the parameters of the spectra in the different pixels of the window so that all the windows within an area with a particular label are represented by a similar set of inputs. Thus, one is basically trying to obtain the smallest window that through some function that combines the spectra of its pixels provides a vision of the area that is spectrally homogeneous.

In terms of the neural networks employed, several types have been employed in our studies. Previous experiences show that Gaussian based ANN's produce very good results in classifying and segmenting hyperspectral images [7] [8], although in some cases simple multilayer perceptrons were used for the sake of simplicity.

## 2 Experimental Set-Up

The hyperspectral sensor used in this work was developed in the frame work of a research project on the developing of a complete, automatic, flexible, reliable, fast and precise real time detection and cartographic representation system for marine spills. The system's core is an instrument designed to be used from light aircraft or AUVs and was constructed by merging a lightweight hyperspectrograph with a series of positioning sensors for determining its geographical position and orientation. This push-broom type light imaging hyperspectrograph is an independent module developed to be used as a portable device on its own. Figure 1 displays its configuration.

It was made of commercially available components allowing a low-cost product. The slit allows just one line of the subject image which is decomposed into a spatial dimension plus a spectral dimension in the image plane. A 12 bit CCD camera acquires these images corresponding to a section of the hyperspectral cube. The instrument has an acceptable signal to noise ratio for wavelengths in the range between 430 and 1000 nm. The measured spectral resolution during calibration tests was of 5 nm while spatial resolution was about 30  $\mu\text{m}$  in the image plane. More details of the instrument can be found in [9]. In this paper the focus is on the processing methodology, thus the particular sensor used is not important, except in terms of the characteristics that affect the measuring precision.



**Fig. 1.** Hyperspectrometer schematics: (1) subject, (2)(4)(6) objectives, (3) slit, (5) grating, (7) CCD

A virtual instrument has been developed to implement the ANN based processing module plus a user friendly graphical interface for visualization, control and post-processing. Currently, the system is not installed on the industrial processing line. All measurements were carried out in our lab on different samples of andalusite provided by a mining company. Figure 2 displays a screenshot of the graphical interface. In this particular case four samples of different qualities of andalusite are being analyzed. The color image of the subject is reconstructed from the hyperspectral data and represented on a frame. Another frame represents the spectrum of the sample at the position of the cursor on the color image.

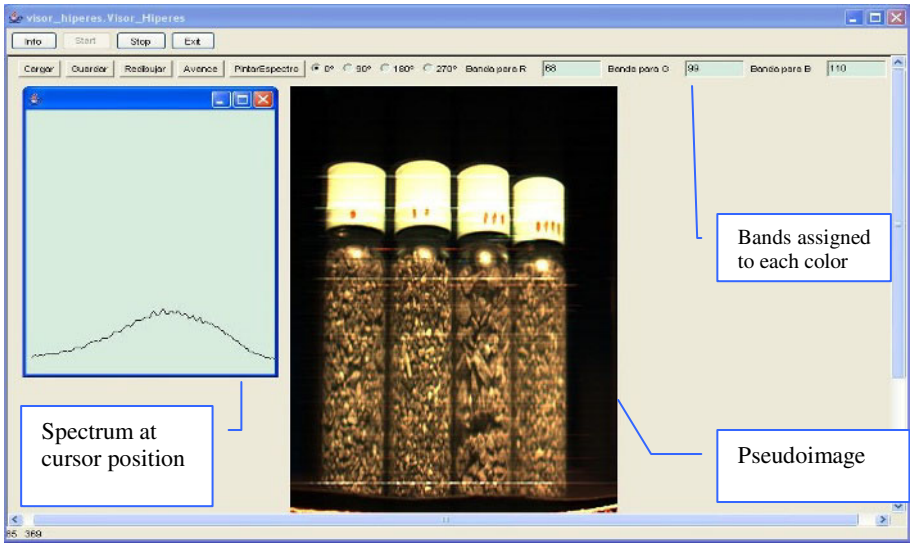


Fig. 2. Virtual instrument interface

### 3 Hyperspectral Image Segmentation Through ANNs

As in this particular application, we have initially relied on a supervised classification and endmember extraction system using different types of Gaussian based ANNs and a multilayer perceptron. Gaussian functions have been widely applied in ANN based image processing due to their noise filtering capabilities. A comparison of the results provided by three different types of these networks is presented in [5]. These types of networks are the radial basis function (RBF), the radial basis function with multiple deviation (RBFMD) neural networks and the Gaussian synapse based networks (GSBN). The first two types are structurally similar; the Gaussian function is applied over the inputs by using one parameter per synapse (center) and one parameter per neuron or per synapse in the hidden layer (deviation). The last one (GSBN) presents a multilayer perceptron structure, but replacing the simple weights used in the synapses

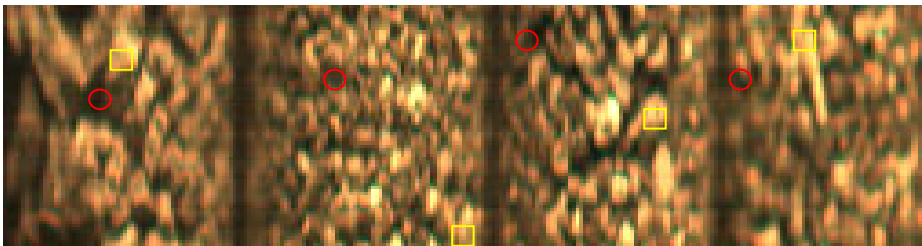
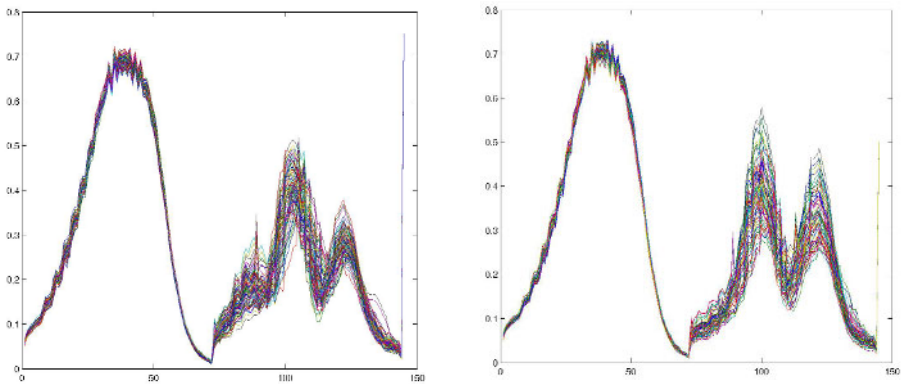


Fig. 3. Four areas with different types of andalusite. Pixelwise, all four areas have the same sets of spectra, just different distributions of them. The circles and the squares show areas where the spectra are the same.

by Gaussian functions in order to filter the inputs. The main difference between these three types of networks is the number of parameters to be trained, being the RBF the one requiring fewer parameters and the GSNB the one requiring most. This number of parameters determines the complexity of the learning process and the generalization capability.

These networks provide more complex and flexible decision boundaries, better adapted to the high dimensionality of the search space and the complexity of the subspace to discriminate in hyperspectral unmixing problems. The application of these networks increases the number of parameters that must be trained and, consequently, the computational cost but it should be compensated by the decrease of the minimum necessary network size and increases the speed of the network training stage while using fewer training samples.



**Fig. 4.** Average spectrum and deviation spectra for the correct andalusite composition and grain sample (right) and a different one (left)

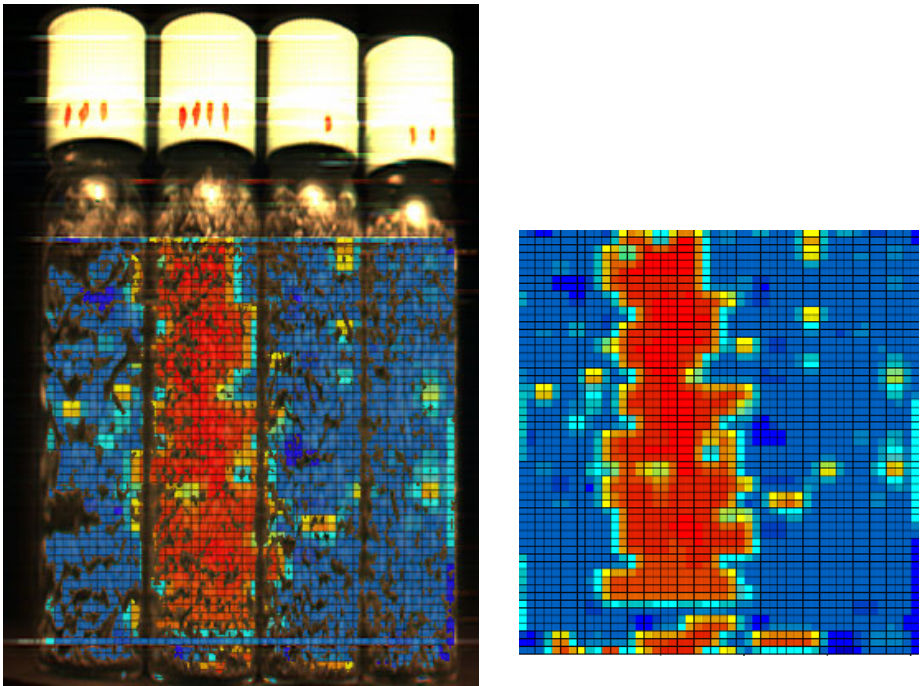
Thus, as an initial test, the andalusite images were processed in a pixelwise manner using these types of networks. The results obtained were quite poor, well below 40% classification accuracy. After analysis of the results it was concluded that the problem lies in the images themselves. The andalusite samples all contained the same elements (as shown in figure 3 for four samples) the only difference between them was how they were mixed. Thus, pixel by pixel and with the resolution of the images available there was no way of discriminating between the four samples. The experts would say that sample two was andalusite with the right granularity and the appropriate proportion of andalusite to the other minerals present and the other two had different proportions of andalusite with a different granularity. But experts were talking about an areawise appreciation and not a pixelwise one. In terms of training the neural networks, what was happening was that they were being presented with spectra that corresponded to the same compound and told to train towards different labels. As a result, the networks ended up producing an average of the labels which led to very low classification accuracy. The conclusion of the study was that spatial or areawise information must be taken into account.

As a first approximation to test this hypothesis, instead of processing individual pixels, the images were divided into windows of different sizes, and it was determined

that the smallest window size that allowed a correct classification was 20x20 pixels. Consequently, in order to train the networks the images were transformed in the spatial dimension by taking 20x20 pixel blocks (with their corresponding spectral dimension). These windows were overlapping by 10 pixels.

Obviously, dividing the image into windows is useless if one does nothing to combine the information in the pixels that make it up. That is, one would like to obtain from the information present in the spectra corresponding to the window pixels some type of descriptor that uniquely defines them as a set. The most obvious approach would be to average the spectra for all the pixels. Averaging is a way of losing information present in the signals, and that is what happened, most areas ended up having average spectra that were so similar that it was very difficult to discriminate them. This can be appreciated for the spectra of figure 4 where the left side of each graph displays the corresponding average spectrum.

After this we studied the deviations of the different points with respect to the average spectrum (right side of the graphs of figure 4). Here it can be seen that the shape of these deviations is different for different samples. Consequently, we took the average deviation spectrum in the window together with the average spectrum and used this as inputs to the networks. What is relevant here is that these descriptors are discriminant for different classes and similar enough for the same class for a network to be able to classify it easily.



**Fig. 5.** Classification results of a test case having 4 samples of andalusite of different quality. The system is able to detect the right one (red).

One of the most important aspects of this work is that once this type of spatial integration preprocessing was carried out, a simple multilayer perceptron with 144 inputs (the average and deviation spectra), 2 hidden layers with 20 neurons each and one output could obtain on a test set of images better than 98% classification accuracy.

The test case presented consists of four samples of andalusite of different quality extracted at four different points along the processing line. Only one of them (the second one from the left) is of the right purity and granulometry. Figure 5 displays a pseudocolor image where red represents positive match for the appropriate andalusite grain and composition and blue no match. This image clearly shows that the samples are classified correctly and there are only a few false negatives in the border of the bottle where there are reflections in the glass from the illuminating light source that induce false spectra.

## 4 Conclusions

An ANN based hyperspectral segmentation system has been developed to perform a quality control procedure in an andalusite processing line. The problem of the spectral inhomogeneity of the areas labeled by experts as ground truth has been addressed in a statistical manner through the determination of the smallest window that holds an appropriate combination of the spectra of the pixels it contains. Thus, the use of the average spectra and the average deviation spectra provides an input information to the networks that is homogeneous enough in each area for the networks to be easily trained. The behavior of different types of ANNs has been analyzed. The preprocessing procedure allows a large reduction in the complexity of the networks that make up the classifiers in addition to making the training procedure more reliable and faster. This classifier operator has been implemented in a virtual instrument with a graphical interface for the representation of results and for easily performing some basic post-processing operations. After training, the ANN based segmentation system provides on the test set classification accuracy better than 98%.

We are now working on the automation of the determination of the optimal window and spectra combination function through the creation of a training algorithm that includes these terms through a supervising ANN that regulates window size and momentum combination as training takes place.

## Acknowledgements

This work was funded by the MEC of Spain through project VEM2003-20088-C04-01 and Xunta de Galicia through project PGIDIT03DPI099E.

## References

- [1] D.L. Glackin and G.R. Peltzer (1999). Civil, Commercial, and International Remote Sensing Systems and Geoprocessing. American Institute of Aeronautics and Astronautics.
- [2] G. Shaw and D. Manolakis. (2002) "Signal processing for hyperspectral image exploitation," IEEE Signal Process. Mag., vol. 19, pp. 12 - 16, Jan. 2002.



- [3] C.T. Willoughby, M.A. Folkman, and M.A. Figueroa. (1996) "Application of hyperspectral imaging spectrometer systems to industrial inspection". In *Three-Dimensional and Unconventional Imaging for Industrial Inspection and Metrology*, Proceedings of SPIE, Volume 2599, pp. 264-272, January 1996.
- [4] McCracken, W.H., and Kendall, T., 1996, *Andalusite review 1995: Industrial Minerals*, no. 346, p. 53-54.
- [5] Landgrebe, D., *Indian Pines AVIRIS Hyperspectral Reflectance Data: 92av3c*, 1992. available at <http://makalu.jpl.nasa.gov/>.
- [6] Guatieri, J.A., Bechdol, M., Chettri, S., Robinson, J.W., Garegnani, J., Vermeulen, A. and Antonille, S. *From Spectra to Classification*. Extended paper of the presentation at 1st International Symposium on Hyperspectral Analysis, Sept. 20-22, 2000, Caceres, Spain.
- [7] A. Prieto, F. Bellas, R.J. Duro, and F. Lopez-Peña. "A Comparison of Gaussian Based ANNs for the Classification of Multidimensional Hyperspectral Signals". *Lecture Notes in Computer Science*. Vol. 3512, pp. 829-836, June, 2005
- [8] J.L. Crespo, R.J. Duro, & F. López Peña, "Gaussian Synapse ANNs in Multi and Hyperspectral Image Data Analysis". *IEEE Transactions on Instrumentation and Measurement*. Vol. 52, Issue 3, pp. 724-732, Jun. 2003.
- [9] F. López Peña & R.J. Duro, "A Hyperspectral Based Multisensor System for Marine Oil Spill Detection, Analysis and Tracking". *Lecture Notes in Artificial Intelligence*, Vol. 3213, pp. 669-676, September, 2004

# A Windowing/Pushbroom Hyperspectral Imager

B. Couce, X. Prieto-Blanco, C. Montero-Orille, and R. de la Fuente

Universidade de Santiago de Compostela, Departamento de Física Aplicada,  
Escola Universitaria de Óptica e Optometría E-15782 Santiago de Compostela, Spain  
faloraul@usc.es

**Abstract.** We show that any pushbroom hyperspectral imager can be converted into a windowing one by only attaching in front of it a simple dispersive element. The resulting device displays the same spectral resolution than the former one. To test this idea we have built a pushbroom imaging spectrograph in our laboratory. A plane transmission grating has been used as the dispersive element. We point out the main characteristics of these devices and we show some illustrative results.

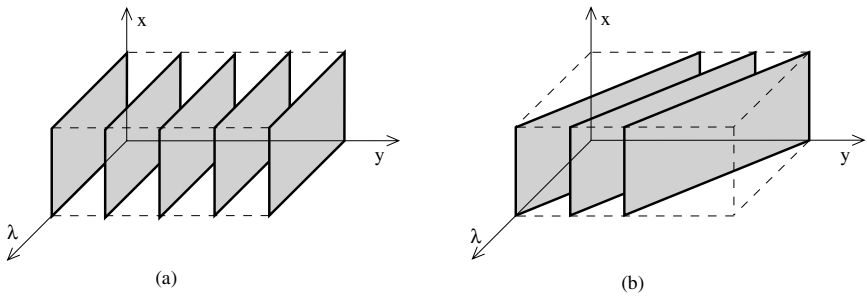
## 1 Introduction

Hyperspectral imagers or imaging spectrometers are radiation sensors that provide a continuous collection of spectral images of an inhomogeneous scene, allowing to obtain a spectral signature of each point in the scene. They can be applied to perform many different tasks such as accurate mapping of wide areas, object identification and recognition, target detection, process monitoring and control, clinical diagnosis imaging and environment assessment and management. Application areas include forestry, geology, agriculture, medicine, security, manufacturing, colorimetry, oceanography, ecology and others[1].

A standard color picture can be regarded as three superimposed gray-level images taken at three broad bands of the visible spectrum: red, green and blue light. Hyperspectral images on the other hand consist of a set of tens or hundreds separated images taken at narrow ( $<10$  nm) and contiguous spectral bands. A hyperspectral image is therefore a data cube with two spatial dimensions and one spectral or wavelength dimension.

As three-dimensional detectors do not exist, a hyperspectral imager has to take a stack of images to build the data cube. Different techniques have been envisaged and different devices have been developed for this purpose[2]. In static applications a two dimensional image can be taken at a given wavelength with a suitable spectral filter. A collection of images in different spectral bands can be obtained using a large set of filters or a tunable one[3]. This staring technique can be applied for remote sensing but for large scene or moving objects other techniques that scan the object are also used[4]. Scanning methods can be classified according to the dimension of the object imaged in a single frame: they are whiskbroom, pushbroom and windowing. In hyperspectral systems known as whiskbrooms, an image of one point is dispersed by a spectrometer and detected by a linear array detector. Scanning has to be performed in two spatial dimensions. In most frequently used systems known as pushbrooms, a stripe of the

sample is imaged into the spectrometer. This stripe is dispersed and detected by a 2-D array detector. Doing so a two dimensional image with the spectral dimension and one spatial dimension is acquired. The other spatial dimension has to be scanned to complete the data cube. In airborne or space-borne systems the platform movement itself can provide the scanning. Another alternative is to use a scan mirror. Finally, a windowing instrument scans also the scene in one dimension as the pushbroom ones do but it acquires a two dimensional image in a single frame. A typical windowing imager is very simple: it employs a conventional camera with a wedged or linear variable filter placed just in front of the 2D detector head. The wavelength of light transmitted by the filter varies with position providing for the spectral dimension. In fig. 1a and 1b we sketch two-dimensional cuts of the data cube corresponding to a single image taken with both pushbroom and windowing imagers. In fig. 1a a frame corresponds to a cut of the data cube with a plane containing one spatial dimension and the spectral dimension. In this cut there is not mixing of spatial and spectral information: we have the spectrum of a line (dimension  $y$  is constant in each cut). In fig. 1b we visualize a slice corresponding to an individual image obtained using windowing imaging. There is still a spatial axis in each frame but the other axis is a combination of spectral and spatial dimensions. In contrast with pushbroom windowing imaging provides an individual image which contains information of a two dimensional portion of the sample. That is, several lines of the sample are imaged simultaneously with each line imaged in a different spectral band.



**Fig. 1.** The data cube and cuts corresponding to single frames for a pushbroom spectral imager (a) and for a windowing one (b)

In this paper we are concerned with a new type of windowing imager. It can be built with a slight modification of a pushbroom imager. The proposed system includes a dispersive element located just in front of the imaging optics of the pushbroom system. This dispersive element mixes 2D spatial and spectral information coming from the scene into the entrance slit of the spectrometer. Such one-dimensional signal is further decomposed in the spectrometer so data are displayed in a two dimensional array, as usual.

We are aware of the complexity of this windowing imager compare to a conventional one. Nevertheless it presents some advantages. First, our instrument

can be easily transformed into a pushbroom imager by removing the dispersive element. Second, as conventional windowing imagers include a variable spectral filter located very near the sensor head of the instrument to perform wavelength discrimination. From a technological point of view the manufacture of such an element and its attachment to the detector is a hard task. In contrast our system uses ordinary dispersive elements such as gratings or prisms that are commercially available at a low cost.

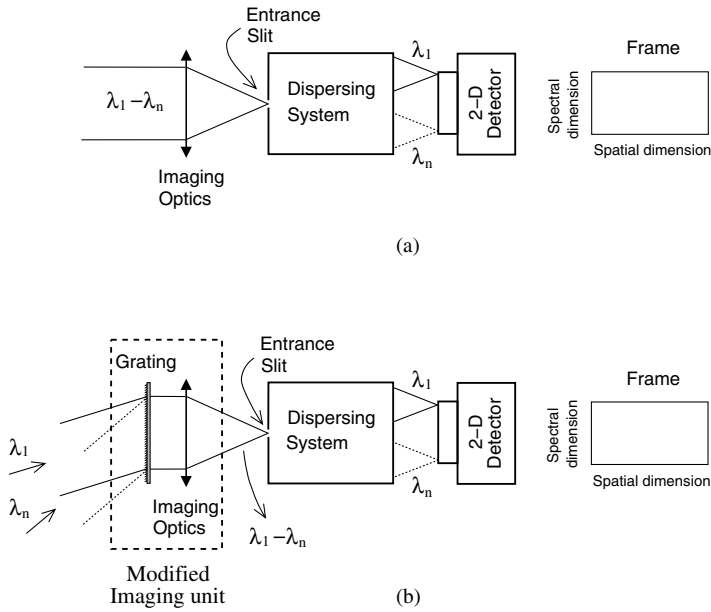
## 2 Description of the Windowing Imager

In fig. 2 we show our proposed windowing imager as a modification of a typical pushbroom imager. In fig. 2a we show the latter. A stripe of a distant object is imaged into the entrance slit of the spectrometer. Light entering the spectrometer is first collimated, spectrally separated by a dispersive element (such as a diffraction grating or a prism) and finally focused into a two dimensional detector. If the spectrometer slit is vertical, dispersion occurs in the horizontal direction. The spatial and spectral dimension are separated so that a cut of the data cube corresponding to an image is as in fig. 1a. In our proposal we include another dispersive element just before the imaging system (see fig. 2b). Now light from a given line of the sample and with a given wavelength is refracted so it goes to the slit. But light with a different wavelength and coming from the same line fails to pass the slit. Furthermore if we consider another line in the sample, light reaches the dispersive element from a different direction with respect to the first line. This means that light of a different wavelength will be deflected to the slit. If we consider now light on the slit, it will be white light with different spectral bands and coming from different lines on the sample. This light is dispersed in the spectrometer and focused to the detector, as usual. Since light from different stripes on the sample has a different spectral content it will be decomposed and directed to different positions on the detector according to its wavelength. In this way we get a image with a spectral and spatial mixture as discussed and sketched in fig. 1b.

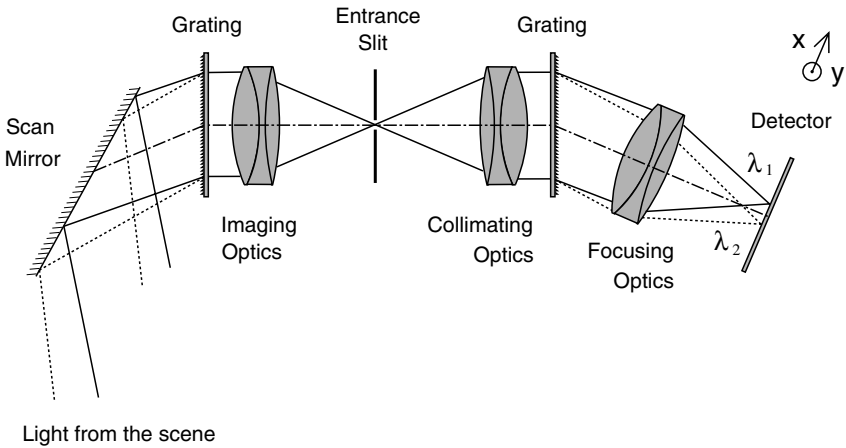
Despite that both pushbroom and windowing techniques fill the whole data cube the windowing cut could be advantageous for some applications. For example, consecutive scanned frames can be easily compared to detect and correct unexpected displacements, rotations or scale changes.

## 3 Laboratory Set-Up

We have tested our proposal with a simple device built in our laboratory. The set-up of the developed spectral imager in the windowing mode is shown in fig. 3. A rotating plane mirror is used to scan the sample. Light from this mirror is reflected to a transmission diffraction grating with a groove density of 300 grooves/mm. An objective lens with a focal length of 17 mm images onto a slit 20  $\mu\text{m}$  wide the portion of the sample diffracted by the grating in the proper



**Fig. 2.** Scheme of a typical pushbroom imager (a) and a windowing imager obtained by attaching a grating to the previous one (b)



**Fig. 3.** Lay-out of our experimental device in windowing mode

direction. The slit acts as a field stop allowing only light from a thin stripe to enter the spectrometer. Light is collimated by another objective lens ( $f = 35 \text{ mm}$ ) and passes through a second diffraction grating (with 300 grooves/mm). A third objective lens, identical to the previous one, images the light on a CCD camera (with pixel size of  $6.45 \mu\text{m}$  and  $1040 \times 1396$  pixels). For each height ( $y$  axis) in the camera a spectrum is recorded along the  $x$ -axis (short dimension in our sensor).

Each of these spectra corresponds to only one point on the slit but to different points on the sample. Alternatively, each vertical line on the CCD corresponds to an image of one stripe on the sample for a given wavelength. Different vertical lines onto CCD correspond to different wavelength and different stripes on the sample. Finally, a personal computer controls the acquisition process of the camera while the mirror rotation is performed by a manual controller.

## 4 Results and Discussion

First we have characterized the pushbroom configuration. The spectral band that fits on the camera in our current layout goes from 400 and 1000 nm. This band correspond to the working range of the camera sensor. Likewise the objective lenses are corrected to minimize chromatic aberration in this spectral range. However they have the disadvantage of a small aperture diameter. This implies vignetting on the image (decreasing of illumination at the detector borders). To reduce vignetting the diaphragm of the first objective lens (the aperture stop of the total system) has been closed to f-number 4.

For spectral calibration mercury and argon lamps were used. Both linear and parabolic fit of the position on the camera versus wavelength were carried out showing only little discrepancies in the extremes of the spectral range. Spatial calibration was performed by imaging a narrow dark line at different camera positions. Factors that determine spatial and spectral resolution are pixel size, aberrations and defocusing. For the spatial dimension along the scanning direction the resolution is also determined by the slit size and scanning speed. The slit size and the number of illuminated grooves in the gratings also affect spectral resolution.

Both spatial and spectral resolutions have been measured. The resolution in the spatial dimension is three pixels in the image plane whereas the spectral resolution is four pixels. This means for example that we can image a 1 Km distant sample resolving a square of 1.3 m side and with wavelength resolution better than 2.5 nm.

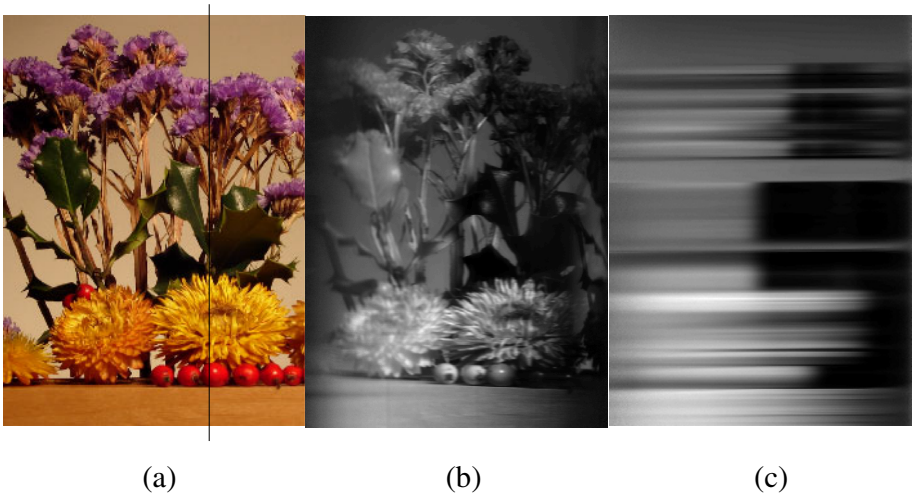
Imaging spectrographs are also commonly affected by two distortions known as "smile" and "keystone"[5]. Smile is a change in dispersion with field position, i.e. a monochromatic image of the slit in the camera is not a straight line but a curved line. Using spectral lines of Hg and Ar lamps the maximum difference in position for a single wavelength was measured to be less than 7 pixels in the infrared region. Keystone refers to the change of magnification with wavelength. Measured keystone was less than one pixel.

Note that we usually take an image with binning  $4 \times 4$  to increase the acquisition speed and to limit the memory space used. That is, we have a super pixel with an area equal to  $4 \times 4$  pixels of the current camera.

When the grating is introduced the spectral resolution remains unchanged. If the object can be considered to be at an infinite distance as in remote sensing applications this grating does not introduce any spatial resolution loss. So it is expected a spatial resolution of  $3 \times 4$  pixels. The picture at the sensor surface

shows a modified aspect ratio as a result of a different focal length of first and second lens of our system. By using a symmetric configuration with respect to the slit the image distortion would be only affected by the Seidel distortion of the third lens.

A still life was prepared to illustrate the operation of our imaging spectrometer. An image of this scene taken with a digital photographic camera is seen in fig. 4a. The scene includes violet (top) and yellow (middle) dry flowers, holly leaves (amongst the flowers) and holly berries (bottom). A hyperspectral image of the same scene was snapped with our system in windowing mode and normalized respect to the image of an uniform white surface to reduce vignetting (see fig. 4b).



**Fig. 4.** Scene discussed in main text (a) and pictures taken with our spectral imager in windowing mode (b) and pushbroom mode (c). In (b) the aspect ratio has been modified to easily compare with (a). Spectra shown in (c) correspond to the vertical line displayed in (a).

Left and right borders in fig. 4b correspond to wavelengths of 1000 nm and 400 nm respectively, whereas the limit between visible and infrared regions is a vertical stripe located about the center of the image. The scene was located near the instrument so the image sharpness is limited by aberrations introduced by the first grating. In fig. 4b we also observe an undesirable ghost image located at the right side of the picture (the blue end). This ghost appears because of second order diffraction in the first grating. A similar effect is generated in the second grating of pushbroom spectrometers. In both cases it can be eliminated with a suitable software or using more efficient gratings or replacing them by prisms.

In spite of the above considerations, it can be easily appreciated in fig. 4b some details that reveals the characteristics of the imaged scene. First, the white and uniform background in the scene appears in the whole image including the IR region, despite vignetting. Second, objects in the scene show more reflectance in

the IR band than in the visible band. This is most outstanding in the case of holly leaves and violet flowers which display a dark color. Finally, we note that in the visible range holly berries appears brighter in the spectral band between 600-700 nm and yellow flowers appears brighter in the spectral band between 500-600 nm, in perfect agreement with their color. These spectra are seen more clearly in fig. 4c taken in pushbroom mode.

## 5 Conclusions

A pushbroom hyperspectral imager can be easily converted into a windowing one by attaching to it a suitable dispersive element. We shown the feasibility of the method with a simple imaging spectrograph built in our laboratory. This device has a spatial resolution about 4 pixels in the image plane and a spectral resolution less than 2.5 nm in the spectral range from 400 to 1000 nm.

This work has being carried out in the framework of the contracts VEM2003-20088-C04-03 and PGIDIT04PXIC22201PN. We thanks the Spanish Ministry of Education and Science (MEC) and the Autonomous Government of Galicia (Xunta de Galicia) for its financial support.

## References

1. Shaw, G.A., Burke, H.K.: Spectral Imaging for remote sensing. Lincoln Laboratory Journal **14** (2003) 3-28
2. Sellar, R.G., Boreman, G.D.: Classification or imaging spectrometers for remote sensing applications. Opt. Eng. **44** (2005) 1-3
3. Gat, N.: Imaging spectroscopy using tunable filters: a review, Vol. 4056. Proc. of SPIE, (2000) 50-64
4. Nieke, J., Schwarzer, H., Neumann, A., Zimmermann, G.: Imaging Spaceborne and Airborne Systems in the Beginning of the Next Century. Conference on Sensors, Systems and Next Generation Satellites III, Vol. 3221. Proc. of SPIE, (1997) 581-592
5. Mouroulis, P., Green, R.O., Chrien, T. G.: Design of pushbroom imaging spectrometers for optimum recovery of spectroscopic and spatial information. Appl. Opt. **39** (2000) 2210-2220



# Genetic Programming of a Microcontrolled Water Bath Plant

Douglas Mota Dias<sup>1</sup>, Marco Aurélio C. Pacheco<sup>1</sup>, and José Franco M. Amaral<sup>2</sup>

<sup>1</sup> ICA - Applied Computational Intelligence Lab  
Electrical Engineering Department

PUC-Rio - Pontifícia Universidade Católica do Rio de Janeiro  
R. Marquês de São Vicente 225 Gávea, Rio de Janeiro, CEP 22453-900  
RJ, Brazil

{douglassm, marco}@ele.puc-rio.br

<sup>2</sup> Department of Electronics Engineering

UERJ - Rio de Janeiro State University  
R. São Francisco Xavier, 524, Maracanã, Rio de Janeiro, CEP 20550-013  
RJ, Brazil  
franco@uerj.br

**Abstract.** Typically, control system design leads to a higher-order non-linear function of the system's state variables. As a result, it is very hard to find a satisfactory mathematical solution. On the other hand, considering a microcontroller based implementation, another difficulty is to program it to carry out the desired control algorithm. This paper presents the application of linear genetic programming in the automatic synthesis of a microcontroller assembly program, which performs an optimized control of a water bath plant. The synthesis starts from the plant's mathematical modeling and supplies directly a assembly code for the microcontroller platform. When comparing the control performance of the synthesized program with that of a neuro-fuzzy based controller, the synthesized program proved to perform slightly better.

## 1 Introduction

The difficulties encountered in the course of traditional developing a digital control system may be divided according to two stages: design and implementation.

At the designing stage, problems arise during the mathematical development of a solution to the control problem. Optimal performance normally requires a higher-order nonlinear function of the system's state variables. As a result, it is often not possible to find a satisfactory mathematical solution.

As for the implementation stage, the difficulty is regarded to developing a program which is able to perform the control strategy. In the so-called embedded systems, it is somewhat difficult to program the microcontroller (MC) because of its small instruction set and the little memory space available.

In this work, we have considered the application of linear genetic programming (LGP) in the automatic synthesis of MC assembly programs, which would control a water bath plant in as optimized as possible way, only based on its

mathematical modeling through its dynamic equation. The aim was to show that this approach would be able to automate the developing process, overcoming the cited traditional development difficulties, by synthesizing a satisfactory control strategy for a typical industrial plant example.

Section 2 presents the MC adopted as platform, section 3 exposes some aspects about evolution of assembly programs, while section 4 briefly describes the considered water bath plant. Section 5 explains the operation of the evolutionary system. Section 6 exposes the final results. Finally, section 7 presents conclusions and possible future work.

## 2 Microcontroller Platform

The Microchip's PIC18F452 [1] has been adopted as the microcontroller platform. It is an 8-bit RISC MC, with a simple architecture and largely employed in several control systems.

This MC has a set of 75 instructions, 22 of which were selected for this work, the same ones of our earlier work [2]. This choice includes arithmetical and conditional operations, as required for this control problem.

In addition, there are five input/output ports, three of them with 8 bits. All these bits can be configured individually as input or output pins.

## 3 Evolution of Assembly Language Programs

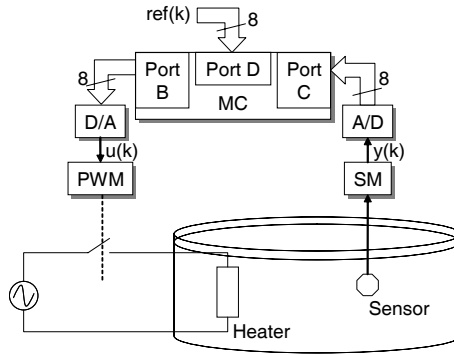
The emergence of GP in the scientific community arose with the use of the tree-based representation, in particular with the use of LISP, a functional language, in the work of Koza [3]. However, GP systems manipulating linear structures do exist [4], which have shown experimental performances equivalent to Tree GP. The so called linear genetic programming (LGP) evolves imperative language programs, i.e. assembly.

Assembly programming is frequently used when efficient solutions are necessary, as is the case of applications that are strongly restricted in terms of run time and memory use. In general, the reasons for evolving assembly as opposed to high-level languages are similar to the reasons for the use of manual programming in assembly. This is the level at which the most efficient optimization of code is obtained.

## 4 Water Bath Plant

This system is used as a case study in several works, like in [5] and [6]. The system simulated in this work is based in a real plant, described in [6] and is shown in figure 1.

In this hypothetical experiment, the MC reads the 7 liters bath's temperature through the sensor module and the 8-bit A/D converter. The MC also reads the 8-bit reference value. The plant's input signal, provided by the MC and the 8-bit



**Fig. 1.** The water bath plant controlled by a MC

D/A converter, ranges between 0 and 5 volts and controls the duty cycle of the 1.3kW heater, through the PWM module.

The objective of this experiment is to evolve a program that controls the heater so the bath's temperature follows a reference (set point) profile as near as possible. This profile rates 35°C ( $0 \leq t < 20$  minutes), 50°C ( $20 \leq t < 40$  minutes), 65°C ( $40 \leq t < 60$  minutes) and 80°C ( $60 \leq t \leq 90$  minutes).

The mathematical modeling demonstrated in [6] results on equation 1.

$$y(k+1) = a(T_s)y(k) + \frac{b(T_s)}{1 + e^{0.5y(k)-\gamma}}u(k) + [1 - a(T_s)]Y_0 \quad (1)$$

where  $a(T_s) = e^{-\alpha T_s}$  and  $b(T_s) = (\beta/\alpha)(1 - e^{-\alpha T_s})$ . The parameters in this simulation are  $\alpha = 1.00151 \times 10^{-4}$ ,  $\beta = 8.67973 \times 10^{-3}$ ,  $\gamma = 40$ ; and  $Y_o = 25^\circ\text{C}$  (environment temperature). The sampling period is  $T_s = 30$  seconds and  $k$  is the discrete time index.

## 5 Evolutionary System

### 5.1 Representation and Population Initialization

Each gene of the linear genotype represents an instruction that is composed of the operation and of one or two operands.

The population initialization is the first stage of the GP. First, an individual is created through the random selection of a length between the minimum and the maximum initial length parameters. Next, all its genes are filled in with randomly generated instructions based on the syntax defined for assembly language. This process is repeated for all the individuals of the population.

### 5.2 Genetic Operators

Linear crossover consists of an exchange of code segments between two parents. It operates selecting a random segment of genes in each parent. These genes are

exchanged between the two parents and as a result, the descendant individuals are generated.

After the crossover, each offspring is submitted to mutation. When an individual is selected for mutation, this operator initially selects one of the individual's genes and changes its content. The type of change is randomly selected and may consist of changing the operation or one of the two operands. This approach was based on [4], where it is called "micromutation" and its purpose is to smoothen the effect of this operator during evolution.

### 5.3 Selection and Evolution

This system has adopted tournament selection and steady-state evolutionary algorithm based on [7].

This algorithm works by randomly selecting four individuals from the population and grouping them in two pairs. For each pair, the two individuals are evaluated and their resulting fitness are compared with one another. The two winners, one from each pair, are probabilistically submitted to mutation and crossover and their offsprings replace the two losers. This process repeats until the maximum number of cycles is achieved.

### 5.4 Operation

The evolutionary process begins with the population initialization, as detailed in section 5.1. Next, the evolutionary kernel selects the individuals that are to be evaluated, evaluates them and submits the best ones to the genetic operators.

When the evaluation of an individual begins, the plant simulator is placed in a given initial state. The individual is then repeatedly executed, once for each plant simulation time step, while the set point's value is varying. Section 5.5 details the evaluation function. This process is repeated until the last time step for this simulation is achieved. In this manner, an individual's fitness is obtained.

This evolutionary process repeats until the last cycle is reached. The program that presented the best fitness among all the executed cycles is considered the evolved one.

### 5.5 Evaluation Function

Firstly, the individual's fitness is initialized to zero, since it will accumulate the absolute error along the  $k$  samples. Next, the reference ( $ref(k)$ ) and temperature ( $y(k)$ ) variables are A/D converted ( $ref_d(k)$  and  $y_d(k)$ ). Then, the individual is executed by the MC simulator, which takes  $ref_d(k)$  and  $y_d(k)$ , and returns the control action ( $u_d(k)$ ). Next,  $u_d(k)$  is D/A converted ( $u(k)$ ). The plant simulator takes  $u(k)$  and  $y(k)$ , and returns the updated plant's temperature for the  $k + 1$  sample ( $y(k + 1)$ ), as a result of the control action applied to the heater. Finally, the fitness related to the  $k$  sample is computed. This whole described process is repeated until the last  $k$  sample is achieved.

In order to avoid evolving programs that would activate the heater prematurely, it was necessary to multiply the difference  $ref(k) - y(k)$  by 10 when it was negative. Therefore, function  $f(ref(k) - y(k))$  rates 10 times its argument's value when it is negative and rates its same argument's value when it is positive ((2) in table 1).

## 6 Results

The training cases are 180 points  $(k, ref(k))$  that represent the temperature profile. Since  $k = 30$  seconds, the complete evaluation represents 90 minutes of the system's simulation. Table 1 resumes this experiment, where M is the population's size, and  $p_c$  and  $p_m$  are the crossover and mutation probabilities, respectively.

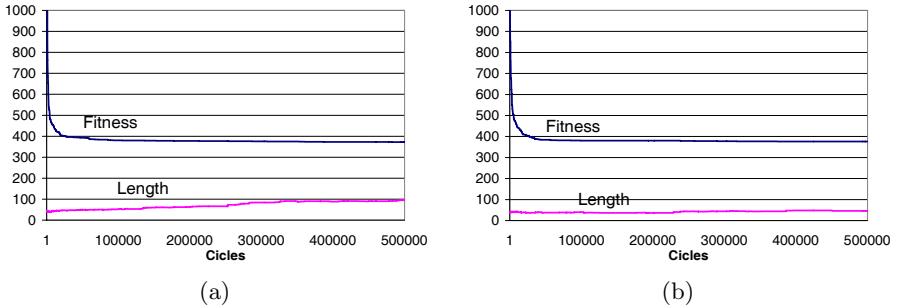
**Table 1.** Summary of the evolutionary experiment

Objective:	To synthesize an assembly program that enables a PIC18F452 MC to control a water bath plant's temperature by following a given temperature profile as near as possible.
Terminal set:	R (reference), Y (temperature) and U (control); A1 and A2 (auxiliary); PRODH and PRODL (multiplier hardware registers); STATUS (flag register).
Function set:	The 22 PIC instructions described in [2].
Training cases:	The 180 points $(k, ref(k))$ of the temperature profile.
Fitness:	$\sum_{k=1}^{180}  f(ref(k) - y(k)) , f(x) = \begin{cases} 10x & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \quad (2)$
Evaluation:	The lowest fitness. As tie-breaking criterion, the shortest length of the individual.
Main parameters:	M = 500. Steady state. Number of individuals processed: 400,000. $p_c = 0.8$ , $p_m = 0.4$ .

Like in [2], a heuristic was used for the purpose of encouraging the evolution of individuals with optimized code sizes. When two individuals have the same fitness (same control performance), their lengths are compared and the shortest one wins the tournament.

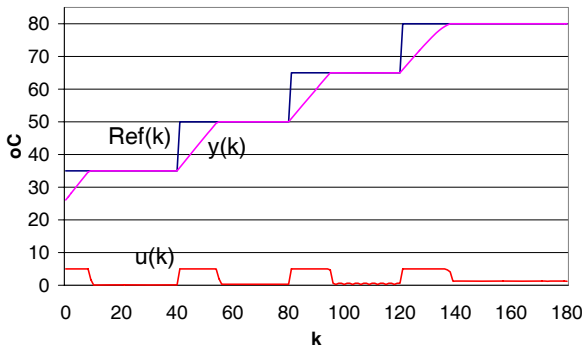
Figures 2(a) and 2(b) show the evolutionary graph, averaged over ten independent runs, without and with individuals' length control, where it is clear the tendency of quick and undefined growing of the code's length. This is a familiar phenomenon in GP called "bloat" [8]. It may be noticed that the bloat problem was satisfactorily controlled with the use of this heuristic without hampering the fitness' evolutionary performance.

The best evolved individual is a ready-to-use assembly code for the microcontroller platform which presented a fitness of about 370. This result is quite



**Fig. 2.** Evolution of the experiment, without (a) and with (b) individuals’ length control

similar to the one in [5], where an ANFIS controller obtained a value of 374 for the same quality parameter. Figure 3 shows the resulting graph of the plant being controlled by the evolved individual.



**Fig. 3.** Result of the plant being controlled by the evolved program

The total evolution time for each run (500,000 cycles) was approximately 50 minutes (Pentium IV / 2.8GHz). Although ANFIS training would require less computational effort, one should note that the GP approach has the advantage of supplying a ready-to-use solution, increasing the automation level of the control system’s design as a whole.

The evolved code, having 33 instructions, was inserted in the program shown in listing 1.1. The objective of this program is to create a main loop that continuously reads the MC’s input ports and updates its output port. Its operation was validated by using the Microchip MPLAB simulator program.

Usually, results of evolutionary synthesis are hard to understand, even by specialists. However, when analyzing the evolved code, one can note that the evolution was parsimonious by using just one of the auxiliary registers made available: A1. It is also interesting to note that the instruction BTFSC (Bit Test reg. F, Skip if Clear), in all occurrences in the code, is always testing bit 7 of

its argument register. This shows the code's capability to analyze the register's numeric signal when deciding whether to skip or not the next instruction, since negative values of registers are represented by its two's complement, where bit 7 indicates its signal. This feature could be important for the code to detect whether the temperature is below or above the set point.

**Listing 1.1.** Listing of the program that contains the evolved code

```

#include <P18F452.INC>
CBLOCK 0x08          ; Start of user memory
R                   ; Reference
Y                   ; Temperature
U                   ; Control
A2                  ; Auxiliary
ENDC                 ; End of memory block

ORG 0x00            ; Reset vector
GOTO BEGIN          ; Initial execution address
BEGIN
SETF TRISD          ; TRISD:=11111111b. Set port D pins as inputs
SETF TRISC          ; TRISC:=11111111b. Set port C pins as inputs
CLRF TRISB          ; TRISB:=00000000b. Set port C pins as outputs
LOOP
CLRF W              ; W:=0. Initialize W
CLRF PRODH          ; PRODH:=0. Initialize PRODH
CLRF PRODL          ; PRODL:=0. Initialize PRODL
MOVFF PORTD, R     ; R:=PORTD. Read port D (reference)
MOVFF PORTC, Y     ; Y:=PORTC. Read port C (temperature)

INCF Y, W           ; Start of evolved code
MOVF U, W
NEGF R
RLNCF Y, F
RRCF Y, F
ADDWFC PRODL, F
ADDWFC R, W
BTFSC Y, 7
RLNCF Y, F
CPFSEQ R
INCF Y, W
INCF R, F
SUBWF U, F
INCF U, F
BTFSC PRODL, 7
RLCF R, F
BTFSC PRODL, 7
CPFSGT R
BTFSC Y, 7
RLCF R, F
SUBWF PRODH, F
INCF A2, F
CPFSEQ Y
BTFSC Y, 7
MOVWF U
MULWF U
INCF U, F
RLCF R, W
SUBWF PRODH, F
BTFSC R, 7
SETF U
CPFSGT R
ADDWFC U, F        ; End of evolved code

MOVFF U, PORTB     ; PORTB:=U. Write port D (control)
GOTO LOOP          ; Return to beginning of loop
END

```

## 7 Conclusions and Future Work

The evolution of programs directly in assembly language has proved to be a promising approach for two main reasons:

1. Since it represents a device's lowest possible programming level, it enabled evolution to find solutions with optimized codes since the evaluation functions excelled in programs with smaller amounts of instructions;

2. In terms of evolution time, it made the real application of the system a viable alternative, since evolution in a higher-level language, such as C, for example, would require the compilation stage, which is executed at each evaluation, and this would represent a significant increase in the total evolution time and make it impossible to use the system.

As for the final results, the experiments demonstrated that the evolutionary system yields a good performance since it was able to automatically synthesize a control program whose performance was at least as good as the one found in the literature. The important aspect of this approach is that the solutions found for the control problems were obtained in the system's final implementation format, that is to say, an assembly program for a MC. This methodology may then be able to help system's designer to find an optimal solution bypassing the conventional programming stage.

A future work would be to investigate the use of MC instructions that make it possible to create execution loops that might contribute in some additional way to the quality of the evolved programs. This would also involve studying a manner by which to deal with undesirable infinite loops that might occur.

Another future work would be to subject the dynamic equation 1 parameters to a stochastic process which would approximate the real environment thus allowing the evolution of a microcontroller code able to cope with the deviations of the real system.

## References

1. Microchip: PIC18FXX Data Sheet. (2002) <http://www.microchip.com>.
2. Mota Dias, D., Pacheco, M.A.C., Amaral, J.F.M.: Automatic synthesis of microcontroller assembly code through linear genetic programming. In Nedjah, N., Abraham, A., de Macedo Mourelle, L., eds.: Genetic Systems Programming: Theory and Experiences. Volume 13 of Studies in Computational Intelligence. Springer, Germany (2006) 195–234
3. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA, USA (1992)
4. Brameier, M., Banzhaf, W.: Effective linear genetic programming. Technical report, Department of Computer Science, University of Dortmund, 44221 Dortmund, Germany (2001)
5. Lin, C.T., Lee, C.S.G.: Neural fuzzy systems: a neuro-fuzzy synergism to intelligent systems. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1996)
6. Tanomaru, J., Omatu, S.: Process control by on-line trained neural controllers. IEEE Transactions on Industrial Electronics **39** (1992) 511–521
7. Nordin, P.: Evolutionary Program Induction of Binary Machine Code and its Application. Krehl-Verlag, Münster, Germany (1997)
8. Banzhaf, W., Nordin, P., Keller, R.E., Francone, F.D.: Genetic Programming – An Introduction; On the Automatic Evolution of Computer Programs and its Applications. Morgan Kaufmann, dpunkt.verlag (1998)



# Symbiotic Sensor Networks in Complex Underwater Terrains: A Simulation Framework

Vadim Gerasimov, Gerry Healy, Mikhail Prokopenko, Peter Wang, and Astrid Zeman

CSIRO Information and Communication Technology Centre  
Locked bag 17, North Ryde, NSW 1670, Australia  
mikhail.prokopenko@csiro.au

**Abstract.** This paper presents a new multi-agent physics-based simulation framework (DISCOVERY), supporting experiments with self-organizing underwater sensor and actuator networks. DISCOVERY models mobile autonomous underwater vehicles, distributed sensor and actuator nodes, as well as multi-agent data-to-decision integration. The simulator is a real-time system using a discrete action model, fractal-based terrain modelling, with 3D visualization and an evaluation mode, allowing to compute various objective functions and metrics. The quantitative measures of multi-agent dynamics can be used as a feedback for evolving the agent behaviors. An evaluation of a simple simulated scenario with a heterogeneous team is also described.

## 1 Introduction

This paper presents a software simulation system for self-organizing underwater sensor and actuator networks. The simulation system provides a test-bed for co-evolution of distributed and mobile sensors and actuators, and required communication topologies. The broad aim is to develop *symbiotic* sensor/actuator networks which include agents recognizing and forming relationships of mutual benefit across various types: e.g., network nodes may assist navigation of submersible robots, while being powered by the robots.

The underwater sensor/actuator networks are intended to protect critical marine infrastructure and water resources. Examples of such safety-critical structures include offshore oil platforms, deep-ocean well heads, tankers, dams, bridges, pipelines, etc. Typical protection and response tasks comprise tracking oil spills to their sources, source identification and diagnostics (e.g., measurement of oil slick thickness), and actions such as burning, skimming, and dispersing. Offshore hydrocarbon exploration by sensor arrays for the identification of petroleum systems is another potential domain of interest.

Simulation of the underwater sensor/actuator networks should account for possible off-shore deployment on demand or in advance, as autonomous devices (nodes or submersible robots) with compound-specific chemical sensing, propulsion and actuation, acoustic and optical communication, and multi-agent self-organizing teamwork capabilities. An important scenario considered in this study is the deployment of a heterogeneous team with some primary agents (leaders) having different or more advanced sensors, and secondary agents with more powerful actuators following the leader(s) as

a backup team until there is a need for distributed actuation. A prey-and-predator variant of this scenario is a search and containment task with the primary agent being the target pursued by the secondary agents. In either case, the employed agents have to deal with a problem that changes concurrently with the problem-solving processes, and cooperate in solving tasks which are distributed over space (3D) and time, across complex underwater terrains. The development of self-organizing strategies, when an incremental loss of a portion of the network leads to an incremental loss in quality, rather than a catastrophic failure, is our main focus.

The following Section describes the simulation system created by the CSIRO DISCOVERY<sup>1</sup> project, developed to study symbiotic behavior in underwater self-organizing sensor and actuator networks. It is followed by preliminary experimental results (Section 3) and conclusions.

## 2 Simulation Platform

### 2.1 Architecture

The main requirements of the physics-based simulation of distributed multi-agent systems include simulation of mobile autonomous underwater vehicles, as well as distributed sensor and actuator nodes, and multi-agent *data-to-decision* (D2D) integration, ranging from data validation by individual agents to decision integration and action coordination by self-organizing sub-networks and networks of agents. An important part of the D2D integration is quantitative measures of multi-agent dynamics [13,14,6,17] which can be used as a feedback for evolving the agent behaviors across multiple runs. Such measures can use either full information on agents' states and their interconnections or work with partial information, obtained locally: *localizable* measures [16]. Of course, localizable measures can be embedded in the agents themselves and be accessible to local "hierarchs" (e.g., cluster-heads [10,15]), controlling agent behaviors during run-time via an adaptive feedback.

The DISCOVERY multi-agent physics-based simulation platform supports the following components: 1) Simulator; 2) Visualizer; 3) Agent; 4) Metric-Analyzer. A simulation session is carried out in client-server style. The Simulator (server) provides a domain (a virtual environment), simulates all the actions of objects in this domain and controls a scenario according to a set of rules. This is a well-known approach to simulation, used, for example, in the RoboCup Simulation League [9,2,1]. The characteristics of the Simulator are specified by a set of server parameters, e.g., the amount of noise added to sensory perceptions and the maximum speed of an agent.

Agents are controlled by autonomous client programs which connect to the server through a specified port. Each client program can control a single agent. All communication between the server and the clients is done via TCP/IP sockets. Using these sockets, client programs send requests to the server to perform an action (e.g. "thrust"). When the server receives such a message it handles the request and updates the environment accordingly. Upon an agent's request, the server also sends sensory information

<sup>1</sup> CSIRO: Commonwealth Scientific and Industrial Research Organisation, Australia.

DISCOVERY: Distributed Intelligence, Sensing and Coordination in Variable Environments.

about the agent's neighbourhood to the agent. Clients communicate with each other indirectly, via the server, using messaging protocols which restrict the communication.

The server is a real-time system using a discrete action model, i.e., working with discrete time intervals (cycles) of a specified duration, e.g., 10 ms. During this period clients can send requests for agent actions to the server. At the end of a cycle the server executes the actions and updates the state of the world. Sending no request during a given cycle means that the agent misses an opportunity to affect its current dynamics. Sensing and acting are asynchronous: clients can send action requests to the server once every cycle, but they receive information on requests. This information is fragmented, limited and degrades with the distance.

Simulator also supports an evaluation mode, allowing to compute objective functions and metrics. In this mode, each agent regularly updates Simulator with a predefined set of agent's internal parameters, enabling calculations on both local and global levels. For example, it is possible to compute entropy of agents' states and characterize diversity of their behaviors. In addition, the evaluation mode includes computation of spatiotemporal distribution of agents, etc. The collected data can be used by Metric-Analyzer offline, and contribute to genetic algorithms evolving agents between experimental runs.

Visualizer displays the virtual world, being connected to the Simulator via TCP/IP. Although similar to an agent, it has no physical representation in the simulated environment and uses a different set of commands. The Simulator sends information to the Visualizer each cycle or upon request, containing the current state of the world. Visualizer also provides a visual interface to the server in order to specify, start, pause and stop a scenario.

The environment Simulator is the most computationally-intensive part of our software system where the performance is critical. There is a range of libraries available for physical simulations, however, we were not able to identify any library or simulator that would adequately cover a required combination of fluid dynamics and kinematics, as well as the right balance between the precision and performance of the simulation system. While we adopted some design and implementation ideas from the available 3D simulation systems such as ODE, Gazebo/Stage, Juice, Webots, the Simulator was developed based on our own set of routines satisfying own coding requirements and standards. A socket-based multi-agent sever-client communication suite (DBP-MAP: Deep Behaviour Projection Multi-Agent Platform), developed earlier by CSIRO [11,12], was our starting point in developing communication architecture for the simulation system. The project also builds on the expertise developed within CSIRO Underwater Robotics [5], CSIRO Directed Self-Assembly in Multi-Agent Networks [8], and CSIRO-NASA Ageless Aerospace Vehicles [17,13,14] projects. Considering the requirement of cross-platform compatibility, Java3D is a good choice for the development of the combination of the rendering engine and user interface. As the Visualizer relays the data incoming from the Simulator to the native graphics engine without any significant calculations, Java performance penalty is not significant.

## 2.2 Terrain and Collision Modelling

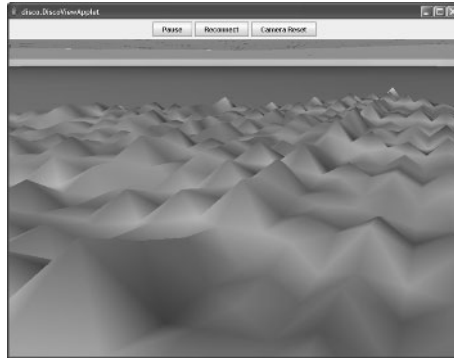
Terrain of the seabed is modelled using a rectangular mesh. The terrain can be randomly generated or user-defined by creating a Height Map, a gray scale raster image

file (pixel, not vector based), in which the RGB value of each pixel is mapped to a corresponding vertex height, thus creating a three dimensional mesh which represents the seabed.

In our terrain model vertices are equidistant along the  $x$  and  $z$  axis, while the height, or  $y$ , coordinate is either derived from a height map, or generated with fractals, or produced as a combination of both these methods. The fractal technique used is the Diamond-Square algorithm [7], which is a form of mid-point displacement using a square base. Starting with the outside corner vertices, the height of the mid-point vertices is generated by averaging the surrounding vertices and adding a random displacement, first in diamond and then square step. The algorithm was modified to allow the seeding of the grid with values from a height map, thus resulting in a user defined terrain, smoothed and modified by noise. There are three methods of loading a terrain into the Simulator: a user-defined height map, a smoothed height map and a randomly generated terrain.

A user-defined height map can be of any width and height in pixels and is stretched to a fixed size in the Simulator. The RGB values are scaled by a factor of 0.2, with a value of 240 indicating sea level. For example, a value of 0 will be 48 meters below sea level, while a value of 255 will be 3 meters above sea level.

A smoothed height map must have a width and height of the form  $2x + 1$ , where  $x$  is an integer, i.e., height maps with odd-numbered widths and heights would be valid. A smoothed height map is handled similarly to a user-defined height map, except that the number of vertices used in the terrain mesh is expanded. The heights of these additional points are then generated using the Diamond-Square algorithm [7]. Figure 1 shows a resulting terrain. Finally, a random terrain can be generated, by taking four random vertices which are used as input into the Diamond-Square algorithm.



**Fig. 1.** An example underwater terrain obtained with a smoothed height map

Terrain collision detection is implemented using Coldet, an Open Source (LGPL) Collision Detection library [3]. The library provides a number of collision detection techniques: intersecting polygons, spheres and rays. We model the terrain as a set of triangles (polygons), and predict collisions between agents and the terrain using a sphere centered on the agent's centre of mass.

When a collision is detected, the physical movement of the agent is altered. Collisions can be modelled as either elastic, where kinetic energy is preserved or non-elastic, where some of the kinetic energy of the colliding objects, is transformed into another form during the collision. In DISCOVERY collisions are modelled as elastic, although this can be modified to provide more realistic behaviour in the future.

The agent's velocity is then calculated at the impact point, considering the reduction in time taken to reach it. The agent's velocity is reflected against the surface plane of the terrain polygon where the impact took place.

### 2.3 Physics-Based Simulation

A physically realistic dynamic modelling of volume of water, both soluble and insoluble contaminants present in water (crude oil, salts, etc.), insoluble contaminants on water surface (crude oil), and underwater robots can be achieved only as a balance between physical accuracy and computational performance. We selected an appropriate adjustable scale of the model, including the simulation grid step, the time step, and an option of selectively disabling certain simulation features to speed up some experiments: e.g., a purely underwater simulation can be done without water surface calculations. If an experiment does not require precise water dynamics, the system can assign constant parameters, such as current direction and rotation, to the water volume grid. At the moment, insoluble liquid contaminants moving in the water are simulated as solid objects affected by buoyancy, gravity, and currents in a normal force-momentum-position cycle. The water surface and surface-bound contaminants, such as crude oil slicks, can be either linked to the water grid or simulated separately (the method currently used in DISCOVERY) from the body of water as weight-spring meshes (updated in a normal force-momentum-position cycle of the physical simulation).

### 2.4 Simulator-Agent Sensor Protocol

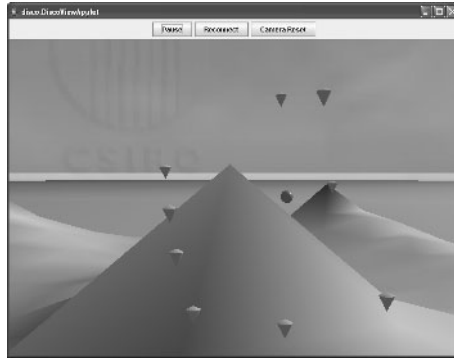
In principle, every environment variable maintained by the Simulator may be perceivable by an agent. One of the DISCOVERY objectives is to design, select and verify a correct set of relevant sensors suitable for a multi-agent task, and couple them with available actuators. In the initial setup, an agent has a number of sensors: e.g., a chemical sensor, a temperature sensor, a pressure sensor, a conductance sensor, a flow sensor, an internal battery sensor, an accelerometer, a compass, a collision sensor, a sonar sensor and a communication (acoustic and optical) sensor. The Simulator provides sensory data to agents upon request.

Actuators include thrusts, fins, a sonar, and communication devices (acoustic and optical). As a result of experiments, new sensors and actuators may be added and some of the listed sensors and actuators may be suppressed. Thrust allows the agent to accelerate both positively and negatively in a particular direction. A specified force provides acceleration in the direction of the thrust actuator. Three thrust actuators need to be constructed for movement within the  $(x, y, z)$  plane. Fin allows a moving agent to change its current direction. Each agent has a particular buoyancy value which allows it to float towards the surface of the water when there is no downward thrust.

### 3 Preliminary Experiments

The main difficulty in tracking and identification of an underwater source of contamination is that the insoluble contaminants (e.g., oil bubbles) can be sensed mostly only locally and may rise to the surface quite a distance away, being shifted by currents, winds, etc. The problem is complicated by non-trivial dynamics of underwater plumes and bubbles, in particular tracking of their gradients, as well as complexities of the underwater terrain, and may require heterogeneous teams of agents with distributed sensing and actuation.

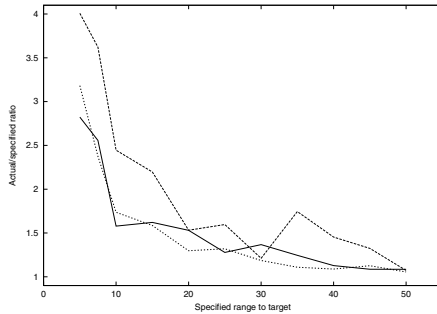
In the context of offshore hydrocarbon exploration, identification of petroleum systems presents an additional challenge. It is well-known that hydrocarbon can be entrapped in sub-terrain reservoirs formed in non-porous rock. Exploring features of interest in such complex environments by a sensor/actuator network is likely to require heterogeneous teams: not only to spread across a large area, but also to combine different sensing and actuation modalities (from sonars to magnetic and seismic surveys).



**Fig. 2.** An emergent spherical surface around the Explorer (a small sphere in the center), with several Supporter agents (conically shaped)

To illustrate the simulator capabilities, we designed a scenario with a heterogeneous team (one Explorer and ten Supporter agents). The Explorer is a task-oriented agent which has a goal of finding a proxy to a feature of interest (e.g, shallowest point in the local terrain, a chemical anomaly sensed by a combination of chemical sensors, etc.). While moving in the environment, it also emits an omnidirectional signal which can be used by Supporter agents to detect its direction, if there is an unobstructed line of signal between the Explorer and the receiving agent. The Supporter agents have the same actuators as the Explorer, with addition of sensors which are able to detect the signal emitted by the Explorer. The Supporters follow two simple rules: (i) move along the direction in which the Explorer is sensed until a specified range is reached within a tolerance limit; (ii) when the distance to the Explorer is less than the specified range, move away from the Explorer. If the signal from the Explorer is not received, a Supporter agent treats the last direction as valid, and may use extra-thrust in following this direction (“extra-thrust behaviour”). We also experimented with the third

behaviour: (iii) within the tolerance limit of the specified range, maintain it by moving orthogonally to the direction of the Explorer. The rules (i)-(iii) lead to emergence of a spherical surface around the Explorer, on which the Supporter agents randomly move in order to maintain the range (Figure 2). There is a difference between two kinds of emergence: pattern formation and intrinsic emergence, distinguished by Crutchfield [4]: a) pattern formation refers to an external observer who is able to recognize how unexpected features (patterns) 'emerge' or 'self-organize' during a process (e.g., spiral waves in oscillating chemical reactions) — these patterns may not have specific meaning within the system, but obtain a special meaning to the observer when detected; b) intrinsic emergence refers to the emergent features which are important within the system because they confer additional functionality to the system itself, like supporting global coordination and computation (e.g., the emergence of coordinated behaviour in a flock of birds allows efficient global information processing through local interaction, which benefits individual agents). To verify whether the emergence of a spherical surface is intrinsic and contributes to the quality of the supporting task, we run a number of experiments.



**Fig. 3.** The actual/specified ratio: sphere-tracing behavior (solid line), extra-thrust behaviour (dashed line), simple behavior with rules (i) and (ii) only (dotted line)

The Metric-Analyzer automates the scenario by setting simulation parameters, repeatedly running the Simulator and agents for each experiment, logging the relevant information (velocities, distances, etc.), and collating the results (the overall experiment runs for several days). In the Explorer-Supporters scenario it is important to have the Supporter agents maintaining the specified range over time, i.e. the actual achieved range should be as close as possible to the specified one. The actual/specified ratio, averaged over the team of the Supporter agents and over time after a certain initial interval, is plotted in Figure 3 against different specified ranges. The results of multiple experiments deployed in the same terrain indicate the difficulty of maintaining close ranges (within [5; 10] meters), for a team of Supporter agents. The Supporter agent closest to the Explorer always attains the specified range (the ratio is close to 1.0). It is worth pointing out that it is not the same agent but rather the one which is the closest at a given time, i.e. there is always at least one Supporter agent directly observing the Explorer. We observed that a) the extra-thrust behavior does not attain better quality, and b) the emergence of a spherical surface is not intrinsic: it does not contribute to the quality of

the supporting task, measured in terms of the actual/specified ratio. A spherical surface may, however, be beneficial if an equidistant spatial spread is important.

## 4 Conclusions

This paper presented a new multi-agent physics-based simulation framework (DISCOVERY), supporting experiments with self-organizing underwater sensor and actuator networks. The simulation system provides a test-bed for co-evolution of distributed and mobile sensors and actuators, and required communication topologies in challenging scenarios. In order to illustrate its capabilities, we briefly described a simple simulated scenario with a heterogeneous team (the Explorer-Supporters scenario), and its evaluation by Metric-Analyzer. Some of the tasks for future experiments include: identification of hazards to safety-critical structures; response to contamination of water supplies such as oil spills; perimeter formation for absorption barriers and traffic exclusion zones; collection, storage, transportation and analysis of contaminated items/evidence. The DISCOVERY Simulator is intended to be a flexible tool for simulation, quantitative analysis and design of complex underwater sensor and actuator networks, with varying degrees of autonomy, decentralized control, and data-to-decision integration.

## References

1. de Boer, R. and Kok, J.R. The Incremental Development of a Synthetic Multi-Agent System: The UvA Trilearn 2001 Robotic Soccer Simulation Team. Master's Thesis, University of Amsterdam, The Netherlands, 2002.
2. Butler, M., Prokopenko, M., Howard, T. Flexible Synchronisation within RoboCup Environment: a Comparative Analysis. In P. Stone, T. R. Balch, G. K. Kraetzschmar (eds.) *RoboCup 2000: Robot Soccer World Cup IV*, LNCS, Vol. 2019, 119–128, Springer, 2001.
3. Coldet: an Open Source Collision Detection library. <http://www.photoneffect.com/coldet/>
4. Crutchfield J. The Calculi of Emergence: Computation, Dynamics, and Induction. *Physica D*, 75, 11–54, 1994.
5. Dunbabin, M., Roberts, J., Usher, K., Winstanley, G., Corke, P. A Hybrid AUV Design for Shallow Water Reef Navigation. IEEE International Conference on Robotics and Automation, 2117–2122, Barcelona, Spain, 2005.
6. Foreman, M., Prokopenko, M., Wang, P. Phase Transitions in Self-organising Sensor Networks. In Banzhaf, W., Christaller, T., Dittrich, P., Kim, J.T. Ziegler, J. (eds.) *Advances in Artificial Life - Proceedings of the 7th European Conference on Artificial Life*, LNCS, Vol. 2801, Springer, 781–791, 2003.
7. Fournier, A., Fussell, D., Carpenter, L. Computer Rendering of Stochastic Models. *Communications of the ACM*, 25(6), 371–384, 1982.
8. Gerasimov, V., Guo, Y., James, G. C., and Poulton, G. T. Physically Realistic Self-assembly Simulation System. In A. Abraham, C. Grosan and V. Ramos (eds.), *Stigmergic Optimization, Studies in Computational Intelligence*, 117–130, Springer, 2006.
9. Kitano, H., Tambe, M., Stone, P., Veloso, M., Coradeschi, S., Osawa, E., Matsubara, H., Noda, I. and Asada, M. The RoboCup Synthetic Agent Challenge. In Proceedings of the 15th International Joint Conference on Artificial Intelligence, 1997.



10. Mahendra, P., Prokopenko, M., Wang, P., Price, D.C. Towards Adaptive Clustering in Self-monitoring Multi-Agent Networks. In R. Khosla, R. J. Howlett, L. C. Jain (eds.) *Knowledge-Based Intelligent Information and Engineering Systems, 9th International Conference, KES 2005, Melbourne, Australia, Proceedings, Part II*, LNCS, Vol. 3682, 796–805, 2005.
11. Prokopenko, M., Wang, P., Howard, T. Cyberooos'2001: 'Deep Behaviour Projection' Agent Architecture. In A. Birk, S. Coradeschi, S. Tadokoro (eds.) *RoboCup 2001: Robot Soccer World Cup V*, LNCS, Vol. 2377, 507–510, Springer, 2002.
12. Prokopenko, M., Wang, P. Relating the Entropy of Joint Beliefs to Multi-Agent Coordination. In G. A. Kaminka, P. U. Lima, Raúl Rojas (eds.) *RoboCup 2002: Robot Soccer World Cup VI*, LNCS, Vol. 2752, 367–374, Springer, 2003.
13. Prokopenko, M., Wang, P., Price, D.C., Valencia, P., Foreman, M., Farmer, A.J. Self-organising Hierarchies in Sensor and Communication Networks. *Artificial Life*, Special issue on Dynamic Hierarchies, Vol. 11(4), 407–426, 2005.
14. Prokopenko, M., Wang, P., Foreman, M., Valencia, P., Price, D., Poulton, G. On connectivity of reconfigurable impact networks in ageless aerospace vehicles. *Journal of Robotics and Autonomous Systems*, Vol. 53, 36–58, 2005.
15. Prokopenko, M., Mahendra, P., Wang, P. On Convergence of Dynamic Cluster Formation in Multi-Agent Networks. In M. S. Capcarrère, A. A. Freitas, P. J. Bentley, C. G. Johnson, J. Timmis (eds.) *Advances in Artificial Life, 8th European Conference, ECAL 2005, Canterbury, UK, September 5-9, 2005, Proceedings*, LNCS, Vol. 3630, 884–894, 2005.
16. Prokopenko, M., Wang, P., Price, D. Complexity Metrics for Self-monitoring Impact Sensing Networks, Proceedings of 2005 NASA/DoD Conference on Evolvable Hardware (EH-05), Washington D.C., USA, 2005.
17. Prokopenko, M., Poulton, G. T., Price, D. C., Wang, P., Valencia, P., Hoschke, N., Farmer, A. J., Hedley, M., Lewis, C., and Scott, D. A. Self-organising impact sensing networks in robust aerospace vehicles. In Fulcher, J. (ed.) *Advances in Applied Artificial Intelligence*, 186-223, Idea Group Inc., 2006.

# Predicting Cluster Formation in Decentralized Sensor Grids

Astrid Zeman and Mikhail Prokopenko

CSIRO Information and Communication Technology Centre  
Locked bag 17, North Ryde, NSW 1670, Australia  
mikhail.prokopenko@csiro.au

**Abstract.** This paper investigates cluster formation in decentralized sensor grids and focusses on predicting when the cluster formation converges to a stable configuration. The traffic volume of inter-agent communications is used, as the underlying time series, to construct a predictor of the convergence time. The predictor is based on the assumption that decentralized cluster formation creates multi-agent chaotic dynamics in the communication space, and estimates irregularity of the communication-volume time series during an initial transient interval. The new predictor, based on the auto-correlation function, is contrasted with the predictor based on the correlation entropy (generalized entropy rate). In terms of predictive power, the auto-correlation function is observed to outperform and be less sensitive to noise in the communication space than the correlation entropy. In addition, the preference of the auto-correlation function over the correlation entropy is found to depend on the synchronous message monitoring method.

## 1 Introduction

There is a distinction between “Sensor Networks” and “Sensor Grids”, as pointed out in recent literature (e.g., [3]): “whereas the design of a sensor network addresses the logical and physical connectivity of the sensors, the focus of constructing a sensor grid is on the issues relating to the data management, computation management, information management and knowledge discovery management associated with the sensors and the data they generate”. One significant issue addressed by sensor grids is dynamic sensor-data clustering, aimed at grouping entities with similar characteristics together so that main trends or unusual patterns may be discovered. This is investigated as decentralized clustering in multi-agent Systems [9], dynamic cluster formation in mobile ad hoc networks [7] and decentralized sensor arrays [8,13,10]. The latter studies describe dynamic cluster formation as *self-organisation* of dynamic hierarchies, with multiple cluster-heads emerging as a result of inter-agent communications, and indicates that decentralized clustering algorithms deployed in multi-agent systems are “hard to evaluate precisely for the reason of the diminished predictability brought about by self-organisation”. The results presented in [13] identified a predictor for the convergence time of dynamic cluster formation, based on the traffic volume of asynchronous inter-agent communications. Following this study, we attempt to adapt a decentralized clustering algorithm to a specific topology (a rectilinear grid) and replace a complicated predictor with a more simple measure, based on synchronized aggregation of multi-agent communications.

Our goal is predicting when the cluster formation will converge to a stable configuration. In achieving this goal, we consider an underlying time series, the traffic volume of inter-agent communications, and relate its irregularity during an initial interval to the eventual convergence time. Clearly, the shorter the initial interval is, the more efficient is the prediction: e.g., when a predicted value exceeds a threshold, agents may adjust parameters and heuristics used in the clustering process.

A simplified version of a decentralized adaptive clustering algorithm developed for evaluation purposes is described in the next section. Section 3 presents the proposed predictor for the convergence time of cluster formation, followed by a discussion of the obtained results.

## 2 Dynamic Cluster Formation Algorithm

A sensor grid node communicates only with immediate neighbours: all data are processed locally, and only information relevant to other regions of the structure is communicated as a multi-hop message. A cluster-head may be dynamically selected among the set of nodes and become a local coordinator of transmissions within the cluster. Clusters may re-form when new data is obtained on the basis of local sensor signals. Importantly, a cluster formation algorithm should be robust to such changes, failures of individual nodes, communication losses, etc.

As pointed out earlier, our main goal is an analysis of a representative clustering technique in a dynamic and decentralized multi-agent setting, exemplified by a rectilinear sensor grid, *in terms of predictability of its convergence time*. We represent a node sensory reading with a single aggregated value, define “differences” between cells in terms of this value, and cluster nodes while minimizing these “differences”.

The algorithm input is a series of events detected at different times and locations, while the output is a set of non-overlapping clusters, each with a dedicated cluster-head (a network node) and a cluster map of its followers in terms of their sensor-data and relative grid coordinates. The algorithm is described elsewhere [8] and involves a number of inter-agent messages notifying agents about their sensory data, and changes in their relationships and actions. For example, an agent may send a recruit message to another agent, delegate the role of cluster-head to another agent, or declare “independence” by initiating a new cluster. Most of these and similar decisions are based on the clustering heuristic described by Ogston et al. [9], and a dynamic offset range [8]. This heuristic determines if a cluster should be split in two, and the location of this split. Each cluster-head (initially, each agent) broadcasts its *recruit* message periodically, with a broadcasting-period, affecting all agents with values within a particular dynamic offset of the sensor reading detected by this agent. Every *recruit* message contains the sensor-data of all current followers of the cluster-head with their relative coordinates (a cluster map). Under certain conditions, an agent, which is not a follower in any cluster, receiving a *recruit* message becomes a follower, stops broadcasting its own *recruit* messages and sends its information to its new cluster-head indicating its relative coordinates and the sensor reading. However, there are situations when the receiving agent is already a follower in some cluster and cannot accept a recruit message by itself — a recruit disagreement. In this case, this agent *forwards* the received recruiting request to its present

cluster-head. Every cluster-head waits for a certain period, collecting all such *forward* messages, at the end of which the clustering heuristic is invoked on the union set of present followers and all agents who *forwarded* their new requests [8,13]. The cluster-head which invoked the heuristic notifies new cluster-heads about their appointment, and sends their cluster maps to them: a *cluster-information* message.

Here we consider an important variant of this algorithm, obtained by modifying both the message passing mechanism and the message monitoring method. First of all, instead of sending a *forward* message by broadcasting or “flooding” which makes the system quite resilient to noise, we use point-to-point messages incorporating relative grid coordinates, routed through the grid using these coordinates. Secondly, given a reduction in the communication traffic resulting from point-to-point messages, we employ a message monitoring method which allows to more precisely count inter-agent messages for each relative unit of system time. This essentially means that, instead of counting messages asynchronously (separately for each node) and aggregating these amounts for an abstract unit of time, we synchronize the system and precisely aggregate all messages for each time point. This is not always feasible and may incur a high cost, but the expected tradeoff is the simplicity and performance of new predictors.

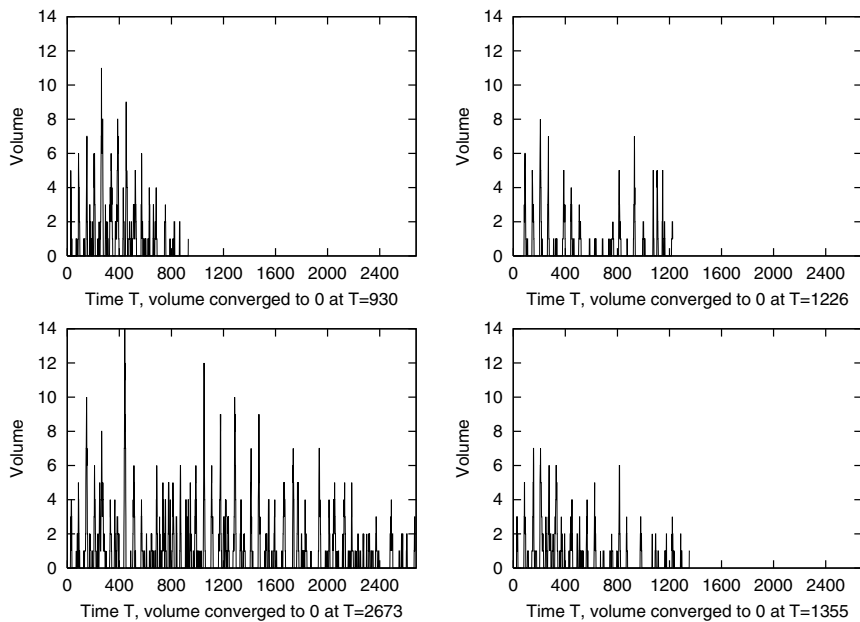
In addition, using point-to-point messages is a less reliable method, and we specifically introduced errors in the message-passing mechanism, simulating noise in the communication space — in order to verify robustness of the predictors. The new point-to-point messages significantly reduce the communication traffic, without affecting quality (measured by the weighted average cluster diameter [18]) and convergence (measured by the number of times the clustering heuristic was invoked before stability is achieved). While the simulation results show that the algorithm robustly converges and scales well in all cases, the convergence time varies significantly (Figure 1 and Figure 2) — highlighting the need for its better prediction.

The cluster formation is driven by three message types: *recruit*, *cluster-information*, and *forward* messages. The first two types are periodic, while the latter type depends only on the degree of disagreements among cluster-heads. The number of *forward* messages traced in time — the traffic volume of inter-agent communications — provides the underlying time series  $\{v(t)\}$  for our predictive analysis.

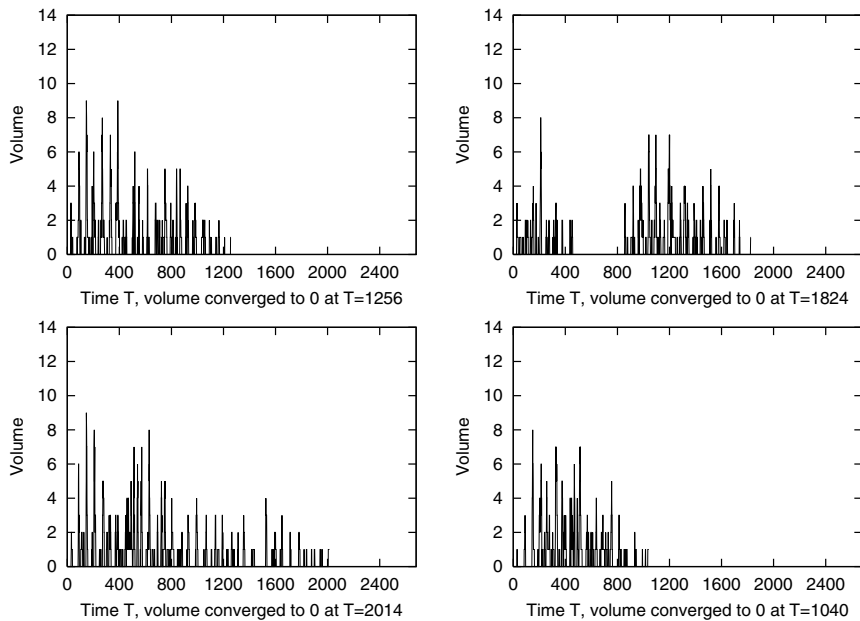
### 3 Regularity of Multi-agent Communication-Volume

In this section, we focus on our main objective: prediction of the convergence time  $T$ , based on regularity of an initial segment  $0, \dots, \mathcal{D}$  of the “communication-volume” series  $\{v(t)\}$ , where  $\mathcal{D} < T$  and  $v(t)$  is the number of *forward* messages at time  $t$ .

It is known that in many experiments, time series often exhibit irregular behavior during an initial interval before finally settling into an asymptotic state which is non-chaotic [1] — in our case, eventually converging to a fixed-point ( $v(T) = 0$ ). The irregular initial part of the series may, nevertheless, contain valuable information: this is particularly true when the underlying dynamics is deterministic and exhibits *transient chaos* [1,5]. It was conjectured and empirically verified [13] that the described algorithm for dynamic cluster formation creates *multi-agent transient chaotic dynamics*.



**Fig. 1.** Varying convergence times  $T_s$  for 4 different experiments,  $1 \leq s \leq 4$ , without noise



**Fig. 2.** Varying convergence times  $T_s$  for 4 different experiments, with noise

We intend to follow the same path as the previous study [13], but streamline the predictor estimation by using a simple auto-correlation function as a measure of regularity during the initial interval. For each experiment  $s$ , we a) select an initial segment of length  $\mathcal{D}$  of the time series; and b) compute the regularity predictor: the auto-correlation function  $\gamma(\mathcal{D}, \tau)_s$  for a range of integer delays  $\tau$ :

$$\gamma(\mathcal{D}, \tau)_s = \sum_{t=\tau+1}^{\mathcal{D}} [v_s(t - \tau) - \overline{v_s}] [v_s(t) - \overline{v_s}] / \sum_{t=1}^{\mathcal{D}} [v_s(t) - \overline{v_s}]^2, \quad (1)$$

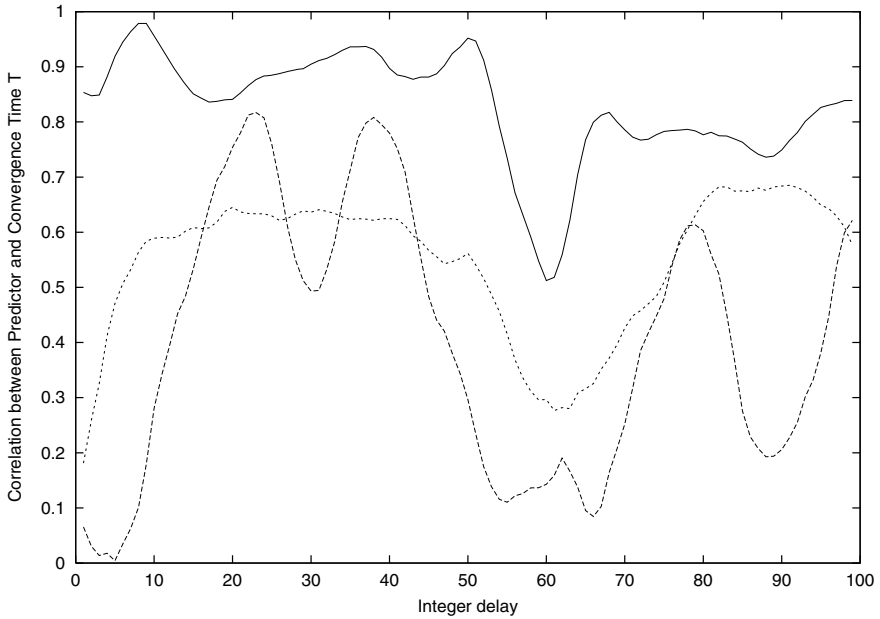
where  $\overline{v_s}$  is the series mean. Then, c) given the estimates  $\gamma(\mathcal{D}, \tau)_s$  for all the experiments, correlate them with the observed convergence times  $T_s$  by using a linear regression  $T = a + b\gamma(\mathcal{D}, \tau)$  and the correlation coefficient  $\rho(\tau)$  between the series  $\{T_s\}$  and  $\{\gamma(\mathcal{D}, \tau)_s\}$ . This would allow us to predict the time  $T_s$  of convergence to  $v_s(T_s) = 0$ , as  $T_s = a + b\gamma(\mathcal{D}, \hat{\tau})_s$ , for the delay  $\hat{\tau}$  providing the best fit: the maximum of  $\rho(\tau)$ .

The auto-correlation is obviously limited to measuring only linear dependencies, and the study [13] considered a more general and elaborate approach, based on the Kolmogorov-Sinai entropy  $K$ , also known as metric entropy [6,16], and its generalization to the order- $q$  Rényi entropy  $K_q$  [15]. The entropy  $K$  or  $K_q$  is an entropy per unit time, or an “entropy rate”, and is a measure for the rate at which information about the state of the system is lost in the course of time. In particular, the predictor estimated the “correlation entropy”  $K_2$  using Grassberger and Procaccia algorithm [4]. The predictor based on  $K_2$  uses the initial segment of length  $\mathcal{D}$  of the observed time series  $\{v(t)\}$  in “converting” or “reconstructing” the dynamical information in one-dimensional data to spatial information in the  $\tau$ -dimensional embedding space [17], and also depends on the length  $\mathcal{D}$  and the embedding dimension  $\tau$ .

The auto-correlation function  $\gamma(\mathcal{D}, \tau)$ , equation (1), was reported to be not sufficient for predictive purposes: the highest correlation coefficient  $\rho(\tau)$  between convergence times  $T_s$  and auto-correlations  $\gamma(\mathcal{D}, \tau)_s$ , for a range of delays  $\tau$ , was only 0.52, while the predictor based on the entropy  $K_2$  attained the maximum  $\rho = 0.90$ . In the following section we shall contrast these two measures,  $\gamma(\mathcal{D}, \tau)$  and  $K_2(\mathcal{D})$ , for the new communication and monitoring mechanisms, with and without noise.

## 4 Experimental Results

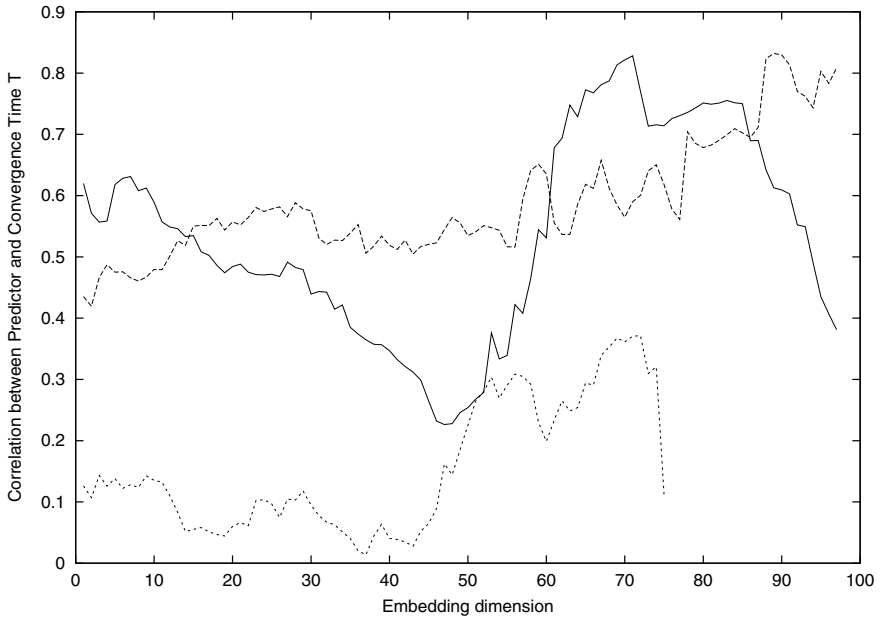
The experiments included three scenarios: (i) noiseless communications; (ii) 1% loss of messages; and (iii) 2% loss of messages. Each scenario included 20 runs of the clustering algorithm on an  $8 \times 8$  grid with 50 events, tracing the communication-volume time series  $\{v(t)\}$ . We then selected an initial segment  $\mathcal{D} = 800$ , and carried out the steps b) and c) described in the previous section. Given data of  $s = 1, \dots, 20$  experiments: the 2-dimensional array  $\gamma(\mathcal{D}, \tau)_s$  for varying  $\tau$  and each  $s$ , the correlation coefficient  $\rho(\{T_s\}, \{\gamma(\mathcal{D}, \tau)_s\})$  was determined for the range of  $\tau$ , based on the auto-correlation predictor  $\gamma(\mathcal{D}, \tau)$ . The data are plotted in Figure 3 for the scenarios (i), (ii) and (iii). The corresponding maximum values of  $\rho(\hat{\tau})$  degrade with noise as expected: from (i)  $\rho(8) = 0.98$  to (ii)  $\rho(23) = 0.82$  to (iii)  $\rho(91) = 0.69$ . As the level of noise grows, the maximums are attained at increasing delays  $\tau$ : (i)  $\hat{\tau} = 8$ ; (ii)  $\hat{\tau} = 23$ ; and (iii)  $\hat{\tau} = 91$ .



**Fig. 3.** The correlation coefficient  $\rho$  between the series  $\{T_s\}$  and predictor  $\{\gamma(\mathcal{D}, \tau)_s\}$ , for the scenarios (i) solid lines, (ii) dashed lines, and (iii) dotted lines

At the same time, the predictor based on  $K_2(\mathcal{D})$  was sensitive to the higher noise levels: the best obtained correlation values were: (i)  $\rho(71) = 0.83$ , (ii)  $\rho(89) = 0.83$ , and (iii)  $\rho(72) = 0.37$ , as shown in Figure 4. Without noise it performed as expected, maintained the performance under 1% loss of messages, but the increase in the noise by an extra percent resulted in more than 50% loss in predictive power. This can simply be explained by the fact that the extra noise made the underlying dynamics unstructured in the phase space created by the considered embedding dimensions [13], and this can be recovered by increasing their number. Nevertheless, from a practical point of view, it is rarely feasible, and the alternative predictor based on a simple auto-correlation function, is preferable as it is less sensitive to noise in the communication space.

We would like to point out that the noise in communication space (missed messages) considered in this paper should be distinguished from the noise in the traffic monitoring method created by its own asynchrony. The preference of the auto-correlation function over the correlation entropy, as the convergence time predictor, is conditional on the synchronous message monitoring method. If the underlying communication traffic is estimated asynchronously, then the observations reported in [13] indicate that the correlation entropy is preferred to the auto-correlation function (even in the presence of noise in the communication space). The reason for this difference is the calculation of correlations: the auto-correlation function simply “matches” separate time points (therefore, it is sensitive to shifts in the time series brought about by asynchronous monitoring), while the correlation entropy “matches” patterns or templates in the time series and is, hence, resilient to possible shifts due to asynchronous monitoring.



**Fig. 4.** The correlation coefficient  $\rho$  between the series  $\{T_s\}$  and predictor  $\{K_2(\mathcal{D})_s\}$ , for the scenarios (i) solid lines, (ii) dashed lines, and (iii) dotted lines (the last scenario did not have a sufficiently long time series to embed in higher dimensions)

## 5 Conclusions

We considered decentralized and dynamic cluster formation in multi-agent sensor grids, proposed and experimentally evaluated a new predictor for the convergence time of cluster formation. The new predictor, based on the auto-correlation function  $\gamma(\mathcal{D}, \tau)$ , was contrasted with the predictor  $K_2(\mathcal{D})$  based on the generalized correlation entropy of the volume of the inter-agent communications [13].

The results indicate that either predictor can be well correlated with the time of cluster formation. However, their applicability depends on the type of the communication traffic's monitoring: if the employed measure is asynchronous then  $K_2(\mathcal{D})$  is preferred, otherwise, if messages can be aggregated synchronously, the auto-correlation function  $\gamma(\mathcal{D}, \tau)$  should be preferred. In addition, the correlation entropy  $K_2(\mathcal{D})$  was shown to be adversely affected, as a predictor, by noise in the communication space.

Efficient and reliable algorithms for cluster formation in sensor grids may include a convergence predictor as a feedback to the algorithms. Such predictors are unlikely to implement measures with a global view, when full information on nodes' states and their inter-connections is available. Instead, a more promising approach is to develop measures that can work with partial information, obtained locally: *localizable* measures [14,11,12,2]. The analysis and results presented here and in [13] make a step towards localizable measures defined on the inter-agent communication space, and highlights



their applicability in decentralized, dynamic and asynchronous sensor grids. Another direction of future research is scale-free sensor grids.

**Acknowledgements.** The authors are grateful to Piraveenan Mahendra and Peter Wang for helpful discussions and valuable contributions to earlier versions of the algorithm.

## References

1. Dhamala, M., Lai, Y.C., Kostelich, E.J. Analyses of transient chaotic time series. *Physical Review E*, 64, 056207, 1–9, 2001.
2. Foreman, M., Prokopenko, M., Wang, P. Phase Transitions in Self-organising Sensor Networks. Banzhaf, W., Christaller, T., Dittrich, P., Kim, J.T. Ziegler, J. (Eds.) *Advances in Artificial Life - Proceedings of the 7th European Conference on Artificial Life*, 781–791, LNAI 2801, Springer, 2003.
3. Sensor Grid for Air Pollution Monitoring, 2004, Ghanem, M., Guo, Y., Hassard, J., Osmond, M., and Richards, M. In Proceedings of the 3rd UK e-Science All-hands Conference AHM 2004, 106–113, Nottingham UK, 2004.
4. Grassberger, P. and Procaccia, I. Estimation of the Kolmogorov entropy from a chaotic signal. *Physical Review A*, 28(4):2591, 1983.
5. Jánosi, I.M., and Tél, T. Time series analysis of transient chaos. *Physical Review E*, 49(4):2756–2763, 1994.
6. Kolmogorov, A.N. Entropy per unit time as a metric invariant of automorphisms. *Doklady Akademii Nauk SSSR*, 124:754–755 (Russian), 1959.
7. Lin R., and Gerla, M. Adaptive Clustering for Mobile Wireless Networks. *IEEE Journal on Selected Areas in Communications*, 1265–1275, September 1997.
8. Mahendra, P., Prokopenko, M., Wang, P., Price, D.C. Towards Adaptive Clustering in Self-monitoring Multi-Agent Networks. In R. Khosla, R. J. Howlett, L. C. Jain (eds.) *Knowledge-Based Intelligent Information and Engineering Systems, 9th International Conference, KES 2005, Melbourne, Australia, September 14-16, 2005, Proceedings, Part II*, Lecture Notes in Computer Science, Vol. 3682, 796–805, 2005.
9. Ogston E., Overeinder, B., Van Steen, M., and Brazier, F. A Method for Decentralized Clustering in Large Multi-Agent Systems. The 2nd International Joint Conference on Autonomous Agent and Multi Agent Systems, 798–796, 2003.
10. Olsson, L., Nehaniv, C. L., Polani, D., Sensory Channel Grouping and Structure from Uninterpreted Sensor Data, 2004 NASA/DoD Conference on Evolvable Hardware (EH'04), 153-160, 2004.
11. Prokopenko, M., Wang, P., Price, D.C., Valencia, P., Foreman, M., Farmer, A.J. Self-organising Hierarchies in Sensor and Communication Networks. *Artificial Life*, Special issue on Dynamic Hierarchies, Vol. 11(4), 407–426, 2005.
12. Prokopenko, M., Wang, P., Foreman, M., Valencia, P., Price, D., Poulton, G. On connectivity of reconfigurable impact networks in ageless aerospace vehicles. *Journal of Robotics and Autonomous Systems*, Vol. 53, 36–58, 2005.
13. Prokopenko, M., Mahendra, P., Wang, P. On Convergence of Dynamic Cluster Formation in Multi-Agent Networks. In M. S. Capcarrère, A. A. Freitas, P. J. Bentley, C. G. Johnson, J. Timmis (eds.) *Advances in Artificial Life, 8th European Conference, ECAL 2005, Canterbury, UK, September 5-9, 2005, Proceedings*, Lecture Notes in Computer Science, Vol. 3630, 884–894, 2005.
14. Prokopenko, M., Wang, P., Price, D. Complexity Metrics for Self-monitoring Impact Sensing Networks, Proceedings of 2005 NASA/DoD Conference on Evolvable Hardware (EH-05), Washington D.C., USA, 2005.

15. Rényi, A. *Probability theory*. North-Holland, Amsterdam, 1970.
16. Sinai, Ya.G. On the concept of entropy of a dynamical system. *Doklady Akademii Nauk SSSR*, 124:768-771 (Russian), 1959.
17. Takens, F. Detecting strange attractors in turbulence. *Dynamical systems and turbulence*, LNM Vol.898, Springer, Berlin, 1981.
18. Zhang, T., Ramakrishnan, R., Livny, M. BIRCH:A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery*, 1(2), 141-182, 1997.

# Making Sense of the Sensory Data – Coordinate Systems by Hierarchical Decomposition

Attila Egri-Nagy and Chrystopher L. Nehaniv

BioComputation Research Group  
School of Computer Science  
University of Hertfordshire  
College Lane, Hatfield, Hertfordshire AL10 9AB, United Kingdom  
{A.Egri-Nagy, C.L.Nehaniv}@herts.ac.uk

**Abstract.** Having the right sensory channels is an important ingredient for building an autonomous agent, but we still have the problem of making sense of the sensory data for the agent. This is the basic problem of artificial intelligence. Here we propose an algebraic method for generating abstract coordinate system representations of the environment based on the agent's actions. These internal representations can be refined and regenerated during the lifespan of the agent.

## 1 Introduction

According to the so-called Good Old-Fashioned Artificial Intelligence approach we have to build a system with a reasonably accurate representation of its environment to make it behave intelligently. But this just does not work. The hard-wired model is rigid, even cannot cope with small changes of the environment, or it the representation should contain all details with all the possible changes, thus combinatorial explosions pop up. Moreover, the basic assumption itself that we have complete knowledge about the environment beforehand can hardly be defended. Therefore the Artificial Intelligent (AI) community has come up with the counterintuitive idea that we do it better without any representation [1]. Clearly, this is a fruitful method showing that one can have complex behavior without complex inner structure. But it is also clear that we cannot get too far without representations [2].

Here we adopt the viewpoint that we often need representations of the environment in order to realize artificial intelligence, but the representation should be flexible and dynamically changing over time and obtained by the artificial system on its own by recognizing regularities of the real world around. We propose an algebraic method for generating abstract coordinate system representations of the environment based on the agent's actions. Sampling the transitions through the sensory channels after the actions of the agent allows us to build a finite state automaton description, from which we can generate abstract coordinate systems using the algebraic hierarchical decomposition of finite state automata.

The general ideas of applying automata decompositions as formal models of understanding were proposed several times [3,4], but now they are closer to

fulfilment. The mathematical theory behind this is the algebraic hierarchical decomposition of finite state automata, the so-called Krohn-Rhodes Theory. For forty years there was no computational implementation for the hierarchical decomposition of automata. However, in the electronic circuit industry there are many different decomposition methods and implementations, but they are not hierarchical since there are several physical constraints on circuit design and the cascaded composition appears not to be the most efficient in terms of power consumption, area and delay minimization [5]. Though it may be very appropriate for understanding such systems [3]. Recently the authors have implemented two methods for the holonomy decompositions [6,7,8].

## 2 Hierarchical Decomposition: The Krohn-Rhodes Theory

Here we present the very basic underlying ideas of algebraic hierarchical decomposition of finite state automata. We use the minimum amount of mathematical notation here. For precise definitions see [9,4].

### 2.1 Reversible and Irreversible Processes

Roughly speaking, we have two different kinds of computational operations: reversible and irreversible ones. For instance, if we move some content of the memory to another empty location, that is reversible, since we can move it back. But if we overwrite a nonempty part of the memory, then it is irreversible, since there is no way to restore the previously stored data. Closer to a formal definition we can say that irreversible processes reduce the size of the set of possible future states, while reversible ones do not. A map  $f : A \rightarrow A$  of a set  $A$  is called a permutation (reversible) if it is a bijection, otherwise it does collapse elements ( $a \in A$  is an image of more than one element), therefore it is irreversible.

Algebraically the distinction is more immediate. A *permutation group* is a set  $G$  of bijective mappings together with the *state set*  $A$  on which the mappings act. A *transformation semigroup*  $(A, S)$  has a similar structure, but  $S$  consists of general functions, not only bijective maps. Roughly speaking we consider finite automata as transformation semigroups. The elements of the semigroup are the transformations of the state set induced by the input symbols. This way the problems in automata theory are transferred into the algebraic domain.

### 2.2 The Prime Decomposition Metaphor

For explaining the Krohn-Rhodes Theory, the best way is to present it by a metaphor. Basically we do the same as the prime decomposition for integers, but instead of numbers we do it for more complicated structures, namely finite state automata (considered as transformation semigroups). The similarities can be summarized the following way:

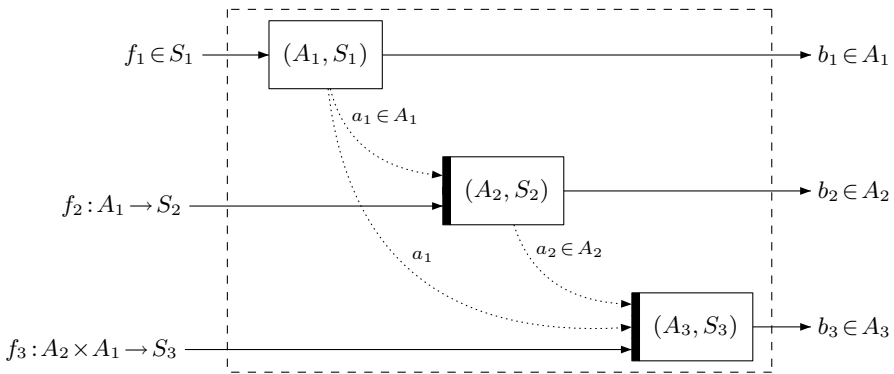
	<b>Integers</b>	<b>Automata</b>
<b>Factors</b>	Primes	Flip-flop Automaton Permutation Automata
<b>Composition</b>	Multiplication	Wreath Product
<b>Precision</b>	Equality	Division, Emulation

The basic building blocks are the simple<sup>1</sup> permutation groups (for the reversible computation) and only one component for the irreversible computation, the so-called flip-flop automaton. It is like a one-bit sized memory.

The way of putting together the components, the so-called *cascaded* or *wreath* product, is hierarchical and no feedback is allowed from deeper levels to upper levels (see Fig. 1). The reason, why we choose this special way of composition, is that the following special properties of hierarchy render the composed structure more comprehensible.

- Information flow between levels is restricted enabling modularity (also within one level with parallel components).
- Generalization and specialization are natural operations realized by taking subsets of levels in either direction up or down the hierarchy

Note that we allow parallel components on one hierarchical level.



**Fig. 1.** State transition in the wreath product  $(A_3, S_3) \wr (A_2, S_2) \wr (A_1, S_1)$ . An input determines a transformation  $(f_3, f_2, f_1)$  which maps the state  $(a_3, a_2, a_1)$  yielding the new state  $(b_3, b_2, b_1) = (a_3 \cdot f_3(a_2, a_1), a_2 \cdot f_2(a_1), a_1 \cdot f_1)$ . The black bars denote the applications of functions  $f_2, f_3$  according to hierarchical dependence. Note that the applications of these functions happen exactly at the same moment since their arguments are the previous states of other components, therefore there is no need to wait for the other components to calculate the new states. We use the state as the output of the automaton. It is often misunderstood to have some time delay during the state transition. This is not the case, the state transition in the wreath product is instantaneous.

<sup>1</sup> This has a well defined meaning in group theory.

### 2.3 Coordinates, Hierarchical Dependence

Hierarchical decompositions provide a coordinate system for the original phenomenon described as an automaton. For each coordinate position we have transformation semigroup components and their state sets are the possible values for that position. Due to its hierarchical nature the order of the coordinates does matter. What happens on deeper levels is determined by the states of the levels above. The simplest example to describe *hierarchical dependence* is a bidirectional counter. Imagine a device which keeps track how many times you press a button, and you have two other buttons set the operating mode. You start from zero in adding mode then as a check whether the resulting number is the correct value, you switch to subtracting mode and count again, but this time downwards, until you reach zero again. For instance to count the number of passengers on an airplane while walking along the aisle. The operation of this device can be represented with the following simple coordinate system:  $(n, \text{mode})$ , where modes are  $+$  and  $-$  corresponding to adding and subtracting. The mode coordinate is the top level of the hierarchy. There are three operations: counting  $c$ , switching to adding mode  $m_+$ , and switching to subtracting mode  $m_-$ . For instance

$$(9, +) \cdot c = (10, +)$$

$$(9, +) \cdot m_- = (9, -)$$

$$(9, +) \cdot m_+ = (9, +)$$

$$(9, -) \cdot c = (8, -)$$

Hierarchical dependence: the counting operation does different things depending on the top level coordinate.

### 2.4 Computational Implementations

Now we have available implementations for Krohn-Rhodes Theory [7] and we can start exploring the vast space of the decomposition of computational structures including actions and sensory activity of agents. However before applying the method to large-scale problems we need to solve some scalability issues. Currently we are working on a new incremental version of the algorithm, which starts at the top level and goes down to decompose further levels when they are feasible. This way we get some information about the hierarchical structure immediately, instead of trying to calculate the first phase of the whole decomposition, which may fail due to combinatorial complexity.

## 3 Building Coordinate Systems from Sensory Data Based on Actions

For acting meaningfully in a complex environment an agent may need a representation, a model of that environment. The model is used to predict the

outcome of certain actions. But where does this representation come from? The widely accepted answers are evolution or learning, since having a predefined and fixed representation often exhibits very unintelligent behaviour. If we want to make representation hardwired, we might not have complete knowledge (if we have any at all) about the environment, and also the environment can be changing. Moreover, the most important thing is that the agent needs a model from its viewpoint (not from our viewpoint), i.e. that is appropriate for its ‘Umwelt’ (cf. [10]). Therefore we can conclude that agent should build its representation of the environment primarily based on the data coming through the sensors.

### 3.1 Experimenting with the Environment

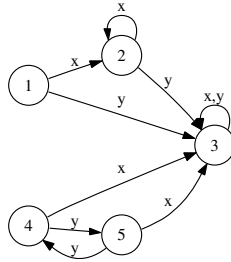
In order to apply hierarchical decompositions, first we need to build a finite state automata description. The state set of the automaton consists of the possible states of the environment from the viewpoint of the agent, i.e. the perception, the data coming from the sensory channel. The accurate definition of the state depends on the actual hardware setup of the sensors. The input symbols are the actions of the agent. That is why we say that the automaton model is built from the agent’s perspective. The agent carries out basic experiments: first it determines the current sensory state, next it carries out an action, then determines the resulting sensory state again. This elementary experiment is recorded as one state transition in the finite state automata being built.

After finite many repeats of the basic experiments, we have a sequence

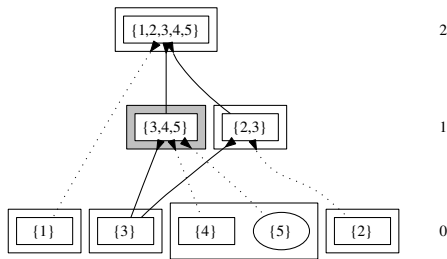
$$\begin{aligned} &\text{perception} \rightarrow \text{action} \rightarrow \text{perception} \rightarrow \text{action} \rightarrow \text{perception} \rightarrow \\ &\dots \rightarrow \text{action} \rightarrow \text{perception}. \end{aligned}$$

This sequence is usually called the perception-action loop, and can be studied by using information theoretical tools [11]. The state transitions define the automaton, and we can do the decomposition.

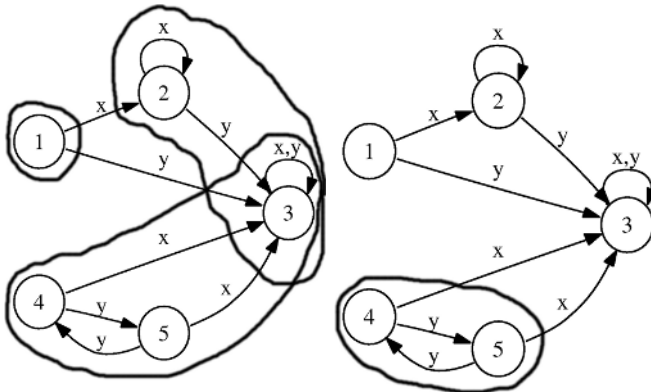
Let’s suppose we have a perception-action automaton and we do not really know what it is doing (though by knowing its generators we fully describe it implicitly), as in Fig. 2. Is it doing some complex computation? Calculating its holonomy decomposition we find that it can be emulated by a cascaded automaton with two levels (for details and visualisation see Figures 3-4). Now if we ask the question, ‘What is the automaton doing roughly?’, then we can answer very easily just by looking at the top level (Fig. 4). We have three states there and the component is not a reversible one, which means that there are actions of the agent that induce decisive changes in the environment, and those changes cannot be undone. Going further down to the second level we find that depending on the state above we either have a reversible component or another irreversible change. The actual reversible component is a permutation of two states of the original automaton, corresponding to actions that can be repeated. This illustrates the idea of having a coordinate system for understanding.



**Fig. 2.** An example perception-action automaton  $\mathcal{A}$  for an agent with two actions  $x$  and  $y$  showing the transitions these actions induce on sensory states. The actions determine the state transformations  $x = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 2 & 3 & 3 & 3 \end{pmatrix}$ ,  $y = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 3 & 3 & 5 & 4 \end{pmatrix}$ .



**Fig. 3.** The structure of the holonomy decomposition of  $\mathcal{A}$ . The numbers on the right denote the hierarchical levels (the level 0 is present just to show the states of the components on the first level, it does not appear as a hierarchical level in the decomposition). The nodes are subsets of the state set, rectangular nodes represent the components of the decomposition. Shaded components denote the existence of some reversible computation. The arrows going into the component come from the component's states. On the first level we have parallel components.



**Fig. 4.** The states of the top (level 2) component of the decomposition of  $\mathcal{A}$  are overlapping subsets of the state set of the transformation semigroup being decomposed (on the left). The identified reversible computation at the second level of the decomposition of  $\mathcal{A}$  (on the right).



### 3.2 Integrating Sensory Channels

Depending on the granularity level on which we define states, we can either have composite states by somehow integrating sensory channels, or we can build the hierarchical model for each channel. The latter seems to be more interesting since we have different models for different modalities and we have the possibility of comparing the different coordinate systems.

### 3.3 Hidden States

It may be very well possible that starting from a state  $a$  the same action  $x$  applied many times yields different results. That is the indication that the sensor does not capture some important aspects of the process in the environment. There are some “hidden states”, that the agent cannot see. This clearly complicates the finite state automata description by making it nondeterministic<sup>2</sup> and may be converted into a deterministic one (with the cost of introducing more states, for instance the states that are not detected by the sensors). As an extreme case we can consider an agent only with one action, the observation. For this problem of hidden states we have more sophisticated approaches, like the  $\epsilon$ -machine reconstruction [12], where the state transitions of the automaton are based on histories, not just on single states of the environment.

### 3.4 Stochasticity

Our basic assumption is that the environment of the agent can be described by finite state automata. It is debatable whether our assumption holds for agent put into real world situations, or by using discrete non-stochastic models we abstract away important layers of the real processes. However, our approach is justified by the very idea of a model, which should be simpler than the modeled phenomenon.

## 4 Future Work and Discussion

We presented a framework for using sensory channels to build models of the environment on the fly. We did not mention many difficult issues that are expected to come up for real experiments (e.g. with physically built robots). The difficulties can be the definition of state for each sensory channels, the resolution of time for actions and perceptions, the number of actions needed for building the model, etc., these should be solved by future attempts. The next steps are to automate the construction of perception-action automata and the related  $\epsilon$ -machines arising from real-world examples. By having a computational implementation we are getting closer to those very promising and possibly successful applications.

---

<sup>2</sup> Nondeterminism in the context of automata theory means only that in a given state the same action may have several outcomes, whereas stochasticity concerns the assignment of probabilities to such transitions.

## References

1. Brooks, R.A.: *Cambrian Intelligence: The Early History of the New AI*. MIT Press (A Bradford Book) (1999)
2. Steels, L.: Intelligence with representation. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* **361**(1811) (2003) 2381–2395
3. Rhodes, J.L.: *Applications of Automata Theory and Algebra via the Mathematical Theory of Complexity to Finite-State Physics, Biology, Philosophy, Games, and Codes*. University of California at Berkeley, Mathematics Library (1971)
4. Nehaniv, C.L., Rhodes, J.L.: The evolution and understanding of hierarchical complexity in biology from an algebraic perspective. *Artificial Life* **6** (2000) 45–67
5. Devadas, S., Newton, A.R.: Decomposition and factorization of sequential finite state machines. *IEEE Transactions on Computer-Aided Design* **8**(11) (1989) 1206–1217
6. Egri-Nagy, A., Nehaniv, C.L.: Algebraic hierarchical decomposition of finite state automata: Comparison of implementations for Krohn-Rhodes Theory. *Conference on Implementations and Applications of Automata CIAA 2004, Lecture Notes in Computer Science* **3317** (2004) 315–316
7. Egri-Nagy, A., Nehaniv, C.L.: *GrasperMachine, Computational Semigroup Theory for Formal Models of Understanding*. (<http://graspermachine.sf.net>). (2003)
8. Egri-Nagy, A.: *Algebraic Hierarchical Decomposition of Finite State Automata – A Computational Approach*. PhD thesis, University of Hertfordshire, School of Computer Science (2005)
9. Krohn, K., Rhodes, J.L., Tilson, B.R.: 5, The Prime Decomposition Theorem of the Algebraic Theory of Machines. In: *Algebraic Theory of Machines, Languages, and Semigroups* (M. A. Arbib, ed.). Academic Press (1968) 81–125
10. von Uexküll, J.: Environment [Umwelt] and inner world of animals. In Burghardt, G.M., ed.: *Foundations of Comparative Ethology*. Van Nostrand Reinhold, New York (1985) 222–245
11. Klyubin, A.S., Polani, D., Nehaniv, C.L.: Organization of the information flow in the perception-action loop of evolved agents. In Zebulum, R.S., Gwaltney, D., Hornby, G., Keymeulen, D., Lohn, J., Stoica, A., eds.: *Proceedings of 2004 NASA/DoD Conference on Evolvable Hardware*, IEEE Computer Society (2004) 177–180
12. Crutchfield, J.P.: The calculi of emergence: Computation, dynamics, and induction. *Physica D* **75** (1994) 11–54

# Biologically-Inspired Visual-Motor Coordination Model in a Navigation Problem

Jacek Jelonek and Maciej Komosinski

Poznan University of Technology, Institute of Computing Science  
Piotrowo 2, 60-965 Poznan, Poland  
{jjelonek, mkomosinski}@cs.put.poznan.pl

**Abstract.** This work presents a biologically-inspired coordination model which associates motor actions with visual stimuli. The model is introduced and explained, and navigation experiments are reported that verify the implemented visual-motor system. Experiments demonstrate that the system can be trained to solve navigation problems consisting in moving around a 3D object to reach a specific location based on the visual information only. The model is flexible, as it is composed of an adjustable number of modules. It is also interpretable, i.e. it is possible to estimate the influence of visual features on the motor action.

## 1 Introduction

In the real world, creatures face a complex, changing environment and need to handle large amounts of information to survive and reproduce. Robots produced by humans should possess analogous qualities if we want them to autonomously make decisions and perform successfully in natural environments. However, this is not yet accomplished; there is still a huge difference between efficiency of performance between creatures and modern robots. If one could develop robots that are as robust, flexible and adaptive as living organisms, a lot of time and money would be saved that is spent on designing and developing robots that are highly specialized in performing a specific task in specific conditions.

One of the factors that play an important role in the success of living organisms is the way they acquire information from the environment. Their senses are interfaces between neural systems and the outer world. Living organisms exhibit a vast number of sensor types, including olfactory, tactile, auditory, visual, electric and magnetic ones. In this work we focus on visual sensing as the one that provides a lot of information about the environment and is therefore popular in natural systems and often used in artificial designs.

In the area of machine vision, problems that are considered are usually related to object recognition and classification. This work adds the aspect of active exploration of the environment based on information that is perceived. The information considered here is visual and much more complex than the information perceived by light-following robots that mimic simple organisms like *Paramecium*.

This paper focuses on a visual-motor model that facilitates stimulus–reaction performance, as it is the basic schema in functioning of living organisms. The stimulus is

visual, and motor reaction is movement of an agent. The purpose of building this biologically-inspired model is twofold. First, it helps in understanding cognitive processes in living organisms. Second, implementations of such models can cope with the complexity of real-world environments because these models are inspired by solutions that proved to be successful in nature.

The experiments with visual-motor model are performed using simulated, artificial agents. The software environment is the Framsticks simulator [1] equipped with a new *vector eye* sensor. We consider navigation and target approaching tasks [5] with an agent moving along a circular path around some scene, observing a three-dimensional object positioned in the center. The agent decides whether it wants to move left or right, and adjusts speed of its movement.

Analogously to natural environments, some locations around the object are advantageous (“life zones” that living organisms try to reach) while others are adverse (avoided “death zones”). The goal for the agent is to reach some optimal location (using motor actions) based only on the visual information that it perceives looking at the object. The visual-motor system inside the agent needs to be trained to accomplish this task.

The next section describes in detail the architecture of the visual-motor system, presents biological inspirations for the model, and introduces its three components: vector eye, visual cortex, and the motor area. Section 3 reports experiments that were performed, and Section 4 summarizes this article and points out directions of future research.

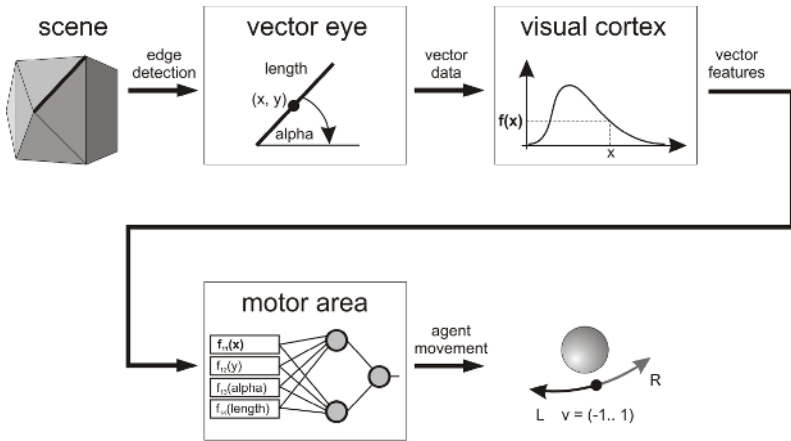
## 2 Architecture of the Visual-Motor System

The architecture of the proposed visual-motor system consists of three components – *vector eye*, *visual cortex* and *motor area*. Vector eye captures edges, basic visual elements of a scene, as observed by an agent. Each edge is characterized by four attributes – length, angle and coordinates of its center. The attributes of all edges (vector data) are transformed and aggregated by the visual cortex and fed to the *motor area* module. The motor area controls agent’s movements in the virtual environment. The data flow is illustrated in Fig. 1.

### 2.1 Vector Eye

Vector eye is a high-level sensor that provides a list of edges in a scene that are visible from some location in space. This information is accurate, i.e. it has no noise or imperfections which would exist if these edges were detected in a raster picture.

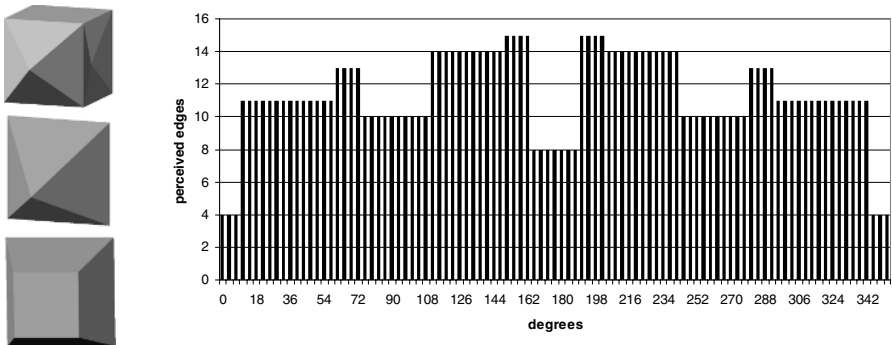
The sensor is implemented in Framsticks, a simulation environment that models three-dimensional bodies of agents and their neural control systems [1]. Framsticks allows users to perform predefined experiments, but the software also supports user-defined experiment definitions, fitness functions, and neuron types. In this work, only basic Framsticks functionality was used. Features like evolutionary optimization, neural control and embodiment can be utilized in future research.



**Fig. 1.** Data flow in the model of visual-motor system

For the purposes of this work, the agent that travels around the object is considered as a point equipped with a single eye sensor that observes the centre of the scene and perceives a single, three-dimensional shape, as shown in Fig. 2 on the left.

Many simple sensors that are commonly used in robotics (touch, proximity, 3D orientation) provide single-valued outputs, and therefore do not need special post-processing of this information in order to make it useful. Vector eye, on the other hand, is a complex sensor that provides a variable amount of information depending on what shape is perceived and what is the relative position and orientation of the sensor with respect to the shape (see Fig. 2, right).



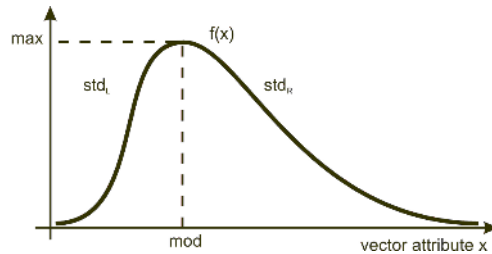
**Fig. 2.** Left: three-dimensional shapes used for experiments. Right: histogram of the number of edges perceived when observing a sample object from different angles.

## 2.2 Visual Cortex

The proposed model of visual cortex is inspired by its biological counterpart. It has been shown that the primary visual cortex consists of cells that are selectively responsive to

different features of the visual stimulus [3, 4]. The cells are called feature detectors because they analyze the visual image to find specific features (such as a bar or line of a specific orientation, length, etc.). For example, there are *simple cells* for which stimulus that maximally excites the cell is a line at a specific angle of orientation across the retina. Deviations from that preferred angle excite the cell less and less, until a line perpendicular to its receptive field has no effect. *Complex cells* show orientation specificity like simple cells, but additionally, they manifest highest sensitivity to stimuli that are lines moving in a specific direction across the visual field. Finally, *hypercomplex cells* are most sensitive to lines of specific length and specific angle of orientation that move in a specific direction.

The idea of feature detectors is employed in the proposed model of visual cortex. Data coming from the vector eye sensor contain a set of geometrical attributes. Four geometrical attributes are considered for each edge – angle, length and location (expressed by two coordinates, X and Y). Values of these attributes are transformed by parameterized Gaussian functions (see Fig. 3). Each function is defined by the following parameters: *modal* (preferred) value of associated attribute that excites the neuron most, *extreme* value of excitation (positive or negative), and *left* and *right standard deviation* that shape the function. The final excitation is a sum of excitations invoked by all edges provided by the vector eye. This solves the problem of aggregating variable numbers of attributes (depending on a number of edges) coming from the eye sensor.



**Fig. 3.** Parameterized Gaussian function

### 2.3 Motor Area

We started from a simple architecture of the motor component, employing the OWA (ordered weighted averaging) method proposed by Yager [6] to aggregate values of features from the visual cortex. However, this approach was not sufficient as such a component could not be trained well. Therefore, more OWA modules were added but it was still difficult to minimize error sufficiently. Finally, neural networks were introduced instead of the OWA operator, which let the system learn the navigation task successfully.

The motor area component is implemented as a set of motor modules. Each module consists of a two-layer feed-forward neural network. First layer neurons are fully connected to outputs of the visual cortex (parameterized feature detectors included in Fig. 4). Each neuron is defined by  $3+w$  parameters, where  $w$  is the number of inputs (weights) and the other three parameters characterize the shape of the sigmoid transfer

function. Motor output is computed as the total activation of all motor modules. A behavioral adaptation of an agent to virtual environment consists in adjusting all the parameters, and can be achieved by various optimization and/or learning techniques.

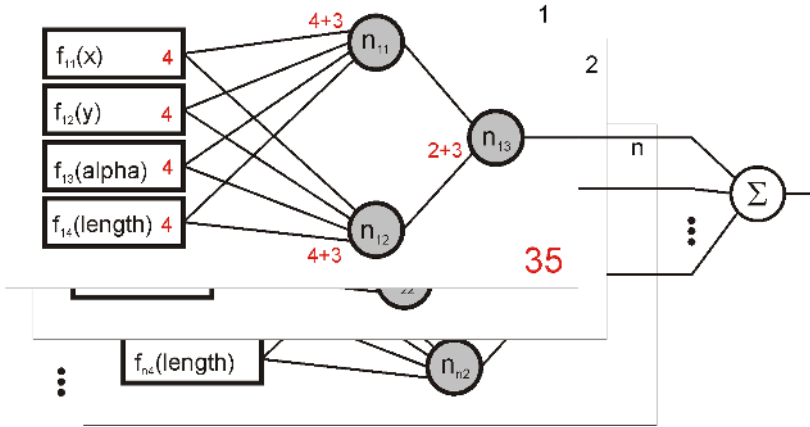


Fig. 4. The topology of the motor component. Number of parameters shown for each part of a module (total of 35 parameters per module).

### 3 Experiments: Training and Analyzing the Visual-Motor System

To train the visual-motor system and adjust its parameters, various training methods and regimes could be used, including evolutionary learning. In this work we employ direct, supervised learning, which is common in nature, where parents teach their children particular behaviors by rewarding or punishing them for specific actions.

A greedy gradient optimization algorithm [2] was selected as the training method, because it is relatively simple and quick. In the beginning, all system parameters are initialized randomly, and a small delta value is associated with each parameter. Then, for each parameter, its delta value is added, and if the new set of parameters is better than the old one, the new parameter value is accepted and the corresponding delta value is increased. Otherwise, if this change was not beneficial to the whole system, the old value for the parameter is retained and the corresponding delta value is negated and decreased. The process is repeated as long as a change in any parameter causes improvement of the whole system (i.e. the global error is decreased).

As we expect the agent to be able to navigate to some specified position around the object and stop there, its target speed should have positive values for one half and negative values for the other half of the circular path. The speed should be zero for the specified position where the agent should stop. There are many speed functions that satisfy these requirements, which causes problems for some learning algorithms, as there is no clear and continuous information on the expected direction of changes for system parameters.

The obvious error measure is the number of mistakes of the sign of speed value for all positions around the object. For example, if we assume positive speed as “move

right” and negative speed as “move left”, the total error is the number of locations where the agent moves in a wrong direction, no matter how fast. Although such error formula is good for evaluating agent behavior, it is not very helpful during gradient optimization. It only yields a limited number of discrete values which does not provide sufficient (continuous) information to minimize the error, especially that there are many feasible speed functions. For this reason we decided to use a specific target speed function that the system is trained. It is a sinus function, and the zero angle is adjusted to match the specified stop position, where the target speed is zero.

The first experiment described in this section concerns selecting the appropriate error function to be minimized, and the second experiment looks for the optimal complexity of the visual-motor system. Other experiments are mentioned in the last section of this article.

### 3.1 Error Functions for the Learning Process

For the learning process, some error formula (or fitness function) is needed that will evaluate the visual-motor system configuration by observing the behavior of an agent. As agent’s behavior is deterministic and fully determined by the visual sensor input, the error function can take into account a finite set of positions (angles) around the object. Let us introduce some variables:

- $x_i$  – target (optimal) speed for image viewed from angle  $i$ , i.e.  $x_i = \sin(i)$
- $y_i$  – speed that is generated by the visual-motor system (Fig. 4) for angle  $i$
- $e_i$  – difference between target and actual speed for angle  $i$ ;  $e_i = y_i - x_i$
- $e$  – total error,  $e = \sum_i |e_i|$

In the experiments, the full circle path around the central object was sampled with 100 angles, so  $i \in \{0^\circ, 3.6^\circ, 7.2^\circ, 10.8^\circ, \dots, 356.4^\circ\}$ . The most obvious error function is  $e$ , the sum of individual errors for each angle. However, minimizing this error function resulted in a ragged speed characteristics (see the white line in Fig. 5). Therefore, another component of the error function was introduced, the standard deviation  $\sigma$  of individual errors  $e_i$ . Table 1 summarizes results obtained for minimizing three error functions, and Fig. 5 shows speed characteristics.

The results of this experiment show that minimizing both  $e$  and  $\sigma$  is advantageous, as it reduces raggedness of speed that is output by the visual-motor system (compare white and bold black lines in Fig. 5). Moreover, it helped the hill-climbing learning algorithm to minimize error  $e$ : the value of this error is actually smaller when minimizing  $e+\sigma$  than when only minimizing  $e$  (see Table 1).

**Table 1.** Performance obtained for various error functions

Minimized error function	Trained performance	
	$e$	$\sigma$
$e$	7.63	0.12
$\sigma$	19.93	<b>0.05</b>
$e+\sigma$	<b>6.17</b>	0.08





Fig. 5. Speed versus angle of observation

### 3.2 Adjusting the Number of Motor Modules

The important advantage of the proposed model is its scalability: the number of motor modules can be easily adjusted to match the difficulty of the problem at hand. To investigate the influence of the number of modules (each with 35 parameters) on the training ability, we tested five visual-motor systems. For each system, 20 training experiments were performed, and the results are summarized in Fig. 6.

Although there is little improvement in the best-trained system among 20 trials, it can be clearly seen that both average errors and standard deviations are decreasing as subsequent modules are added. This means that the more modules there are in the system, the better it can cope with transforming visual stimuli into appropriate motor actions, and the stability of results increases.

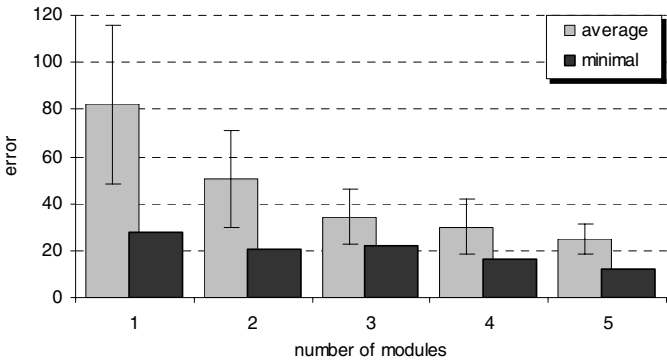


Fig. 6. Error values for visual-motor systems of increasing complexity

## 4 Summary and Future Work

This work presented a biologically-inspired visual-motor coordination model. The model has been verified in a number of navigation tasks and perceived shapes, and proved to be flexible and appropriate for such tasks.

Other interesting experiments with this model were performed as well. It was verified that minor changes in the shape of the 3D object do not deteriorate agent behavior. Changing the size of the object, and changing the distance of the agent from the object did not increase the error much, so the system proved to be robust to minor changes in the environment. The specific place where the agent should stop was also set in various locations around the central object, and training was always successful.

An interesting feature of the proposed model is that the system performance can be visualized and interpreted (explained). It is possible to estimate the influence of each edge on the output speed value, and to visualize it (e.g. edges that cause the agent to move right are red, and those causing the agent to move left are green). Moreover, it is possible to perform such analysis for individual sensory features (like edge angles, lengths, etc.).

Future works concern adding more degrees of freedom for the agent (i.e. moving in the 3D space, and near/far from the object), using evolutionary algorithms for more complex navigation tasks, introducing more complex shapes and real-world objects, using many eyes and visual-motor systems simultaneously, embodiment of such a system within a virtual body, and finally, perceiving realistic, raster camera images. Most of these experiments are work in progress.

## Acknowledgment

This work has been supported by the State Committee for Scientific Research, from KBN research grant no. 3 T11C 050 26, and by the Foundation for Polish Science, from subsidy no. 38/2004.

## References

1. Adamatzky, A., Komosinski, M. (eds): *Artificial Life Models in Software*, chapter 2. Springer-Verlag (2005).
2. Baldi, P.: Gradient descent learning algorithm overview: a general dynamical systems perspective. *IEEE Transactions on Neural Networks* 6:1 (1995) 182–195.
3. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology* 160, London (1962) 106–154.
4. Hubel, D.H., Wiesel, T.N.: Receptive fields and functional architecture of the monkey striate cortex. *Journal of Physiology* 195, London (1968) 215–243.
5. Trullier, O., Wiener, S., Berthoz, A., Meyer, J.-A.: Biologically-based Artificial Navigation Systems: Review and Prospects. *Progress in Neurobiology*, Vol. 51 (1997) 483–544.
6. Yager, R., Rybalov, A.: Uninorm Aggregation Operators. *Fuzzy Sets and Systems* 80 (1996) 111–120.

# A Self-organising Sensing System for Structural Health Management

N. Hoshcke, C.J. Lewis, D.C. Price, D.A. Scott, G.C. Edwards, and A. Batten

CSIRO Industrial Physics, P.O. Box 218, Lindfield, NSW 2070, Australia  
nigel.hoshcke@csiro.au

**Abstract.** This paper describes a new approach to structural health monitoring and management (SHM) that aims to diagnose and respond to damage using the self-organization of a complex system of distributed sensors and processing cells. To develop and evaluate the approach, an experimental SHM test-bed system has been developed, with the aim of detecting and characterising the damage from high-velocity impacts such as those due to micrometeoroids on a space vehicle. An important new feature of the system is an ability to support mobile (robotic) agents that can roam the exterior surface of the test-bed, obtaining additional damage information and providing a crude repair capability. The focus of this paper is the development of a self-organised approach to the operation of such a robotic agent, for which it obtains local information by direct communication with the fixed agents embedded in the underlying structure.

## 1 Introduction

Structural health monitoring and management (SHM) is a new approach to maintaining the operational performance of critical structures, which will initially reduce the need for, and ultimately replace, the current regime of periodic non-destructive inspection and evaluation (NDE). The SHM approach employs sensors built into a structure to continuously monitor its state. The benefits of SHM will be improved safety, reduced costs of maintenance, and, most significantly, it will allow the use of more efficient structural designs. This paper outlines an approach to the sensing and diagnosis of structural damage that utilises principles of self-organisation. Earlier published reports outlining aspects of this work may be found in [1,2,3,4,5,6,7,8,9].

The ultimate functional requirements of an SHM system are that it should be able to detect the onset of damage (or the likelihood of damage), evaluate the nature, severity and extent of damage, and develop a diagnosis of the condition of the structure. On the basis of this diagnosis, it should form a prognosis for the damage and its effect on structural performance, and determine and initiate remedial actions, and monitor their effectiveness. These requirements have been outlined earlier [2].

There are further requirements for effective SHM system operation. Firstly, a viable SHM system must be reliable and robust. It must be less prone to failure than the structure it is monitoring and, since its purpose is to monitor damage, it must be capable of operating effectively in the presence of structural damage: the performance of the system as a whole must not degrade significantly when individual components fail or are damaged. Secondly, SHM is an essentially dynamic problem, since the state of the

structure and environmental conditions may change concurrently with the system response. Thirdly, both damage and required responses are inherently multi-scalar, both spatially and temporally. Small, localised damage modes can be treated locally, perhaps even at a molecular level, while more extensive damage may require a global response from the system. Finally, SHM systems are likely to cover large areas and contain a large number of sensors, so it is necessary to seek a system architecture that is scalable.

The system architecture chosen to satisfy these stringent requirements is that of the complex multi-agent system [10]. Processing capability is distributed throughout the structure, such that sensor data is processed in the local neighbourhood of the sensor, and only compact information is communicated around the system. Each agent consists of a local group of sensors and a controlling processor, which forms a network node. An agent is therefore capable of obtaining only local information, and the system response is developed cooperatively as a result of the interactions between the agents. There is no central controller: damage diagnosis and response are produced by self-organisation. While SHM systems are the subject of a significant world-wide research effort, to the best of our knowledge this is the only reported approach whose operation is based on self-organisation in a complex multi-agent system.

Work reported to date has been based only on fixed sensing agents embedded in the structure. This paper reports progress towards the development of an evolved, self-organised response to guide a mobile (robotic) sensing agent to a detected damage site to provide evaluation and diagnosis of the damage.

## 2 Experimental Test-Bed and Concept Demonstrator

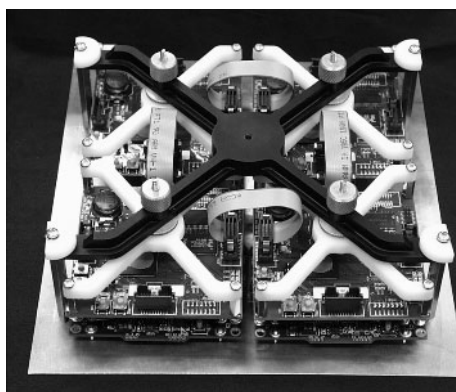
CSIRO, with support from NASA, has been developing an experimental SHM test-bed system for the detection of high-velocity impacts, such as may occur due to the impact of a micrometeoroid on a space vehicle (see [1,2] for some background to this work). The test-bed has been built as a tool for research into sensor design, sensing strategies, communication protocols, and distributed processing using multi-agent systems. High-velocity impacts are simulated using short laser pulses and/or steel spheres fired using a light-gas gun. The SHM system has been designed to be highly flexible: by replacing the sensors and their associated interface and data acquisition electronics, the system can be readily reconfigured for other applications.

The physical structure of the test-bed is a hexagonal prism with an aluminium skin (Figure 1). The initial goal of the test-bed is to detect and characterise impacts on the skin, and to form a diagnosis of the accumulated damage. The skin consists of 48 aluminium panels (eight on each side of the hexagon), each of which contains four “cells”. Cells are the fundamental building blocks of the system: they are the electronic modules containing the sensing, processing and communication electronics. Each cell is an agent of the distributed multi-agent system. It communicates with its four immediate neighbours.

Each cell occupies an area of 100 mm x 100 mm of the skin, mounted on the inside of which are four piezoelectric polymer (PVDF) sensors to detect the acoustic waves that propagate through the skin as a result of an impact. Thus the complete test-bed contains 192 cells. One of the panels, and its four cells, is shown in Figure 2.



**Fig. 1.** The hexagonal prism physical implementation of the test-bed, lying on its side with the end open to reveal the cellular internal structure of the electronics



**Fig. 2.** Aluminium panel containing four cells. Each cell consists of a data acquisition sub-module (DAS) below a network application sub-module (NAS). Each cell is connected to its four immediate neighbours, via the ribbon cables that can be seen in the photograph, to form a square network array.

The cell electronics are constructed as two sub-modules, each 80 mm  $\times$  80 mm and mounted directly on top of each other as shown in Figure 2. One of the sub-modules, called the network application sub-module (NAS), contains the communications and processing hardware, while the data acquisition sub-module (DAS) contains the analogue electronics and digitization hardware specific to the attached sensors. A benefit of this division is that the NAS is flexible enough for almost any SHM sensor network application, and only the DAS needs to be changed to accommodate the different sensors that may be required in different applications. Further details of the electronics can be found in [3,5].

An important feature of the system is an ability to support mobile (robotic) agents that can roam the exterior surface of the test-bed, communicating with the fixed agents embedded in the underlying structure. The function and operation of such an agent will

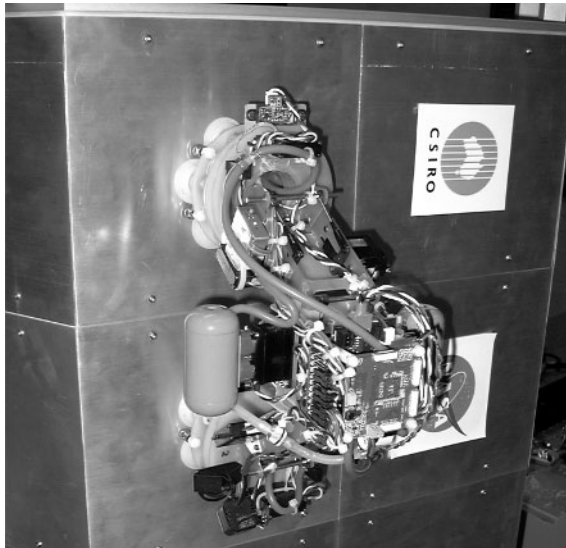
be described in the next section, but it should be emphasised that it is not controlled centrally, but cooperatively with the network of fixed local agents with which it communicates.

It should be emphasized that the system described here is no more than a test-bed, whose primary purpose is for investigation of the practicality of the self-organised complex system approach to damage diagnosis and response. Thus, details of the specific hardware implementation (such as the use of suction for the robot's attachment, which is obviously inconsistent with a space-based application) are not considered to be important at this stage. While the present implementation of the robot is bulky and represents a single point of failure, the eventual aim is to develop a swarm of very small robots that can perform internal or external tasks cooperatively. The work described in this paper represents a first step towards that ultimate goal.

### 3 Description of the Inch-Worm Robot

When sensing impacts using passive sensors, the information received may be insufficient to characterize the damage, and where damage is detected it may need to be repaired. One approach to obtaining additional damage data, and to providing a crude repair capability, is the development of a mobile robot that can move around the outside skin (see Figure 3).

The robot design is based on an articulated box section with six degrees of freedom, and moves rather like an inch-worm. The joints are driven by commercial model aircraft servos and have no position feedback to the controlling processor. The robot is equipped with six suction cups on each of its two feet, and a pneumatic system with a variable

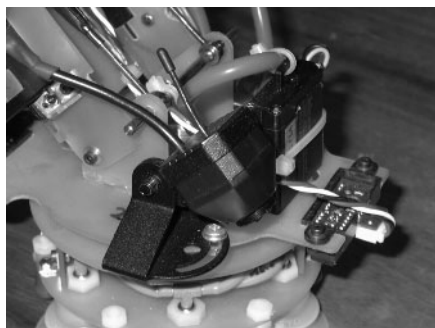


**Fig. 3.** The inchworm robot on a vertical face of the test-bed

speed vacuum pump that allows it to selectively attach and detach its feet to and from the surface. To allow the robot to find and attach to the surface reliably there are two optical range-finders on each foot that measure the distance to, and the angle of the surface. A lithium polymer battery supplies power to the robot for approximately 30 minutes of operation before recharging is necessary.

The robot attaches itself to the skin using suction cups in both feet and has two modes of locomotion. The first mode is very much like an inch-worm: to move forward the robot alternately stretches out and contracts whilst detaching and attaching its feet in sequence. The second mode requires the robot to detach one foot, pivot 180 around the other (attached) foot and then reattach the first. It can change direction by pivoting through 90 . Initially the robot will carry a small video camera (Figure 4), which will send images back to the network for further analysis. In the future other sensors may be included, such as an active ultrasonic transducer that can interact with the piezoelectric receivers embedded in the skin for ultrasonic damage evaluation.

Communication and navigation software for the robot is not yet complete, but it will communicate with the fixed agents in the network using piezoceramic (PZT) ultrasonic transducers in both feet (Figure 5) to pass messages through the aluminium skin to the underlying cells: the fixed agents will receive messages via the four piezoelectric polymer sensors that are used for detecting impacts. A fifth transducer, in this case a piezoceramic, has been added at the centre of each cell for transmission of messages from the cell to the robot. The robot's navigation and functions will be determined cooperatively with the local agents embedded in the test-bed skin, with which it is in contact, as outlined in the next section.



**Fig. 4.** Close-up of one foot of the robot, showing the inspection video camera (foreground), and one of the optical range finders (right)



**Fig. 5.** The ultrasonic communications transducer mounted on the robot foot

Communication through the skin of the test-bed is via an acoustic data link that employs a 937.5 kHz carrier signal introduced into the aluminium skin. When the robot is transmitting, the carrier is generated by the robot's transducer and is received by the four PVDF impact sensors in the cell to which the relevant robot foot is attached. When the cell is transmitting the carrier is generated by the cell's centre transducer and received by the robot's transducer.

The carrier is Binary Phase Shift Key-modulated (BPSK) at a rate of 5.04 kBaud, which is slow enough to allow ultrasonic reflections within the panel to decay before the next symbol is sent. The effective data rate for the channel is approximately 3 k bits/s, the reduced rate compared to the baud rate being due mainly to the error correction encoding used.

The robot will navigate around the surface of the test-bed using data available from the underlying cell to which it is attached at the time. This data is specific to the cell's local neighbourhood, and does not contain any global information about the system. Because the robot has no global navigation capabilities and can only move from one cell to the next using dead reckoning, large position errors could rapidly accumulate as the robot moves over the surface.

To correct such position errors, the robot will use information from the underlying cell to which it has just attached. The robot's position relative to the centre of this cell will be determined by the cell using a modification to the triangulation technique that is used to find the location of impacts: the impact-generated signal will be replaced by a chirp transmitted by the robot's communications transducer. The relative arrival time of the chirp signal at the four receiving sensors of the cell is determined by correlating the signals with a copy of the original digital chirp.

## 4 Impact Detection and System Operation

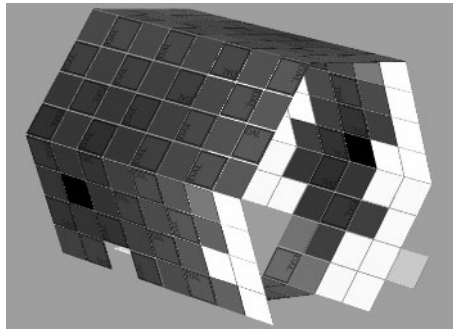
Impacts are detected when the elastic waves they generate are received by an agent's sensors. Some typical signals are shown in [3,4]. The impact location is determined from the arrival times of the signals at the sensors [4].

A general discussion of the approach to damage diagnosis by self-organisation is given in [11]. In this case self-organising maps (Kohonen neural networks [12,13]) have been implemented to classify impact severity, distinguishing critical impacts for which the skin has been ruptured, from non-critical impacts. Electronic failures, which are detected when a cell loses its communication capability, are distinguished from critical impacts that have damaged the electronics by the absence of an impact recorded by neighbouring cell.

There are two related methods by which critical issue of navigation and function of the robot may be achieved, directed by self-organisation. Firstly, algorithms based on ant colony optimisation (ACO) [14,15] have been developed in earlier work to link sub-critical impact locations by a simulated pheromone trail and a dead reckoning scheme (DRS) that form a minimum spanning tree [8]. The decentralised ACO-DRS algorithm has low communication cost, is robust to information loss within any individual cells, and allows navigation around critically damaged regions in which communication capability has been lost. An alternative scheme evaluated in [8] is a distributed dynamic programming algorithm, employing a gradient field set by each impact cell.

While the concept of the robot following the self-organised pheromone trails produced by ACO is appealing, there is a trade-off between the low communication cost of the ACO-DRS algorithm and a better quality of the minimum spanning tree approximation computed by the gradient-based algorithm [8].





**Fig. 6.** An image of the gradient produced by two non-critical impacts that occurred at the black cells. The squares indicate the cells on the surface of the hexagonal prism test-bed. The gradient values are shown as shades of green, with a higher gradient being darker. The white cells are those with which the robot cannot communicate. Absent cells in the image have been physically removed.

In this work, gradients are initiated by an impacted cell, and propagated successively to neighbouring cells. Figure 6 shows the gradient produced from two non-critical impacts, located in the black cells. The gradient value at each cell is indicated by the shade of green. White cells are those with which the robot cannot communicate. This may be due to electronic failure or, as is the case here, sensors have not yet been fitted.

In addition, the gradient-based algorithm incorporates multiple fields to enable an optimal prioritisation to account for the relative importance of visiting severe impact sites rapidly versus the time saved by visiting nearby sites. A gradient field is also established to guide the robot back to its base, in the event that battery recharging is required. The robot decides which gradient to follow, according to circumstances. Dynamic modifications of the gradients occur when new impacts are recorded, or when inspections and/or repair of impact sites is completed.

## 5 Discussion and Future Work

The eventual objective is to extend this system to employ a number of small robots that could perform inspection and repair tasks cooperatively. The introduction of more mobile robotic agents to the system will add a significant degree of complexity to the task of robot navigation. In this scenario, as in the single robot case, the navigation of the mobile agents must be coordinated without the use of a centralised processor. Another complicating factor is that the current hardware does not provide a direct communications channel for robot-to-robot communications. This means that, unless such a channel is provided (possibly by wireless), even detecting and avoiding another robot will require co-operation with the embedded agents.

The development of such a system poses a number fascinating research questions.

- At what level should the decisions be made?
  - Should the robots simply interrogate the network and co-operate with each other to navigate?

- Should the sensor network decide on navigation and direct the robots?
- Can the problem be broken down and solved on different levels?
- Is global knowledge required or can a good solution be formed with only local observation and communication?
- How does the solution scale?
- How does the system deal with conflicts that may be caused by noise, sensor faults, or outdated information being sent through the network?

**Acknowledgements.** It is a pleasure to thank Drs. Ed Generazio and Bill Prosser, NASA Langley Research Center, for their continued encouragement and support for this work, and Dr Mikhail Prokopenko of the CSIRO ICT Centre for his fundamental contributions to this work.

## References

1. D. Abbott, B. Doyle, J.B. Dunlop, A.J. Farmer, M. Hedley, J. Herrmann, G.C. James, M.E. Johnson, B. Joshi, G.T. Poulton, D.C. Price, M. Prokopenko, T. Reda, D.E. Rees, D.A. Scott, P. Valencia, D. Ward and J.G. Winter (2003). Concepts for an Integrated Vehicle Health Monitoring System, *Review of Progress in Quantitative Nondestructive Evaluation*, Vol. 22, pp.1606-14 (eds. D.O. Thompson and D.E. Chimenti), American Institute of Physics Conference Proceedings Vol. 657.
2. D.C. Price, D.A. Scott, G.C. Edwards, A. Batten, A.J. Farmer, M. Hedley, M.E. Johnson, C.J. Lewis, G.T. Poulton, M. Prokopenko, P. Valencia and P. Wang (2003). An Integrated Health Monitoring System for an Ageless Aerospace Vehicle, *Structural Health Monitoring 2003*, Proceedings of the 4th International Workshop on Structural Health Monitoring, Stanford, CA, September 2003. pp. 310-18. (ed. F-K. Chang) DEStech Publications.
3. A. Batten, J. B. Dunlop, G. C. Edwards, A. J. Farmer, B. Gaffney, M. Hedley, N. Hoschke, P. Isaacs, M. E. Johnson, C. J. Lewis, A. Murdoch, G. T. Poulton, D. C. Price, M. Prokopenko, I. Sharp, D. A. Scott, P. Valencia, P. Wang and D. F. Whitnall (2004). *Development and Evaluation of Sensor Concepts for Ageless Aerospace Vehicles. Report 5: Phase 2 Implementation of the Concept Demonstrator*. CSIRO Telecommunications and Industrial Physics. Internal Report No. TIPP 2056, April 2004. (Available on request).
4. D.A. Scott, A. Batten, G.C. Edwards, A.J. Farmer, M. Hedley, N. Hoschke, P. Isaacs, M.E. Johnson, A. Murdoch, C.J. Lewis, D.C. Price, M. Prokopenko, P. Valencia and P. Wang (2005). An Intelligent Sensor System for Detection and Evaluation of Particle Impact Damage. *Review of Progress in Quantitative Nondestructive Evaluation*, Vol. 24, pp. 1825-32 (eds. D.O. Thompson and D.E. Chimenti), American Institute of Physics Conference Proceedings Vol. 760.
5. M. Hedley, M.E. Johnson, C.J. Lewis, D.A. Carpenter, H. Lovatt, D.C. Price (2003). Smart Sensor Network for Space Vehicle Monitoring, *Proceedings of the International Signal Processing Conference (Dallas, Tx.), March 2003*.
6. M. Foreman, M. Prokopenko, P. Wang. (2003). Phase Transitions in Self-organising Sensor Networks. *Advances in Artificial Life - Proceedings of the 7th European Conference on Artificial Life*, LNCS, Vol. 2801, (eds. W. Banzhaf, T. Christaller, P. Dittrich, J.T. Kim, J. Ziegler) Springer, pp. 781-791.
7. M. Prokopenko, P. Wang, D.C. Price, P. Valencia, M. Foreman, A.J. Farmer. (2005). Self-organising Hierarchies in Sensor and Communication Networks. *Artificial Life, Special Issue on Dynamic Hierarchies*, Vol. 11(4), 407-426.

8. M. Prokopenko, P. Wang, M. Foreman, P. Valencia, D. Price, G. Poulton. (2005). On Connectivity of Reconfigurable Impact Networks in Ageless Aerospace Vehicles. *Journal of Robotics and Autonomous Systems*, Vol. 53, 36-58.
9. M. Prokopenko, P. Wang, D. Price. (2005). Complexity Metrics for Self-monitoring Impact Sensing Networks, *Proceedings of 2005 NASA/DoD Conference on Evolvable Hardware (EH-05)*, Washington D.C., USA. pp. 239-46.
10. J. Ferber (1999). Multi-Agent Systems. *An Introduction to Distributed Artificial Intelligence*. Addison-Wesley, London, UK.
11. D. C. Price, A. Batten, G. C. Edwards, A. J. D. Farmer, V. Gerasimov, M. Hedley, N. Hoschke, M. E. Johnson, C. J. Lewis, A. Murdoch, M. Prokopenko, D. A. Scott, P. Valencia and P. Wang (2004). Detection, Evaluation and Diagnosis of Impact Damage in a Complex Multi-Agent Structural Health Management System, *Proceedings of the 2nd Australasian Workshop on Structural Health Monitoring*, Melbourne, Australia. December 2004. pp. 16-27.
12. T. Kohonen (2001). *Self-organizing Maps*. Springer (3rd Edition).
13. T. Kohonen (2003). Self-organized Maps of Sensory Events, *Phil. Trans. Roy. Soc. Lond. A* Vol. 361, 1177-1186.
14. M. Dorigo and G. Di Caro (1999). Ant Colony Optimization: A New Meta-Heuristic, *Proc. 1999 Congress on Evolutionary Computation*, pp. 1470-1477, Washington DC, July 1999.
15. M. Dorigo and T. Sttzle (2004). *Ant Colony Optimization*. MIT Press (Cambridge, Mass.).

# On Models in Fuzzy Propositional Logic

Jorma K. Mattila

Lappeenranta University of Technology  
Laboratory of Applied Mathematics  
jorma.mattila@lut.fi

**Abstract.** A *model theory* of fuzzy propositional logic is considered. The basic frame for fuzzy propositional logics are Zadeh-algebras, i.e., special quasi-Boolean algebras, where valuation functions are universes of these algebras. There are two levels of truth-values, numerical (usually the unit interval  $[0, 1]$ , or in general, a lattice  $L$ ) and linguistic. Linguistic truth-values are fuzzy subsets of the set of numerical truth-values. Fuzzy model is defined based on numerical truth-values, i.e. it is the set of designated truth-values. Its linguistic label is *true*. Truth conditions and the concepts validity, satisfiability, refutability, and invalidity are considered.

**Keywords:** Fuzzy Propositional Logic, Fuzzy Truth-values, Model, Modifier.

## 1 Many-Valued and Fuzzy Propositional Languages

### 1.1 Many-Valued Propositional Language

We consider here a many-valued propositional language with the usual syntax, i.e. alphabets and formation of expressions are quite usual in our language. We define our language as follows.

**Definition 1.** A propositional language  $\mathcal{L}$  consists of

1. a set of propositional letters  $p_0, p_1, \dots, p_k, \dots$  and
2. the truth-functional connectives ' $\wedge$ ', ' $\vee$ ', and ' $\neg$ '.

The connectives can be interpreted as follows: negation  $\neg$  is a strong negation, conjunction  $\wedge$  is *glb*, and disjunction  $\vee$  is *lub*.

**Definition 2.** Well-formed formulas of  $\mathcal{L}$  are given as follows:

$$\alpha ::= p_k \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi. \quad (1)$$

Metavariables are used in usual way to refer to well-formed formulas. We use Greek lower case letters as metavariables.

Definitions 1 and 2 are usually similar for any propositional languages. Differences between propositional languages appear in the ways of defining connectives, especially, how truth conditions are given to them, and what are the truth-values associated with each language. We define the logical connectives to be many-valued in the basic situation, i.e. in the first level. In the second level we use linguistic truth-values being fuzzy sets of the set of many-valued truth-values. The connectives are set operations for fuzzy sets.

When we use linguistic truth-values, we call the corresponding logic *fuzzy propositional logic*, i.e., it is a second level logic.. Because the consideration of truth-values in many-valued logics and in fuzzy logics belongs to the level of meta-language, we need to distinguish linguistic truth-values from similar words in metalevel. Hence, when we refer to linguistic truth-values, we use italic fonts. Thus, for example, the word 'true' is a usual word in meta-language, and '*true*' is a linguistic truth-value.

We must mention that strong negations, *glb*'s (as t-norms), and *lub*'s (as t-conorms) are associated with fuzziness, but these concepts has been introduced much earlier than fuzzy sets and fuzzy logic (cf. for example [10] or [11]). Thus we can consider our object language not to be actually fuzzy, but it is many-valued because of the form of connectives.

### 1.2 Zadeh-Algebras

The algebraic frame for fuzzy propositional logics considered here is *Zadeh-algebra*. It is a certain quasi-Boolean algebra, called also Morgan-algebra (cf. e.g. [9], [10], or [12]). Zadeh-algebras are considered in [4], [5], and [6].

We give the definition of the concept *Zadeh-algebra*. Let  $L$  be a lattice and  $X$  a non-empty set, then the expression  $L^X$  is the set of all functions from  $X$  to  $L$ ,

$$L^X = \{f \mid f : X \longrightarrow L\}.$$

**Definition 3.** *Suppose  $L$  is a distributive complete lattice. Let  $\wedge$  and  $\vee$  be binary operations (lub and glb, respectively) and  $neg$  a strong negation on  $L^X$ , such that the following axioms are satisfied:*

- (Z1) *the operations  $\wedge$  and  $\vee$  are commutative on  $L^X$ ;*
- (Z2) *there exist constant functions  $\mathbf{0}$  and  $\mathbf{1}$ , such that for all  $\mu \in L^X$ ,  $\mu \vee \mathbf{0} = \mu$  and  $\mu \wedge \mathbf{1} = \mu$ ;*
- (Z3) *the operations  $\wedge$  and  $\vee$  are distributive on  $L^X$ ;*
- (Z4) *for any  $\mu \in L^X$ , there exists  $\mu' \in L^X$ , such that  $\mu' = neg(\mu)$ ;*
- (Z5)  *$\mathbf{0} \neq \mathbf{1}$ .*

Then  $\mathcal{Z} = (L^X, \wedge, \vee, neg, \mathbf{0}, \mathbf{1})$  is a Zadeh-algebra.

Because in  $\mathcal{Z} = (L^X, \wedge, \vee, neg, \mathbf{0}, \mathbf{1})$ ,  $L$  is a lattice of the truth-values and  $X$  is the set of propositional variables of the logic, a function  $\mu : X \longrightarrow I$  associates propositional variables with truth-values. Thus  $\mu$  is a *numerical valuation function* corresponding a certain truth-value distribution for formulas. Especially,  $\mathbf{0}$  and  $\mathbf{1}$  assign the values 0 and 1 with formulas, respectively.

When we choose  $L = I = [0, 1]$ , and the standard fuzzy set opetations  $\min$  as  $\wedge$ ,  $\max$  as  $\vee$ , and  $\mathbf{1} - \mu$  as the negation, we have the Zadeh algebra

$$\mathcal{Z}_{I^X} = (I^X, \min, \max, \mathbf{1} - \mu, \mathbf{0}, \mathbf{1}), \tag{2}$$

( $\mu \in I^X$ ), where  $\mathbf{0}$  and  $\mathbf{1}$  are the constant functions  $\mathbf{0} : x \mapsto 0$  and  $\mathbf{1} : x \mapsto 1$ . This is an algebraic approach to the first fuzzy settheory due to L. A. Zadeh (cf. [13]). The set

$I^X$  is the set of all membership functions  $\mu : X \rightarrow I$ . This algebra defines the standard connectives of the second level logic.

Corresponding to the algebra (2), the algebra of truth-functions is

$$Z_I^n = (I^n, \min, \max, \mathbf{1} - \mu, \mathbf{0}, \mathbf{1}), \tag{3}$$

i.e., the elements of the algebra are truth-functions mapping  $n$ -tuples of truth-values to truth-values,  $(i_1, \dots, i_n) \mapsto i_k, (i_1, \dots, i_n, i_k \in I)$ . The constant functions  $\mathbf{0}$  and  $\mathbf{1}$  can be interpreted to constant truth-functions, where  $\mathbf{0}$  assigns the truth-value 0 to any logically false formula, and  $\mathbf{1}$  assigns the truth-value 1 to any logically true formula.

For example, we can associate Łukasiewicz many-valued logics with corresponding Zadeh-algebras. Further, it is possible to extend these logics by adding further operations, like modifiers, to them. Relations between Łukasiewicz many-valued logics, Zadeh-algebras, and modifiers are presented in [8].

### 1.3 Fuzzy Truth-Values

Many-valued truth-values, i.e., numerical values, are often used as fuzzy truth-values. In general, the unit interval  $I = [0, 1]$  is used as a continuous set of truth-values, and also in discrete cases, subsets of  $I$ , where the values are rational numbers including 0 and 1 as the extreme values, are often used (for example in Łukasiewicz logics, cf. [11]). It is clear that usually numerical truth-values form a linearly ordered set, and hence this kind of set is a lattice. Here we use the unit interval  $I$  as the set of numerical truth-values, i.e., the truth-values of the logic in the first level.

Especially in many-valued logics the concept *designated truth-values* are used (cf. for example [11]). These truth-values are those ones we think to indicate the truth. In fact, we may think that the linguistic label of a set of designated truth-values is "true". Thus the numerical set of the designated truth-values is a mathematical model of the linguistic truth-value "true". Hence, the truth-value "true" can be regarded as a fuzzy subset of a set of numerical truth-values, for example,  $true = ]a, 1] \subset [0, 1]$  where  $a$  is chosen from the interval  $0.5 \leq a \leq 1$ , where  $[0, 1]$  is the set of numerical truth-values. Similarly, we define a set of *non-designated truth-values* consisting of the points of an interval  $[0, 1 - a[, 0.5 \leq a \leq 1$ .

Besides numerical truth-values, linguistic truth-values are used in fuzzy logic. A good starting point for this is the concept designated truth-values together with its linguistic interpretation *true*. For example, a set of linguistic truth-values may be

$$\{very\ false, false, more\ or\ less\ false, middle, more\ or\ less\ true, true, very\ true\}. \tag{4}$$

where *middle* is the "mean value" between falsity and truthness. Following the above mentioned leading idea that the designated truth-values form a fuzzy set *true*, the other linguistic truth-values in (4) can be regarded as suitable fuzzy subsets of a given numerical set of truth-values, for example the unit interval  $[0, 1]$ . When we use  $[0, 1]$  as the set of numerical truth-values, we may agree that in the set of linguistic counterparts, *very true* represents the numerical value 1 and *very false* the numerical value 0. The linguistic truth-value *false* refers to the set of non-designated truth values  $D^* = [0, 1 - a[, a \geq 0.5$ .

We combine linguistic truth-values using the operations of Zadeh-algebra, usually those of (2) that is the actual algebra of fuzzy sets. Hence we have linguistic truth-values like "middle or more or less true" and "more or less false and more or less true". It is obvious that, as being fuzzy sets, the last combined linguistic value can be a subset of *middle*. We can modify linguistic fuzzy truth-values of fuzzy propositional logic, like any fuzzy sets, by modifiers. They may be special set modifiers, or simple modifiers for fuzzy sets, i.e., for membership functions. About set modifiers, their topological properties, and relations to fuzzy sets, see [1] and [2]. About simple modifiers, see [3], [5], and [7].

## 2 Definition and Properties of Fuzzy Models

### 2.1 Idea and Definition of Fuzzy Model

In classical propositional logic a *model* is a non-empty set of atomic propositional symbols. It can be considered to be maximal consistent, i.e. adding a new atom to a model it becomes inconsistent. This means that any model depends on a given truth distribution. We can approach the case by considering a model to be connected with a *possible world* where certain atomic states of affairs holds true. Thus, according to the correspondence theory of truth, a true expression indicates a state of affairs to hold in a given possible world. Thus a model is a collection of expressions describing states of affairs holding in the corresponding possible world.

A set of designated truth-values in the set  $I$  is the key concept for defining fuzzy model. We may motivate this by its linguistic counterpart, the linguistic truth-value 'true'. Thus we have the following

**Definition 4.** Let  $a \geq 0.5$  and  $D = ]a, 1]$  be a subset of the set of numerical truth-values  $[0, 1]$ , and let  $\mathcal{M} = \{p_1, \dots, p_k\}$ , such that

$$\forall i = 1, \dots, k, \exists t_i \in D : p_i := t_i.$$

Then  $\mathcal{M}$  is a *fuzzy model* of fuzzy propositional logic. Hence we say that a formula  $\varphi \in \mathcal{M}$  if and only if  $\varphi$  is *true* in  $\mathcal{M}$ .

Note that we sometimes write  $p := t$  instead of  $\mu(p) = t$ .

A connection of fuzzy models to Zadeh-algebras is, that a given valuation  $\mu$  determines a model in such a way that the numerical truth-values of the propositional variables belonging to the model belong to the set of designated truth-values. It is clearly possible that more than one valuations can determine the same model.

**Definition 5.** Let  $\mathcal{M} = \{p_1, \dots, p_k\}$  be a fuzzy model, then the *dual* of  $\mathcal{M}$ , denoted by  $\mathcal{M}^*$ , is a fuzzy model, such that  $\mathcal{M}^* \models \neg p_i, 1 \leq i \leq k$ .

For any  $i = 1, \dots, k$ , let  $\mu(p_i) \in D = ]a, 1] \iff \mu(p_i) > a \iff 1 - a > 1 - \mu(p_i) \iff \mu(\neg p_i) \in [0, 1 - a[ = D^*$  by the axioms of Zadeh-algebras, Definitions 4 and 5. Hence we have got the result

**Theorem 1.**  $\mathcal{M} \models \alpha$  if and only if  $\mu(\neg \alpha) \in D^*$ .

From Theorem 1 it immediately follows

$$\mathcal{M} \models \neg\alpha \iff \mathcal{M}^* \models \alpha, \tag{5}$$

This is the case because of using a strong negation '¬'.

### 2.2 Some Properties of Fuzzy Models

Consider a fuzzy propositional logic with the numerical set of truth-values  $[0, 1]$ , and where the truth-values of connected formulas can be evaluated by the rules

$$\mu(\neg\varphi) =_{df} 1 - \mu(\varphi) \tag{6}$$

$$\mu(\varphi \wedge \psi) =_{df} \min\{\mu(\varphi), \mu(\psi)\} \tag{7}$$

$$\mu(\varphi \vee \psi) =_{df} \max\{\mu(\varphi), \mu(\psi)\}, \tag{8}$$

where  $\mu$  is a valuation associating truth-values with formulas. Let us denote it  $\mathcal{L}_f$ . These connectives come from the Zadeh-algebra  $Z_{fX}$  in (2).

*Example 1.* Let  $\mathcal{M} = \{\varphi, \psi, \omega\}$  and the set of designated truth-values  $D = ]0.7, 1]$ . This means that  $0.7 < \mu(\varphi), \mu(\psi), \mu(\omega) \leq 1$  and the linguistic truth-value of  $\varphi, \psi,$  and  $\omega$  are true in  $\mathcal{M}$ , i.e.,  $\mathcal{M} \models \varphi, \mathcal{M} \models \psi, \mathcal{M} \models \omega$ , or, alternatively, we may write  $\mathcal{M} \models \varphi, \psi, \omega$ .

The set of non-designated truth-values is the interval  $D^* = [0, 0.3[$ . Hence, the linguistic truth-value of  $\neg\varphi, \neg\psi,$  and  $\neg\omega$  is clearly false in the model  $\mathcal{M}$ . In this case, we write  $\mathcal{M} \not\models \neg\varphi$  etc. Also,  $\mathcal{M} \not\models \gamma$  because  $\gamma \notin \mathcal{M} = \{\varphi, \psi, \omega\}$ .

In order to have different fuzzy propositional logics, we may define suitable implication operations which are in accordance with the other connectives. There are several ways to define implications, for example t-conorms give the way to create *S-implications* defined by

$$x \rightarrow y =_{df} S(n(x), y), \quad \text{or} \quad p \rightarrow q \equiv \neg p \vee q.$$

where  $S$  is a t-conorm and  $n$  is a negation on  $[0, 1]$ . The other way is to create *R-implications* by residuation of continuous t-norm  $T$ , i.e.

$$x \rightarrow y =_{df} \sup\{z \in [0, 1] \mid T(x, z) \leq y\}.$$

There also exist some other implications. For example, in Łukasiewicz many-valued logic the connectives negation and implication are the original primitive connectives. Disjunction is defined by the condition

$$p \vee q \equiv (p \rightarrow q) \rightarrow q \tag{9}$$

and conjunction from this by DeMorgan's law

$$p \wedge q \equiv \neg(\neg p \vee \neg q). \tag{10}$$

The truth condition in Łukasiewicz logic to negation is the same as above in (6). The truth condition to implication is

$$\mu(p \rightarrow q) \equiv \min\{1, 1 - \mu(p) + \mu(q)\}. \tag{11}$$



Hence, using (11), (9) and (10) we have the same truth conditions for Łukasiewicz connectives as above in (8) and (7). From the equivalences (9), (10), and (11) together with the negation (6) it follows that in Łukasiewicz logic, the conjunction and disjunction are  $\min\{\mu(p),\mu(q)\}$  and  $\max\{\mu(p),\mu(q)\}$ , respectively (cf., for instance, Rescher [11], or Mattila [8]).

*Example 2.* Consider a fuzzy model  $\mathcal{M} = \{\alpha, \beta\}$  and the set of designated truth-values is  $D = ]a, 1]$ , where  $a$  is chosen under the condition  $0.5 \leq a \leq 1$ . Does  $\mathcal{M} \models \alpha \rightarrow \beta$  in any designated truth-value combinations? If  $\mu(\alpha) \leq \mu(\beta)$  then clearly  $\mathcal{M} \models \alpha \rightarrow \beta$ . Suppose  $\mu(\alpha) > \mu(\beta)$ . The "worst" case is when  $\mu(\alpha) = 1$  and  $\mu(\beta) = b$ , the least designated truth-value. Thus  $b = a + \varepsilon$  where  $\varepsilon > 0$ . Consider the case  $\mu(\beta) = a$ . Then we have

$$\mu(\alpha \rightarrow \beta) = 1 - 1 + a = a = \mu(\beta).$$

Now, if  $\mu(\beta) > a$  then also  $\mu(\alpha \rightarrow \beta) > a$ , i.e.  $\mathcal{M} \models \alpha \rightarrow \beta$ . So, the answer to the question is *yes*.

### 2.3 Truth of a Formula in a Fuzzy Model

In Example 1 we considered only the truth of propositional variables or actually corresponding metavariables and their negations in a given model. Now the connectives determined by the Zadeh-algebra give a motivation to give truth rules for other combined formulas. We collect all the cases here. Let  $p$  be a propositional variable, and  $\alpha$  and  $\beta$  be any formulas. Then

1.  $\mathcal{M} \models p$  iff  $\mu(p) \in D = ]a, 1]$ ,  $0.5 \leq a \leq 1$ . Linguistically it means that  $p$  is true in  $\mathcal{M}$  iff  $p$  is *true* (i.e., *true* is a linguistic truth-value) in  $\mathcal{M}$ ;
2.  $\mathcal{M} \models \neg\alpha$  iff  $\mathcal{M}^* \models \alpha$  (as we showed in (5)). Linguistically,  $\neg\alpha$  is *true* in  $\mathcal{M}$  iff  $\alpha$  is *true* in  $\mathcal{M}^*$ ;
3.  $\mathcal{M} \models \alpha \wedge \beta$  iff  $\mathcal{M} \models \alpha$  and  $\mathcal{M} \models \beta$ . Linguistically,  $\alpha \wedge \beta$  is *true* in  $\mathcal{M}$  iff both  $\alpha$  and  $\beta$  are *true* in  $\mathcal{M}$ ;
4.  $\mathcal{M} \models \alpha \vee \beta$  iff  $\mathcal{M} \models \alpha$  or  $\mathcal{M} \models \beta$ . Linguistically,  $\alpha \vee \beta$  is *true* in  $\mathcal{M}$  iff at least one of  $\alpha$  and  $\beta$  is *true* in  $\mathcal{M}$ .

No doubt, there are some formal similarities between these truth conditions and those of classical propositional logic. This is of course natural, because analyzing the linguistic truth-values *true* and *false* the consideration is similar to the linguistic analysis of classical truth-values.

### 2.4 Validity, Satisfiability, Refutability, and Invalidity

The definitions of the concepts *validity*, *satisfiability*, *refutability*, and *invalidity* are quite similar as those in classical logic.

**Definition 6.** A formula  $\alpha$  is

- (1°) *valid* iff it is *true* in every fuzzy model;
- (2°) *satisfiable* iff there is at least one fuzzy model where it is *true*;

- (3°) *refutable* iff there is at least one fuzzy model where it is *false*;  
 (4°) *invalid* iff it is *false* in every model.

These concepts are very similar to those in classical propositional logic. The reason for this is the same as already mentioned in the subsection 2.3.

### 3 Few Concluding Remarks

Zadeh-algebras and fuzzy models in fuzzy propositional logics are related very closely. Thus Zadeh-algebras serve a good frame and fuzzy models a good tool to fuzzy propositional logics. Many-valued logics as the first level logic is here in the role of auxiliary tools for the second level logics which are actual fuzzy logics with linguistic truth-values. Fuzzy models have an important role in the formal semantics of fuzzy propositional logic with linguistic truth-values. They are very well applicable to the both levels, numerical truth-values based on a given set of designated truth-values and linguistic truth-values based on the truth-value *true* determined by the given numerical designated truth-values.

It is possible to create different fuzzy propositional logics beginning a case where there are only two truth-values *true* and *false* being suitable fuzzy sets of the unit interval  $I$ . This means that we have a two-valued fuzzy propositional logic. We can extend the number of truth-values either by modifying these two truth-values by means of suitable modifiers, or by taking additional truth-values in the similar way as in the set (4). This kind of research work is going on in the Fuzzy Systems Research Group at the Laboratory of Applied Mathematics, Lappeenranta University of Technology.

### References

1. J. Kortelainen, On relationship between modified sets, topological spaces and rough sets, *Fuzzy Sets and Systems* 61 (1994) p. 91-95, North-Holland
2. J. Kortelainen, *A Topological Approach to Fuzzy Sets*, Acta Universitatis Lappeenrantaensis 90, 1999 (Doctoral thesis)
3. J. K. Mattila, On modifier logic, in: L. A. Zadeh, J. Kacprzyk (eds.), *Fuzzy Logic for Management of Uncertainty*, John Wiley & Sons, Inc., New York, 1992
4. J.K. Mattila, "Aristotelian Tradition of Science and Fuzzy Revolution", *Proceedings of International Conference on Fuzzy Systems*, IEEE Catalog No. 04CH37542C, ISBN: 0-7803-8354-0 (invited), 2004.
5. J.K. Mattila, *On Simple Modifiers, Zadeh-Algebras and Modifier Algebras*, Research Report 92, Lappeenranta University of Technology, Department of Information Technology, Lappeenranta, 2004, ISBN 951-764-939-8, ISSN 0783-8069.
6. J.K. Mattila, "Zadeh-Algebras as a Syntactical Approach to Fuzzy Sets", - in: De Baets, De Caluwe, De Tré, Fodor, Kacprzyk, Zadrozny (eds.), *Current Issues in Data and Knowledge Engineering*, Problemy Współczesnej Nauki Teoria I Zastosowania, Informatyka, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2004, (Selected papers presented at EUROFUSE'2004, Warszawa, Poland on September 22-25, 2004) p. 343-349. ISBN 83-87674-71-0.
7. J.K. Mattila, "Modifiers based on some t-norms in fuzzy logic" - *Soft Computing* Vol. 8, No. 10, November 2004, Springer-Verlag, DOI 10.1007/S00500-003-0323-X. Published online: 23 September 2003. ISSN: 1432-7643 (Paper) 2004, ISSN: 1433-7479 (Online) 2003.

8. J.K. Mattila, "On Łukasiewicz Modifier Logic" - *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 9 No. 5, 2005
9. C. V. Negoita, D. A. Ralescu, *Applications of Fuzzy Sets to Systems Analysis*, Birkhäuser, 1975.
10. H. Rasiowa, *An Algebraic Approach to non-classical Logics*, North-Holland, 1974.
11. N. Rescher, *Many-valued Logic*, McGraw-Hill, 1969.
12. H.T. Nguyen, E.A. Walker (2000), *A First Course in Fuzzy Logic*, Chapman & HALL/CRC, Boca Raton, London, New York, Washington, DC.
13. L.A. Zadeh, *Fuzzy Sets*, *Information and Control*, 8, 1965.

# Is Soft Computing in Technology and Medicine Human-Friendly?

Rudolf Seising and Jeremy Bradley

Medical University of Vienna,  
Core Unit for Medical Statistics and Informatics,  
Section of Medical Expert and Knowledge-Based Systems  
Spitalgasse 23, 1090 Vienna, Austria  
rudolf.seising@meduniwien.ac.at,  
jeremy.bradley@gmx.net

**Abstract.** In this paper, which is both historical and philosophical, we regard soft computing (synonymous with computational intelligence) to be an approach in artificial intelligence research. Soft computing and fuzzy sets, its core, were founded by Lotfi A. Zadeh. We show his contributions to the establishment of this field of artificial intelligence since 1950 and we propose that it provides proper methods for the development of human-friendly technology and medicine in the 21<sup>st</sup> century.

## 1 Introduction

In the last decades of the 20<sup>th</sup> century the “technological revolution” of information and communication manifested itself in personal computers, the Internet, e-mail and the World Wide Web. In the 21<sup>st</sup> century, alongside the continuing advance of these developments, another technological revolution is taking place – the revolution of intelligent systems. This label refers to man-made systems that can reason and learn from experience, using machines that can make rational decisions without human intervention. A constituent part of this is the emergence of another new generation of technical systems that have to interact with human beings and their psychological, physiological, and physical behaviour. Such systems of human-friendly technology will be used in the areas of technology, research, industry, business, entertainment, medicine and health care to achieve a higher quality of life.

In the history of artificial intelligence (AI), we distinguish between two research directions, which in AI philosophy are called *strong* AI and *weak* AI [1]. *Weak* AI refers to the use of software to study intelligent systems, but the protagonists of weak AI do not believe in truly intelligent man-made systems. Research in this field consists of the investigation of natural intelligence – of the study of the human mind and its (cognitive) simulation. Examples of this research carried out in the 1950s and 1960s are Allen Newell and Herbert A. Simon’s *Logical Theory Machine* [2] and their *General Problem Solver* [3-5], Herbert Gelernter’s *Geometry Theorem Prover* [6], and game-playing software, e.g. chess programs. These and all the newer systems are devices that do not have a conscious mind.

In *strong* AI scientists and engineers are committed to creating AI-systems that are really able to reason and solve problems, and have self-awareness – in short, to constructing an artificial mind.

Another approach to 20<sup>th</sup> century research programs in AI is to differentiate between conventional AI, on the one hand, and computational intelligence (CI) or soft computing (SC), on the other. In conventional AI the methods of symbolic logic and statistical analysis are used. In this line of research the first expert systems of the 1970s were SHRDLU [7], DENDRAL [8], and MYCIN [9]. The other orientation, CI or SC, is based on non-symbolic AI that is comprised of fuzzy sets and systems, artificial neural networks, and biologically inspired concepts such as evolutionary computation, and collective and swarm intelligence (e.g. ant algorithms). This “non-symbolic” direction of AI is preparing the ground for a technology that seems to be similar to problem-solving procedures as developed in nature and, what’s more, these soft computing methods seem to be a candidate for human-friendly technology.

## **2 Information, Communication, and Thinking Machines: A New Field of Electrical Engineering**

Information theory, the mathematical theory of communication, and cybernetics, developed during the Second World War by Claude E. Shannon, Norbert Wiener, Andrej N. Kolmogorov, Ronald A. Fisher and many others, became well-known in the late 1940s and early 1950s. When Shannon and Wiener went to New York to give lectures on their new theories at Columbia University in 1946, they introduced these new milestones in science and technology to the young doctoral student of electrical engineering Lotfi A. Zadeh. In 1950 Zadeh moderated a discussion between Shannon, Edmund C. Berkeley (author of the book *Giant Brains or Machines That Think* [10] published in 1949), and the mathematician and IBM consultant, Francis J. Murray. “Can machines think?” was Alan Turing’s question in his famous *Mind* article “Computing Machinery and Intelligence” in the same year [11]. Turing proposed the imitation game, now called the “Turing Test”, to decide whether a computer or a program could think like a human being or not. Inspired by Wiener’s *Cybernetics* [12], Shannon’s *Mathematical Theory of Communication* [13], and the new era of digital computers, which began in the 1940s with the Electronic Numerical Integrator and Computer (ENIAC) and the Electronic Discrete Variable Computer (EDVAC) – both designed by J. P. Eckert and J. W. Mauchly – Zadeh wrote the article “Thinking Machines – A New Field in Electrical Engineering” in the student journal *The Columbia Engineering Quarterly* in 1950 [14].

As a prelude, Zadeh quoted some of the headlines that had appeared in newspapers throughout the USA during 1949: “Psychologists Report Memory is Electrical”, “Electric Brain Able to Translate Foreign Languages is Being Built”, “Electronic Brain Does Research”, “Scientists Confer on Electronic Brain,” and he asked, “What is behind these headlines? How will “electronic brains” or “thinking machines” affect our way of living? What is the role played by electrical engineers in the design of these devices?” ([14], p. 12.)

Zadeh was interested in “the principles and organization of machines which behave like a human brain. Such machines are now variously referred to as ‘thinking

machines', 'electronic brains', 'thinking robots', and similar names." In a footnote he added that the "same names are frequently ascribed to devices which are not 'thinking machines' in the sense used in this article" and specified that "The distinguishing characteristic of thinking machines is the ability to make logical decisions and to follow these, if necessary, by executive action." ([14], p. 12) In the article, moreover, Zadeh gave the following definition: "More generally, it can be said that a thinking machine is a device which arrives at a certain decision or answer through the process of evaluation and selection."

On the basis of this definition, he decided that the MIT differential analyzer was not a thinking machine, but that both of the large-scale digital computers that had been built at that time, UNIVAC and BINAC, were thinking machines because they both were able to make non-trivial decisions. ([14], p. 13) Zadeh explained "how a thinking machine works" (Fig. 1) and stated that "the box labeled *Decision Maker* is the most important part of the thinking machine".

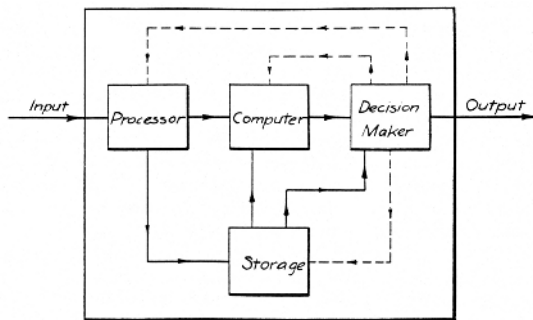


Fig. 1. Zadeh's "diagram with the basic elements of a thinking machine [14]"

### 3 Intelligent Systems, Fuzzy Sets, and Subsequent Developments

"Intelligence" is a very complex concept and it is very difficult to say exactly what it is, not to mention how the intelligence of a natural being can be measured. There has been lively discussion on these issues ever since the French psychologist Alfred Binet created the first intelligence test in 1905 and the German psychologist William (Wilhelm) Stern introduced the intelligence quotient (IQ) in 1912. The concept of artificial intelligence with its various dimensions is at least twice as complex. After the Second World War, when the new digital computers proved to be spectacularly successful, their ability to carry out rational decision-making played a crucial role in case scenarios of future technology, including social implications and potential conflicts.

In his 1950 paper, Zadeh prophesized that these intelligent systems "can be made as crude or as elaborate as it may be desired. It is not as fantastic as it may appear. In fact, such machines may be commonplace in anywhere from ten to twenty years hence. Furthermore, it is absolutely certain that thinking machines will play a major role in any armed conflict that may arise in the future." ([14], p. 30) The world's experience in the Cold War and many military conflicts has made it clear that the

omnipresence of such intelligent systems is a reality. Recently, Zadeh took thinking machines to a higher level, when he described measurement of their intelligence and coined the term MIQ (machine intelligence quotient). From this perspective, an intelligent system is a system which has a high MIQ. ([15], p. 899). Of course, Zadeh does not equate machine intelligence with human intelligence. Far from that, his position includes a separate concept of product-specific and time variant MIQs.

In the mid 1960s Zadeh proposed the theory of fuzzy sets, a mathematically exact theory of sets with non-exact boundaries. Today, it lies at the heart of soft computing and computational intelligence. Zadeh's approach to this "radically different kind of mathematics, the mathematics of fuzzy or cloudy quantities which are not describable in terms of probability distributions" ([16] p. 857) was not a contribution to the philosophy of mathematics, science and technology, but a generalized approach to system theory [17-21]. When he analyzed the ability of conventional mathematical tools in engineering, he saw serious shortcomings: The framework was not adequate for the treatment of systems as complex as those in modern information and communication technology and even less so for those in biology and medicine.

In 1965 Zadeh presented new entities called fuzzy sets, some basics of their properties and algebraic operations of fuzzy sets. In the theory of fuzzy sets it is not only possible for an object to be either an element of a set (membership value 1) or a non-element (membership value 0), it can also have a membership value between 0 and 1. Therefore Zadeh defined fuzzy sets by their *membership function*  $\mu$ , which is allowed to assume any value in the interval [0,1], instead of by their *characteristic function*, under which they can only assume the values 0 or 1 [22].

With respect to the fuzzy sets  $A, B$  in any universe of discourse  $X$  (and for all  $x \in X$ ), Zadeh defined *equality* ( $A = B \Leftrightarrow \mu_A(x) = \mu_B(x)$ ), *containment* ( $A \subseteq B \Leftrightarrow \mu_A(x) \leq \mu_B(x)$ ), *intersection* ( $A \cap B \Leftrightarrow \mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))$ ), *union* ( $A \cup B \Leftrightarrow \mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))$ ), and the complement  $\neg A$  with  $\mu_{\neg A}(x) = 1 - \mu_A(x)$ .

In 1973 Zadeh defined fuzzy relations: If  $L(A \times B)$  is the set of all fuzzy sets in the Cartesian product  $A \times B$  of crisp sets  $A$  and  $B$ , then a fuzzy relation is an element of  $L(A \times B)$  [23]. In the case of three sets  $A, B$ , and  $C$  with the fuzzy relations  $Q \subseteq L(A \times B)$  and  $R \subseteq L(B \times C)$ , in order to get another fuzzy relation  $T \subseteq L(A \times C)$ , Zadeh introduced the combination rule of a *max-min-composition*:  $T = Q * R$  is defined by the following membership function

$$\mu_T(x, y) = \max_{y \in B} \min \{ \mu_Q(x, y) : \mu_R(y, z) \}, x \in A, y \in B, z \in C.$$

Later, in the 1970s, Zadeh introduced the concepts of linguistic variables and fuzzy algorithms [24]. The potential of these new techniques in the theory of fuzzy sets prompted Ebrahim Mamdani, a professor of electrical engineering in London, to develop a trial in which a fuzzy control system could be realized under laboratory conditions. He proposed this plan to his doctoral student Sedrak Assilian, who – in the course of a few days – designed a fuzzy algorithm to control a small steam engine with the input variables *heat* and *throttle* and the output variables *pressure* and *speed* (Fig. 2) by a fuzzy rule based system [25, 26].

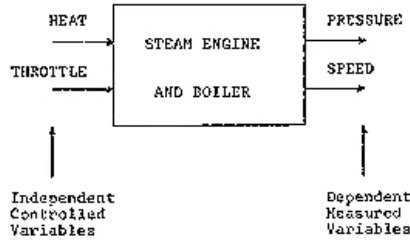


Fig. 2. Schema of the real “steam engine“ system [25]

Some 10 years after the foundation of the theory of fuzzy sets and systems, the new mathematic theory had thus embarked onto the terrain of real world applications in science and technology. This was the prelude to the “fuzzy boom” that began in the 1980s in Japan with cameras, rice cookers, washing machines, etc., and later spread to the west as well.

Zadeh proposed the medical sciences as a field of application as early as 1969, when he wrote that “a human disease, e.g., diabetes, may be regarded as a fuzzy set”, and he proposed to characterize it by its relation to various symptoms which in themselves are fuzzy in nature. For example, in the case of diabetes a fuzzy symptom might be, for example, hardening of the arteries. If this fuzzy set in  $X$  [a collection of human beings] is denoted by  $A$ , then we can speak of the fuzzy inclusion relation between [the fuzzy set “diabetes”]  $D$  and  $A$  and assign a number in the interval  $[0,1]$  to represent the “degree of containment” of  $A$  in  $D$ . In this way, we can provide a partial characterization of  $D$  by specifying the degrees of containment of various fuzzy symptoms  $A_1, \dots, A_k$  in  $D$ .” ([27], p. 205)

A fully developed theory to model relationships of symptoms and diseases came from Elie Sanchez in Marseille in 1979. “In a given pathology, we denote by  $S$  a set of symptoms,  $D$  a set of diagnoses and  $P$  a set of patients. What we call ‘medical knowledge’ is a fuzzy relation, generally denoted by  $R$ , from  $S$  to  $D$  expressing associations between symptoms, or syndromes, and diagnoses, or groups of diagnoses.” ([28], p. 438) Sanchez adopted Zadeh’s *max-min-composition rule* as an inference mechanism. It accepts fuzzy descriptions of the patient’s symptoms and infers fuzzy descriptions of the patient’s diseases by means of the fuzzy relationships described above. If a patient’s symptom is  $S_i$ , then his/her state in terms of the diagnosis is a fuzzy set  $D_j$  with the following membership function:

$$\mu_T(p, d) = \max_{s \in S} \min \{ \mu_Q(p, s) : \mu_R(s, d) \}, s \in S, d \in D, p \in P.$$

$\mu_R(s, d)$  is the membership function of the fuzzy relation “medical knowledge”.

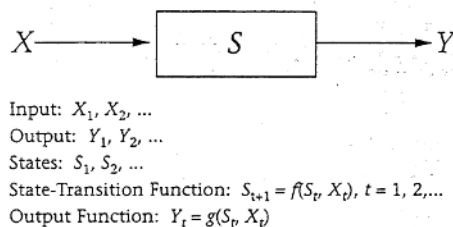
In the 1980s, Zadeh’s visions of the future applications of his theory of fuzzy sets went beyond domestic appliances, cameras and the like, and even exceeded fuzzy process controllers and fuzzy expert systems. His call then was “*Making Computers Think like People*” [29]. In the 1990s he established “*Computing with Words*” (CW) [30, 31] as a methodology for reasoning and computing with perceptions based on the theory of fuzzy sets. In 1996 he stated that “the main contribution of fuzzy logic is a



methodology for computing with words. No other methodology serves this purpose” ([30], p. 103). Three years later he clarified that “the computational theory of perceptions, or CTP for short, is based on the methodology of CW. In CTP, words play the role of labels of perceptions and, more generally, perceptions are expressed as propositions in natural language” ([31], p. 105).

#### 4 Perception-Based Technology = Human-Friendly Technology

Soft computing with its core theory of fuzzy sets became a new dimension of artificial intelligence in the final years of the 20<sup>th</sup> century. Zadeh was inspired by the “remarkable human capability to perform a wide variety of physical and mental tasks without any measurements and any computations. Everyday examples of such tasks are parking a car, playing golf, deciphering sloppy handwriting and summarizing a story. Underlying this capability is the brain’s crucial ability to reason with perceptions – perceptions of time, distance, speed, force, direction, shape, intent, likelihood, truth and other attributes of physical and mental objects.” ([15], p. 903). Zadeh observed “that progress has been, and continues to be, slow in those areas where a methodology is needed in which the objects of computation are perceptions – perceptions of time, distance, form, direction, color, shape, truth, likelihood, intent, and other attributes of physical and mental objects.” ([16], p. 73.) He set out his ideas on “A New Direction in AI” in the *AI Magazine* in the spring of 2001. [16] In this article he presented perception-based system modeling: “A system,  $S$ , is assumed to be associated with temporal sequences of input  $X_1, X_2, \dots$ ; output  $Y_1, Y_2, \dots$ ; and states  $S_1, S_2, \dots$ .  $S_2$  is defined by state-transition function  $f$  with  $S_{t+1} = f(S_t, X_t)$ , and output function  $g$  with  $Y_t = g(S_t, X_t)$ , for  $t = 0, 1, 2, \dots$ . In perception-based system modeling, inputs, outputs and states are assumed to be perceptions, as state-transition function,  $f$ , and output function,  $g$ .” ([16], p. 77.)



**Fig. 3.** Perception-based system modeling, [16]

Man-made systems that can compute with words and can therefore reason with human perceptions could be the next generation on the way to more human-friendly technology. The human brain can be regarded as a fundamentally fuzzy system. Only in very few situations do people reason in binary terms, as machines classically do. This human characteristic is reflected in all natural languages, in which very few terms are absolute. The use of language is dependent on specific situations and is very seldom 100% certain. For example, the word “thin” cannot be defined in terms of

numbers and there is no measurement at which this term suddenly stops being applicable. Human thinking, language and reasoning can thus indeed be called fuzzy. The theory of fuzzy sets has created a logical system far closer to the functionality of the human mind than any previous logical system.

Fuzzy sets enable computers and human beings to communicate in terms that enable them to express uncertainty regarding measurements, diagnostics, evaluations, etc. In theory, this should put the methods of communication used by machines and human beings on levels that are much closer to each other. In practice, however, any medical decision includes a certain degree of subjectivity on the part of the physician, which is generally based on facts determined by various means. Some practitioners might see fuzzy sets that include uncertainties of other parties as an intrusion into their area of responsibility. They want to review the hard facts themselves and base their decisions on what they consider right and relevant. They might see being subjective as their job, rather than that of a computer that lacks feelings and common sense. In spite of its advances, artificial intelligence cannot compete with human intelligence on many levels and will not be able to do so in the very near future.

However, fuzzy representation would not exclude the availability of exact facts that would be at the disposal of the user with other applications – any knowledge-based system must be able to justify its reasoning. Soft computing should provide proper tools for the development of human-friendly systems in medicine and technology. Studies regarding the willingness of human beings to accept systems of this sort in every-day hospital life is a topic of great interest for the near future.

## References

1. Searle, John R.: *Minds, Brains and Programs*, *The Behavioral and Brain Sciences*, vol. 3, pp. 417-458, (1980).
2. Newell, Allen, Simon, Herbert A.: *The logic theory machine*. *IRE Transactions on Information Theory*, vol. IT-2 (3), pp. 61-79, (1956).
3. Newell, Allen, Shaw Clifford, J., Simon, Herbert A.: *Report on a General Problem-Solving Program*, *Proc. of the Int. Conf. Info. Process.*, pp. 256-264, (1960).
4. Feigenbaum, Edward A., Feldman, Julian(eds): *Computers and Thought*, McGraw-Hill, New York, (1963).
5. Newell, Allen, Simon, Herbert A.: *GPS, a program that simulates human thought*. In: 4., pp. 279-293.
6. Gelernter, Herbert: *Realization of a Geometry Theorem Proving Machine*. In: *Proc. Int. Conf. Info. Process.*, pp. 273-282. Also in 4., pp. 134-152.
7. Winograd, Terry: *Understanding Natural Language*, *Cognitive Psychology*, vol. 3, 1 pp. 1-191, (1972).
8. Buchanan, B. G., Feigenbaum, E. A.: *DENDRAL and Meta-DENDRAL: Their applications dimension*, *Journal of Artificial Intelligence*, vol. 11, 5, (1978).
9. Shortliffe, Edward Hance: *Computer-based medical consultations: MYCIN*. New York: Elsevier, (1976). Based on a PhD thesis, Stanford University, Stanford, CA, (1974).
10. Berkeley, Edward C.: *Giant Brains or Machines That Think*, New York: John Wiley & Sons, (1949).
11. Turing, Alan Mathison: *Computing machinery and intelligence*, *Mind*, vol. LIX, no. 236, pp. 433-460, (1950).

12. Wiener, Norbert: *Cybernetics or Control and Communication in the Animal and the Machine*, Cambridge-Mass., and New York: Hermann & Cie., and John Wiley & Sons, (1948).
13. Shannon, Claude E.: A mathematical theory of communication, *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, (1948).
14. Zadeh Lotfi A.: Thinking machines – a new field in electrical engineering, *Columbia Engineering Quarterly*, January, pp. 12-13, 30-31, (1950).
15. Zadeh Lotfi A.: The Birth and Evolution of Fuzzy Logic – A Personal Perspective, *Journal of Japan Society for Fuzzy Theory and Systems*, Vol. 11, No. 6, pp. 891-905, (1999).
16. Zadeh, Lotfi A.: From Circuit Theory to System Theory. *Proceedings of the IRE*, vol. 50, pp. 856-865, (1962).
17. Seising, Rudolf: Noninferiority, Adaptivity, Fuzziness in Pattern Separation: Remarks on the Genesis of Fuzzy Sets, *Proc. of the Annual Meeting of the North American Fuzzy Information Processing Society: (NAFIPS 2004)*, Banff, Alberta, Canada, pp. 2002-2007.
18. Seising, Rudolf: 40 years ago: 'Fuzzy Sets' is going to be published, *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks / International Conference on Fuzzy Systems (FUZZ-IEEE 2004)*, Budapest, Hungary, CD.
19. Seising, Rudolf: 1965 – 'Fuzzy Sets' appear – A Contribution to the 40th Anniversary. *Proceedings of the Conference FUZZ-IEEE 2005*, Reno, Nevada, CD.
20. Seising, Rudolf: The 40th Anniversary of Fuzzy Sets – A New View on System Theory". In: Hao Ying, Dimitar Filev (eds.): *Proc. Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS 2005)*, Ann Arbor, Michigan, USA, pp. 92-97.
21. Seising, Rudolf: Die Fuzzifizierung der Systeme. Die Entstehung der Fuzzy Set Theorie und ihrer ersten Anwendungen, Stuttgart: Franz Steiner (2005).
22. Zadeh, L. A.: Fuzzy Sets, *Information and Control*, 8, pp. 338-353, (1965).
23. Zadeh Lotfi A.: Similarity relations and fuzzy orderings, *Information Science*, vol. 3, pp. 177-200, (1971).
24. Zadeh Lotfi A.: Outline a New Approach to the Analysis of Complex Systems and Decision Processes, *IEEE Transactions on Systems Theory*, vol. 3, pp. 28-44, (1973).
25. Assilian Sedrak: *Artificial Intelligence in the Control of Real Dynamic Systems*. Ph. D. Thesis Nr. DX193553, University London, August (1974).
26. Mandani Ebrahim H., Assilian, Sedrak: An experiment in linguistic synthesis with a Fuzzy logic controller, *Int. Journal of Man-Machine Studies*, vol. 7, Nr. 1, pp. 1-13. (1975).
27. Zadeh, Lotfi A.: Biological Applications of the Theory of Fuzzy Sets and Systems. *Proc. Intern. Symp. on Biocybernetics of the Central Nervous System*. Little, Brown and Company: Boston, pp. 199-206, (1969).
28. Sanchez, Elie: Medical Diagnosis and Composite Fuzzy Relations. In: Gupta, M. M., Ragainde, R. K., Yager Ron R. (eds.): *Advances in Fuzzy Set Theory and Applications*. Amsterdam, New York, Oxford: North-Holland, 437-444, (1979).
29. Zadeh Lotfi A.: Making Computers Think like People, *IEEE Spectrum*, 8, pp. 26-32, (1984).
30. Zadeh Lotfi A.: Fuzzy Logic = Computing with Words, *IEEE Transactions on Fuzzy Systems*, Vol. 4, No. 2, pp. 103-111, (1996).
31. Zadeh Lotfi A.: From Computing with Numbers to Computing with Words – From Manipulation of Measurements to Manipulation of Perceptions, *IEEE Trans. on Circuits And Systems-I: Fundamental Theory and Applications*, Vol. 45, No. 1, pp. 105-119, (1999).
32. Zadeh, Lotfi A.: A New Direction in AI. Toward a Computational Theory of Perceptions. *AI-Magazine*, vol. 22, No. 1, pp. 73-84, (2001).

# From Vague or Loose Concepts to Hazy and Fuzzy Sets – Human Understanding Versus Exact Science

Rudolf Seising and Jeremy Bradley

Medical University of Vienna,  
Core Unit for Medical Statistics and Informatics,  
Section of Medical Expert and Knowledge-Based Systems  
Spitalgasse 23, 1090 Vienna, Austria  
rudolf.seising@meduniwien.ac.at,  
jeremy.bradley@gmx.net

**Abstract.** In this paper we examine important research carried out on vagueness, haziness, and fuzziness in 20<sup>th</sup> century philosophy, logic, and science. Whereas vagueness was avidly discussed in the fields of logic and philosophy during the first decades of the century – particularly in Vienna and at Cambridge – haziness and fuzziness were concepts of interest in mathematics and engineering during the second half of the 1900s. Our logico-philosophical history covers the work of Bertrand Russell, Max Black, and Ludwig Wittgenstein. The mathematical-technical history deals with the theories founded by Karl Menger and Lotfi Zadeh. We note interesting connections between these two protagonists and their findings as well as their preparatory work for the establishment of human-friendly technology.

## 1 Introduction

Since the 17<sup>th</sup> century, beginning with Galileo (1564-1642) and Descartes (1596-1650), the tools of logics and mathematics have given modern science its exactness. It has been possible to formulate axioms, definitions, theorems, and proofs in the language of mathematics. Moreover, the ascendancy of modern science achieved through the works of Newton (1643-1727), Leibniz (1646-1716), Laplace (1749-1827) and many others made it seem that science was able to represent all the facts and processes that people observe in the world. But scientists are human beings, who also use natural languages that clearly lack the precision of logic and mathematics. Human language – and possibly also human understanding – are not strictly defined systems. On the contrary, they include a great deal of uncertainty, ambiguities, etc. They are, in fact, vague.

“Vagueness” is also part of the vocabulary of modern science. In his *Essay Concerning Human Understanding* (1689) John Locke (1632-1704) complained about the “vague and insignificant forms of speech” and in the French translation (1700), the French word “vague” is used for the English word “loose”. Nevertheless, “vague” did not become a technical term in philosophy and logic during the 18<sup>th</sup> and

19<sup>th</sup> century. In the 20<sup>th</sup> century, however, philosophers like Gottlob Frege (1848-1925), Bertrand Russell (1872-1970), Ludwig Wittgenstein (1889-1951), and Max Black (1909-1988) focused attention on and analyzed the problem of “vagueness” in modern science.

In the first third of the 20<sup>th</sup> century another group of philosophers who concerned themselves with the interrelationships between logic, science, and the real world was the so-called Vienna Circle. These scholars regularly debated these issues over a period of years until the annexation of Austria by Nazi Germany in 1938 marked the end of the group. One member of the Vienna Circle was Karl Menger (1902-1985), who later became a professor of mathematics in the USA. As a young man in Vienna, Menger raised a number of important questions that culminated in the so-called principle of logical tolerance. In addition, in his work after 1940 on the probabilistic or statistical generalization of metric space, he introduced the new concepts “hazy sets” (ensembles flous), t-norms and t-conorms, which are also used today in the mathematical treatment of problems of vagueness in the theory of fuzzy sets.

This new mathematical theory to deal with vagueness was established in the mid 1960s by Lotfi A Zadeh, who was then a professor of electrical engineering at Berkeley. In 1962 he described the basic necessity of a new scientific tool to handle very large and complex systems in the real world: “we need a radically different kind of mathematics, the mathematics of fuzzy or cloudy quantities which are not describable in terms of probability distributions. Indeed, the need for such mathematics is becoming increasingly apparent even in the realm of inanimate systems, for in most practical cases the *a priori* data as well as the criteria by which the performance of a man-made system are judged are far from being precisely specified or having accurately-known probability distributions” ([2], p. 857). In the two years following the publication of this paper, Zadeh developed the theory of fuzzy sets [3-5], and it has been possible to reconstruct the history of this process [6-12].

Very little is known about the connectivity between the philosophical work on vagueness and the mathematical theories of hazy sets and fuzzy sets. In this paper we will show that there is common ground in the scientific developments that have taken place in these different disciplines, namely, the attempt to find a way to achieve a more human-friendly science. This is needed because the exactness of modern science does not correspond to human understanding and language.

## 2 Vagueness

In the early 20<sup>th</sup> century, when the German philosopher and mathematician Gottlob Frege (1848-1925) published his *Grundgesetze der Arithmetik* (Foundations of Arithmetic), he called for concepts with sharp boundaries, because otherwise we could break logical rules and, moreover, the conclusions we draw could be false [7]. Frege’s specification of vagueness as a particular phenomenon influenced other scholars, notably the British philosopher and mathematician Bertrand Russell (1872-1970), who published the first article on “Vagueness” in 1923 [8]. Russell stated “that every proposition that can be framed in practice has a certain degree of vagueness; that is to

say, there is not one definite fact necessary and sufficient for its truth, but a certain region of possible facts, any one of which would make it true. And this region is itself ill-defined: we cannot assign to it a definite boundary.” ([8], p. 88).

Russell emphasized that there is a difference between what we can imagine in theory and what we can observe with our senses in reality: “All traditional logic habitually assumes that precise symbols are being employed. It is therefore not applicable to this terrestrial life, but only to an imagined celestial existence.” ([8], p. 88 f). He proposed the following definition of accurate representations: “One system of terms related in various ways is an accurate representation of another system of terms related in various other ways if there is a one-one relation of the terms of the one to the terms of the other, and likewise a one-one relation of the relations of the one to the relations of the other, such that, when two or more terms in the one system have a relation belonging to that system, the corresponding terms of the other system have the corresponding relation belonging to the other system.” ([8], p. 89). And in contrast to this, he stated that “a representation is *vague* when the relation of the representing system to the represented system is not one-one, but one-many.” ([8], p. 89) He concluded that “Vagueness, clearly, is a matter of degree, depending upon the extent of the possible differences between different systems represented by the same representation. Accuracy, on the contrary, is an ideal limit.” ([8], p. 90).

The Cambridge philosopher and mathematician Max Black responded to Russell’s article in “Vagueness. An exercise in logical analysis”, published in 1937. [9]. Influenced by Russell and Wittgenstein (and the other famous analytical philosophers at Cambridge, Frank P. Ramsey (1903-1930) and George E. Moore (1873-1958), he continued Russell’s approach to the concept of vagueness and differentiated vagueness from ambiguity, generality, and indeterminacy. He emphasized “that the most highly developed and useful scientific theories are ostensibly expressed in terms of objects never encountered in experience. The line traced by a draughtsman, no matter how accurate, is seen beneath the microscope as a kind of corrugated trench, far removed from the ideal line of pure geometry. And the ‘point-planet’ of astronomy, the ‘perfect gas’ of thermodynamics, or the ‘pure species’ of genetics are equally remote from exact realization.” ([9], p. 427)

Black proposed a new method to symbolize vagueness: “a quantitative differentiation, admitting of degrees, and correlated with the indeterminacy in the divisions made by a group of observers.” ([9], p. 441) He assumed that the vagueness of a word involves variations in its application by different users of a language and that these variations fulfil systematic and statistical rules when one symbol has to be discriminated from another. He defined this discrimination of a symbol  $x$  with respect to a language  $L$  by  $DxL (= Dx \neg L)$ .

Most speakers of a language and the same observer in most situations will determine that either  $L$  or  $\neg L$  is used. In both cases, among competent observers there is a certain unanimity, a preponderance of correct decisions. For all  $DxL$  with the same  $x$  but not necessarily the same observer,  $m$  is the number of  $L$  uses and  $n$  the number of  $\neg L$  uses. On this basis, Black stated the following definition: “We define *the consistency of application of  $L$  to  $x$*  as the limit to which the ratio  $m/n$  tends when the number of  $DxL$  and the number of observers increase indefinitely. [...] Since the

consistency of the application,  $C$ , is clearly a function of both  $L$  and  $x$ , it can be written in the form  $C(L, x)$ .” ([9], p. 442)

In his 1963 article “Reasoning with loose concepts”, Black labelled concepts that do not have precise boundaries “loose concepts” rather than “vague” ones, in order to avoid misleading and pejorative implications [10]. In this, he once again expressly rejected Russell’s assertion that traditional logic is “not applicable” as a method of conclusion for vague concepts: “Now, if all empirical concepts are loose, as I think they are, the policy becomes one of abstention from any reasoning from empirical premises. If this is a cure, it is one that kills the patient. If it is always wrong to reason with loose concepts, it will, of course, be wrong to derive any conclusion, paradoxical or not, from premises in which such concepts are used. A policy of prohibiting reasoning with loose concepts would destroy ordinary language – and, for that matter, any improvement upon ordinary language that we can imagine.” ([10], p. 7)

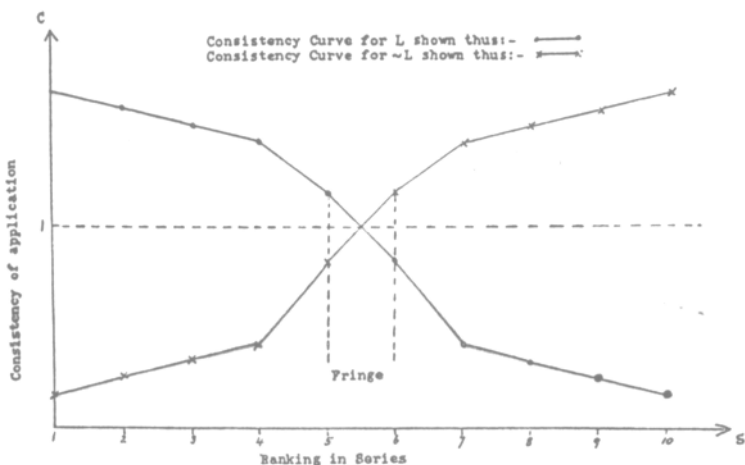


Fig. 1. Consistency of application of a typical vague symbol, ([9], p. 443)

### 3 Haziness – Ensembles Flous

Karl Menger (1902-1985), a mathematician in Vienna and later in the USA, was one of the first to begin laying the groundwork for human-friendly science, i.e. for the development of scientific methods to deal with loose concepts. In his work, he never abandoned the framework of classical mathematics, but used probabilistic concepts and methods.

Menger was a professor of geometry at the University of Vienna from 1927 onwards, but in 1937 – one year before the annexation of Austria by Nazi Germany – he immigrated to the USA, where he was appointed to a professorship at the University of Notre Dame. In 1948 Menger moved to the Illinois Institute of Technology, and he remained in Chicago for the rest of his life. With the intention of generalizing the theory of metric spaces more in the direction of probabilistic concepts, he introduced

the term “statistical metric” in 1942: A *statistical metric* is “a set  $S$  such that with each two elements (‘points’)  $p$  and  $q$  of  $S$ , a probability function  $\Pi(x; p, q)$  is associated satisfying the following conditions:

- 1.  $\Pi(0; p, p) = 1$ .
- 2. If  $p \neq q$ , then  $\Pi(0; p, q) < 1$ .
- 3.  $\Pi(x; p, q) = \Pi(x; q, p)$
- 4.  $T[\Pi(x; p, q), \Pi(y; q, r)] \leq \Pi(x+y; p, r)$ .

where  $T[\alpha, \beta]$  is a function defined for  $0 \leq \alpha \leq 1$  and  $0 \leq \beta \leq 1$  such that

- (a)  $0 \leq T(\alpha, \beta) \leq 1$ .
  - (b)  $T$  is non-decreasing in either variable.
  - (c)  $T(\alpha, \beta) = T[\beta, \alpha]$ .
  - (d)  $T(1, 1) = 1$ .
  - (e) If  $\alpha > 0$ , then  $T(\alpha, 1) > 0$ .”
- ([11], p. 535f)

Menger called  $\Pi(x; p, q)$  the *distance function of  $p$  and  $q$* , which bears the meaning of the *probability* that the points  $p$  and  $q$  have a distance  $\leq x$ . Condition 4, the “triangular inequality” of the statistical metric  $S$  implies the following inequality for all points  $q$  and all numbers  $x$  between 0 and  $z$ :

$$\Pi(z; p, r) \geq \text{Max } T[\Pi(x; p, q), \Pi(z-x; q, r)].$$

In this paper Menger used the name *triangular norm (t-norm)* to indicate the function  $T$  for the first time.

In another paper the same year [12], Menger addressed the difference between the mathematical and the physical continuum. Regarding  $A, B$ , and  $C$  as elements of a continuum, he referred to Henri Poincaré’s claim “that only in the mathematical continuum do the equalities  $A = B$  and  $B = C$  imply the equality  $A = C$ . In the observable physical continuum, ‘equal’ means ‘indistinguishable’, and  $A = B$  and  $B = C$  by no means imply  $A = C$ . »The raw result of experience may be expressed by the relation  $A = B, B = C, A < C$ , which may be regarded as the formula for the physical continuum.« According to Poincaré, physical equality is a non-transitive relation.” ([18], p. 178.) Menger suggested a realistic description of the equality of elements in the physical continuum by associating with each pair  $(A, B)$  of these elements the probability that  $A$  and  $B$  will be found to be indistinguishable. He argued: “For it is only very likely that  $A$  and  $B$  are equal, and very likely that  $B$  and  $C$  are equal – why should it not be less likely that  $A$  and  $C$  are equal? In fact, why should the equality of  $A$  and  $C$  not be less likely than the inequality of  $A$  and  $C$ ?” ([12], p. 178.) To solve “Poincaré’s paradox” Menger used his concept of probabilistic relations and geometry: For the probability  $E(a, b)$  that  $a$  and  $b$  would be equal he postulated:

- (1)  $E(a, a) = 1$  for every  $a$ ;
- (2)  $E(a, b) = E(b, a)$ , for every  $a$  and  $b$ ;
- (3)  $E(a, b) \cdot E(b, c) \leq E(a, c)$ , for every  $a, b, c$ .

If  $E(a, b) = 1$ , then he called  $a$  and  $b$  *certainly equal*. (In this case we obtain the ordinary equality relation.) “All the elements which are certainly equal to  $a$  may be united to an ‘equality set’,  $A$ . Any two such sets are disjoint unless they are identical.” ([12], p. 179.)



In 1951, as a visiting lecturer at the Sorbonne University, Menger presented similar ideas in the May session of the French Académie des sciences: “Ensembles flous et fonctions aléatoires”. He proposed to replace the ordinary element relation “ $\in$ ” between each object  $x$  in the universe of discourse  $U$  and a set  $F$  by the probability  $\Pi_{F(x)}$  of  $x$  belonging to  $F$ . In contrast to ordinary sets, he called these entities “ensembles flous” [13]. Later, he also used the English expression “hazy sets” [14].

At the 1966 Mach symposium held by the *American Association for the Advancement of Science* to honour the Viennese physicist and philosopher Ernst Mach (1838-1916), Menger presented a paper on “Positivistic Geometry” [14]. He began with a quotation from Mach’s chapter on the continuum in *Die Prinzipien der Wärmelehre* (Principles of Thermodynamics) published in 1896: “All that *appears* to be a continuum might very well consist of *discrete* elements, provided they were sufficiently small compared with our smallest practically applicable units or sufficiently numerous.” [15] Again he described Poincaré’s reflections on the physical continuum and he summarized his own work on statistical metrics, probabilistic distances, indistinguishability of elements, and ensembles flous. He believed it would be important in geometry to combine these concepts with that “of lumps, which can be more easily identified and distinguished than points. Moreover, lumps admit an intermediate stage between indistinguishability and apartness, namely that of overlapping. It is, of course, irrelevant whether the primitive (i.e. undefined) concepts of a theory are referred to as points and probably indistinguishable, or as lumps and probably overlapping. All that matters are the assumptions made about the primitive concepts. But the assumptions made in the two cases would not be altogether identical and I believe that the ultimate solution of problems of microgeometry may well lie in a probabilistic theory of hazy lumps. The essential feature of this theory would be that lumps would not be point sets; nor would they reflect circumscribed figures such as ellipsoids. They would rather be in mutual probabilistic relations of overlapping and apartness, from which a metric would have to be developed.” [14]

It seems that Menger could not envisage a mathematical theory of vagueness that was different from probability theory. He had taken notice of Zadeh’s postulation of fuzzy sets in an article that had appeared the same year [14], and he had compared his microgeometry with the theory of fuzzy sets, in which one speaks “of the degree rather than the probability of an element belonging to a set.” [14] Menger did not see that this “slight difference” between “degrees” (of fuzziness) and “probabilities” is a difference not only in terminology, but also in the meaning of the concepts.

## 4 Fuzzy Sets – An Exact Science for Non-exact Phenomena

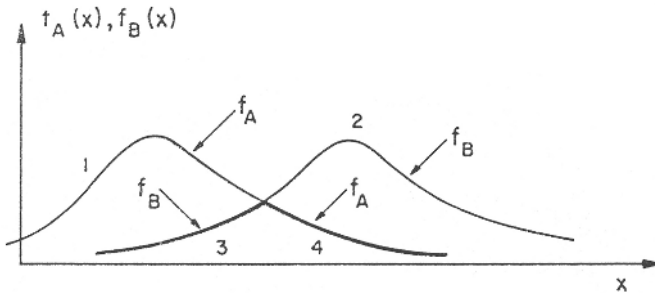
After the Second World War information and communication theory evolved into successful scientific disciplines. Lotfi A. Zadeh was a young electrical engineer who was deeply involved in these developments. In March 1950 he gave a lecture on *Some Basic Problems in Communication of Information* at the New York Academy of Sciences [16], in which he represented signals as ordered pairs  $(x(t), y(t))$  of points in a

signal space  $\Sigma$ , which is embedded in a function space. For instance, he discussed the *multiplex transmission of two or more signals*: A system has two channels and the sets of signals assigned to their respective channels are  $X = \{x(t)\}$  and  $Y = \{y(t)\}$ . If we are given the sum signal  $u(t) = x(t) + y(t)$  at the receiving end, how can we extract  $x(t)$  and  $y(t)$  from  $u(t)$ ? – We have to find two filters  $N_1$  and  $N_2$  such that, for any  $x$  in  $X$  and any  $y$  in  $Y$ ,

$$N_1(x + y) = x \quad \text{and} \quad N_2(x + y) = y. \quad (1)$$

But in reality filters don't do exactly what they are supposed to do in theory. Therefore, in later papers Zadeh started (e.g. in [17]) to differentiate between ideal and optimal filters. Ideal filters are defined as filters that achieve a perfect separation of signal and noise, but in practice such ideal filters do not exist. Zadeh regarded optimal filters to be those that give the “best approximation” of a signal and he noticed that “best approximations” depend on reasonable criteria. At that time he formulated these criteria in statistical terms. In the late 1950s and the 1960s he noticed that there are many situations in applied science (primarily those involving complex, large-scale systems, or complex or animated systems) in which we cannot compute exact solutions and therefore we have to be content with unsharp – fuzzy – solutions. To handle such problems it turned out that reasoning with loose concepts was very successful and in the 1960s Zadeh developed a mathematical theory to formalize this reasoning with vague – or fuzzy – concepts: the theory of fuzzy sets [4-6].

In his seminal article on “Fuzzy Sets” Zadeh introduced these new mathematical entities as classes or sets that “are not classes or sets in the usual sense of these terms, since they do not dichotomize all objects into those that belong to the class and those that do not.” He introduced “the concept of a fuzzy set, that is a class in which there may be a continuous infinity of grades of membership, with the grade of membership of an object  $x$  in a fuzzy set  $A$  represented by a number  $f_A(x)$  in the interval  $[0,1]$ .” [3]



**Fig. 2.** Zadeh’s Illustration of fuzzy sets in  $R^1$ : “The membership function of the union is comprised of curve segments 1 and 2; that of the intersection is comprised of segments 3 and 4 (heavy lines).” ([2], p. 342).

He maintained that these new concepts provide a “convenient way of defining *abstraction* – a process which plays a basic role in human thinking and communication.” ([3], p. 29)

As is manifest in all natural languages, communication between human beings rarely takes place in absolute terms. Few words are defined so precisely that one can pinpoint exact boundaries at which they stop being applicable. By contrast, computers classically handle data in a binary manner – a fact is either true or false. The tension between these two modes of operation has increasingly been recognized as a significant problem. The application of fuzzy sets opens up opportunities to make machines more human-friendly by moving away from a system of strictly binary representation. We examine this approach more closely in our other contribution, along with questions regarding implementation in technology and medicine. [18]

## References

- [1] Lotfi A. Zadeh, “From Circuit Theory to System Theory”, *Proceedings of the IRE*, vol. 50, No. 5, pp. 856-865, May 1962.
- [2] Lotfi A. Zadeh, “Fuzzy Sets”, *Information and Control*, vol. 8, pp. 338-353, 1965.
- [3] Lotfi A. Zadeh, “Fuzzy Sets and Systems”. In: Fox, Jerome (ed.): *System Theory*. Microwave Res. Inst. Symp. Ser. XV, Brooklyn, New York: Polytechnic Press 1965, pp. 29-37.
- [4] Rudolf Seising, “40 years ago: ‘Fuzzy Sets’ is going to be published”, *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks / International Conference on Fuzzy Systems (FUZZ-IEEE 2004)*, July 25-29, 2004, Budapest, Hungary, CD.
- [5] Rudolf Seising, “1965 – ‘Fuzzy Sets’ appear – A Contribution to the 40th Anniversary”. *Proceedings of the Conference FUZZ-IEEE 2005*, Reno, Nevada, May 22-25, 2005, CD.
- [6] Rudolf Seising, “The 40<sup>th</sup> Anniversary of Fuzzy Sets – A New View on System Theory”. In: Hao Ying, Dimitar Filev (eds.): *Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society Soft Computing for Real World Applications (NAFIPS)*, 22-25 June, 2005, Ann Arbor, Michigan, USA, pp. 92-97.
- [7] Gottlob Frege, *Grundgesetze der Arithmetik*. 2 vols., Jena: Hermann Pohle, 1893-1903.
- [8] Bertrand Russell, “Vagueness”, *The Austral. J. Psych. and Phil.*, vol. 1, 1923, pp 84-92.
- [9] Max Black, “Vagueness. An exercise in logical analysis”, *Philosophy of Science*, vol. 4, 1937, pp. 427-455.
- [10] Max Black, Reasoning with loose concepts, *Dialogue*, vol. 2, (1963), pp 1-12.
- [11] Karl Menger, “Statistical Metrics”, *Proc. Nat. Acad. of Sci.*, vol. 28, 1942, pp 535-537.
- [12] Karl Menger, “Probabilistic Theories of Relations”, *Proc. Nat. Acad. of Sci.*, vol. 37, 1951, pp 178-180.
- [13] Karl Menger, «Ensembles flous et fonctions aléatoires», *Comptes Rendus Académie des sciences* Vol. 37, 1951, pp 2001-2003.
- [14] Karl Menger, “Geometry and Positivism. A Probabilistic Microgeometry”, *Selected Papers in Logic and Foundations, Didactics, Economics*, Karl Menger (ed.), Vienna Circle Collection, 10, D. Reidel Publ. Comp., Dordrecht, Holland, 1979, pp 225-234.
- [15] Ernst Mach, *Die Prinzipien der Wärmelehre. Historisch-kritisch entwickelt*. Leipzig, Verlag von Johann Ambrosius Barth, 1896.
- [16] Lotfi A. Zadeh, “Some Basic Problems in communication of information”, *The New York Academy of Sciences*, March 1952, Series II, Vol.14, No. 5, pp. 201-204.

- [17] Lotfi A. Zadeh, "Theory of filtering", *J. Soc. Ind. Appl. Math.*, Vol. 1, 1953, pp. 35-51.
- [18] Rudolf Seising, Jeremy Bradley, "Are Soft Computing and Its Applications in Technology and Medicine Human-Friendly?", *Proc. KES2006 10<sup>th</sup> Intern. Conf. on Knowledge-Based & Intelligent Inf. & Eng. Systems* Bournemouth, 9, 10 and 11 October 2006, (to appear).

# New Classifier Based on Fuzzy Level Set Subgrouping

Paavo Kukkurainen and Pasi Luukka

Lappeenranta University of Technology, P.O. Box 20, Lappeenranta 53851, Finland  
paavo.kukkurainen@lut.fi, pasi.luukka@lut.fi

**Abstract.** We present new classification system which is based on fuzzy level sets subgrouping. This new classification system allow us fast classification method with quite accurate results.

## 1 Introduction

Theoretical work in subgrouping in fuzzy level set has been interesting and it is bound to develop interesting applications in the future. So far to our knowledge it has not yet been applied before. Especially in a field of pattern recognition which is one quite likely application area for fuzzy level sets.

In this article we have derived a classifier which is using in fuzzy subgrouping. Fuzzy subgroups are considered in [1]. Both fuzzy sets and fuzzy level sets are considered in [2]. The subgrouping is based on the level set  $\mathcal{A}(x) = \langle x \rangle$  introduced also in [1]. However, in this paper, it is used in the different way. We are testing how well these suites in the task of classification. We also form fuzzy level subgroups  $\mathcal{A}_{\langle x \rangle} = \langle x \rangle$ . This is obtained by combining the results in [1], p.86 and in [2], p.75.

Fuzzy set theory is an active research area, highly mathematical in its nature. It can provide a robust and consistent foundation for information processing, including pattern-formatted information processing. It plays at least two roles in pattern recognition. In one role, it serves as an interface between the linguistic variables seemingly preferred by humans and the quantitative characterizations appropriate for machines. In this role, it might also serve as a bridge between symbolic processing of artificial intelligence and the parallel distributed processing approaches favored by adaptive pattern recognition. In the another role, it emphasizes the possibility-distribution interpretation of the concept of fuzziness. The value of this role is that it legitimizes and provides a meaningful interpretation for some distributions that we believe to be useful, but that might be difficult to justify on the basis of the objective probabilities. The two roles are not distinct, but the differences are interesting and worth noting [3].

Classification is often very important part of process in many different fields like medicine and manufacturing. Disease must be recognized before proper treatment can be given or faulty products or materials must be removed from pipeline. Traditionally this kind of job has been made by humans but nowadays one of course wants automize this kind of work as much as possible. Reasons for that are very simple. First, usually classification can be made much more faster by computers than by humans and computers are cheaper to use than humans. Also with easy classification tasks one can usually found an algorithm for computer which do not make classification errors, humans always make some human errors. Some classification tasks can be also simply

too difficult for humans because there can be too many parameters to measure and/or they can be heavily correlated. Sometimes situation can be contrary, for example human might sense much more easier than any algorithm if some individual might be violent.

Here we have derived an classification algorithm that constructs an ideal matrix of each class in a learning phase of the algorithm. In a testing phase we compute the differences between the samples and the ideal matrices. The decision criteria which we use to decide which class which sample belongs is based on minimal distance between sample and ideal matrices which represents the class best. Algorithm is quite fast but problem is in distinguishing which class has smallest difference between the sample. Here we have used information provided by fuzzy level set subgrouping to add this knowledge to bring better classification accuracy. Fuzzy level set subgrouping also provides nice mathematical background for this method.

Data sets as diverse as possible were chosen so that the properties of the classifier would be apparent. The data sets were taken from the UCI-Repository of Machine Learning Database [4] so that they were differently distributed and their dimensions varied. The classifier was implemented with the *MATLAB<sup>TM</sup>*-software.

## 2 Mathematical Background

Let  $(P, \leq)$  be a partially ordered set and  $G$  a nonempty set. Mapping  $\mathcal{A} : G \rightarrow P$  is said to be a  $P$ -fuzzy set and set  $\mathcal{A}_p = \{x \in G \mid \mathcal{A}(x) \geq p\}$  is a  $p$ -level set (shortly a level-set) of  $\mathcal{A}$ . Let  $\mathcal{G} = (G, \cdot)$  be a group and  $(\mathcal{A}_p, \cdot)$  a subgroup of  $\mathcal{G}$  for any  $p \in P$  defining  $\mathcal{A} : G \rightarrow P$  to be a  $P$ -fuzzy subgroup of  $\mathcal{G}$  and  $\mathcal{A}_p$  a level subgroup of  $\mathcal{A}$ .

B. Šešelja and A. Tepavčević presented the following proposition in [1], p.86:

**Proposition 1.** *Let  $\mathcal{F}$  be a family of subgroups of a group  $\mathcal{G} = (G, \cdot)$  such that  $\bigcup \mathcal{F} = G$ . For every  $g \in G$ , let the intersection of subgroups in  $\mathcal{F}$  containing  $g$  belong to  $\mathcal{F}$ . If  $\mathcal{F} = (\mathcal{F}, \leq)$  is a partially ordered set dual to  $\mathcal{F} = (\mathcal{F}, \subseteq)$ , then  $\mathcal{A} : G \rightarrow \mathcal{F}$ , such that for  $g \in G$*

$$\mathcal{A}(g) = \bigcap (p \in \mathcal{F} \mid g \in p)$$

*is a  $\mathcal{F}$ -fuzzy subgroup of  $\mathcal{G}$ .*

Let  $x$  be a square matrix,  $G = \langle x \rangle$  a group generated by  $x$  and  $\mathcal{F} = \{\langle x^q \rangle \mid q \text{ is an integer}\}$  a set of subgroups of  $\mathcal{G}$ . We apply  $G$  and  $\mathcal{F}$  to Proposition 1 leading to a construction of the mathematical background for empirical testing.

Let  $x^p \in G$ . It is known that  $\langle x^p \rangle$  is the smallest subgroup of  $G$  containing  $x^p$  and therefore

$$\bigcap (r \in \mathcal{F} \mid x^p \in r) = \langle x^p \rangle \in \mathcal{F}.$$

Also

$$\bigcup \mathcal{F} = \bigcup (r \mid r \in \mathcal{F}) = \langle x \rangle = G,$$

where subgroups  $r$  are understood as sets. Proposition 1 yields

$$\mathcal{A}(x^p) = \bigcap (r \in \mathcal{F} \mid x^p \in r) = \langle x^p \rangle .$$

---

<sup>1</sup> The order of authors is alphabetical. Paavo Kukkurainen is the corresponding author of the section 2 and Pasi Luukka of the sections 3 and 4.

Further, for level subgroups  $\mathcal{A}_{\langle x^p \rangle} = \langle x^p \rangle$  [1], p.86, [2], p.75 together. Define a relation  $\sim$  in  $G$  as follows:

$$x^p \sim x^q \quad \text{iff} \quad \mathcal{A}(x^p) = \mathcal{A}(x^q) \quad \text{iff} \quad \langle x^p \rangle = \langle x^q \rangle$$

Then  $\sim$  is an equivalence relation. Let  $[x^p]$  be an equivalence class determined by  $x^p$ . We see immediately that there is a one-to-one correspondence between  $[x^p]$  and  $\langle x^p \rangle$ , and therefore between  $x^p$  and  $\langle x^p \rangle$ .

Assuming there are  $N$  different classes  $C_j$ , let  $v_i$  be the ideal matrix of class  $C_i$ . We set for any sample matrix  $x$  and a positive integer  $p$

$$\|x^p - v_i^p\| = \min\{\|x^p - v_j^p\|, 1 \leq j \leq N\}$$

iff  $x$  belongs to  $C_i$ .

Thus every  $x$  is classified to some class  $C_j$ ,  $j = 1, \dots, N$ . Let  $p$  be a number which yields the proportional part of matrices  $x$  belonging to their right classes. This part is obtained by dividing the sum of numbers of matrices  $x$  classified to these classes correctly by the whole number of matrices  $x$ .

Because there is a one-to-one correspondence between  $x^p$  and  $\langle x^p \rangle$ , the number of matrices  $x$  belonging to their right classes is the same as the number of the subgroups  $\langle x^p \rangle$  for the fixed  $p$ . Clearly, the numbers of  $x$  and  $\langle x \rangle$  are the same for all the sample matrices. Consequently, the quotient of the sum of the numbers of subgroups  $\langle x^p \rangle$  and the whole number of groups  $\langle x \rangle$  is the classification accuracy in figures for the  $p$ . The smaller  $p$  is the greater accuracy is obtained. The empirical results support this conclusion. Since  $\mathcal{A}_{\langle x^p \rangle} = \langle x^p \rangle$  we can interpret the result as level sets.

### 3 Algorithm

The problem of classification is basically one of partitioning the feature space into regions, one region for each category. Ideally, one would like to arrange this partitioning so that none of the decisions is ever wrong [5].

We would like to classify a set  $X$  of objects to  $N$  different classes  $C_1, \dots, C_N$  by their features. We suppose that  $D$  is the number of different kinds of features  $f_1, \dots, f_D$  that we can measure from objects. We assume that the values for the magnitude of each feature is normalized so that it can be presented as a value between  $[0, 1]$ .

The first thing is to determine for each class the ideal matrix  $v_i = (v_i(f_1), \dots, v_i(f_D))$  that represents class  $i$  as well as possible. This matrix can be user defined or created from some sample set  $X_i$  of vectors  $\mathbf{x} = (x(f_1), \dots, x(f_D))$  which are known to belong to class  $C_i$ . We can basically use all the sample set vectors to create this matrix. We can take a mean vector from the samples and make a diagonal matrix from it.

Once the ideal matrices have been determined, then the decision to which class an arbitrarily chosen  $\mathbf{x} \in X$  belongs to is made by comparing it to each ideal matrix. Before comparison can be made it must create its diagonal matrix. This sample matrix must have same dimensions as the ideal matrices. After this is done we can use the fuzzy level set subgrouping in a manner presented in previous section and find the proper subgroups. After subgroups have been found we can make the decision to which

class the sample belongs by using minimum distance between the sample matrix and the ideal matrices created for each class.

In the algorithmic form, a classifier would be:

---

```

Require:  $test, learn[1..n], weights, dim$ 
scale  $test$  between  $[0, 1]$ 
scale  $learn$  between  $[0, 1]$ 
for  $i = 1$  to  $n$  do
     $idealmatrix[i] = IDEAL[learn[i]]$ 
     $dist[i] = \sum_{j=1}^{m1} \sum_{k=1}^{m2} |(idealmatrix[i][j][k])^p - (test[j][k])^p|$ 
end for
 $class = \arg \min_i dist[i]$ 

```

---

## 4 Empirical Results

### 4.1 Data Sets

Next data set used in classification are introduced shortly. Three different data sets were used and they are all freely available in [4]. The fundamental properties of the data sets are shown in Table 1. All data sets were splitted in half; one half was used for training and the other for testing the classifier.

**Table 1.** Data sets and their properties.

<b>Name</b>	<b>classes</b>	<b>Dimension</b>	<b>cases</b>
<b>Iris data set</b>	3	4	150
<b>Thyroid data</b>	3	5	215
<b>Wine data</b>	3	15	178

Database Iris consists of three different types of iris plant. One class is linearly separable from the other two; the latter are nonlinearly separable from each other. The predicted attribute is the class of iris plant.

In the thyroid data set the purpose is to predict whether a patient’s thyroid belongs to the class euthyroidism, hypothyroidism or hyperthyroidism (see [6]).

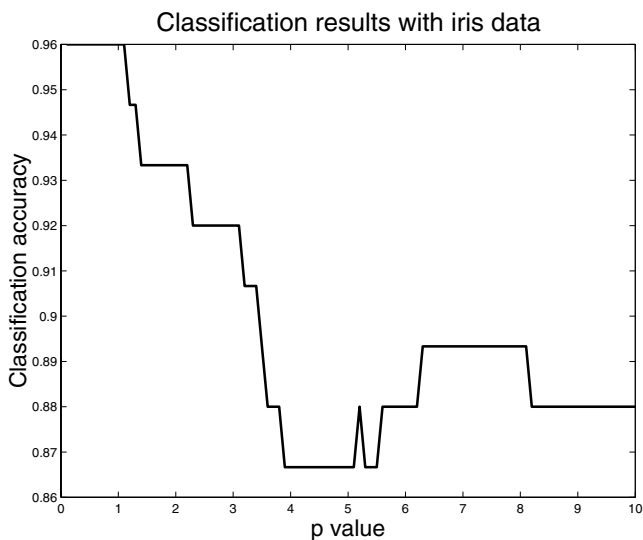
The wine data is the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars [4].

### 4.2 Classification Results

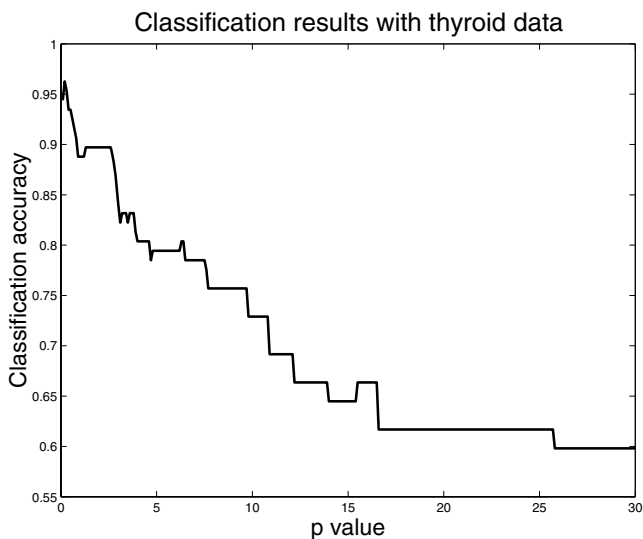
**Iris data:** With iris data classification results were overall quite good. Best classification accuracy was 96% and it was achieved with  $p$  value of  $p = 1$ . Classification results with respect to  $p$  value changes can be seen in Figure 1.

**Thyroid data:** With thyroid data results depend very much on correct  $p$  value. With small  $p$  values classification results are quite good and when we go to higher  $p$  values worse results are achieved. In Figure 2 one can see the  $p$  values effect on classification





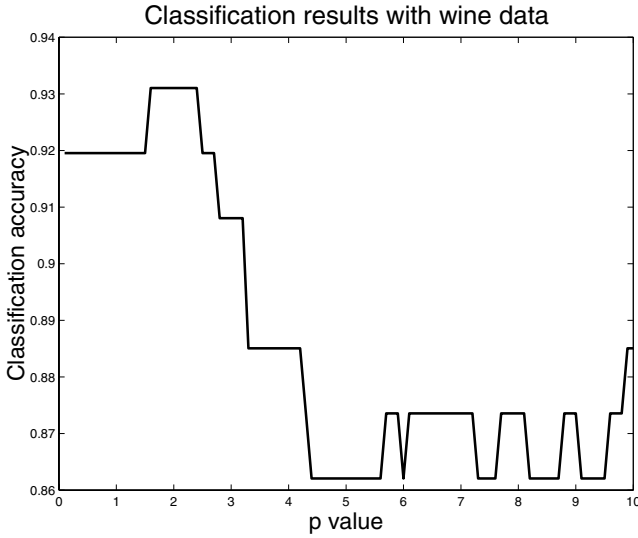
**Fig. 1.** Classification results with iris data



**Fig. 2.** Classification results with thyroid data

results. Best classification accuracy was 96.26% and it was achieved with  $p$  value of  $p = 0.2$ .

**Wine data:** With wine data set classification results with respect to  $p$  value changes can be seen in Figure 3. Best classification result was 93.10% and it was achieved with  $p$  value  $p = 2.0$ .



**Fig. 3.** Classification results with wine data

### 4.3 Discussion of the Method

Desirable properties that a pattern classifier should possess according to [7](Simpson 1992): A successful pattern classifier should be able to:

1. Learn the required task quickly
2. Learn new data without having to retrain with old data (on-line adaptation)
3. Solve non-linearly separable problems
4. Provide the capability for soft and hard decisions regarding the degree of membership of the data within each class
5. Offer explanations of how the data are classified, and why the data are classified as such
6. Exhibit performance that is independent of parameter tuning
7. Function without knowledge of the distributions of the data in each class
8. For overlapping pattern classes, create regions in the space of the input parameters that exhibit the least possible overlap

If we go into detail of these properties and compare our classifier to the properties listed we can see that our classifier satisfies most of the properties. The method is capable to learn the required task quickly. On-line adaptation is not included in our algorithm but is one area of future work. Classifier is able to solve non-linearly separable problems. It provides a partial membership for each class through the distance matrices but so far capability for both soft and hard decisions is not implemented to the classifier. It offers the explanations of how the data are classified, and why the data are classified as such. As can be seen from the results classifier can exhibit performance that is independent of parameter tuning by just setting suitable  $p$ -level set which seems to work well.

In classifier algorithm the data goes through fuzzyfication and distribution of the data is not needed. For overlapping pattern classes, regions in the space of the input parameters that exhibit the least possible overlap is created by finding the suitable  $p$  level set for functioning for each data separately.

## 5 Conclusions

In this paper, we have and utilized fuzzy level set subgrouping and introduced a new classifier based on the comparison of the matrices and derived information from the fuzzy level set subgrouping so that they can be used in classification. We have tested this new classifier with three different data sets.

Classification results were very good in overall. The use of fuzzy level set subgrouping in classification seems to be well-stated.

Authors acknowledge that work done here is still in an early stages and in future more effort must be devoted in development of the classifier and how information from the fuzzy level set subgrouping can be even more efficiently used in the field of pattern recognition.

Another area for future development of the method is to build the possibility of on-line adaptation and provide a way to efficiently utilize the capability for both soft and hard decision regarding the degree of membership of the data within each class.

## Acknowledgments

This work was supported by the laboratory of applied mathematics, Lappeenranta University of Technology.

## References

1. B.Šešelja, A. Tepavčević: *Fuzzy groups and collection of subgroups*, Fuzzy Sets and Systems 83 pp.85-91, 1996.
2. B.Šešelja, A. Tepavčević: *On a construction of codes by  $p$ -fuzzy sets*, Ser. Mat. 20, 2, pp.71-80, 1990.
3. Y.H. Pao, *The Pattern Recognition and Neural Networks*. Addison-Wesley Publishing Company, Inc, 1989.
4. UCI Repository of Machine Learning Databases network document. Referenced 4.11.2004. Available: <ftp://ftp.ics.uci.edu/pub/machine-learning-databases>
5. Duda, R. and Hart, P.: *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973
6. Coomans, D., Broeckaert, M. Jonckheer M. and Massart D.L.: *Comparison of Multivariate Discriminant Techniques for Clinical Data - Application to the Thyroid Functional State*, Meth. Inform. Med. 22 pp. 93-101, 1983.
7. Simpson, P.K.: *Fuzzy min-max neural networks - Part 1: Classification* IEEE Transactions on Neural Networks, Vol. 3. No. 5. pp. 776-786, 1992.

# Fast Shape Index Framework Based on Principle Component Analysis Using Edge Co-occurrence Matrix

Zhiping Xu, Yiping Zhong, and Shiyong Zhang

Department of Computing and Information Technology  
Fudan University, Shanghai, China, 200443  
dr.bennix@gmail.com

**Abstract.** The shape of an object is one of the most important features in content based image retrieval. However, the statistical feature of edge is rarely used as a feature that codes local spatial information. This paper presents an approach to represent spatial edge distributions using principal component analysis (PCA) on the edge co-occurrence matrix (ECM). The ECM is based on the statistical feature attained from the edge detection operators which applied on the image. The eigenvectors obtained from PCA of the ECM can preserve the high spatial frequencies components, so they are well suited for shape as well as texture representation. Projections of the ECM from the image database to the local PCs serve as a compact representation for the search database. The framework presented in the paper grantee the accuracy and speed of the content based image retrieval in our work.

**Keywords:** edge co-occurrence matrix, content based image retrieval, compact representation.

## 1 Introduction

Massive image databases are used in multimedia applications in the fields such as entertainment, business, art, engineering and science. Retrieving images by their content rather than external annotations has become an important operation. A fundamental ingredient for content based image search is the method used for comparing image features attained from each image stored in the image database. There are two general approaches for image comparison: intensity based and shape based. L Schomaker *et al.* [1] mentioned users of Content Based Image Retrieval (CBIR) were more interested in retrieval by shape than by color and texture. However, retrieval by shape is still considered one of the most difficult aspects of content based search. Some other system like IBM's Query By Image Content (QBIC) [2] is relatively successful in retrieving by color and texture, but performs poorly when searching on shape. The similar behavior is exhibited in some other systems.

Shape matching is a central problem of computer vision, visual information systems, pattern recognition, and robotics. Shape matching can be applied in the industrial inspection, fingerprint matching, and content-based image retrieval.

The matching process deals with transforming features into representations, and compares with the query representations using some dissimilarity measures. However,

the shape of a pattern is the pattern under all transformations in a transformation group. The matching problem is studied in various forms. The problems encountered by the researchers in this scope are listed below:

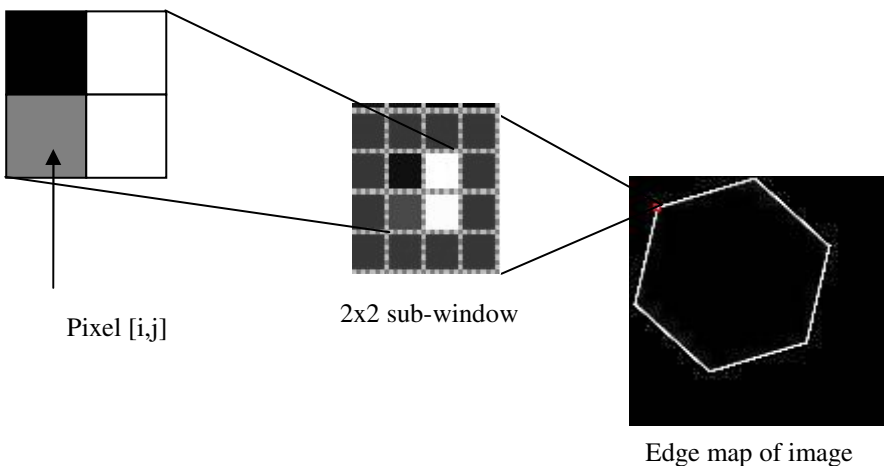
- **High dimension data reduction problem.** The dimension of the feature vectors attained from the image is extremely high, the reduction of feature dimension can achieved high performance in retrieval.
- **Computation problem.** Compute the dissimilarity between the two patterns.
- **Decision problem.** For a given threshold, decide whether the dissimilarity is smaller than the threshold.

From many researches, the time complexities in retrieval are rather great, in order to solve this problem we proposed a framework that can fast index shape based on the Principle Component Analysis (PCA) using edge co-occurrence matrix (ECM).

The rest of paper is organized as follows, section 2 shows the concept of edge co-occurrence matrix, Section 3 illustrates the principle component analysis applied in our work, Section 4 demonstrates the measure of dissimilarity of query image and stored images, Section 5 shows the architecture of Fast Shape Index Framework, Section 6 gives the experiments results in the framework, Section 7 concludes the works we have done.

## 2 Edge Co-occurrence Matrix

The edge co-occurrence matrix is motivated by the Gray Level Co-occurrence Matrix (GLCM) [3,4]. The edge co-occurrence matrix (ECM) estimates image properties related to second-order statistics. Each entry  $(i,j)$  in ECM corresponds to the number of occurrences of the pair of edge block style  $i$  and  $j$  which are a distance  $d$  apart in original image. The edge block is generated from the edge map attained from edge detection algorithms, like Sobel, Canny detectors [5].



**Fig. 1.** The Edge Block of the Edge map

As is shown in Fig. 1, the edge block of the edge map is taken from the 2x2 sub-window in the edge map, the style of each edge block can be calculated according to Eq. (1).

$$Style = \sum_{(i,j) \in EdgeBlock} 2^{2i+j} \cdot P[i, j] \quad (1)$$

where  $EdgeBlock$  represents the 2x2 sub-window in the edge map,  $i, j$  respectively represents the coordination in the sub-window,  $P[i, j]$  represents the intensity value or the binary value applied by the threshold of current coordination  $(i, j)$ .

In our work, we only consider the binary form of  $P[i, j]$ . According to enumeration of the edge block, there are 16 types of edge blocks. The combination of current edge block and neighbor edge block forms the matrix of edge block style based on the Eq. (2).

$$M_{16 \times 16} = [S]$$

$$S = \sum p, p = \begin{cases} 1 & Style_{i,j} = Style_{i,j+1} \\ 0 & Style_{i,j} \neq Style_{i,j+1} \end{cases} \quad (2)$$

This matrix show all the statistical information about the style of edge block; however we found in the experiments that a large portion of the elements' values were extremely small, and the dimension of the edge block style matrix was very high.

### 3 Principle Component Analysis

Principle component analysis [6,9,10] is a useful statistical technique that has been found in application in many fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension. It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data are hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data. As above mentioned the dimension of the edge block style matrix is extremely high, in our work PCA is applied to analyze the edge block style matrix feature to gain the important feature vector from the image according to Eq. (3).

$$F = f \cdot D \quad (3)$$

where  $F$  is the final data vector of the image it represented,  $f$  is the matrix with the eigenvectors in the columns transposed so that the eigenvectors of the matrix of edge block style are now in the rows, with the most significant eigenvector at the top, and  $D$  is the mean-adjusted data transposed, i.e. the data items are in each column, with each row holding a separate dimension.

## 4 The Measure of Dissimilarity

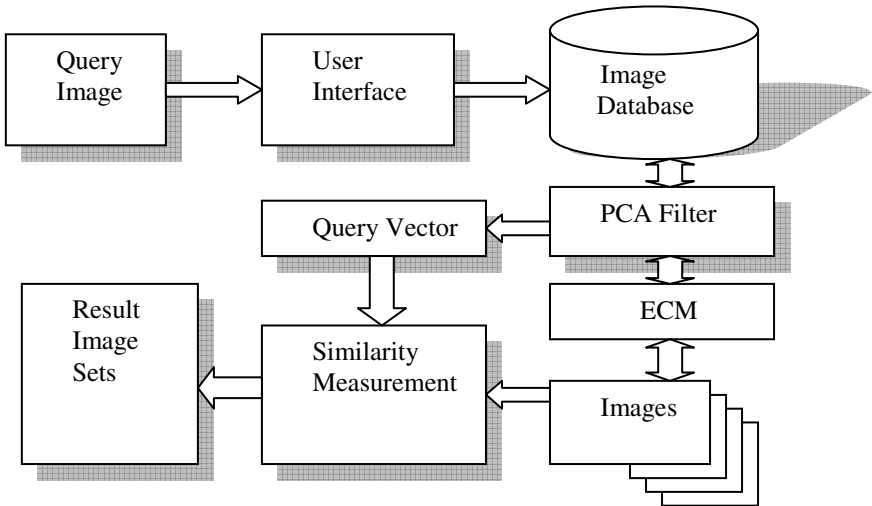
We gained the final data vector of its represented image, and these data are stored in the image database. When the user issue a query to the CBIR system, the query image was firstly applied with the edge detection operation to gain the edge map, then formulated the query image's ECM matrix using above mentioned approaches, then using the  $f$  multiplied the query image's ECM to gain the query vector  $Q$ . In our work, the measure of dissimilarity to each image stored in the image database can be attained by Eq. (4):

$$d(Q, I_i) = \min \|Q - I_i\| \quad (4)$$

where  $d(Q, I_i)$  is the Euclidian distance between two vectors,  $I_i$  represents each image's final PCA transformed vectors stored in the image database. The criterion of two images' similarity is the minimum distance between the two images.

## 5 Architecture of Fast Shape Index Framework

The architecture of Fast Shape Index Framework (FSIF) is illustrated below:



**Fig. 2.** The architecture of the FSIF

As illustrated in Fig. 2, the system first calculates the ECM of each image in the image database, and then uses PCA to transform the ECM into feature vectors to be stored in the image database. When the users issue the query image through the user interface, the query image applied the same steps that applied to each image in the image database to form the query vector. After the query vector is generated, the

framework will do the similarity measurement between query vector and feature vectors of the image to find out the minimum distance image result set. Fig.3 shows the user interface of our framework.



Fig. 3. The user interface of the Fast Shape Index framework

## 6 Experiments

### 6.1 Shape Based Query Experiments

Our work is focused on the shape similarity query in the image database. To illustrate the process of the experiments, we give the test image sets listed as follows:



Fig. 4. The test image sets for shape based query



As illustrated in Fig.4, the images in the first row are same in size and shape but different in tilt and rotation angle, the images in the second row are same in shape and size but different in position, the third row images are neither same in size nor same position, but the outside shape is same, the forth row images are neither same in size rotation and position nor same in shape.

Then the user issues the query image to the CBIR system to gain the results.

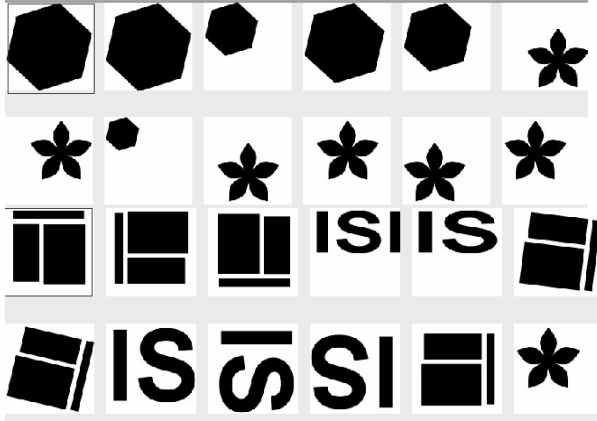


Fig. 5. Shape Query Examples

The rectangle box which enveloped hexagon and assembled shape in the Fig.5 indicated the query sample given by the user, the return image sets are ordered by the similarity according to the query image. The framework shows high precision rate in the rotation, shift and scale according to the experiments. In our case, the precision is calculated by terms of Eq. (5)

$$precision = \frac{relevant\_images}{returned\_images} \times 100\% \quad (5)$$

The average precision results of this type of experiments are shown in **Table 1**.

**Table 1.** The Result of Shape based query experiments

Number of Images	Average Precision (%)
20	84.3
50	80.4
100	79.4
500	68.4

## 6.2 Natural Image Query Experiment

To date, we have tested our retrieval algorithm on a general-purpose image database with 1000 images from the Corel [7]. These images are stored in JPEG format with size 384x256, 256x384. The entire database has 10 categories with 100 images in each category. Most categories contain distinct semantics including building, flowers, vehicles, etc. When we do the nature image query experiment, we choose the same image categories used in Xiaojun Qi *et al.*'s work [8] for comparison.

**Table 2.** Comparison of the average retrieval precision of each category

	Proposed	Ref. 8
Architecture	0.7061	0.5467
Beach	0.3274	0.2800
Vehicle	0.6044	0.5600
Flowers	0.8333	0.8233
Food	0.3926	0.5067
Horse	0.8872	0.8900
<b>Average</b>	<b>0.6252</b>	<b>0.5650</b>

From the Table 2, we can find that our method has better retrieval accuracy than the Xiaojun Qi's approach in 4 categories and a little worse retrieval accuracy for the food and horse queries. It improves the overall average retrieval accuracy by 10.64%. Xiaojun Qi *et al.*'s work [8] used a combination of features to ensure the precision; however in our work we only used only one feature to gain such results, the reason of little worse retrieval accuracy for the food and horse queries is the interfere from statistical noise from the background object. The better way to solve this limitation is to combine some other features or using some filters to reduce the noise.

## 6.3 Speed

The algorithm has been implemented using Delphi 6 on a Pentium IV 1.4 GHz, 752MB memory LENOVO Laptop running Windows XP Media Center operating system. Computing the feature vectors for 8000 color images of size 384x256 requires around 10 min. In average, 7ms is needed to compute all the features and apply the PCA for each image.

## 7 Conclusions

A fast approach to CBIR is proposed in this paper. An image is first transformed into ECM generated from edge detectors. Then by using PCA method on the ECM data entry of each image in the image database, our system gains the final feature vectors. Users can query image by a sample image, which will be transformed into query vector. Such methods proposed by the paper have been proven to be more accurate and efficient than some other CBIR methods. This approach also shows the in born immunity to the scale and rotation of the shape.

In order to gain higher accuracy in CBIR, our approach can be combined with some other features like textures and colors, also the artificial neural network and other machine learning methods will be plausible to enhance to precision of CBIR.

**Acknowledgments.** The authors would like to acknowledge the comments of the two unknown reviewers whose insightful comments helped to improve this paper.

## References

1. Schomaker L., Leau E. D., Vuurpijl L.: Using Pen-Based Outlines for Object-Based Annotation and Image-Based Queries, Visual Information and Information Systems: Third International Conference, VISUAL'99, Amsterdam, The Netherlands, (1999)
2. QBIC Project, <http://www.qbic.almaden.ibm.com>
3. Partio M., Cramariuc B., Gabbouj M.: Rock Texture Retrieval Using Gray Level Co-occurrence Matrix, Proc. of 5th Nordic Signal Processing Symposium, (2002)
4. Haralick R. M., Shanmugam K., Dinstein I.: Textural Features for Image Classification, IEEE Trans. on Systems, Man, and Cybernetics, Vol. 6. (1973) 610–621.
5. Ziou D., Tabbone S.: Edge Detection Techniques-an Overview, Pattern Recognition And Image Analysis C/C Of Raspoznavaniye, 1998
6. Daffertshofer A., Lamoth C. J., Meijer O. G., Beek P. J.: PCA in studying coordination and variability: a tutorial, Clin Biomech (Bristol, Avon), (2004)
7. Corel, <http://www.corel.com>
8. Qi X., Han Y.: A novel fusion approach to content-based image retrieval, Pattern Recognition Vol. 38. (2005), 2449–2465
9. Pearson, K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine, Vol. 2. (1901) 559–572.
10. Hotelling, H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, Vol. 24. (1933) 417–441, 498–520.

# Stock Index Modeling Using Hierarchical Radial Basis Function Networks

Yuehui Chen<sup>1</sup>, Lizhi Peng<sup>1</sup>, and Ajith Abraham<sup>1,2</sup>

<sup>1</sup> School of Information Science and Engineering  
Jinan University, Jinan 250022, P.R. China  
yhchen@ujn.edu.cn

<sup>2</sup> IITA Professorship Program, School of Computer Science and Engg.  
Chung-Ang University, Seoul, Republic of Korea  
ajith.abraham@ieee.org

**Abstract.** Forecasting exchange rates is an important financial problem that is receiving increasing attention especially because of its difficulty and practical applications. This paper proposes a Hierarchical Radial Basis Function Network (HiRBF) model for forecasting three major international currency exchange rates. Based on the pre-defined instruction sets, HRBF model can be created and evolved. The HRBF structure is developed using the Extended Compact Genetic Programming (ECGP) and the free parameters embedded in the tree are optimized by the Degraded Ceiling Algorithm (DCA). Empirical results indicate that the proposed method is better than the conventional neural network and RBF networks forecasting models.

## 1 Introduction

Exchange rates are affected by many highly correlated economic, political and even psychological factors. These factors interact in a very complex fashion. Exchange rate series exhibit high volatility, complexity and noise that result from an elusive market mechanism generating daily observations [1]. Much research effort has been devoted to exploring the nonlinearity of exchange rate data and to develop specific nonlinear models to improve exchange rate forecasting, i.e., the Autoregressive Random Variance (ARV) model [2], Autoregressive Conditional Heteroscedasticity (ARCH) [3], self-exciting threshold autoregressive models [4]. There has been growing interest in the adoption of neural networks, fuzzy inference systems and statistical approaches for exchange rate forecasting problem [5][13][14].

For a recent review of neural networks based exchange rate forecasting, please consult [7]. The input dimension (i.e. the number of delayed values for prediction) and the time delay (i.e. the time interval between two time series data) are two critical factors that affect the performance of neural networks. The selection of dimension and the number of time delays has great significance in time series prediction.

Hierarchical neural networks consist of multiple neural networks assembled in different level or cascade architecture. Mat Isa et al. used Hierarchical Radial Basis Function (HiRBF) to increase RBF performance in diagnosing cervical cancer [17]. HiRBF cascading together two RBF networks, where both networks have different structure but using the same learning algorithms. The first network classifies all data and performs a filtering process to ensure that only certain attributes to be fed to the second network. The study shows that HiRBF performs better than the single RBF model. HiRBF has been proved effective in the reconstruction of smooth surfaces from sparse noisy data points [18]. In order to improve the model generalization performance, a selective combination of multiple neural networks by using Bayesian method was proposed in [19].

In this paper, an automatic method for constructing HiRBF network is proposed. Based on the pre-defined instruction sets, a HiRBF network can be created and evolved. The HiRBF network also allows input variables selection. In our previous studies, in order to optimize Flexible Neural Tree (FNT) the hierarchical structure of FNT was evolved using Probabilistic Incremental Program Evolution algorithm (PIPE) [8][9] and Ant Programming with specific instructions. In this research, the hierarchical structure is evolved using the Extended Compact Genetic Programming (ECGP), a tree-structure based evolutionary algorithm. The fine tuning of the parameters encoded in the structure is accomplished using the degraded ceiling algorithm [16]. The proposed method interleaves both optimizations. The novelty of this paper is in the usage of HiRBF model for selecting the important inputs and/or time delays and for forecasting foreign exchange rates.

## 2 The Hierarchical RBF Model

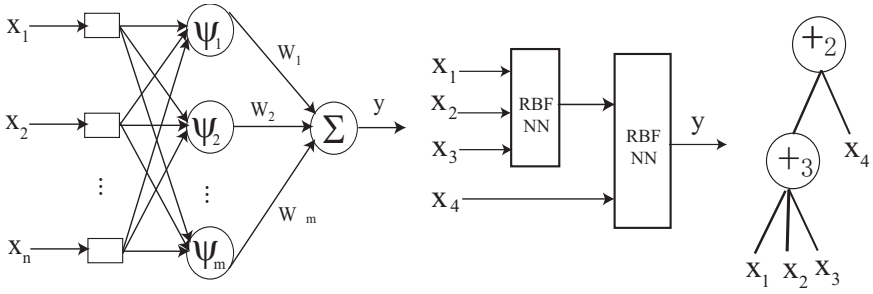
A function set  $F$  and a terminal instruction set  $T$  used for generating a hierarchical RBF model are described as  $S = F \cup T = \{+_2, +_3, \dots, +_N\} \cup \{x_1, \dots, x_n\}$ , where  $+_i (i = 2, 3, \dots, N)$  denote non-leaf nodes' instructions and taking  $i$  arguments.  $x_1, x_2, \dots, x_n$  are leaf nodes' instructions and taking no other arguments. The output of a non-leaf node is calculated as a RBF neural network model (see Fig.1). From this point of view, the instruction  $+_i$  is also called a basis function operator with  $i$  inputs.

The basis function operator is shown in Fig.1(left). In general, the basis function networks can be represented as

$$y = \sum_{i=1}^m \omega_i \psi_i(x; \theta) \tag{1}$$

where  $x \in R^n$  is input vector,  $\psi_i(x; \theta)$  is  $i$ th basis function, and  $\omega_i$  is the corresponding weights of  $i$ th basis function and  $\theta$  is the parameter vector used in the basis functions. In this research, Gaussian radial basis functions are used,

$$\psi_i(x; \theta) = \prod_{j=1}^n \exp\left(-\frac{\|x_j - b_j\|^2}{a_j^2}\right) \tag{2}$$



**Fig. 1.** A RBF network (left), a hierarchical RBF network (middle), and a tree-structural representation of the HRBF (right)

and the number of basis functions used in hidden layer is same with the number of inputs, that is,  $m = n$ .

In the creation process of HiRBF tree, if a nonterminal instruction, i.e.,  $+_i (i = 2, 3, 4, \dots, N)$  is selected,  $i$  real values are randomly generated and used for representing the connection strength between the node  $+_i$  and its children. In addition,  $2 \times n^2$  adjustable parameters  $a_i$  and  $b_i$  are randomly created as Gaussian radial basis function parameters. The output of the node  $+_i$  can be calculated by using (1) and (2). The overall output of HiRBF tree can be computed from left to right by depth-first method, recursively.

**Tree Structure Optimization.** Finding an optimal or near-optimal HiRBF is formulated as a product of evolution. In our previously studies, the Genetic Programming (GP), Probabilistic Incremental Program Evolution (PIPE) have been explored for structure optimization of the FNT [8][9]. In this paper, the Extended Compact Genetic Programming (ECGP) [11] is employed to find an optimal or near-optimal HiRBF structure.

ECGP is a direct extension of ECGA to the tree representation which is based on the PIPE prototype tree. In ECGA, Marginal Product Models (MPMs) are used to model the interaction among genes, represented as random variables, given a population of Genetic Algorithm individuals. MPMs are represented as measures of marginal distributions on partitions of random variables. ECGP is based on the PIPE prototype tree, and thus each node in the prototype tree is a random variable. ECGP decomposes or partitions the prototype tree into sub-trees, and the MPM factorises the joint probability of all nodes of the prototype tree, to a product of marginal distributions on a partition of its sub-trees. A greedy search heuristic is used to find an optimal MPM mode under the framework of minimum encoding inference. ECGP can represent the probability distribution for more than one node at a time. Thus, it extends PIPE in that the interactions among multiple nodes are considered.

**Parameter Optimization with Degraded Ceiling Algorithm.** Simulated annealing is one of the most widely studied local search meta-heuristics. It was proposed as a general stochastic optimization technique in 1983 [15] and has been applied to solve a wide range of problems including the weights optimization of

```

Set the initial solution S
Calculate initial fitness function f(s)
Initial ceiling B=f(s)
Specify input parameter dB
While not some stopping condition do
    define neighbourhood N(s)
    Randomly select the candidate solution s* in N(s)
    If ( f(s*) < f(s) ) or ( f(s*) <= B )
    Then accept s*
    
```

**Fig. 2.** The Degraded ceiling algorithm

a neural network. The basic ideas of the simulated annealing search are that it accepts worse solutions with a probability  $p = e^{-\frac{\delta}{T}}$ , where  $\delta = f(s^*) - f(s)$ , the  $s$  and  $s^*$  are the old and new solution vectors,  $f(s)$  denotes the cost function, the parameter  $T$  denotes the temperature in the process of annealing. Originally it was suggested to start the search from a high temperature and reduce it to the end of the process by an equation:  $T_{i+1} = T_i - T_i * \beta$ . However, the cooling rate  $\beta$  and initial value of  $T$  should be carefully selected since it is problem dependent.

The degraded ceiling algorithm also keeps the acceptance of worse solutions but with a different manner [16]. It accepts every solution whose objective function is less than or equal to the upper limit  $B$ , which is monotonically decreased during the search. The procedure for the degraded ceiling algorithm is given in Fig.2.

**Procedure of the General Learning Algorithm.** The general learning procedure for constructing the HiRBF network can be described as follows.

- 1) Create an initial population randomly (HiRBF trees and its corresponding parameters);
- 2) Structure optimization is achieved by using ECGP algorithm;
- 3) If a better structure is found, then go to step 4), otherwise go to step 2);
- 4) Parameter optimization is achieved by the DCA algorithm as described in subsection 2. In this stage, the architecture of HiRBF model is fixed, and it is the best tree developed during the end of run of the structure search. The parameters (weights and flexible activation function parameters) encoded in the best tree formulate a particle.
- 5) If the maximum number of local search is reached, or no better parameter vector is found for a significantly long time then go to step 6); otherwise go to step 4);
- 6) If satisfactory solution is found, then the algorithm is stopped; otherwise go to step 2).

**Variable Selection using Hierarchical RBF Paradigms.** It is often a difficult task to select important variables for a classification or regression problem, especially when the feature space is large. Conventional RBF neural network usually cannot do this. In the perspective of hierarchical RBF framework, the

nature of model construction procedure allows the HiRBF to identify important input features in building an HiRBF model that is computationally efficient and effective. The mechanisms of input selection in the HiRBF constructing procedure are as follows. (1) Initially the input variables are selected to formulate the HiRBF model with same probabilities; (2) The variables which have more contribution to the objective function will be enhanced and have high opportunity to survive in the next generation by a evolutionary procedure; (3) The evolutionary operators i.e., crossover and mutation, provide a input selection method by which the HiRBF should select appropriate variables automatically.

### 3 Exchange Rates Forecasting Using HiRBF Paradigms

#### 3.1 The Data Set

We used three different datasets in our forecast performance analysis. The data used are daily forex exchange rates obtained from the Pacific Exchange Rate Service [12], provided by Professor Werner Antweiler, University of British Columbia, Vancouver, Canada. The data comprises of the US dollar exchange rate against Euros, Great Britain Pound (GBP) and Japanese Yen (JPY). We used the daily data from 1 January 2000 to 31 October 2002 as training data set, and the data from 1 November 2002 to 31 December 2002 as evaluation test set or out-of-sample datasets (partial data sets excluding holidays), which are used to evaluate the good or bad performance of the predictions, based on evaluation measurements.

The forecasting evaluation criteria used is the normalized mean squared error (NMSE),

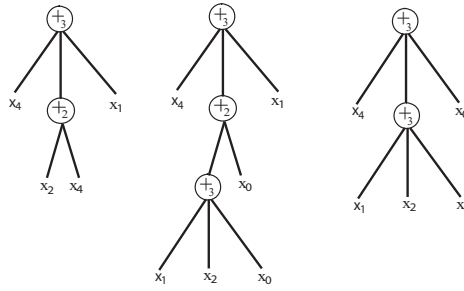
$$NMSE = \frac{\sum_{t=1}^N (y_t - \hat{y}_t)^2}{\sum_{t=1}^N (y_t - \bar{y}_t)^2} = \frac{1}{\sigma^2} \frac{1}{N} \sum_{t=1}^N (y_t - \hat{y}_t)^2, \quad (3)$$

where  $y_t$  and  $\hat{y}_t$  are the actual and predicted values,  $\sigma^2$  is the estimated variance of the data and  $\bar{y}_t$  the mean.

#### 3.2 Feature/Input Selection with HiRBF

It is often a difficult task to select important variables for a forecasting or classification problem, especially when the feature space is large. A fully connected NN classifier usually cannot do this. In the perspective of HiRBF framework, the nature of model construction procedure allows the HiRBF to identify important input features in building a forecasting model that is computationally efficient and effective. The mechanisms of input selection in the HiRBF constructing procedure are as follows. (1) Initially the input variables are selected to formulate the HiRBF model with same probabilities; (2) The variables which have more contribution to the objective function will be enhanced and have high opportunity to survive in the next generation by a evolutionary procedure; (3) The evolutionary operators provide a input selection method by which the HiRBF should select appropriate variables automatically.





**Fig. 3.** The evolved HRBF trees for forecasting euros (left), British pounds (middle) and Japanese yen (right)

**Table 1.** Forecast performance evaluation for the three exchange rates (NMSE for testing)

Exchange rate	Euros	British Pounds	Japanese Yen
MLFN [13]	0.5534	0.2137	0.2737
ASNN [13]	0.1254	0.0896	0.1328
RBF-NN	0.1130	0.0852	0.1182
HRBF-NN (This paper)	0.0240	0.0212	0.0095

### 3.3 Experimental Results

For simulation, the five-day-ahead data sets are prepared for constructing HiRBF models. A HiRBF model was constructed using the training data and then the model was used on the test data set. The instruction sets used to create an optimal HiRBF forecaster is  $S = F \cup T = \{+2, +3\} \cup \{x_1, x_2, x_3, x_4, x_5\}$ . Where  $x_i (i = 1, 2, 3, 4, 5)$  denotes the 5 input variables of the forecasting model.

The optimal HiRBF models evolved for three major internationally traded currencies: British pounds, euros and Japanese yen are shown in Figure 3. It should be noted that the important features for constructing the HiRBF models were formulated in accordance with the procedure mentioned in the previous section.

For comparison purpose, three single-stage RBF networks are also employed with structure of {5-10-1} for forecasting three major internationally traded currencies. The forecast performances of a traditional multi-layer feed-forward network (MLFN) model and an adaptive smoothing neural network (ASNN) model are also shown in Table 1. The actual daily exchange rates and the predicted ones for three major internationally traded currencies are shown in Figure 4. From Tables 1, it is observed that the proposed HiRBF forecast models are better than the considered neural networks models for three major internationally traded currencies.

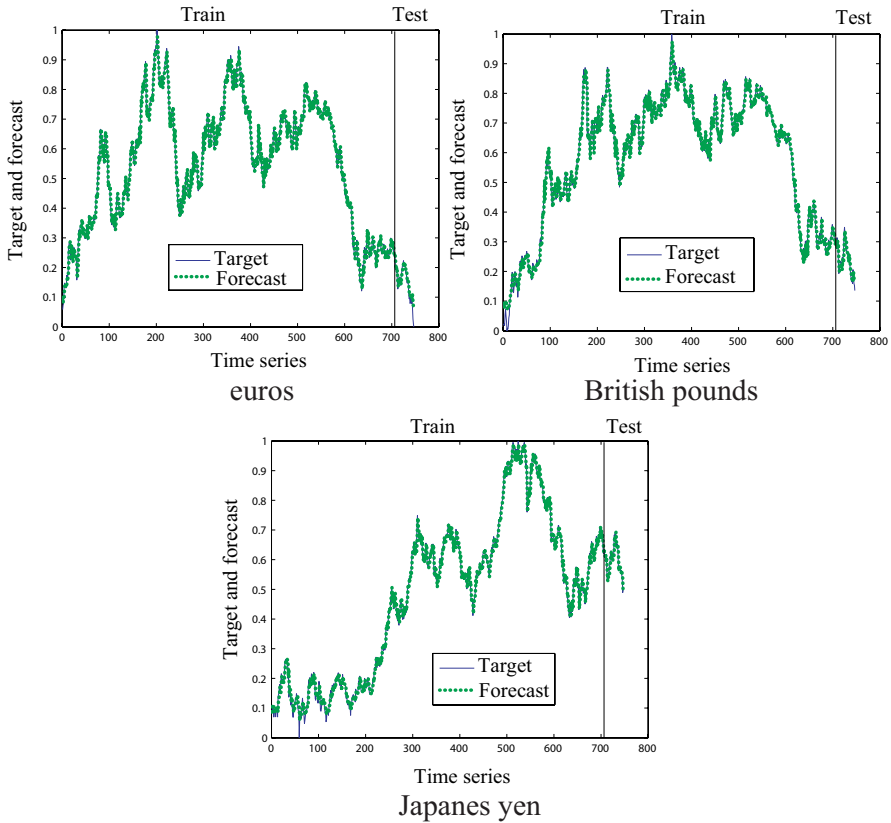


Fig. 4. The actual exchange rate and predicted ones for training and testing data set

## 4 Conclusions

In this paper, we presented a HiRBF model for forecasting three major international currency exchange rates. We have demonstrated that the evolved HiRBF forecasting model may provide better forecasts than the traditional MLFN forecasting model, the ASNN forecasting model and a traditional single RBF network. The comparative evaluation is based on a statistical measures (NMSE). Our experimental analyses reveal that the NMSE for three currencies using the HiRBF model are significantly better than those using the MLFN model, the ASNN model and the RBF model. This implies that the proposed HiRBF model can be used as a feasible solution for exchange rate forecasting.

## Acknowledgment

This research was partially supported the Natural Science Foundation of China under contract number 60573065, and The Provincial Science and Technology Development Program of Shandong under contract number SDSP2004-0720-03.

## References

1. P. Theodossiou, "The stochastic properties of major Canadian exchange rates". *The Financial Review* 29(2):193-221, 1994.
2. M. K. P. So, K. Lam and W. K. Li., "Forecasting exchange rate volatility using autoregressive random variance model", *Applied Financial Economics*, 9:583-591, 1999.
3. D. A. Hsieh. "Modeling heteroscedasticity in daily foreign-exchange rates", *Journal of Business and Economic Statistics*, 7:307C317, 1989.
4. D. Chappel, J. Padmore, P. Mistry and C. Ellis. "A threshold model for French franc/Deutsch mark exchange rate", *Journal of Forecasting*, 15:155-164, 1996.
5. A. N. Refenes, M. Azema-Barac, L. Chen and S. A. Karoussos. "Currency exchange rate prediction and neural network design strategies", *Neural Computing and Application*, 1:46-58, 1993.
6. L. Yu, S. Wang, and K. K. Lai. "Adaptive Smoothing Neural Networks in Foreign Exchange Rate Forecasting", V.S. Sunderam et al. (Eds.): *ICCS 2005, LNCS 3516*, pp. 523-530, 2005.
7. W. Wang, K.K.Lai, Y.Nakamori, S.Wang. "Forecasting Foreign Exchange Rates with Artificial Neural Networks: A Review". *International Journal of Information Technology & Decision Making*, 3(1):145-165, 2004.
8. Y. Chen, B. Yang and J. Dong, "Nonlinear System Modeling via Optimal Design of Neural Trees", *International Journal of Neural Systems*, 14(2):125-137, 2004.
9. Y. Chen, B. Yang, J. Dong, and A. Abraham, "Time-series Forecasting using Flexible Neural Tree Model", *Information Science*, 174(3-4):219-235, 2005.
10. J.T. Yao, C.L. Tan. "A case study on using neural networks to perform technical forecasting of forex", *Neurocomputing*, 34:79-98, 2000.
11. K. Sastry and D. E. Goldberg. "Probabilistic model building and competent genetic programming", In R. L. Riolo and B. Worzel, editors, *Genetic Programming Theory and Practise*, chapter 13, pp. 205-220. Kluwer, 2003.
12. <http://fx.sauder.ubc.ca/>
13. L. Yu, S. Wang, and K. K. Lai, "Adaptive Smoothing Neural Networks in Foreign Exchange Rate Forecasting", V.S. Sunderam et al. (Eds.): *ICCS 2005, LNCS 3516*, 523-530, 2005.
14. A. Abraham, "Analysis of Hybrid Soft and Hard Computing Techniques for Forex Monitoring Systems", *IEEE International Conference on Fuzzy Systems (IEEE FUZZ'02)*, IEEE Press pp. 1616 -1622, 2002.
15. L. Kirkpatrick et al., "Optimization by simulated annealing", *Science*, 220:671-680, 1983.
16. E.K. Burke et al., "A new local search approach with execution time as an input parameter", *Technical Report No. NOTTCS-TR-2002-3*, School of Computer Science and Information Technology, University of Nottingham, (2002).
17. N. A. Mat Isa, Mashor, M. Y., and Othman, N. H., "Diagnosis of Cervical Cancer using Hierarchical Radial Basis Function (HiRBF) Network", In Sazali Yaacob, R. Nagarajan, Ali Chekima (Eds.), *Proc. of the Int. Conf. on Artificial Intelligence in Engineering and Technology*, pp. 458-463, 2002.
18. S. Ferrari, I. Frosio, V. Piuri, and N. Alberto Borghese, "Automatic Multiscale Meshing Through HRBF Networks", *IEEE Trans. on Instrumentation and Measurement*, 54(4):1463-1470, 2005.
19. Z. Ahmad, J. Zhang, "Bayesian selective combination of multiple neural networks for improving long-range predictions in nonlinear process modelling", *Neural Comput & Applic.*, 14:78C87, 2005.

# Mathematical Formulation of a Type of Hierarchical Neurofuzzy System

Omar Sánchez, Sixto Romero, Francisco Moreno, and Miguel A. Vélez

Escuela Politécnica Superior, Huelva University, Ctra. Palos de la Frontera-Huelva,  
21819 La Rábida (Huelva), Spain

omar@uhu.es

<http://www.uhu.es>

**Abstract.** This paper presents a class of hierarchical fuzzy system applied to a cigar classification system. The weight, texture and chromatic characteristics are used to classify multiple cigars using a previous classification based on heuristic knowledge. In the adaptive part, a gradient descent and error backpropagation method is applied for adjusting the parameters. A detailed description of the algorithm is addressed. Copyright © 2005 IFAC.

## 1 Introduction

The miniaturization of electronic systems has increased the number of applications in every day life. The possibility of classification and conservation of cigars for long periods of time (improving their taste) is one of them [1].

Cigar quality depends on some parameters tested by cigar smokers: chromatic variations of wrapper, veins, texture, weight, flavour, age, mildness, strength (or body), smoothness or harshness, inconsistency and other heuristic variables [3]. Cigar experts can classify cigars using these considerations. Neurofuzzy systems are nonlinear dynamic models. Compared with other black box model techniques, this one is unique in its ability to use both qualitative and quantitative information [4]. Qualitative information is a human criteria-based knowledge formalized by fuzzy set, fuzzy logic and fuzzy rules, learned from input-output data. That is, a nonlinear system modeling technique where, after adaptation, the nonlinear function has physical meaning [6]. It is known that the number of fuzzy rules grows exponentially with the number of input variables. To overcome the problem, the fuzzy hierarchical structure is used. The problem with this method is that the knowledge present in the universe of discourse is lost in the hidden layers, and strategies for parametric adjustment of membership functions [5] cannot be applied.

In this paper the input data capture from force sensors and image camera and output data inferred from a cigar expert are used to classify cigars. Given that there is a large number of input variables, a hierarchical neurofuzzy system is used to design an intelligent classifier.

This paper is organized as follows: Section 2 introduces some basic mathematical principles of TSK fuzzy model. Section 3 presents the hierarchical fuzzy

system (HFS) for cigar classification. In Section 4 we derive a gradient descent algorithm for adjusting the HFS parameters to match the input-output pairs, and finally Section 5 and 6 summarize the results and draw the conclusions.

## 2 Mathematical Principle of the TSK Model

A fuzzy rule based on a multi-input single output system consists of a collection of fuzzy if-then rules that can be stated as the following

$$\text{If } x_1 \text{ is } F_1^l \text{ AND } x_2 \text{ is } F_2^l \text{ AND } \dots \text{ AND } x_n \text{ is } F_n^l \text{ then } y = y^l(\mathbf{x}) \quad (1)$$

where  $l = 1, 2 \dots M$  defines the number of rules,  $F_i^l$  are fuzzy sets in the universe of discourse  $U_i \subset R$  of the input variables,  $i = 1 \dots n$  represents the  $n$  input variables,  $y^l(x)$  is a consequent function,  $\mathbf{x} = (x_1, \dots, x_n) \in U_1 \times U_2 \dots \times U_n$  are the input variables and  $y \in V$  is the output variable. The premise if  $x_i$  is  $F_i^l$  in the fuzzy rules, taken here in the Gaussian function forms

$$\mu_{F_i^l} = \exp \left[ - \left( \frac{x_i - \gamma_i^l}{\beta_i^l} \right)^2 \right] \quad (2)$$

where  $\gamma_i^l$  and  $\beta_i^l$  are the centers and widths, respectively, of the function of the  $i$ th input variable  $x_i$ .

Applying the aggregation of the antecedent part of the  $l$ th rule for each input variable with the form of product inference rule, we obtain

$$\xi^l(\mathbf{x}) = \prod_{i=1}^n \mu_{F_i^l}(x_i, \beta_i^l, \gamma_i^l) = \prod_{i=1}^n \exp \left[ - \left( \frac{x_i - \gamma_i^l}{\beta_i^l} \right)^2 \right] \quad (3)$$

The consequent, represented in 1 by  $y^l(\mathbf{x})$ , is used here in Takagi-Sugeno-Kang (TSK) form, where the results must be a linear combination of the input variables, used in the multi-input single output fuzzy system model, as

$$y^l(\mathbf{x}) = f^l(x_1, x_2, \dots, x_n) = a_0^l + \sum_{i=1}^n a_i^l x_i = a_0^l + b^l(\mathbf{x}) \quad (4)$$

where  $a_0^l$  and  $a_i^l$  denote the adaptable parameters of the TSK consequent. The resulting system output signal with center average defuzzifier is represented by:

$$\hat{y}(\mathbf{x}) = f(\mathbf{x}) = \frac{1}{\sum_{l=1}^M \xi^l(\mathbf{x})} \sum_{l=1}^M y^l(\mathbf{x}) \xi^l(\mathbf{x}) \quad (5)$$

In order to adapt the parameters of the Neuro-Fuzzy System (NFS) defined above, a cost function is defined using the euclidean measure for a number of learning pairs  $P$  of the form  $[\mathbf{x}^p, y^p]$

$$J = \frac{1}{2} \sum_{p=1}^P (\hat{y}(\mathbf{x}^p) - y^p)^2 \quad \text{and} \quad e = (\hat{y}(\mathbf{x}) - y) \quad (6)$$

The adaptation of parameters can be done iteratively using the steepest descent method and backpropagation of the error in the iteration  $k$

$$\theta_i^l(k+1) = \theta_i^l(k) - \eta \frac{\partial J(k)}{\partial \theta_i^l(k)} \quad \text{and} \quad \frac{\partial J(k)}{\partial \theta_i^l(k)} = e \frac{\partial \hat{y}(\mathbf{x})}{\partial \theta_i^l(k)} \quad (7)$$

where  $\theta_i^l(k) = (\gamma_i^l(k), \beta_i^l(k), a_i^l(k), a_0^l(k))$ ,  $\eta$  is the learning rate and  $e$  is defined in (6).

### 3 The Hierarchical Fuzzy System

In order to classify cigars, a HFS must be designed [1], as shown in Fig. 1. The HFS is based on the measurement of weight (8 force sensors, that represent any generic sensor used to measure weight), texture (16 force sensors) and image characteristics of the cigar. As shown in Fig. 2, for obtaining the measure of the sensors that will constitute the input-output data, first, the cigar is weighted, then pressed, and finally, frames are captured by a camera. The input data obtained from image processing in order to identify the chromatic characteristics are the statistical approaches described in [7]: mean (M, average intensity), standard deviation (SD, average contrast), smoothness (S, relative smoothness of the intensity in a region), third moment (TM, skewness of a histogram), uniformity (U) and entropy (E, measure of randomness).

The HFS can be represented by

$$\hat{y}_5 = f_{HFS}(\mathbf{W}, \mathbf{T}, \mathbf{S}, \Theta_1, \Theta_2, \Theta_3, \Theta_4, \Theta_5) \quad (8)$$

where  $\Theta_1 - \Theta_5$  represent the parameters to be adapted in the NFS1-NFS5, respectively (for example, in Fig. 1 the output  $y_1$  corresponds to NFS1),  $\mathbf{S} = (M, SD, S, TM, U, E)$  are the statistical approaches obtained from image processing,  $\mathbf{W} = (f_1, \dots, f_8)$  represents the weight component and  $\mathbf{T} = (f_1, \dots, f_{16})$  represents the component of cigar texture.

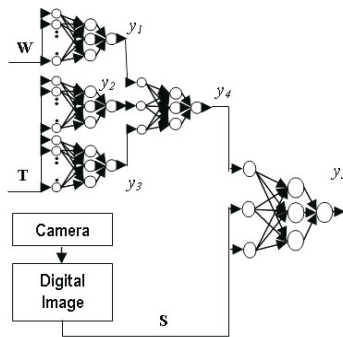
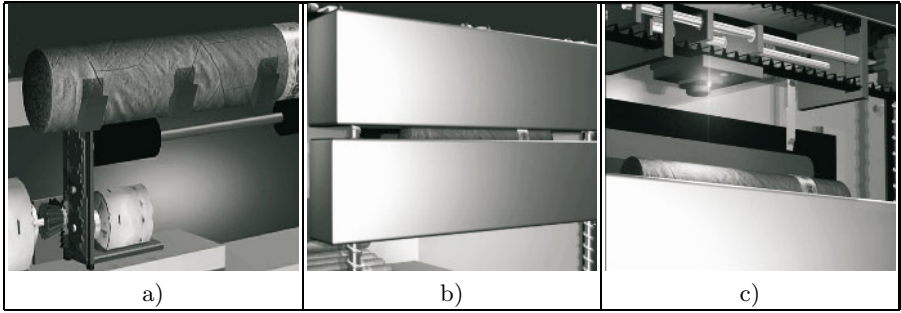


Fig. 1. Hierarchical fuzzy system for cigar classification



**Fig. 2.** Steps to obtain sensorial information of cigars. From top to bottom: a) weighting, b) texture measurement and c) frame capture.

The five NFS present in the cigar classification are defined by (see Fig. 1)

$$\hat{y}_1 = f_1(\mathbf{W}, \Theta_1); \quad \hat{y}_2 = f_2(\mathbf{T}_1, \Theta_2); \quad \hat{y}_3 = f_3(\mathbf{T}_2, \Theta_3) \quad (9)$$

where  $\mathbf{T}_1 = (f_1, \dots, f_8)$  and  $\mathbf{T}_2 = (f_9, \dots, f_{16})$ . Note that  $\mathbf{W}$  and  $\mathbf{T}_1$  have the information obtained from the same sensors, but the first is the result of cigar weight and the second from cigar texture in two different measurements. Note that (9) represents three NFS with 8 known inputs of the form represented by (5), without loss of generality. Equation (9) can be reformulated

$$\hat{y}_r = f_r(\mathbf{x}_r, \Theta_r) \quad \text{where } r = 1, 2, 3. \quad (10)$$

The next layer in the HFS is defined by

$$\hat{y}_4 = f_4(\hat{y}_1, \hat{y}_2, \hat{y}_3, \Theta_4) \quad (11)$$

where  $f_4$  represents the hidden NFS4, that is, their input-output data have no physical meaning. And finally

$$\hat{y}_5 = f_5(\hat{y}_4, \mathbf{S}, \Theta_5). \quad (12)$$

Note that the above NFS5 has known inputs ( $\mathbf{S}$ , as defined in (8)), and one unknown input  $y_4$  (the result of projecting the NFS4).

The more general form of (12) is the result of substituting (9) in (11) and finally in (12), that results in

$$\hat{y}_5 = f_5(f_4(\hat{y}_1, \hat{y}_2, \hat{y}_3, \Theta_4), \mathbf{S}, \Theta_5). \quad (13)$$

## 4 The Adaptive Law of HFS

The input output data are obtained here in the following form:

- a)  $\mathbf{W}$  and  $\mathbf{T}$  are obtained by sensors (see the description of (8)).

- b) Equation (9) is applied in order to obtain the values of  $(f_1, f_2, f_3)$ .
- c) Equation (11) is computed and hence  $f_4$  is obtained.
- d)  $\mathbf{S}$  is obtained from image processing (note how obtaining  $\mathbf{S}$  from image processing after computing (9) and (11) compensates the computation effort).
- e) The cigar quality classified according to the expert's criteria is quantified, defining the values of  $y$  in (6).

The estimated output of HFS ( $\hat{y}_5$ ) is obtained from the application of (12). The error is computed using (6).

#### 4.1 The Adaptation of Parameters

In order to obtain the parameter adjustment law of the NFS5, we must compute (7). The adaptation of the width parameters defined in (2), substituted in (3), and finally in (5), becomes

$$\frac{\partial \hat{y}(\mathbf{x})}{\partial \beta_i^l(k)} = \frac{2 \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{x_i - \gamma_i^l}{\beta_i^l} \right)^2 \right] \right\} \frac{(x_i - \gamma_i^l)^2}{(\beta_i^l)^3} (y^l - \hat{y})}{\sum_{l=1}^M \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{x_i - \gamma_i^l}{\beta_i^l} \right)^2 \right] \right\}} \quad (14)$$

Similarly, the centers can be adapted using

$$\frac{\partial \hat{y}(\mathbf{x})}{\partial \gamma_i^l(k)} = \frac{2 \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{x_i - \gamma_i^l}{\beta_i^l} \right)^2 \right] \right\} \frac{(x_i - \gamma_i^l)}{(\beta_i^l)^2} (y^l - \hat{y})}{\sum_{l=1}^M \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{x_i - \gamma_i^l}{\beta_i^l} \right)^2 \right] \right\}} \quad (15)$$

The affine term  $a_0^l$  of the consequent part (4) also responds to

$$\frac{\partial \hat{y}(\mathbf{x})}{\partial a_0^l} = \frac{\prod_{i=1}^n \exp \left[ - \left( \frac{x_i - \gamma_i^l}{\beta_i^l} \right)^2 \right]}{\sum_{l=1}^M \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{x_i - \gamma_i^l}{\beta_i^l} \right)^2 \right] \right\}} \quad (16)$$

and finally, the linear parameters of the consequent part

$$\frac{\partial \hat{y}(\mathbf{x})}{\partial a_i^l} = \frac{\left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{x_i - \gamma_i^l}{\beta_i^l} \right)^2 \right] \right\} x_i}{\sum_{l=1}^M \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{x_i - \gamma_i^l}{\beta_i^l} \right)^2 \right] \right\}} \quad (17)$$

Note that the above expressions can't be directly applied because the output data of the HFS depend on the precedents layers. For example, because the error of HFS (see (6)) is

$$e = \hat{y}_5(\mathbf{x}) - y^c \quad (18)$$



where  $y^c$  represents the quantification of the cigar expert's criteria and  $\hat{y}_5$  is the equation (13). The adaptive law (14) applied over NFS1, computing the chain rule, becomes

$$\frac{\partial \hat{y}_5(\mathbf{x})}{\partial \beta_{1i}^l(k)} = \frac{\partial \hat{y}_5(\mathbf{x})}{\partial \hat{y}_4(\mathbf{x})} \frac{\partial \hat{y}_4(\mathbf{x})}{\partial \hat{y}_1(\mathbf{x})} \frac{\partial \hat{y}_1(\mathbf{x})}{\partial \beta_{1i}^l(k)} \tag{19}$$

In the same way, the parameters of NFS2 and NFS3 must be adapted. Then, using (10) and (7), we can generalize

$$\frac{\partial \hat{y}_5(\mathbf{x})}{\partial \theta_{ri}^l(k)} = \frac{\partial \hat{y}_5(\mathbf{x})}{\partial \hat{y}_4(\mathbf{x})} \frac{\partial \hat{y}_4(\mathbf{x})}{\partial \hat{y}_r(\mathbf{x})} \frac{\partial \hat{y}_r(\mathbf{x})}{\partial \theta_{ri}^l(k)} \tag{20}$$

where  $\theta_{ri}^l = (\gamma_{ri}^l(k), \beta_{ri}^l(k), a_{ri}^l(k), a_{r0}^l(k))$ . The number of parameters to be adapted in the first 3 NFS will be  $3(3Mn + M)$ .

The derivative term with respect to the input variable is similar to (15) but with respect to  $x_i(\hat{y}_4$  in (12)), then

$$\frac{\partial \hat{y}_5(\mathbf{x})}{\partial \hat{y}_4} = \frac{-2 \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{x_i - \gamma_{5i}^l}{\beta_{5i}^l} \right)^2 \right] \right\} \frac{(x_i - \gamma_{5(i=\hat{y}_4)}^l)}{(\beta_{5(i=\hat{y}_4)}^l)^2} (y_5^l - \hat{y}_5)}{\sum_{l=1}^M \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{x_i - \gamma_{5i}^l}{\beta_{5i}^l} \right)^2 \right] \right\}} \tag{21}$$

where  $\beta_{5(i=\hat{y}_4)}^l$  represents the width of Gaussian MFs of NFS5 corresponding to rule  $l$  and input variable  $i$  in which the variable  $\hat{y}_4$  has been placed (for example,  $i = 1$  if  $x_i = (\hat{y}_4, \mathbf{S})$ ). In the same way,

$$\frac{\partial \hat{y}_4(\mathbf{x})}{\partial \hat{y}_r(\mathbf{x})} = \frac{-2 \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{y_{ri} - \gamma_{4i}^l}{\beta_{4i}^l} \right)^2 \right] \right\} \frac{(y_{ri} - \gamma_{4(i=\hat{y}_r)}^l)}{(\beta_{4(i=\hat{y}_r)}^l)^2} (y_4^l - \hat{y}_4)}{\sum_{l=1}^M \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{y_{ri} - \gamma_{4i}^l}{\beta_{4i}^l} \right)^2 \right] \right\}} \tag{22}$$

The differentials represented by equations (14)-(17) are valid for one NFS, but (20) is expressed for adapting the parameters of 3 NFS ( $r = 1, 2, 3$ ). The parameters of NFS4 to be adapted can be given by

$$\frac{\partial \hat{y}_5(\mathbf{x})}{\partial \theta_{4i}^l(k)} = \frac{\partial \hat{y}_5(\mathbf{x})}{\partial \hat{y}_4(\mathbf{x})} \frac{\partial \hat{y}_4(\mathbf{x})}{\partial \theta_{4i}^l(k)} \tag{23}$$

And finally, the adaptation of NFS5, using (7), responds to

$$\theta_{5i}^l(k + 1) = \theta_{5i}^l(k) - \eta e \frac{\partial \hat{y}_5(\mathbf{x})}{\partial \theta_{5i}^l(k)} \tag{24}$$

The number of parameters to be adapted in each NFS is shown in the table 1, and the total number of parameters of HFS is  $5(3Mn + M)$ .

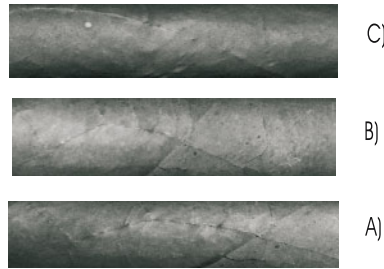
It should be noted that the HSF used here has proven good results in a curvature controller applied to mobile robot [2]

**Table 1.** Parameters adapted in each hierarchical level

Hierarchical Level	Parameters to be adapted	Number of parameters
NFS1, 2, 3	$\gamma_{ri}^l(k), \beta_{ri}^l(k), a_{ri}^l(k), a_{r0}^l(k)$	$3(3Mn+M)$
NFS4	$\gamma_{4i}^l(k), \beta_{4i}^l(k), a_{4i}^l(k), a_{40}^l(k)$	$3Mn+M$
NFS5	$\gamma_{5i}^l(k), \beta_{5i}^l(k), a_{5i}^l(k), a_{50}^l(k)$	$3Mn+M$

## 5 Simulation Results

An expert has classified three cigars from highest Fig. 3A) to lowest quality Fig. 3C). Let us design the HFS in the form of (13).



**Fig. 3.** Cigars classified by an expert. A) Highest quality. C) Lowest quality.

The first step is to obtain the input-output data. The weight component  $\mathbf{W}$  is the values obtained from force sensors placed in the lower structure (see Fig. 2A)). The texture component  $\mathbf{T}$  is the measure of the 16 force sensors while the cigar is being pressed by the upper and lower structures (see Fig. (2B)). The frames to classify the chromatic characteristics of cigars are captured by a digital camera (see Fig. 2) c). Suppose that the result of the cigar measure shown in Fig. 3 is (in Newton):

$$\mathbf{W}_a = [0.12, 0.124, 0.116, 0.123, 0.121, 0.118, 0.121, 0.119]$$

$$\mathbf{W}_b = [0.126, 0.123, 0.125, 0.125, 0.122, 0.127, 0.124, 0.121]$$

$$\mathbf{W}_c = [0.119, 0.122, 0.118, 0.12, 0.119, 0.12, 0.121, 0.118]$$

The texture is measured considering a force until one of the sensors is twice the maximum force obtained when the weight was measured, then

$$\mathbf{T}_a = [0.22, 0.24, 0.248, 0.23, 0.221, 0.231, 0.21, 0.243, 0.215, 0.245, 0.243, 0.227, 0.229, 0.245, 0.223, 0.242]$$

$$\mathbf{T}_b = [0.22, 0.254, 0.24, 0.232, 0.213, 0.223, 0.221, 0.25, 0.205, 0.234, 0.224, 0.212, 0.232, 0.244, 0.243, 0.247]$$

$$\mathbf{T}_c = [0.22, 0.244, 0.214, 0.221, 0.203, 0.231, 0.215, 0.234, 0.205, 0.24, 0.231, 0.202, 0.242, 0.213, 0.224, 0.227]$$

The  $\mathbf{S}$  vector will give information about "image texture" based on statistical approaches, [7]. In this case figure 3 can be represented by

$$\begin{aligned} \mathbf{S}_a &= [105.8889, 28.2299, 0.0121, -0.0285, 0.0098, 6.8186] \\ \mathbf{S}_b &= [122.4414, 26.3556, 0.0106, -0.0705, 0.0105, 6.7118] \\ \mathbf{S}_c &= [98.7088, 29.3887, 0.0131, -0.1656, 0.0100, 6.8101] \end{aligned}$$

Finally, and considering that the cigar of Fig. 3A) has the highest quality, the input-output data becomes

$$\begin{bmatrix} \mathbf{W}_a, \mathbf{T}_a, \mathbf{S}_a; 1 \\ \mathbf{W}_b, \mathbf{T}_b, \mathbf{S}_b; 0.5 \\ \mathbf{W}_c, \mathbf{T}_c, \mathbf{S}_c; 0.1 \end{bmatrix} \tag{25}$$

The model was implemented using 1000 epochs in each NFS, with  $M = 9$  rules, the number of inputs is  $n = 8$  for NFS1, 2 and 3;  $n = 3$  for NFS4 and  $n = 7$  in the case of NFS5. The learning rate (with momentum) was initialized with 0.01.

The same input data used in the learning phase were applied as input data in the final system, resulting in:  $\hat{y}_5^a(\mathbf{x}) = 0.95$ ,  $\hat{y}_5^b(\mathbf{x}) = 0.48$  and  $\hat{y}_5^c(\mathbf{x}) = 0.13$ , very similar to the output data to be learned. Because the data of sensors have been assumed (using as reference the weight of three cigars), the validation has been made with the original data.

## 6 Conclusions

A class of hierarchical fuzzy system has been applied in order to quantify the knowledge of a cigar expert, reducing the "dimensionality" problem of the number of fuzzy rules. The lack of interpretability of the hidden variables in internal layers can be addressed in this application supposing a universe of discourse according to the meaning of the NFS output. For example, the output of NFS can be  $[0, \mathbf{W}_{max}]$  where  $\mathbf{W}_{max}$  represents the maximum weight measured by sensors. This issue will be addressed in the next article.

## Acknowledgment

This research was supported by grant number DPI2004-07310 from the Ministry of Education, Science and Business.

## References

1. Sánchez O.,Romero S.: Robotic Humidor: A Cigar Classification Approach. 8th World Multi-Conference on Systemics, Cybernetics and Informatics. **10** (2004) 294–300
2. Sánchez O.,Ollero A., Heredia G.: Hierarchical fuzzy path tracking and velocity control of autonomous vehicle. Integrated Computer Aided Engineering, v.6(4) (1999) 289–301
3. Freccia D.,Jacobsen J., Kilby P.: Exploring the relationship between price and quality for the case of hand-rolled cigars. The Quartely Review of Economics and Finance **43** (2003) 169–189

4. Ying H.: Fuzzy Control and Modelling. Analytical Foundations and Applications. IEEE Press Series on Biomedical Engineering (2000)
5. Vélez M.A., Sánchez O., Ollero A.: Fuzzy Modelling with Parameters Adjustment Based on Overlap Ratio. 10th Mediterranean Conference on Control and Automation. Lisboa (Portugal) (2002)
6. Tsoukalas L., Uhrig R.: Fuzzy and Neural Approach in Engineering. Wiley. New York (1997)
7. González R., Woods R.: Digital Image Processing (Second Edition). Prentice Hall. (2002)

# Hardware Support for Language Aware Information Mining

Michael Freeman and Thimal Jayasooriya

Department of Computer Science, University of York, UK  
{mjf, thimal}@cs.york.ac.uk

**Abstract.** Information retrieval from text or ‘text mining’ is the process of extracting interesting and non-trivial knowledge from unstructured text. With the ever increasing amounts of information stored on the web or archived within a computing system, high performance data processing architectures are required to process this data in real time. The aim of the work presented in this paper is the development of a hardware text mining IP-Core for use in FPGA based systems. In this paper we will describe the pre-processing engine we have developed for the PRESENCE II PCI card, to accelerate the identification of significant words within a document, logging their frequency and position. The performance of this system is then compared to an equivalent software implementation using the Lucene software package.

## 1 Introduction

The aim of information retrieval systems is to automate the process of identifying and categorising text, informing a user of the existence of documents relating to their request. Information retrieval tasks have a wide variety of applications – in intelligent data mining, search engine technologies and natural language processing tasks. The work presented in this paper examines the possibility of using field programmable gate arrays (FPGA) to construct an application specific co-processor capable of accelerating this process. We show that a scalable, resource efficient hardware architecture for information extraction (IE) and information retrieval (IR) tasks is far faster than the equivalent software for the same task. Using hardware for IR/IE tasks has other benefits, power conservation and efficiency being chief among them. The final aim of this work is to construct a complete hardware based text data mining architecture, combining this initial word pre-processing engine with additional hardware IP-cores e.g. to generate training data for a binary neural network, or other similarity functions [1]. The intended application domains for such a system are those requiring high levels of processing performance to meet real time deadlines. These include real time market analysis of web based stock market or newswire services [2], or “second generation” search technologies [3] to process complex queries on large unstructured databases quickly and effectively.

In the next section, we will introduce a hardware architecture based on hardware hashing as a means of performing the required word count operation. This hardware forms part of a processing pipeline that performs tasks such as word identification,

stemming, stop word removal and word categorisation. Section 3 presents some performance results of this hardware IP-core, compared to an equivalent software implementation using the Lucene software package [4]. Finally we close this paper with conclusions and future work.

### 1.1 Information Extraction and Intelligent Text Analysis

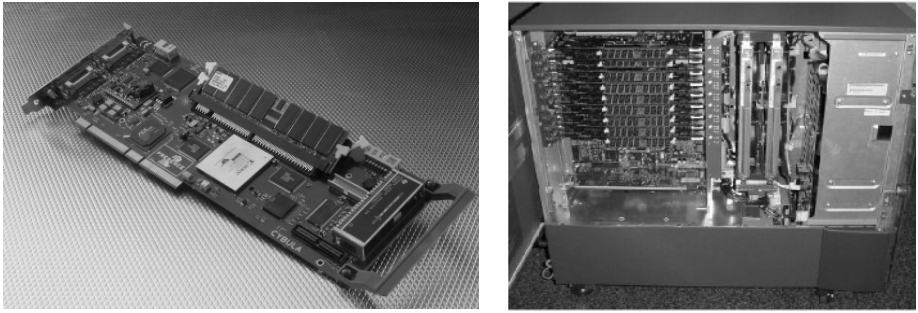
One of the key ideas on which much of automatic text analysis research has been built is the work by Luhn [5]. Keyword frequency and density are used to identify significant sentences within a document, from which its contents can be categorized [6]. This information can now be used to categorise a document, as there is a high probability that such sentences contain information which is highly relevant to the document's content [7].

Identifying significant sentences within a document can have other applications; extraction of key sentences supplements existing feature selection, ranking and clustering techniques for a document corpus [8]. Our intent is to provide a hardware based architecture for common natural language based analysis of document text and also to provide access to common statistics about keyword frequency and density in a given document.

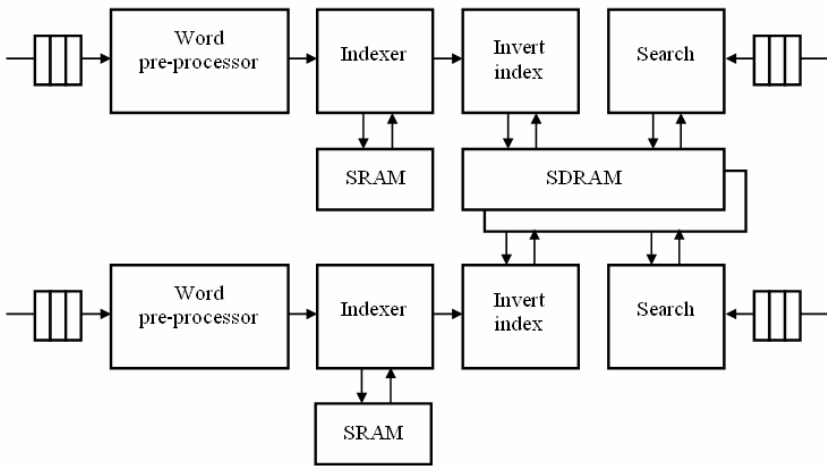
## 2 Hardware Architecture

The hardware architecture developed has been targeted at large volume, high bandwidth application areas e.g. library archives or data stream searches. To accelerate these systems, processing is distributed across an array of PRESENCE II PCI cards [9]. The main components of this card include; Virtex-II FPGA, Texas instruments DSP, up to 4GB of SDRAM (72 bit), two 1MB ZBT SRAM (36 bit) memories and eight bi-directional low voltage differential signal (LVDS) communication channels. Data to be processed will be typically stored on an array of hard disks or streamed across an Ethernet port. In either case this data must be transferred across the PCI bus to each card (transfer speeds 60MB/s – 90MB/s). In a typical workstation 6 – 9 PRESENCE II cards can be supported, an example is shown in figure 1, a Cortex-2 data mining system [10]. These cards can be configured as either a single master, passing data to slave cards across the LVDS communication channels, or multiple independent cards each accessing a block of data in turn, processing it, before returning results back across the PCI bus to main memory. In both cases data to be processed is buffered locally on each card in SDRAM.

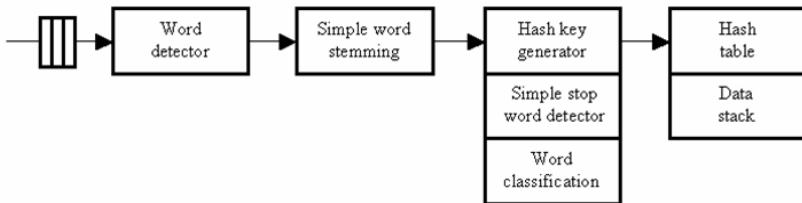
Each PRESENCE-2 card will implement two IE engines, capable of processing independent text input streams, the proposed system architecture is shown in figure 2. This processing pipeline is made up of four distinct stages; word pre-processing, indexing, index inverter and search. In typical operation, a document is streamed into the processing pipeline. Word boundaries are identified and words extracted, with common stems and non-significant words removed. The remaining words are then added to an index, logging their position, type and frequency within the document. As more documents are processed, the resulting indexes are inverted and merged i.e. an common index is formed, detailing what documents contain identified keywords. This



**Fig. 1.** Cortex-2 system, with nine PRESENCE-2 cards



**Fig. 2.** FPGA System architecture



**Fig. 3.** Word pre-processor and indexing hardware

index can then be searched to identify possible documents relating to a users query. The work presented in this paper describes the first two stages of this pipeline, which form the core indexing module. This hardware has been designed to process standard 8 bit ASCII characters and can be divided into four distinct pipeline stages, as shown in figure 3, each written in VHDL.

The first pipeline stage is word identification. In this stage, word boundaries are identified using a defined set of termination characters e.g. *space*, ( ), ? etc. Case information is also removed to convert each character into a 6 bit internal representation. Full stop and comma characters are not considered as word terminating characters to simplify the processing of abbreviations and numerical values e.g. N.A.T.O. or 1,000,000 etc. In these cases a defined set of equivalent characters are replaced with a don't-care representation e.g. 10\_000 is considered to be the same as 10,000 etc. Each of the nine characters in the 72 bit input word, addresses a 128 x 6 bit ROM, R<sub>0</sub>-R<sub>8</sub>, constructed from LUT elements within the FPGA. These perform the required parallel translations into letter, number, don't-care and terminate data types. Decode logic then determines the number of words within this nine character array i.e. 0 – 4, and passes valid data into the output FIFO. Word boundaries are hard coded in logic as a transition from a letter, number or don't-care data type to a terminating data type, removing the need to process strings of terminating characters. To further increase processing performance, this decode window can be increased by chaining decoders, therefore, increasing the probability of identifying and processing a number of words in parallel.

The second pipeline stage is a simple stem decoder. In this stage a set of commonly used word suffixes have been hard coded into an array of string comparators, allowing a common stem to be extracted i.e. matching a common word stem ending with either s, 's, s', es, ed, or ment. This suffix stripping stage helps minimise the number of identified key words, reducing processing and storage requirements. Each word's character string is passed through a delay line, which drives a set of string comparators, encoded to detect the defined suffixes. When the end of the string is detected, a priority decoder decodes the comparator outputs and the correct mask is applied to the delay line. This process converts the matched characters into don't-care characters; such that *adjust*, *adjusts*, *adjusted* and *adjustment* will all be considered equivalent in the next processing stage.

The third pipeline stage is key generation. In this stage the variable length character string i.e. the identified word, is hashed into a fixed hash address. Don't-care characters are not included in this hashing process such that equivalent words hash to the same address e.g. U.S.A. and USA. The variable length hash function is implemented using a simple accumulating multiplication by a constant prime number (C). This approach was chosen to minimise the required hardware to process don't-care characters i.e. removes the need for barrel shifters to reorder input data. However, this solution does introduce a sequential processing element into the design, reducing performance when compared to a parallel implementation. Also included in this phase is a 5 x 30 bit register file as a temporary store, forming the required 36 bit data packets for external memory. This register file, therefore, defines the maximum key size i.e. the characters in a word containing more than 25 characters will still be hashed to form a unique hash address, however, only the first 25 characters from that word will be used to identify it within the hash table. This approximation was considered acceptable for the test data set used, however, the required key size can be increased if required. To minimise the number of hash operations performed, 42 of the most common stop words e.g. a, if, will etc., are hard coded into a simple stop word decoder. In the event that a stop word is detected, the hashing operation is halted and the next identified word processed. In addition to generating a fixed hash key,



each word or phrase is also classified at this stage. Three data bits are used to encode this data, indicating if a word or phrase contains, letters, numbers or don't-care characters e.g. the date 10-05-2005 would be tagged as containing numbers and don't-care characters. This representation can be used in later search operations to minimise retrieval times.

The final pipeline stage is a hardware hashing memory architecture, implementing keyword frequency and position logging. Hardware hashing has been used to implement an associative memory structure to allow fast keyword searches to be performed. The fixed 32bit initial hash address generated by the previous stage is passed through a further hash function to reduce this data width to the external memory's 17 bit address bus. The hashing functions chosen are based on a combination of bit extraction, exclusive OR bit folding and a pseudo random number generator [11]. These schemes were selected to match the available hardware resources within the FPGA, in order to maximise performance and minimise area requirements. The hardware implementation of the hash table is based on synchronous ZBT SRAM devices, with control logic implemented in the FPGA. A 128K open addressing hash table scheme has been chosen, with collisions handled by the pseudo random number generator i.e. the next number from the initial seed value. This implementation allows data to be quickly searched for and inserted into the hash table. However, deleting data from the hash table is not a simple matter of marking slots as empty, since this would cause the search path to terminate at that point. One solution is to use a deleted flag, indicating that an entry can be overwritten, allowing the search to continue until a real empty location is found. However, if a block or the whole hash table needed to be cleared e.g. when a new document is processed, the delay in marking each entry as overwriteable or empty would be unacceptable. To overcome this problem, a stack data structure has been added. Data is no longer stored in the hash table, but replaced by a pointer to this data on a stack. Data in the hash table is now defined as valid if its associated data on the stack is within the current stack frame. The hash table contains the frequency (number instances), size (number of characters, not including wildcards) and position on the stack for each identified word. Each stack frame contains the data (identify word), word / phrase class (contains letters, numbers or wildcards) and a link list of the positions at which this word has occurred (an offset word count). To ensure that stack frames cannot be misinterpreted each data word's type is also stored with this data. Once a document has been processed, the identified keyword data will be stored within the current stack frame. To retrieve this data, the head of each stack frame is accessed, allowing the frequency and size data to be retrieved from the hash table. The link list containing each words position information can then be accessed, with the MSB of the last value being set to one to indicate the end of the list.

### 3 Experimental Results

The text pre-processing pipeline has been benchmarked using text taken from the 1998 New York Times and compared to an equivalent software implementation taken from the Lucene information retrieval library. The chosen documents are constructed from a series of articles, containing approximately 400,000 words, with characteristics

shown in figure 4. For the purposes of this benchmark data transfers to and from disk have been excluded to give a true indication of the indexing performance of each approach. The Lucene software package is an open source, freely available information retrieval library, comprising of several individual components, which can be plugged to form an end to end information retrieval task. Lucene is written in Java, with almost every aspect of its operation being available for extension and customization. This useful flexibility allowed us a wide array of benchmarks in a variety of real world conditions.

Lucene conceptually organizes indexed text into *documents*. A document comprises of multiple *fields*; significant words or phrases that comprise an atomic unit for searches. A field could include almost any chunk of characters and digits, which are considered significant in the language being indexed e.g. high level constructs such as telephone numbers, social security numbers and dates or lower level linguistic constructs such as individual words. The core analysis component of the Lucene information retrieval library is found in *analysers*; a segment of logic responsible for performing basic cleansing and organizing of indexed terms, before they are added into a persistent index. Lucene offers a wide variety of predefined analyzers for use, although in practice most implementations customize and tailor an analyzer that specifically fits the needs of the project. For the purposes of our evaluation we produced a version of Lucene with several customizations, which eliminated writes to disk on successful completion of an indexing run. This is the equivalent of the *RamDirectory* structure in Lucene. Another feature of Java, which creates unpredictable benchmarking results, is the invocation of the garbage collection mechanism. To eliminate this overhead the Java Virtual Machine (JVM) was invoked with a large maximum memory limit, 300MB of usable memory. This allows a complete indexing run for our test set of documents without a call to the garbage collector. The JVM was also invoked with a command line switch to continuously run the garbage collector in the background, instead of invoking the garbage collector only when memory for new allocations runs low. However, we found that there was no difference between executing the code with the specialized switch, as the dataset indexed was sufficiently small so as to not require even a single garbage collection invocation. To negate the relatively large JVM start up overhead the indexer is fed spurious data before the actual benchmarking (the warm up procedure) to allow the JVM adequate instructions to generate optimisations.

For the purposes of our experiments, we used a customized version of the Lucene *IndexWriter* class; which allowed us to programmatically toggle the writing of both indices and segment information. We also set *mergeFactor* to 1000 and *minMergeDocs* to 1000 – essentially instructing Lucene to trade memory for index writes and thus provide the fastest possible indexing operation. The indexing experiments were carried out using the Lucene *StandardAnalyzer* class, which is identical in its functionality to the FGPA IP-Core indexing functions. The graph depicted in Figure 5 shows the benchmark data for the customized Lucene indexing procedure, running on a Pentium 4, 3.4GHz (1GB memory), a 2.4GHz (1.5GB memory) and a Pentium M 1.6GHz (1GB memory). All benchmarked times are averaged over 10 runs. These graphs show that processing performance for small document sizes e.g. 15K lines or 108K words, are still significantly affected by initial software overheads, resulting in an average processing time per word of 2.5  $\mu$ s. This

effect is minimised for larger document sizes, with the average processing time per word for 544K lines or 3990K words falling to 366 nS. An interesting feature of these results is that the Pentium M out performs the 2.4GHz Pentium 4 machine. Possible reasons for this result could be due to the larger L2 cache or that the Pentium III core at the heart of the Dothan variant of the Pentium M processor is better suited to this type of application.

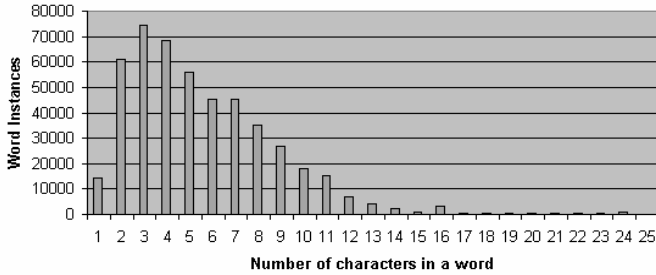


Fig. 4. Test data characteristics

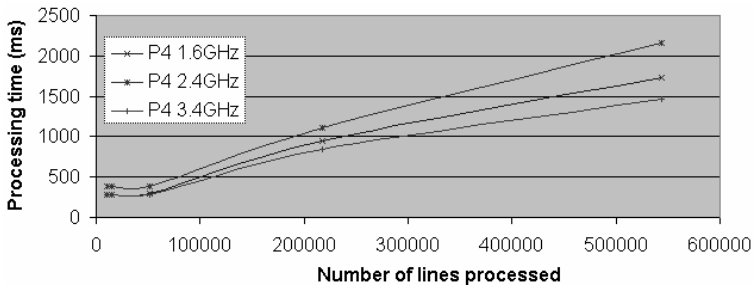


Fig. 5. Lucene software processing performance

The processing performance of the hardware pre-processing engine using a Virtex II FPGA, with a 100MHz system clock, has been measured for 1000 – 15000 lines of text (approximately 7200 – 108000 words). From the graphs shown in figures 6 and 7, it can be seen that processing time and memory usage is proportional to the number of lines, requiring on average 87 nS to process each word. This compares very favourably to the equivalent software implementation, resulting in a 29 times speedup for small documents and a 4 times speedup for large documents. These figures could be further improved if:

- External ZBT SRAM data width could be increased. This would minimise the number of stack operations i.e. from figure 4 it can be seen that the majority of the words will require 1–3 storage location, increasing the external data width would reduce read, write, and compare operations to a single cycle. Reducing the average processing time by 10-20 ns.
- Switch from pipelined to flow-through ZBT SRAM. The current PRESENCE cards are populated with pipelined memory devices. However, owing to the

number of sequentially dependent memory accesses required, the hardware controller must insert at least 3 stall cycles i.e. a 30 ns delay. Flow-through memory devices would remove these.

The main difference between the hardware and software implementations is that the FPGA based system has been designed to process documents in isolation, whereas the software implementation uses a common data structure for a number of documents. As a result the hardware stack memory depth limits the maximum document size that can be processed to approximately 1500 lines of text, requiring 126K words of memory. For the intended application this limitation was considered acceptable as these partial indexes are combined in later stages of the processing pipeline shown in figure 2. To overcome this limit, one solution would be to implement the stack data structure in the onboard SDRAM, instead of ZBT SRAM i.e. allowing up to 4GB of storage. This modification would require a more complex memory controller, but would not affect processing performance.

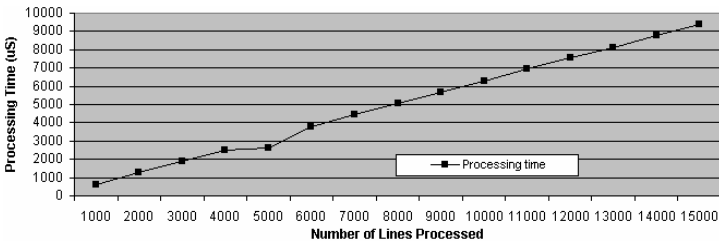


Fig. 6. Hardware processing performance

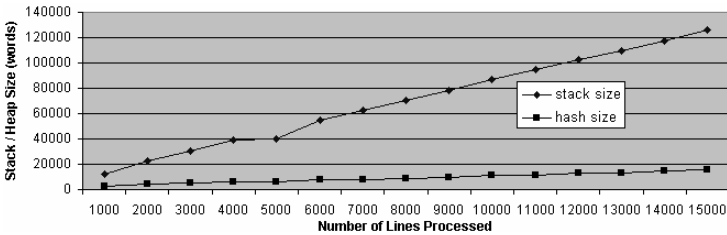


Fig. 7. Hardware memory usage

## 4 Conclusions

The results for a single processing engine running on a PRESENCE II card have shown that it is possible to achieve a speedup of 4 to 29 when compared to an equivalent software implementation running on a top end Pentium 4 system. Switching from pipelined to flow-through ZBT SRAM would double this speedup i.e. 8 to 58, by removing wait states during memory accesses. Distributing this application across an array of PRESENCE II PCI cards can further increase this speedup. The main limiting factor of such a system will be the bandwidth across the

PCI bus. This limitation should not be significant, as all data is locally buffered on each card and the size of the raw text data is an order of magnitude larger than the results produced i.e. data transfers can be hidden. Therefore, the speedup for this processing array should be approximately proportional to the number of card e.g. for the Cortex-2 system this would give a speedup of approximately 36 to 261, compared to a top end workstation. Performance will also be improved with the development of the PRESENCE III PCI card. This card will have a Virtex 4 FPGA and additional memory banks, increasing system clock speeds and allowing multiple processing engines on a single card e.g. assuming that the 200MHz system clock can be used (the main limiting factor would now be the external memory), this would result in a 4 times speedup on the existing card. In conclusion if an application specific variant of the PRESENCE III PCI card could be produce, the possible speedup per card would be 32 to 232 times faster than a top end workstation.

Further work will include the development of additional hardware to complete a full text mining IP-Core, capable of performing complex text search and categorisation functions without the need of software processing layers. It is hoped that this approach will again significantly improve the performance of these types of operations. At present work is currently focused on implementing a Porter stemming function [12] in hardware and the addition of a loadable stop word module based on an additional hardware hash table.

## References

1. Freeman M.J., Weeks, M., Austin, J., (2005), Hardware implementation of Similarity Functions, IADIS International Conference on Applied Computing, Algarve, Portugal
2. Sholom M.W., Naval V.K., (2002), A System for Real-time Competitive Market Intelligence., WWW: [http://www.research.ibm.com/dar/papers/pdf/weiss\\_kdd2002\\_mi.pdf](http://www.research.ibm.com/dar/papers/pdf/weiss_kdd2002_mi.pdf).
3. Sturgeon W., (2005), Interview: Mike Lynch, founder of Autonomy on Google, penguins and the future of search., WWW: <http://software.silicon.com/applications.0,39024653,39152405,00.html>
4. Cutting D. and others, (2005), The Lucene search engine, WWW: <http://lucene.apache.org>
5. Luhn H.P.,(1958) The automatic creation of literature abstracts, IBM Journal, April.
6. Rijsbergen C.J. van, (1979), Information Retrieval 2nd Edition, Butterworths
7. Baeza-Yates, R., Ribiero-Neto, B., (1999), Modern Information Retrieval, Addison Wesley
8. Wang L., Xiuju F., (2005), Data mining with computational intelligence, Springer Verlag
9. ACAG, (2002), AURA - Research into high-performance pattern matching systems, WWW: <http://www.cs.york.ac.uk/aura>
10. Cybula (2005), WWW: <http://www.cybula.com>
11. Chowdhury D.R., Gupta I.S., Chaudhuri P.P., (1995) A low cost high capacity associative memory design using cellular automata., IEEE Transactions on computers, Vol. 44, No 10, pp. 1260 – 1264
12. Porter M.F., (1980), An Algorithm for suffix stripping, Program, 14(3), p130-137

# Measuring GNG Topology Preservation in Computer Vision Applications

José García Rodríguez, Francisco Flórez-Revuelta, and Juan Manuel García Chamizo

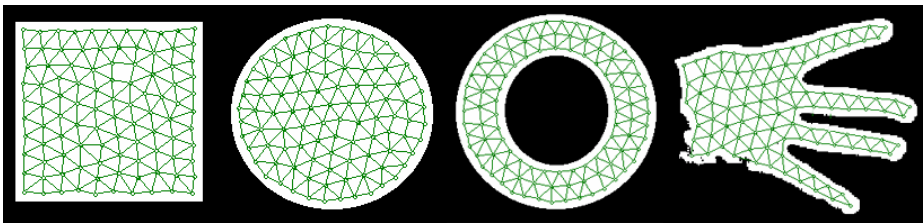
Department of Computer Technology and Computation. University of Alicante,  
Apdo. 99. 03080 Alicante, Spain  
{jgarcia, florez, juanma}@dtic.ua.es

**Abstract.** Self-organizing neural networks try to preserve the topology of an input space by means of their competitive learning. This capacity has been used, among others, for the representation of objects and their motion. In addition, these applications usually have real-time constraints. In this work we have study a kind of self-organizing network, the Growing Neural Gas with different parameters, to represent different objects. In some cases, topology preservation is lost and, therefore, the quality of the representation. So, we have made a study to quantify topology preservation to establish the most suitable learning parameters, depending on the kind of objects to represent and the size of the network.

## 1 Introduction

Self-organizing neural networks, by means of a competitive learning, make an adaptation of the reference vectors of the neurons, as well as, of the interconnection network among them; obtaining a mapping that tries to preserve the topology of an input space. Besides, they are able of a continuous re-adaptation process even if new patterns are entered, with no need to reset the learning.

These capacities have been used for the representation of objects [1] (Figure 1) and their motion [2] by means of the Growing Neural Gas (GNG) [3] that has a learning process more flexible than other self-organizing models, like Kohonen maps [4].



**Fig. 1.** Representation of two-dimensional objects with a self-organizing network

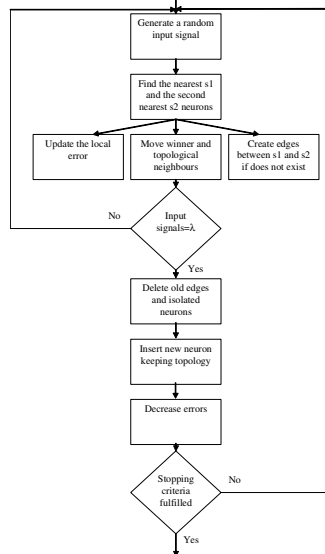
These two applications, representation of objects and their motion, have in many cases temporal constraints, reason why it's interesting the acceleration of the learning process. In computer vision applications the condition of finalization for the GNG

algorithm is commonly defined by the insertion of a predefined number of neurons. The election of this number can affect the quality of the adaptation, measured as the topology preservation of the input space [5].

In this work a study of the representation of two-dimensional objects with GNG has been made measuring the degree of topology preservation depending on the available neurons for their adaptation and the kind of shapes of the objects represented.

## 2 Growing Neural Gas

With GNG a growth process takes place from minimal network size and new units are inserted successively using a particular type of vector quantisation. In figure 2 a flowchart with the algorithm of the GNG is showed.



**Fig. 2.** Flowchart of the GNG algorithm

To determine where to insert new units, local error measures are gathered during the adaptation process and each new unit is inserted near the unit which has the highest accumulated error. At each adaptation step a connection between the winner and the second-nearest unit is created as dictated by the competitive hebbian learning algorithm. This is continued until an ending condition is fulfilled, as for example, a time limit or a predefined numbers of neurons inserted, defined by the evaluation of the optimal network topology based on the topographic product [6]. This measure is used to detect deviations between the dimensionalities of the network and that of the input space, detecting folds in the network and, indicating that is trying to approximate to an input manifold with different dimensions. In addition, in GNG networks learning parameters are constant in time, in contrast to other methods whose learning is based on decaying parameters.

### 3 Measures of Topology Preservation

The adaptation of a self-organizing neural network is made mainly from two points of view: its resolution and its topology preservation of an input space.

The measure of resolution most usually employed is the quantization error [4], expressed like:

$$E = \sum_{\forall \xi \in \mathbb{R}^d} \|w_{s_\xi} - \xi\| \cdot p(\xi) \tag{1}$$

where  $s_\xi$  is the nearest neuron to the input pattern  $\xi$ .

One of the most used measures to evaluate topology preservation is the topographic product [6] that compares the neighbourhood relationship among all pair of neurons of the network with concerning, on one hand to their position inside the map, and on the other hand, according to their reference vectors:

$$P = \frac{1}{\mathcal{N}(\mathcal{N}-1)} \sum_{j=1}^{\mathcal{N}} \sum_{\kappa=1}^{\mathcal{N}-1} \log \left( \left( \prod_{l=1}^{\kappa} \frac{d^{\mathcal{V}}(w_j, w_{n_l^{\mathcal{A}}(j)})}{d^{\mathcal{V}}(w_j, w_{n_l^{\mathcal{V}}(j)})} \cdot \frac{d^{\mathcal{A}}(j, n_l^{\mathcal{A}}(j))}{d^{\mathcal{A}}(j, n_l^{\mathcal{V}}(j))} \right)^{1/2\kappa} \right) \tag{2}$$

where  $j$  is a neuron,  $w_j$  is its reference vector,  $n_l^{\mathcal{V}}$  is the  $l$ -th closest neighbor to  $j$  in the input manifold  $\mathcal{V}$  according to a distance  $d^{\mathcal{V}}$  and  $n_l^{\mathcal{A}}$  is the  $l$ -th nearest neuron to  $j$  in the network  $\mathcal{A}$  according to a distance  $d^{\mathcal{A}}$ . In order to use this measure to non-linear input spaces the geodesic distance [7] is employed as  $d^{\mathcal{V}}$ .

A similar measure, the topographic function [8], compares the resulting neural network with the Delaunay triangulation induced by the input space, measuring the number of neurons that have adjacent receptive fields but are not connected and vice versa.

### 4 Comparative Study

In this section, first we have compared compare the GNG with Kohonen maps [4] and Neural Gas [9], next we have measured and compared topology preservation of the maps generated to represent different objects depending on two aspects: size and shape of the represented object.

The experiments have been made with different objects in the image. For the first aspect we have used an image of a hand with three different sizes: 395x500, 195x247 and 95x120 pixels. For the second one we have used images representing a circle, a ring, a square and a hand (Figure 1) with similar size.

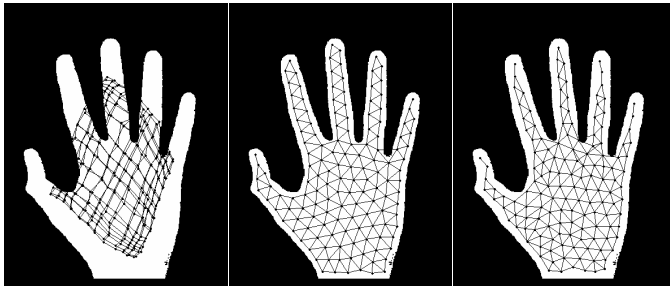
In the experiments we have measured the topology preservation of the maps inserting 10, 25, 49, 100 and 144 neurons. The parameters used to adapt the GNG to the shape of different objects have been fixed:  $\epsilon_1 = 0.1$ ,  $\epsilon_2 = 0.01$ ,  $\alpha = 0.5$ ,  $\beta = 0.0005$ ,  $a_{max} = 250$  and  $\lambda = 1000$ .



#### 4.1 Comparing GNG with Other Self-organizing Models

In this experiment we have used three different self-organizing models to adapt the shape of a hand with 144 neurons.

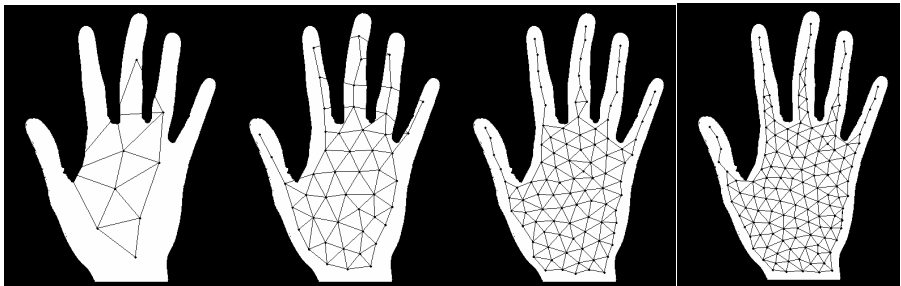
Figure 3 shows the adaptation process using Kohonen maps, NG and GNG models. The topology preservation of the Kohonen maps in comparison with GNG and NG is very poor because of the fixed structure of this model. On the contrary, with NG the topology preservation is well defined but the learning time is more than ten times higher than the time for GNG what is unacceptable for time restrictions of most of computer vision tasks.



**Fig. 3.** Net adaptation with Kohonen maps, NG and GNG algorithms

#### 4.2 Topology Preservation Depending on the Number of Neurons

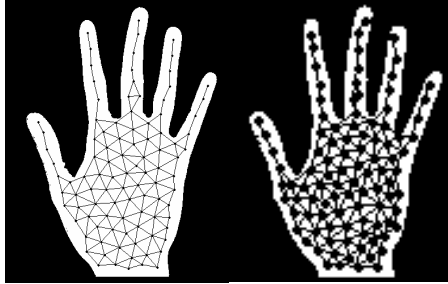
Figure 4 shows topology preservation for different variants of the same object image, depending on the number of neurons that the network finally has. Analysing the obtained results is evident that 10 neurons are not enough to map the object. On the other hand the mapping with 100 neurons is good enough and has no difference, in terms of topographic representation, with the map obtained inserting 144 neurons.



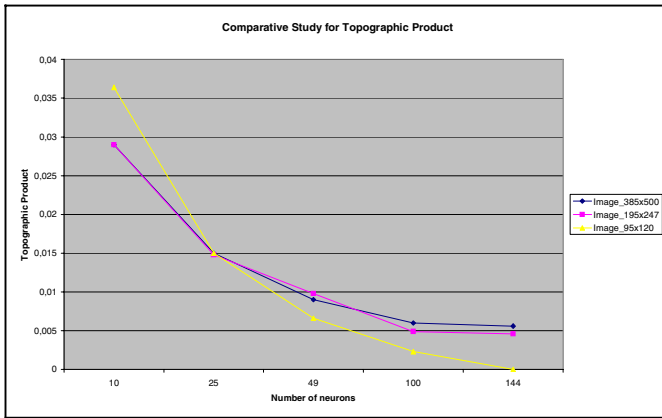
**Fig. 4.** Mapping for images of 385x500 pixels with nets of 10, 49, 100 and 144 neurons

In figure 5 is presented the adaptation with a net of 100 neurons to the same object (same relation of pixels belonging to the object and the background) but with different

size. While for the largest image the adaptation is correct and the topology is preserved, for the small image the number of neurons is excessive and it would have been enough to define the condition of finalization of the algorithm when 49 neurons or less would have been inserted.



**Fig. 5.** Net adaptation to images of 385x500 and 95x120 with a map of 100 neurons



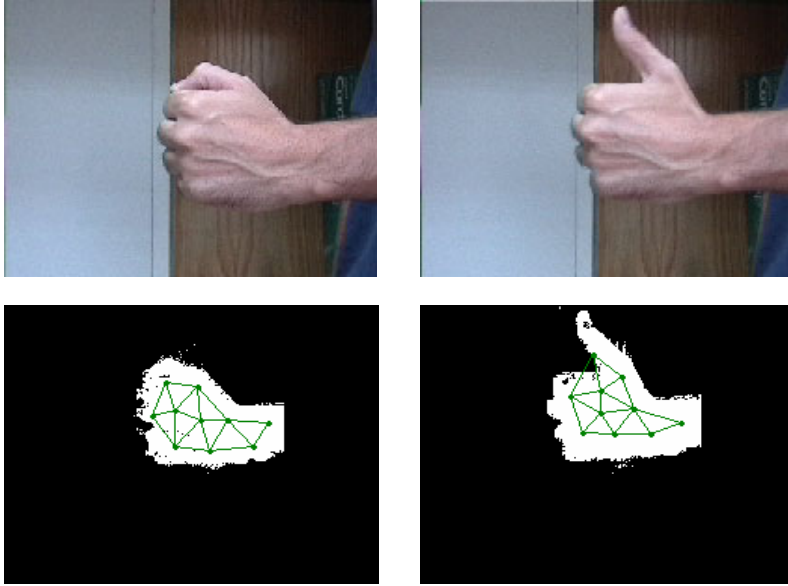
**Fig. 6.** Topographic product for images and network with different sizes

Figure 6 summarizes in a graph the results of the experiments. Topology preservation with 10 and 25 neurons is not good enough in all the cases. On the other hand topology preservation with 100 and 144 neurons is similar. We can extract from that results that with 49 or 100 neurons the topology preservation is good enough for all the images tested.

### 4.3 Topology Preservation Depending on the Objects Shape

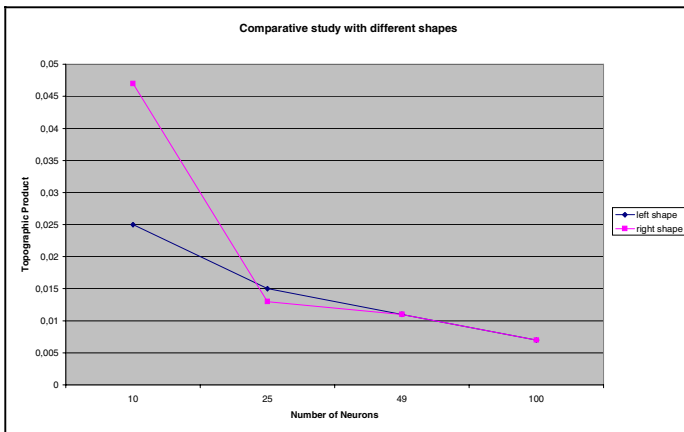
For this study, we have use a set of objects (Figure 1) and two images of a hand with the same size but with different shape, in the first one, on the left, the fist is closed and in the second, on the right, one finger is extended. Figure 7 shows both images and the topology preservation inserting 10 neurons.

Differences in topology preservation of both images are not too significant when measuring with the topographic product. Nevertheless, topology preservation is enough to distinguish the first gesture while is impossible to recognize the other.



**Fig. 7.** Images with different shapes and GNG adaptation with 10 neurons

Shape of objects represented in the images is an important factor for the election of the number of neurons used in the adaptation. In figure 8 is shown a graph with the topology preservation for both images represented with maps containing 10, 25, 49



**Fig. 8.** Comparative study for images with different shapes

and 100 neurons. In this study can be observed that there is a minimum necessary number of neurons not only because the bad topology preservation but because the recognition of the object is impossible. While for shape 1 (close fist) 10 neurons are enough meanwhile for shape 2 (extended finger) more neurons are needed.

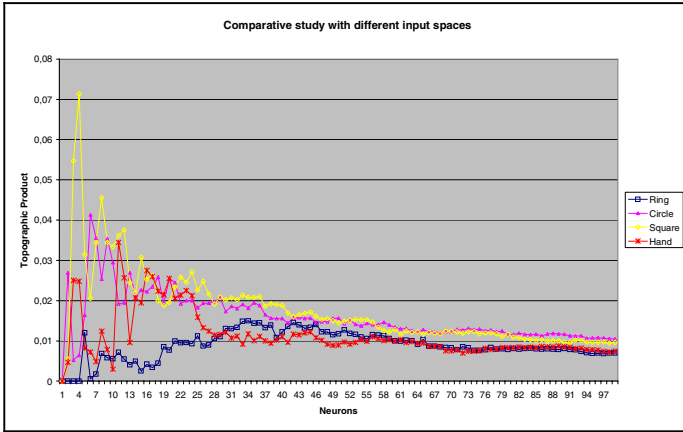


Fig. 9. Comparative study for images of different objects

In figure 9 we present a graph obtained from the study of topographic product for images representing different objects, showing the values of this measure for GNG with different number of neurons. From this graph some conclusions can be extracted; the unstable state in the first steps of the learning, the convergence of the topology preservation when a reduced number of neurons are inserted, and the no need of big networks to represent the objects shape with the consequent low computational cost.

## 5 Conclusions and Further Work

GNG has demonstrated to have better topology preservation than Kohonen maps and the learning algorithm is faster than other self-organizing methods what made this method interesting for vision tasks like represent objects with temporal constraints.

Topology preservation of the Growing Neural Gas is affected by the learning parameters and available time. Faster methods improve resolution but, in many cases, it deteriorates topology preservation, because the relation between number of neurons and input signals by iteration decreases. From the point of view of the implementation, as much software as hardware, it will be more interesting to obtain small networks, with worse resolution but good topology preservation, since calculation and storage requirements will be smaller.

At the moment, we are doing similar studies with other self-organizing models (Neural Gas [9], GWR [10]), studying their degree of topology preservation. We want to extract which are the characteristics of these networks that allow a suitable and fast representation of an input space, in order to the development of new self-organizing neural networks based on the combination of them.

Therefore, the application of these models to the representation of objects and their motion will be possible, adapting the different learning parameters depending on the available time and the quality required.

## References

1. Flórez, F., García, J.M., García, J., Hernández, A.: Representation of 2D Objects with a Topology Preserving Network. In Proceedings of the 2<sup>nd</sup> International Workshop on Pattern Recognition in Information Systems (PRIS'02), Alicante. ICEIS Press (2001) 267-276
2. Flórez, F., García, J.M., García, J., Hernández, A.: Hand Gesture Recognition Following the Dynamics of a Topology-Preserving Network. In Proc. of the 5<sup>th</sup> IEEE Intern. Conference on Automatic Face and Gesture Recognition, Washington, D.C. IEEE, Inc. (2001) 318-323
3. Fritzke, B.: A Growing Neural Gas Network Learns Topologies. In Advances in Neural Information Processing Systems 7, G. Tesauro, D.S. Touretzky y T.K. Leen (eds.), MIT Press (1995) 625-632
4. Kohonen, T.: Self-Organizing Maps. Springer-Verlag, Berlin Heidelberg (1995)
5. Martinetz, T., Schulten, K.: Topology Representing Networks. Neural Networks, 7(3) (1994) 507-522
6. Bauer, H.-U., Pawelzik, K.R.: Quantifying the Neighborhood Preservation of Self-Organizing Feature Maps. IEEE Transactions on Neural Networks, 3(4) (1992) 570-578
7. Flórez, F.; García, J.M.; García, J.; Hernández, A. Geodesic Topographic Product: An Improvement to Measure Topology Preservation of Self-Organizing Neural Networks. Lecture Notes in Artificial Intelligence, vol 3315 (2004) 841-850
8. Villmann, T., Der, R., Herrmann, M., Martinetz, T.M.: Topology Preservation in Self-Organizing Feature Maps: Exact Definition and Measurement. IEEE Transactions on Neural Networks, 8(2) (1997) 256-266
9. Martinetz, T, Schulten, K.: A "Neural-Gas" Network Learns Topologies. In Artificial Neural Networks, T. Kohonen, K. Mäkisara, O. Simula y J. Kangas (eds.) (1991) 1:397-402
10. Marsland, S., Shapiro, J.; Nehmzow, U.: A self-organising network that grows when required. Neural Networks, 15 (2002) 1041-1058

# Outlier Resistant PCA Ensembles

Bogdan Gabrys<sup>1</sup>, Bruno Baruque<sup>2</sup>, and Emilio Corchado<sup>2</sup>

<sup>1</sup>Computational Intelligence Research Group, Bournemouth University, United Kingdom  
bgabrys@bournemouth.ac.uk

<sup>2</sup>Department of Civil Engineering, University of Burgos, Spain  
escorchado@ubu.es, bbaruque@ubu.es

**Abstract.** Statistical re-sampling techniques have been used extensively and successfully in the machine learning approaches for generation of classifier and predictor ensembles. It has been frequently shown that combining so called unstable predictors has a stabilizing effect on and improves the performance of the prediction system generated in this way. In this paper we use the re-sampling techniques in the context of Principal Component Analysis (PCA). We show that the proposed PCA ensembles exhibit a much more robust behaviour in the presence of outliers which can seriously affect the performance of an individual PCA algorithm. The performance and characteristics of the proposed approaches are illustrated on a number of experimental studies where an individual PCA is compared to the introduced PCA ensemble.

## 1 Introduction

Projectionist methods are those based on the identification of "interesting" directions in terms of any one specific index or projection. Such indexes or projections are, for example, based on the identification of directions that account for the largest variance of a data set as in the Principal Component Analysis (PCA) method [1]-[2]. Having identified the interesting projections, the data is then projected onto a lower dimensional subspace in which it is possible to examine its structure visually, which normally involves plotting the projection in two or three dimensions. The remaining dimensions are discarded as they are mainly related to a very small percentage of the information or the data set structure. In that way, the structure identified through a multivariable data set may be easily analyzed with the naked eye. This visual analysis may be distorted by the presence of outliers [3, 4]. Outliers are observations that lie an abnormal distance from other values in a set of data. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. The presence of outliers can be caused by a number of different reasons and usually indicates faulty data, erroneous procedures, or areas where a certain theory might not be valid. In this study we analyse the use of statistical re-sampling theory [7,9,10,12] in generation of PCA ensembles as a way of reducing or removing the influence of outliers on the generated principal components as well as identifying outliers which in themselves could be very interesting for the data analyst. The ideas explored in this paper are similar to those that have been employed in generation of multiple classifier systems (classifier ensembles) [7-13] where the so called unstable classifiers (i.e. classifiers like decision trees or some neuro-fuzzy classifiers, the

performance of which can be significantly affected by the presence of outliers) have been stabilized through the use of classifier ensembles. It has been frequently observed that PCA is also very sensitive to the outliers and the principal directions found can be significantly affected by their presence which in turn can lead to much more difficult analysis of the projected data or wrong conclusions.

The proposed approach is based on voting and averaging with the principal directions selected from the multiple PCA runs on sub-samples of the data set. Firstly the most frequently occurring principal directions are identified and as they can be somewhat different a further stabilizing effect is achieved through the averaging of the relevant eigenvectors. The hypothesis related to the presence or absence of harmful significant outliers is tested through the analysis of the consistency of the generated principal directions and the relative spread of the percentages of the variance explained. The significant shift in the directions of the principal components and large variation of the explained variance by different principal components obtained from different subsets of the original data set are used as indicators of the presence of the possible outliers.

The remaining parts of this paper are organised as follows. Basic PCA algorithm is summarised in section 2. Statistical re-sampling techniques and PCA ensembles are discussed in section 3. This is followed by the experimental analysis and results in section 4. And finally, conclusions and future work are described in section 5.

## 2 Principal Component Analysis

PCA originated in work by Pearson (1901) [1], and independently by Hotelling (1933) [2] to describe the variation in a set of multivariate data in terms of a set of uncorrelated variables each of which is a linear combination of the original variables. Its goal is to derive new variables, in decreasing order of importance, that are linear combinations of the original variables and are uncorrelated with each other. PCA can be implemented by means of some connectionist models [5], [6].

The disadvantage of this technique, both employing statistical or connectionist models is that this process is accomplished in a global way. This means that every data point that is situated far from the majority of the other cases belonging to the dataset can influence the final result, as it introduces a high variance compared with the rest, although it could be very small in number and could be considered as anecdotic or dispensable case. Almost in every mid-size non-artificial dataset a number of these outlier cases appear, distorting its variance and hence hindering its analysis.

## 3 Statistical Re-sampling Techniques and PCA Ensembles

The technique utilised in this study to resist or detect the presence of outliers in a multidimensional dataset, is based on statistical re-sampling theory. One of the most widely known approaches utilizing statistical re-sampling techniques introduced by Breiman [7] is called "bootstrap aggregation" or "bagging".

In our case, the idea is to employ the bagging technique [7, 9] in combination with the PCA analysis in order to have more than one independent analysis performed over the same dataset. It is expected that, if any significant perturbation of the statistical

characteristics of the dataset is produced only by a few of its components it will be more evident in analysis of some data subsets than in others. Firstly, it is necessary to obtain different subsets of the dataset. This is achieved by randomly selecting several cases from the dataset and considering them as if they were a complete dataset. This process simulates the obtaining of several replications of the dataset we are working with. By doing this operation  $n$  times,  $n$  different datasets will be available, although they are really subsets of the main dataset. The next step consists of performing an individual PCA analysis on each one of the  $n$  subsets obtained by re-sampling the original one (Re-sampling PCA or Re-PCA). If the whole dataset does not include elements that alter drastically its statistical properties (i.e. in this case, its second statistical moment: the variance), the set of results obtained on the analysis of different subsets should be similar within a small margin. On the other hand, if few cases that alter these statistical properties are included in the main dataset, it is expected to generate different results in terms of directions of the principal components obtained. While re-sampling the data it is easy to imagine that one of those infrequent outlier data points can be included in a minority of the subsets, but will not be present in a majority of the other subsets. It can also be intuitively expected that the PCA performed on subsets containing outliers will be more influenced by the outliers if the ratio of the outliers to the number of other data points is high.

It is stated in [10] that bagging is especially recommended when applied to unstable algorithms or learning methods. As PCA can be considered as such an unstable algorithm an application of bagging for stabilizing of PCA in presence of outliers is one of the main premises of this investigation.

The description of the Re-PCA model proposed in this work can be summarized in the following two major steps:

*I. Re-sampling and Principal Components Calculation.* In this step first  $n$  subsets of the original data set are generated by re-sampling without replacement. This is followed by application of the standard PCA to each of the subsets. For further analysis the set of eigenvectors representing the directions of the first 3 principal directions and the percentages of variance explained by each of these principal components are recorded.

*II. Voting and averaging.* To perform voting and averaging of directions in order to obtain the final principal components the following steps are performed. A) For each of  $n$  subsets of eigenvectors we first identify the similar directions by performing pair wise similarity test by calculating the scalar product between the eigenvectors; B) All the vectors with their respective scalar products below certain threshold are then clustered together; C) The cluster with the largest number of the eigenvectors is selected and the sum of only these eigenvectors is calculated giving the final averaged direction for a respective principal component.

## 4 Experimental Analysis

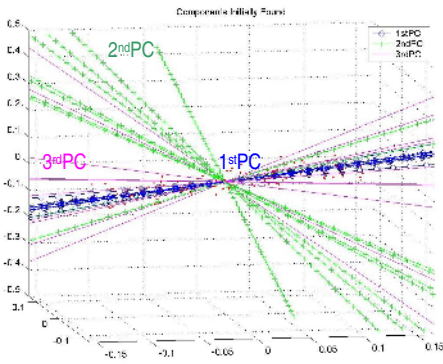
The artificial data set used in this series of experiments is made of one cloud of points and several points spread far above the main cloud of points which will be considered as *outliers*. The main cloud is an elongated cluster which moves within the axis delimited by the line defined by the points [1,1,0], [2,1.6,0], [3,2.2,0] and [4,2.8,0]. By employing this dataset we expect to obtain 3 clear principal components, as the



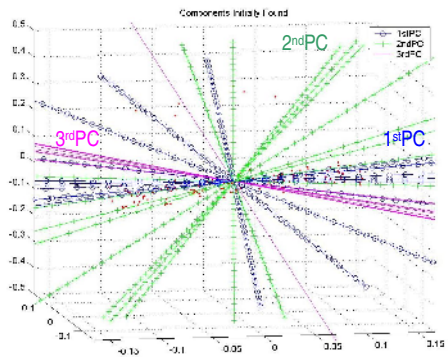
variance of each direction is different in comparison to the other two. The outlier points are spread over the same axis but displaced 5 units above in the vertical axis. There are 118 points in the main cluster and 8 outliers.

In order to test various characteristics of the proposed Re-PCA algorithm with regard to different proportions of outliers to the other data points and various sizes of the data sets, the experiments with 30, 50, 70 and 100 randomly selected points have been carried out. The experiments have been repeated 10 times for each one of those cases and the comparative analysis is presented below.

**Dataset 1.** In this set of experiments 10 subsets of 30 points randomly selected from the entire dataset (without replacement) are generated and PCA performed on each of them. Firstly, the method described above is applied to the dataset formed only by the main elongated data cluster (i.e. without the outliers). The results of PCA obtained from those 10 subsets are represented in Fig. 1.



**Fig. 1.** Projections of Re-PCA using 30 points (excluding outliers)



**Fig. 2.** Projections of Re-PCA using 30 points (including outliers)

Examining Fig. 1 it is easy to observe that the Re-PCA method has found almost the same direction for the first principal component, as it was expected. For the direction of the second and third principal components, they are clearly more dissimilar in the different tests, but still they all follow a consistent direction, except in one case.

The percentage of information (in form of the explained variance) that is represented by each one of the principal components is detailed in Table 1.

**Table 1.** Percentage of information captured by each of the principal components in the first part of the experiment (without outliers) including the maximum and minimum percentage of information (variance) from the analysed 10 subsets

Principal component	Percentage of information captured	
	Max	Min
First	72 %	68 %
Second	18 %	14 %
Third	14 %	11 %

Fig. 2 represents the results obtained performing exactly the same experiment but including now the 8 outliers in the sampled dataset. As it can be seen, the distribution of the directions corresponding to the principal components, produced when outliers are taken into account, are much more spread than in Fig. 1 (data without outliers). This means that the direction found in each case is rather dissimilar to the other corresponding ones. We can even consider that in 3 cases out of 10 (30 % of the cases), the method has found opposed directions for the first and second principal components. Looking at Fig 2, it can be seen that the first principal component appears in an almost horizontal direction on 7 occasions, while it appears in the diagonal that goes from the bottom right corner to the upper left one on the other 3 occasions. This three deviated directions will not be taken in account in the average calculation stage as the majority cluster consists of the 7 cases where the 1<sup>st</sup> principal component appears in the horizontal direction. The “percentage of information” that is represented by each of the principal components in this case is detailed in Table 2.

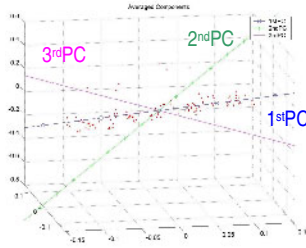
**Table 2.** Percentage of information captured by each one of the principal components in the second part of the experiment (including outliers) including the maximum and minimum percentage of information (variance) from the analysed 10 subsets

<i>Principal component</i>	<i>Percentage of information captured</i>	
	<i>Max</i>	<i>Min</i>
First	69 %	49 %
Second	41 %	17 %
Third	13 %	8 %

The results presented in Table 2 are quite different from the results obtained without including the outliers. Comparing both tables (Table 1 and Table 2), the influence of the presence or absence of outliers in a dataset in terms of the direction of the largest variance and relative difference between the Max and Min values for the principal components can be clearly seen.

The amount of information associated with the first principal component is different depending on whether outliers are included or not into the analysed data. The presence of these outliers makes the amount of information detected by the first component (Table 2) to be inferior to the situation without outliers (Table 1). In this case, due to the shape of the artificial dataset used and due to the fact that the outliers are situated in the direction associated with the second principal component, the amount of information represented by this second component (Table 2) is a lot higher than in the case that does not include outliers (Table1). As it was expected, the inclusion of the outliers brings a great instability to the dataset, making different individual PCAs behave in an inconsistent way and resulting in very different results where really the analysis is made over subsets of the same dataset. The use of PCA ensemble in cases like this is of particular use as the 70% of the cases where "the true" principal component is found represents the majority which is selected and then further stability is added through averaging of the eigenvectors from these 70% of majority similar principal directions. The final averaged directions are shown in Fig 3

**Dataset 2.** In this case the same experiment is performed for 50 points randomly selected from the entire dataset (without replacement). This is also performed 10 times.



**Fig. 3.** Resulting average for each of the principal components by voting between the directions shown in Fig 2 (excluding the 30% of the directions strongly influenced by the outliers)

We have noted that increasing the number of samples included in each of the subsets analysed by Re-PCA, brings stability to the performance of the experiments. The “fans” formed (for the case of 50 points data set) by the directions corresponding to the three principal components of the ten tests are far closer than the ones obtained in an analogous experiment including only 30 samples. The "percentage of information" that is represented by each of the principal components is shown in Table 3.

**Table 3.** Percentage of information captured by each of the principal components (selecting 50 points but excluding outliers) including the maximum and minimum percentage of information from the analysed 10 subsets

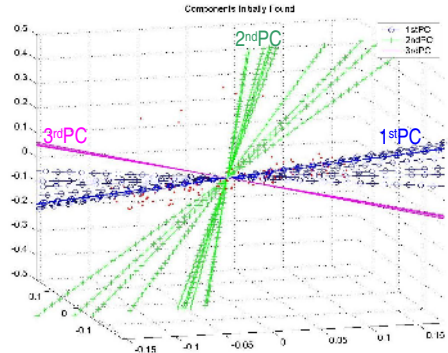
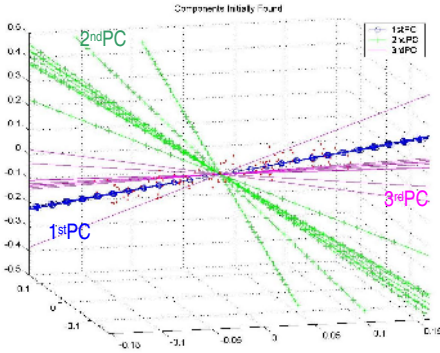
<i>Principal component</i>	<i>Percentage of information captured</i>	
	Max	Min
First	72 %	68 %
Second	16 %	14 %
Third	14 %	12 %

Performing the same operations but including now the 8 outliers, we have obtained the following. Although including the additional 20 data points has had a stabilizing effect on the individual PCAs, there are still two occasions out of ten (20% of the cases) where the first and second principal components appear in an almost perpendicular direction to the other eight occasions, indicating some instability which may be due to the presence of outliers in the dataset. The second principal component is always very unstable because all the outliers are situated in its direction. Table 4 shows the “percentage of information” for each of the principal components.

**Table 4.** Percentage of information captured by each of the principal components (selecting 50 points and including outliers) including the maximum and minimum percentage of information from the analysed 10 subsets

<i>Principal component</i>	<i>Percentage of information captured</i>	
	Max	Min
First	70 %	44 %
Second	44 %	15 %
Third	13 %	9 %

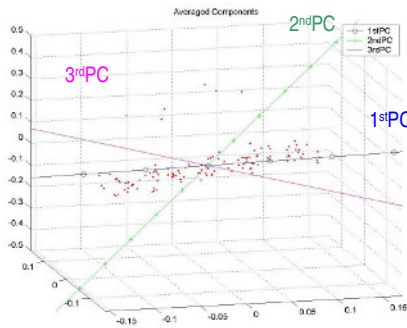
**Datasets 3 and 4.** To generate datasets 3 and 4, again 10 subsets of 70 and 100 points respectively have been used to test the stability of the PCA analysis performed over them. The results obtained for data set 4 are shown in Fig. 4 and Fig. 5.



**Fig. 4.** Projections of Re-PCA using 100 points (excluding outliers)

**Fig. 5.** Projections of Re-PCA using 100 points (including outliers)

As it can be seen in all the above experiments, the more samples are included into the analysis, the more stable behaviour of the individual PCA. Comparing Fig. 2, Fig. 5 (and the results obtained for data sets 2 and 3) gives a visual prove of that, as the directions found using 100 points are slightly more consistent than when using only 30, 50 or 70 points. It can also be seen (Fig. 4 and Fig. 5) that including outliers in the analysed dataset brings a substantial degree of instability, giving as a result more spread “fans” (less consistent results) or even completely different directions for its principal components.



**Fig. 6.** Resulting average for each of the principal components by voting and averaging of the directions shown in Fig5

Calculating the average directions (Fig. 6) as explained above, we can obtain approximately the same main directions for the three principal components, as when

we have only used 30 points for calculations. This can be considered as an empirical proof of the robustness of the proposed Re-PCA method.

## 5 Conclusions and Future Work

In this study we have applied a simple projectionist model (PCA) as a powerful technique to identify the existence of outliers in a dataset by using statistical re-sampling techniques in combination with voting and averaging.

We have observed that in absence of outliers, the re-sampling technique gives very similar Principal Components (PCs) as a result of a number of independent runs. The first principal component is the same almost in 100% of the cases. The second principal component could be different in a larger percentage of cases, due to our particular artificial dataset. However, when outliers are present in the dataset the situation is different. The smaller the number of points included in a subset, the bigger the difference in the response of the variance obtained due to a greater influence of the outliers in the subset. A higher ratio of the outliers to the normal points significantly affects the directions of maximum variance of the dataset and thus the directions of the principal components. The proposed Re-PCA algorithm has shown a very robust behaviour in presence of outliers consistently finding the right principal directions while the individual PCA was significantly affected. The use of re-sampling in the context of PCA has had an additional benefit by allowing analysing the variance and its differences from different runs which in itself proved to be a very useful tool for detection of the presence of outliers.

Future work will also investigate this and other neural and statistical methods, based on higher order statistics, on a larger range of data sets which have been impossible to include in this paper due to the space limitations.

## Acknowledgments

This research has been supported by the MCyT project TIN2004-07033 and the project BU008B05 of the JCyL.

## References

1. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2:559-572. (1901).
2. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417-441,498-520. (1933).
3. Cook, R. D. Detection of influential observations in linear regression. *Technometrics* 19, 15-18. (1977).
4. Dixon, W. J. Analysis of extreme values, *Ann. Math. Stat.*, 21, 488-506. (1950)
5. Oja, E. Neural networks, principal components and subspaces. *International Journal of Neural Systems* 1(1):61-68. (1989).
6. Sanger, D. Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. *Connection Science*, 1:115--138. (1989).

7. Breiman, L. Bagging predictors. *Machine Learning*, 24:123–140. (1996).
8. Schapire, R.E; Freund, Y; Bartlett, P. and Lee, W.S. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.
9. Gabrys, B. Combining neuro-fuzzy classifiers for improved generalisation and reliability. In *Proceedings the Int. Joint Conference on Neural Networks (IJCNN'2002) a part of the WCCI'2002 Congress*, pages 2410–2415, Honolulu, USA, 2002.
10. Kuncheva, L, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.
11. Ruta, D. and B.Gabrys, Classifier Selection for Majority Voting, Special issue of the journal of information fusion on Diversity in Multiple Classifier Systems, vol. 6, issue 1, pp. 63-81, 1 March 2005.
12. Gabrys, B., Learning Hybrid Neuro-Fuzzy Classifier Models From Data: To Combine or not to Combine?, *Fuzzy Sets and Systems*, vol. 147, pp. 39-56, 2004.
13. Ruta, D. and B. Gabrys, A Theoretical Analysis of the Limits of Majority Voting Errors for Multiple Classifier Systems, *Pattern Analysis and Applications*, vol. 5, pp. 333-350, 2002.

# Intelligent Channel Time Allocation in Simultaneously Operating Piconets Based on IEEE 802.15.3 MAC\*

Peng Xue, Peng Gong, and Duk Kyung Kim

Dept. Of Information and Communication Engineering,  
INHA University, Incheon, Korea  
yqssxp@hotmail.com, gongp1981@hotmail.com, kdk@inha.ac.kr  
<http://wireless.inha.ac.kr>

**Abstract.** The IEEE 802.15.3 Medium Access Control (MAC) protocol is designed to support low power and high rate Wireless Personal Area Networks (WPANs). It works on a Time Division Multiple Access (TDMA) basis within a piconet, which guarantees interference-free connections. However, the current protocol is not efficient for Simultaneously Operating Piconets (SOPs), which are linked by the parent/child (P/C) and parent/neighbor (P/N) relationships. It provides interference mitigation but the throughput is limited since the channel time is exclusively allocated. In this paper we propose Intelligent Channel Time Allocation (Intelligent CTA), which includes Public CTA and Normal CTA, to intelligently manage the inter-piconet interference (IPI). The proposed scheme is able to greatly reduce the IPI. Based on the DS-UWB system, the simulation results show that the proposed scheme can effectively support the coexistence of SOPs and it can achieve higher throughput without significant loss of link success probability (LSP) in the SOPs.

## 1 Introduction

Wireless Personal Area Networks (WPANs) are defined as networks that are formed by low power wireless devices with relatively short transmission distances (less than 10 meters). The IEEE 802.15.3a is working on the standard for high-speed WPAN with Ultra Wideband (UWB) as its physical layer. Direct Sequence UWB (DS-UWB) is one of the proposals for the IEEE 802.15.3a [2].

The IEEE 802.15.3 Medium Access Control (MAC) protocol has been developed for high-rate WPANs [1]. It works based on a piconet which allows a number of devices (DEVs) to communicate with each other. The piconet works based on the superframe, which includes the beacon, the contention access period (CAP) and the channel time allocation period (CTAP). There is no interference within single piconet because of the Time Division Multiple Access (TDMA) structure.

When Simultaneously Operating Piconets (SOPs) are in the reachable range of one another, the inter-piconet interference (IPI) can happen at any time. In 802.15.3, the SOPs with parent/child (P/C) and parent/neighbor (P/N) configurations provides interference mitigation but limits the throughput. The piconet coordination was proposed

---

\* This research was supported by University IT Research Center (INHA UWB-ITRC), Korea.

to support coexistence of SOPs and beacon alignment can effectively avoid beacon interference [4]. But the interference in the CTAP has still not been solved.

In this paper we propose the Intelligent Channel Time Allocation (Intelligent CTA) to manage the IPI in the CTAP. Combined with beacon alignment, the proposed scheme can greatly reduce the IPI. It significantly increases the throughput of SOPs compared with P/C and P/N configurations, especially in the small overlap cases. And the performance is also better than the scheme in [4]. The simulation results based on the DS-UWB system show that our scheme can effectively support the coexistence of SOPs with high throughput and high link success probability (LSP).

The paper is organized as follows. Section 2 gives an overview of 802.15.3 MAC protocol, SOPs problems and the DS-UWB proposal. Section 3 introduces the proposed scheme and in Section 4 we evaluate the performance of the proposed scheme by means of simulations. Finally in Section 5 we draw our conclusions.

## 2 Background

In this section, we present relevant background information on IEEE 802.15.3 MAC protocol, SOPs problems and some previous works. And then, we give an introduction of the DS-UWB system and the support of SOPs.

### 2.1 IEEE 802.15.3 MAC Protocol

The IEEE 802.15.3 MAC mainly works within a piconet which allows a number of independent devices (DEVs) to communicate with each other in a short range. One DEV is required to be the piconet coordinator (PNC). The PNC provides the basic timing for the piconet based on the superframe. As shown in Fig. 1, it contains beacon, contention access period (CAP) and channel time allocation period (CTAP).

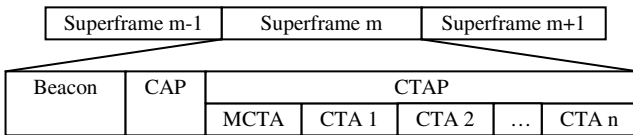


Fig. 1. IEEE 802.15.3 Superframe

At the beginning of the superframe, PNC broadcasts beacon which provides the timing information for the piconet. The CAP is used for commands with the access method of Carrier Sensing Multiple Access with Collision Avoidance (CSMA/CA). The CTAP is composed of channel time allocations (CTAs) and management CTAs (MCTAs). CTAs are used for isochronous and asynchronous data communications. MCTAs are used for communications between DEV and PNC. Channel access in the CTAP is based on TDMA. If one DEV (source DEV) wants to communicate with another DEV (destination DEV), it sends Channel Time Request (CTRq) to PNC during the CAP, which is identified by a <source DEV, destination DEV> link. PNC allocate different CTAs for different links. All the CTA information, including the



start time and duration, is broadcast in the beacon. Following the beacon information, every link works during its own CTA without any interference.

## 2.2 Problems of Simultaneously Operating Piconets

For SOPs, if one piconet is in the reachable range of another one, the transmission quality can be affected because the IPI can happen at any time. Beacon interference can affect the association of DEVs. During the CAP, the collision avoidance rules of CSMA will ensure eventual transmission but with delay. The interference in the CTAP also results in faulty transmissions.

In 802.15.3, if a PNC in one piconet detects the presence of another piconet, it can associate with that piconet by a P/C or P/N relationship. Parent piconet is the piconet with more than one dependent piconet (child or neighbor). The child piconet is used for extending the range while the neighbor piconet is for sharing the same frequency spectrum between different piconets. Each of them exists entirely within a private CTA of the parent piconet. A private CTA is a reserved channel time used for a dependent piconet. The child and neighbor piconets work based on the child and neighbor superframes, which also contain beacon, CAP and CTAP. During their superframes, there are no transmissions in the parent piconet. This provides interference avoidance but the throughput is limited.

The piconet coordination is proposed to support the coexistence of SOPs [4]. An intermediate DEV exchanges beacon and management information for coordination in SOPs with a heartbeat signal. The heartbeat uses Application Specific Information Element (ASIE) to relay information [4]. Beacon alignment is efficient to keep beacon from collisions. But the interference during the CTAP is still not solved.

## 2.3 IEEE 802.15.3a DS-UWB

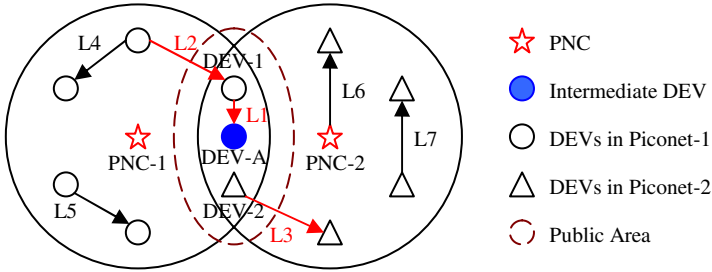
DS-UWB is one of the proposals for the physical layer for IEEE 802.15.3a [2]. It is based on the use of high-rate coded UWB pulses to provide scalable performance. Similar with the conventional DS spread spectrum systems, the spreading codes are used to spread the data bit into multiple chips. The proposal supports data rates of 28, 55, 110, 220, 660 and 1320 Mbps. The nominal chip rate is 1320 Mbps and the spreading codes vary from  $L=24$  (28 Mbps) to  $L=1$  (1320 Mbps).

The DS-UWB proposal provides support for SOPs. The proposal defines two frequency bands for piconet operation: a low band from 3.1 to 4.85 GHz and a high band from 6.2 to 9.7 GHz. Within each band the spread spectrum technique is used to support six piconets with offset chipping rates and separate codes. The IPI still exists because the spreading codes are not ideally orthogonal and the near-far effect is another problem in the DS-UWB system [3]. If SOPs are overlapped, the links of one piconet can be affected by the simultaneous links in the nearby piconets if there is no coordination to avoid interferences, and the performance of the system degrades.

## 3 The Proposed Intelligent Channel Time Allocation Scheme

We consider two SOPs apart far each other and there is no interference between them. Each piconet has a throughput of  $R$ . Since WPAN supports mobility, two piconets

may approach and their coverage areas are partially overlapped as in Fig. 2. The conventional method is the P/C or P/N configuration. So the same superframe is shared, which implies that the overall throughput becomes half, i.e.,  $R$  instead of  $2R$ .



**Fig. 2.** Two Simultaneously Operating Piconets

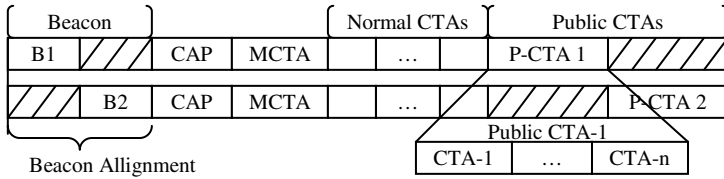
The proposed Intelligent CTA scheme introduces a concept of Public Area, where the DEV can hear more than one beacons. The DEV located in the Public Area is called a Public DEV. If one Public DEV is transmitting or receiving data in a CTA, the transmission can affect the simultaneous one or be affected by that one in another piconet. For example the DEV-1 of the Piconet-1 in the Fig. 2, when it transmits data in the link1 (L1), can make interference in the Piconet-2. When it receives data in the link2 (L2), it can be affected by the active link in the Piconet-2, for example L3. So if the Public DEV wants to transmit or receive data correctly, simultaneous links are not allowed. While the links out of the Public Area can be simultaneously operated without influential interference, for example L4, L5, L6, and L7. Provided that we allocate some special CTAs for all the Public DEVs, and during these times only one link is active, the interference can be avoided. The special CTAs are called Public CTAs, which provide exclusive transmissions. Every piconet has its own Public CTA for its Public DEVs. The CTAs for the links out of the Public Area are called Normal CTAs which allow concurrent transmissions. The combination of Public CTA and Normal CTA is called Intelligent CTA.

Now we give the coordination procedure based on the SOPs in Fig. 2.

1. Provided DEV-A belongs to Piconet-1 and it first hears the beacon from PNC-2. DEV-A can act as an intermediate DEV and send a heartbeat signal to PNC-1 [4].
2. PNC-1 adjusts the superframe duration appropriately, makes beacon alignment, allocates Public CTAs and keeps a MCTA for the intermediate DEV to relay message.
3. Beacon alignment is performed in the same way as in [4]. PNC-1 reserves two Public CTAs; one for itself and another for Piconet-2. And the start time and duration of each Public CTA are determined by PNC-1.
4. PNC-1 sends the beacon with the information of the coordinated superframe.
5. The intermediate DEV copies the beacon information and sends heartbeat.
6. PNC-2 follows the information in the heartbeat to achieve the coordination.

A coordinated superframe is illustrated in Fig. 3. After the beacon is broadcast in the piconet, every DEV scans the channel and the Public DEV should report its status

(Public) to PNC. Every PNC has a list of Public DEVs in its piconet. Any link related to the Public DEVs should be allocated in the Public CTA of its piconet. When the Public DEV sends Channel Time Request (CTRq), it should set the “Public” in the CTRq control frame, as in Fig. 4, to request a Public CTA. If the source DEV has no information about the destination DEV, the PNC should determine the link status when it receives CTRq from the source DEV. The PNC intelligently allocates Public CTA for this link if the destination DEV is a Public DEV. Any DEV which is not Public DEV but finds the channel status is not good can also request to communicate in Public CTA.



**Fig. 3.** Coordinated Superframe with Intelligent Channel Time Allocation

Bits: b7	b6	b5	b4	b3	b2-b0
Target ID list type	CTA rate type	CTA type	CTRq type	Reserved (Public)	Priority

**Fig. 4.** CTRq Control Format modified for support of Public CTA

When the SOPs are fully overlapped, that is, all the area is the Public Area. All the links in the SOPs work in the Public CTA and the throughput is the same with the P/C and P/N piconets since one superframe is shared. When the SOPs are partially overlapped, the exclusive transmissions in the Public CTAs avoid the serious interference and the concurrent links are allowed with slight interference in the Normal CTAs. This translates into an increase in throughput.

The PNC should intelligently change the Public CTA duration according to the number of the Public CTA requests. The PNC can send command to the intermediate DEV to increase or decrease its Public CTA duration and the information is relayed to other PNCs during the MCTA. After the information exchange, the newly allocated Public CTA should be operated from the next superframe. At the next superframe, PNC broadcasts the new beacon to announce the updated CTA information.

## 4 Simulations and Results

We carried out a campaign of simulations to analyze the performance of the proposed scheme based on the DS-UWB system. In this section we describe the simulation model and summarize the results of both the link level simulations (LLS) and the system level simulations (SLS).

#### 4.1 Link Level Simulations

This part describes the LLS results of the DS-UWB system. The transceiver is based on the current DS-UWB proposal [2]. And the receiver is assumed to know perfect channel information and maintain perfect timing and frequency synchronization. The simulation parameters are summarized in Table 1 [5].

It is required that the error rate criterion shall be a packet error ratio (PER) of less than 8% with a frame body length of 1024 octets [2]. Table 2 gives the required SNR value and the reachable distance at 8% PER level in UWB Channel Model (CM) 1.

**Table 1.** Link Level Simulation Parameters

Packet size	1024 bytes
Modulation	BPSK
Channel coding	Convolutional coding (rate=1/2)
Spreading code length	24, 12, 6, 3
Data rate (Mbps)	28, 55, 110, 220
Transmit Power (dBm)	-10.0
Channel Model	Path Loss, UWB Fading Channel [5]
Noise power (dBm)	-174+10lg(Bit Rate)
Noise figure (dB)	6.6
Rake receiver	16 fingers with MRC

**Table 2.** Link Level Simulation Results

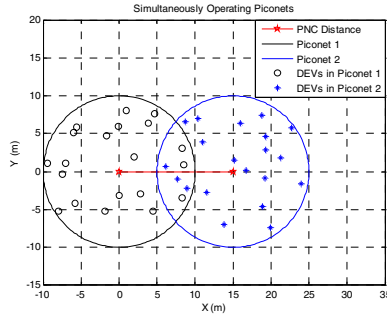
Data Rate (Mbps)	220	110	55	28
SNR (8% PER) in CM1 (dB)	11.6	6.1	4.9	4.5
Distance (8% PER) in CM1 (m)	6.0	12.7	20.0	29.1

#### 4.2 System Level Simulations

The SOPs model is illustrated in Fig 5. Two SOPs are considered for simplicity. The piconet range is 10 meters, that is, the DEV can associate with the piconet when the distance to PNC is less than 10 meters. Each piconet has 20 DEVs and all DEVs are homogeneous. The piconet distance is defined as the PNC distance (D).

In 802.15.3 MAC, any DEV is able to request information about the quality of the link between itself and another DEV with the Channel Status Request command. We consider the destination DEV can feedback the required SNR value (8% PER level) to the source DEV and the highest rate satisfying the error rate criterion is selected. In our simulations, Gaussian approximation is used for the IPI [6].

Table 3 summarizes the SLS parameters. We consider three different scenarios. The SOPs with P/C or P/N configuration are considered in the first scenario. The SOPs with coordination in [4] exist in the second scenario. Finally, the SOPs with the proposed scheme are considered.



**Fig. 5.** SOPs Model

**Table 3.** System Level Simulation Parameters

Piconet channels	1-4 (Low band)
Number of SOPs	2
PNC Distance	0-20m
Superframe time	45 ms
Beacon time	1 ms
CAP time	4 ms
CTAP time	40 ms
MCTA time	0.5 ms
CTA size	1 ms
ASIE time	0.2 ms

The measured performance metrics are the throughput and the link success probability (LSP). The throughput is defined as the total number of the information bits of the successfully received packets in a given time period. The LSP is the percentage of the successful links among the total links, which has been allocated CTAs. A successful link means the link in which all packets can be received with a PER value less than 8%. These two measures give a description about the quantity and the quality of the offered traffic in a network. Our aim is to maximize the throughput of the network without significant loss of LSP.

Fig. 6 compares the throughputs and LSPs in three scenarios with CM 1. In scenario 1 the throughput has a constant value and the LSP is 100% since the piconets work without any interference. When the SOPs are fully overlapped ( $D=0$ ), the throughput in scenario 3 nearly equals to that of scenario 1 and the LSP is also 100%. However, the throughput in scenario 2 is only 71.8% of that of scenario 1 and its LSP is only 60.6%. When the SOPs are mostly overlapped ( $0 < D < 4$ ), the throughput in scenario 3 increases slowly because almost all the links work in the Public CTA. When the PNC distance becomes farther, the throughput increases faster since more concurrent links are allowed. The throughput in scenario 3 is always better than that of scenario 2, till to the overlap boundary ( $D=20$ ) where the coordination of SOPs is not available. And when the proposed scheme is applied, the lowest LSP is 95.3% at the boundary of overlap. That means the proposed scheme can guarantee almost all the links to be successful.

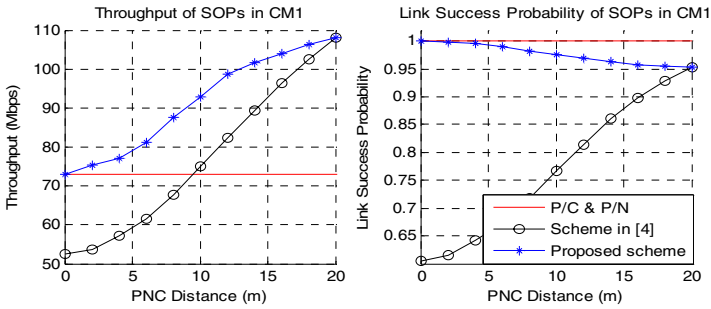


Fig. 6. Comparisons of Throughputs and Link Success Probabilities in three scenarios

## 5 Conclusions

In this paper, we have improved the SOPs coordination scheme with newly-introduced Intelligent CTA in addition to beacon alignment to resolve the IPI problem in IEEE 802.15.3 WPANs. From the simulation results with 2 SOPs based on the DS-UWB system, it is found that the throughput can be significantly increased compared with the P/C and P/N configurations in the current 802.15.3 protocol and even compared with the scheme in [4]. When the SOPs overlap perfectly, the throughput is the same with the P/C and P/N configurations and increases by nearly half compared with the scheme in [4]. The throughput gain increases as the PNC distance increases. Compared with P/C and P/N configurations, the throughput increases by 43.8% with 10% overlap and the LSP has no significant loss. The proposed scheme can be also applied when more than two SOPs are considered.

## References

1. “Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High Rate Wireless Personal Area Networks (WPANs)”, IEEE Standard 802.15.3, 2003
2. R. Fisher, R. Kohno, M. M. Laughlin and M. Welborn, “DS-UWB Physical Layer Submission to 802.15 Task Group 3a”, IEEE 802.15-04/0137r4, January, 2005
3. Matt Welborn, “Multi-User Support in UWB Communication System Designs”, IEEE 802.15-03216r1, May, 2003
4. “Mesh Dynamics and Advanced Cybernetics Group-Dynamic Beacon Alignment”, Available at: <http://www.meshdynamics.com/Publications/MDPBACONALIGNMENT.pdf>
5. S. S. Ghassemzadeh and L. Greanstein, “Parameter assumptions for the simulation of the proposed 802.15.3a PHYs”, DCN# 15-04-0488-00-003a, September, 2004
6. Bo Hu and N. C. Beaulieu, “Accurate Performance Evaluation of Time-Hopping and Direct-Sequence UWB Systems in Multi-User Interference”, *IEEE Transactions on Communications*, VOL. 53, No. 6, pp. 1053-1062, June, 2005

# An Intelligent Synchronization Scheme for WLAN Convergence Systems\*

Sekchin Chang<sup>1</sup> and Jaehak Chung<sup>2</sup>

<sup>1</sup> Dept. of Electrical and Computer Engineering, University of Seoul, Seoul, Korea  
<sup>2</sup> Dept. of Electronics Engineering, Inha University, Incheon, Korea

**Abstract.** Various WLAN standards have been converged into one wireless terminal to support the data services in multi-mode and multi-band environments. In this paper, an intelligent synchronization scheme is presented for the efficient use of WLAN convergence systems. The approach consists of intelligent protocol identification and intelligent frequency synchronization. Simulation results indicate that using the algorithm, the synchronization performance can significantly be improved in WLAN convergence systems.

## 1 Introduction

A lot of attention has recently been paid to wireless convergence systems. Especially, several WLAN standards have been converged into one wireless terminal such as IEEE802.11a/b/g WLAN convergence system [1,2]. The aim of WLAN convergence systems is to effectively support data services in multi-band and multi-mode environments. Usually, accurate synchronization is very crucial to guarantee the aim of WLAN convergence systems. Therefore, an intelligent synchronization scheme is presented in this paper to enhance the synchronization performance in the IEEE802.11a/b/g WLAN convergence system. The intelligent approach consists of intelligent protocol identification and intelligent frequency synchronization. The intelligent protocol identification performs band selection and mode selection to identify the service protocol in the service area. In this paper, the intelligent frequency synchronization scheme is especially proposed for orthogonal frequency division multiplexing (OFDM)-based WLAN in the convergence system since more accurate frequency-offset estimation is required to achieve better receiver performance in OFDM systems [3]. The 802.11a/g standards [4,6] specifies a preamble usage for frequency-offset estimation. However, the usage inefficiently utilizes the 802.11a/g preamble, which results in sub-optimal performance in the frequency-offset estimation. The proposed scheme intelligently exploits the features of the 802.11a/g preamble. Simulations show that the intelligent synchronization scheme yields considerably better synchronization performance than existing schemes. In addition, the simulation reveals that the packet error rate (PER) performance of the OFDM-based WLAN system has significantly been enhanced using the proposed intelligent scheme.

\* This work was supported by Smart (Ubiquitous) City Consortium under Seoul R&BD Program.

## 2 The Intelligent Synchronization Scheme

As stated earlier, the suggested intelligent synchronization approach consists of intelligent protocol identification and intelligent frequency synchronization.

### 2.1 The Intelligent Protocol Identification

Most WLAN convergence systems just utilize signal detection for protocol identification. However, the proposed intelligent scheme utilizes flag detection in addition to signal detection for more accurate protocol identification. In the 802.11a/b/g WLAN convergence system, the suggested approach intelligently utilizes the preambles [4,5,6] defined for IEEE802.11a, 11b, and 11g protocols. Fig. 1 illustrates the preambles. As shown in Fig. 1(a), the pre-defined data of SYNC field is utilized for signal detection and time synchronization in the 802.11b. In Fig. 1(a), SFD stands for start frame delimiter. In addition, in the case of IEEE802.11a/g, 10 identical short symbols (SS) are utilized for signal detection and time synchronization as shown in Fig. 1(b). In Fig. 1(b), GI and LS represent guard interval and long symbol, respectively. As the 1<sup>st</sup> step of the intelligent scheme, the scheme performs band selection since IEEE802.11a and IEEE802.11b/g operate in 5 GHz and 2.4 GHz band, respectively. The band selection intelligently selects the active band according to signal detection and flag detection as illustrated in Fig. 2. In the band selection, the GI is used as the flag. Since two modes of 11b and 11g exist on 2.4 GHz band, the band is initially selected as 5 GHz to simplify the procedure of the band selection. If the time elapsed for signal detection or GI detection exceeds 8 us (the time length of 10 SS's), the band is switched to 2.4 GHz, which enables mode selection. The mode selection intelligently selects the active mode according to signal detection and flag detection as shown in Fig. 3. In the mode selection, the GI and the SFD are used as the additional flags. Since the time length of 10 SS's for 11g is

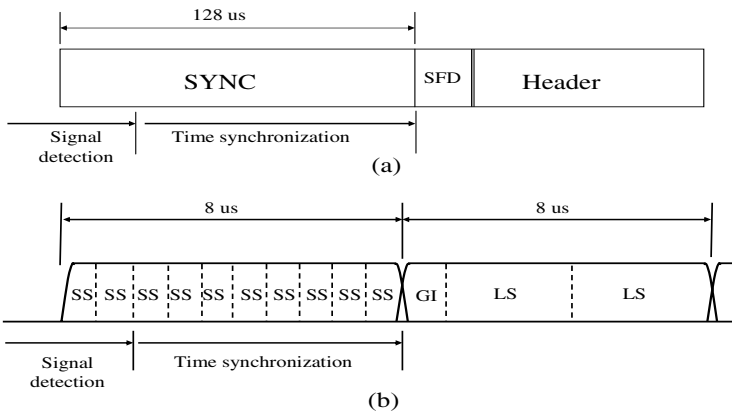


Fig. 1. The preambles for (a) IEEE802.11b and (b) IEEE802.11a/g



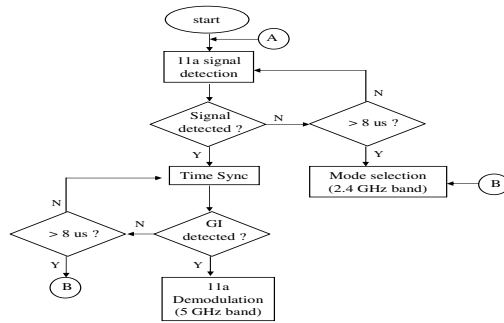


Fig. 2. The band selection in the intelligent protocol identification

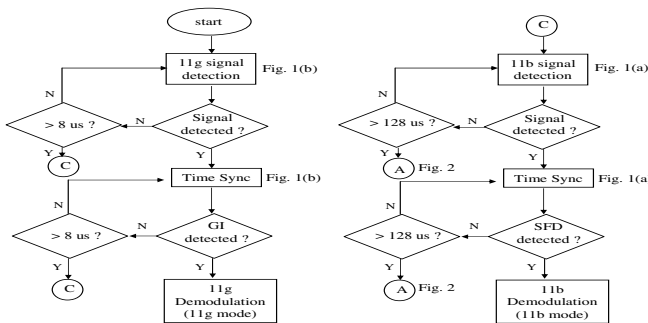
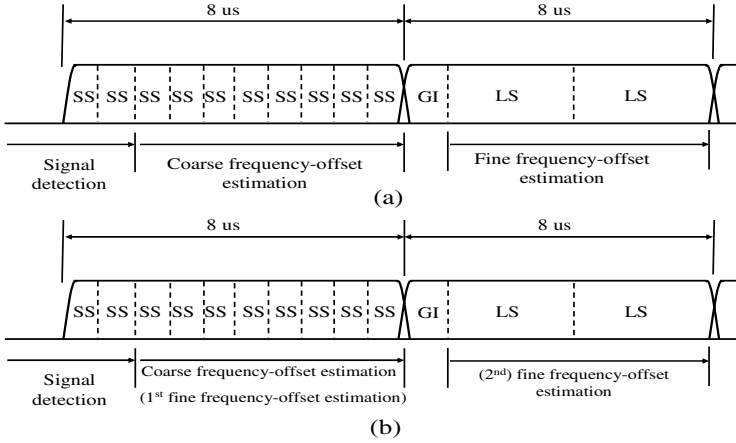


Fig. 3. The mode selection in the intelligent protocol identification

shorter than that of 11b SYNC in Fig. 1, the initial mode is set to 11g in Fig. 3 to simplify the procedure of the mode selection. If the time elapsed for signal detection or GI detection exceeds 8 us, the mode is switched to 11b. If the time elapsed for signal detection or SFD detection exceeds 128 us (the time length of 11b SYNC) in the 11b mode, the band is switched back to 5 GHz, which again enables the detection of 11a signal as shown in Figs. 2 and 3.

### 2.2 The Intelligent Frequency Synchronization

The proposed scheme intelligently utilizes the 802.11a/g preamble to enhance the performance of frequency-offset estimation for OFDM-based WLAN in the WLAN convergence system. Fig. 4(a) and (b) illustrate the conventional scheme and the proposed scheme, respectively for frequency synchronization. In Fig. 4(a), the conventional scheme exploits the SS's and the 2 identical LS's for coarse and fine frequency-offset estimation, respectively. On the other hand, the proposed scheme can intelligently perform the 1<sup>st</sup> fine estimation and the 2<sup>nd</sup> fine estimation on SS's and LS's, respectively as shown in Fig. 4(b). In the proposed scheme, the inclusion of the 1<sup>st</sup> fine estimation depends on the number of the remaining SS's after the signal detection. Usually, the valid signal of 802.11a/g



**Fig. 4.** The frequency-offset estimation on 11a/g preamble using (a) conventional scheme and (b) proposed scheme

is detected at some arbitrary sample point on the 10 SS's duration. After the signal detection, if the number of the available SS's is more than or equal to 8 whose time length is equal to that of the 2 LS's, the 1<sup>st</sup> fine estimation and the coarse estimation can simultaneously be performed on the SS's before the 2<sup>nd</sup> fine estimation on the LS's. Otherwise, only the 2<sup>nd</sup> fine estimation and the coarse estimation are performed on the LS's and the SS's, respectively, which is the same as the conventional scheme of Fig. 4(a). The number of the remaining SS's can be determined during the time synchronization of Fig. 1(b) since the time synchronization usually estimates the SS boundary. The proposed scheme produces more accurate frequency-offset estimate since the fine estimation can intelligently be made twice according to the number of the remaining SS's after signal detection. For optimal estimation when two fine-estimation procedures are available, the proposed scheme generates a fine frequency-offset estimate which minimizes the following cost function:

$$f(\varepsilon, w_1, w_2) = w_1 \cdot f_1(\varepsilon) + w_2 \cdot f_2(\varepsilon) \tag{1}$$

where  $w_k$  denotes an optimal weighting coefficient for the  $k^{th}$  fine estimation. In (1),  $f_k(\varepsilon)$  is expressed as

$$f_k(\varepsilon) = \sum_{n=N_{s,f}}^{N_{e,f}} \|r_k(n + 16 \cdot N_f) - r_k(n)e^{j\frac{2\pi \cdot 16 \cdot N_f \varepsilon}{N}}\|^2 \tag{2}$$

where  $N_{s,f}$  and  $N_{e,f}$  indicate the sample index of starting-point and ending-point for the fine estimation, respectively, and  $N_f$  is set to 4 for the fine estimation. In (2),  $r_k(n)$  is the received signal during the  $k^{th}$  fine estimation, and is formulated as

$$r_k(n) = \sum_{l=0}^{N_h} h_{l,k}s(n-l)e^{j2\pi n\varepsilon/N} + \zeta(n), k = 1, 2 \tag{3}$$

where  $h_{l,k}$  and  $s(n)$  denote the discrete-time complex channel impulse response during the  $k^{th}$  fine estimation and transmitted signal, respectively,  $N_h$  is the channel length,  $N$  is the IFFT size,  $\varepsilon$  is the frequency-offset normalized by sub-carrier spacing, and  $\zeta(n)$  indicates additive white Gaussian noise (AWGN). In (1), the weighting coefficients,  $w_1$  and  $w_2$  are determined to maximize the signal-to-noise ratio (SNR) of  $f(\varepsilon, w_1, w_2)$ . Using Cauchy's inequality, the SNR of  $f(\varepsilon, w_1, w_2)$  can be expressed as

$$\frac{|\sum_{k=1}^2 w_k \sigma_{s,k}^2|^2}{(\sigma_n^2)^2 \sum_{k=1}^2 |w_k|^2} \leq \frac{\sum_{k=1}^2 \sigma_{s,k}^4}{\sigma_n^4} \quad (4)$$

where  $\sigma_{s,k}^2$  and  $\sigma_n^2$  denote signal power during the  $k^{th}$  fine estimation and noise power, respectively. Using the condition for the equality in (4),  $w_k$  is determined as

$$w_k = \frac{\sigma_{s,k}^2}{\sigma_n^2} = SNR_k \quad (5)$$

From the minimization of (1), the frequency-offset estimate  $\hat{\varepsilon}$  can directly be calculated as follows

$$\hat{\varepsilon} = \frac{N}{2\pi \cdot 16 \cdot N_f} \arg\left(\sum_{k=1}^2 w_k G_{f,k}\right) \quad (6)$$

In (6),  $G_{f,k}$  is expressed as

$$G_{f,k} = \sum_{n=N_{s,f}}^{N_{e,f}} r_k(n + 16 \cdot N_f) r_k^*(n) \quad (7)$$

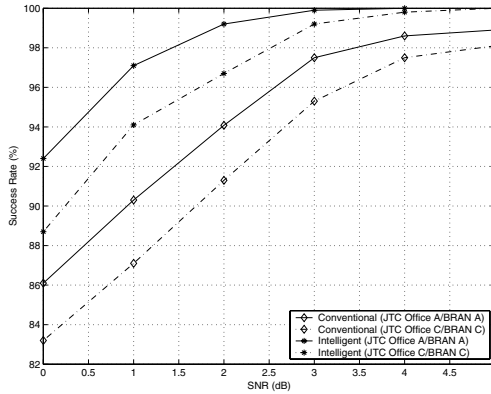
As in [3], the variance of (6) is found as follows

$$\text{var}(\hat{\varepsilon}) = \left(\frac{N}{2\pi \cdot 16 \cdot N_f}\right)^2 \cdot \frac{1}{2(N_{e,f} - N_{s,f} + 1) \cdot \text{SNR}_{\text{ave}}} \quad (8)$$

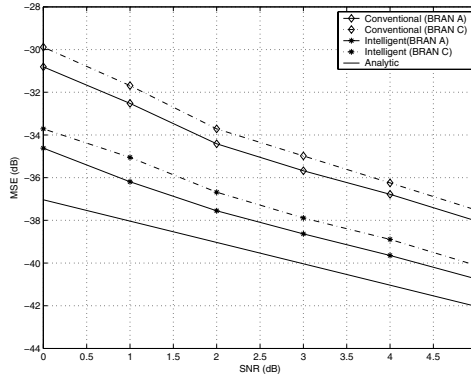
In (8),  $\text{SNR}_{\text{ave}} = (\text{SNR}_1 + \text{SNR}_2)/2$  where  $\text{SNR}_1$  and  $\text{SNR}_2$  denote SNR during the 1<sup>st</sup> and the 2<sup>nd</sup> fine estimation, respectively.

### 3 Simulation Result

Simulation results exhibit the effectiveness of the proposed intelligent scheme. BRAN channels [7] and JTC indoor channels [8] are considered fading channels for 5 GHz 802.11a and 2.4 GHz 802.11b/g systems, respectively. For the performance evaluation of the protocol identification, the success rates are given in Fig. 5. In this simulation, the WLAN convergence system operates in 2.4 GHz 802.11b mode. In other words, when the band selection and the mode selection choose 2.4 GHz and 802.11b mode, respectively, the protocol identification is successful. Note that the interfering signals whose carrier-to-interference (C/I)



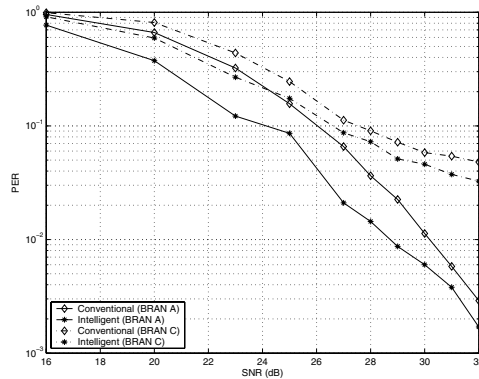
**Fig. 5.** Success rate versus SNR in the protocol identification ( $C/I = 0$  dB at 2.4 GHz and 5 GHz)



**Fig. 6.** MSE versus SNR in the frequency synchronization (802.11a WLAN)

ratio is 0 dB are used at 2.4 GHz and 5 GHz bands in the simulation. Compared to the conventional scheme, the intelligent approach achieves gains of about 1.8 dB and about 1.5 dB at the success rate of 96% in JTC office A/BRAN A and JTC office C/BRAN C, respectively in Fig. 5.

For the performance evaluation of the frequency synchronization in the OFDM-based WLAN, the mean squared error (MSE) values between exact and estimated frequency-offset are given in Fig. 6. In this simulation, the 802.11a WLAN is used as the OFDM-based WLAN since 5 GHz fading channel usually exhibits more deleterious effects on the performance. For this simulation,  $N$  is set to 64 in (2), (3), (6), and (8). In the simulation, a frequency-offset of 73% sub-carrier spacing, which is the allowed maximum value in the 802.11a WLAN [4], is used. Since the number of the samples for the fine estimation is 64 [4],  $N_{e,f} - N_{s,f} + 1 = 64$  in (8). In Fig. 6, the intelligent scheme achieves the improvement of about 2.5 dB over the conventional scheme in both BRAN channels.



**Fig. 7.** PER performance for the 802.11a WLAN (64QAM)

Fig. 7 exhibits the comparison of the PER performances for the OFDM-based WLAN in the cases of the intelligent synchronization and the conventional synchronization. In this simulation, the 802.11a WLAN is also used as the OFDM-based WLAN because of the same reason in the simulation of Fig. 6. In the simulation of Fig. 7, each subcarrier of the OFDM-based WLAN is modulated by 64QAM symbols, and the coding rate for convolutional coder is 3/4. This corresponds to the maximum data rate, 54 Mbps [4]. The packet consists of 1000-byte data as specified for the investigation of minimum sensitivity level [4]. As illustrated in the figure, the intelligent synchronization scheme achieves gains of about 2 dB and about 1 dB at the 10% PER in BRAN A and BRAN C channels, respectively.

## 4 Conclusion

In this paper, an intelligent synchronization scheme is presented to enhance the synchronization performance in WLAN convergence systems. The intelligent scheme consists of intelligent protocol identification and intelligent frequency synchronization. The intelligent identification utilizes flag detection as well as signal detection for more accurate protocol identification. The proposed frequency synchronization intelligently performs the 1<sup>st</sup> fine estimation and the 2<sup>nd</sup> fine estimation to produce a more accurate frequency-offset estimate. The performances of the protocol identification and the frequency synchronization have been evaluated in terms of success rate and MSE, respectively. The simulation results reveal that the proposed intelligent scheme significantly improves the synchronization performance in WLAN convergence systems. In addition, the PER simulation results confirm that the proposed scheme can especially enhance the PER performance of the OFDM-based WLAN in WLAN convergence systems.

## References

1. Three WLAN standards in a single solution. Technical Report. Atheros Communications, Inc. (2003)
2. WLAN solutions for IEEE802.11a/b/g. Technical Report. Texas Instruments, Inc. (2003)
3. Schmidl, T.M., Cox, D.C.: Robust frequency and timing synchronization for OFDM. *IEEE Trans. Comm.* **45** (1997) 1613–1621
4. IEEE Std 802.11a: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: high-speed physical layer in the 5 GHz band. (1999)
5. IEEE Std 802.11b: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: higher-speed physical layer extension in the 2.4 GHz band. (1999)
6. IEEE Std 802.11g: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. (2003)
7. Melbo, J., Schramm, P.: Channel models for HIPERLAN/2 in different indoor scenarios. 3ER1085B. HIPERLAN/2 ETSI/BRAN contribution. (1998)
8. Joint Technical Committee on Wireless Access. Final Report on RF Channel Characterization. (1994)

# Aggressive Sub-channel Allocation Algorithm for Intelligent Transmission in Multi-user OFDMA System

SangJun Ko, Joo Heo, Yupeng Wang, and KyungHi Chang

The Graduate School of Information Technology & Telecommunications  
Inha University, Incheon, Korea  
sj987608@empal.com, heojoo@hanmail.net,  
charles535@hotmail.com, khchang@inha.ac.kr

**Abstract.** This paper solves the problem of finding an intelligent sub-channel allocation method to communicate with multiple users in OFDMA wireless systems. We propose an ASA (Aggressive Sub-channel Allocation) algorithm, which is an intelligent dynamic channel allocation algorithm considering all users' channel state information conditionally to maximize throughput in an OFDMA system. We compare the ASA algorithm with Round-robin, PF (Proportional Fair) and ACG (Amplitude Craving Greedy) in the 2-tier cell environment and then analyze the performance of each algorithm through computer simulation in downlink OFDMA system. Simulation results show that the proposed ASA algorithm gets 24 %, 16 % and 14 % better ring throughput compared with the Round-robin, PF and ACG, respectively.

## 1 Introduction

OFDMA is recognized as one of the most promising multiple access techniques in future wireless communication systems [1]. Recently, a number of algorithms have been developed for channel allocation in OFDMA systems. Most of these algorithms deal with single-cell scenarios [2]-[5] and few studies have focused on the resource allocation for multi-cell OFDMA systems [6], [7]. Some algorithms are resource allocation algorithms for throughput maximization only, i.e., they assign resource unfairly based on allocating more resources to the users with good channel conditions [4], [6], [7], and other algorithms mainly consider fairness when allocating resource [3], [5]. Accordingly, this paper proposes an ASA (Aggressive Sub-channel Allocation) algorithm for an OFDMA system, which schedules traffic of each user considering fairness among all the users as well as the maximization of the total system throughput.

This paper is organized as follows. In Section 2, existing dynamic channel allocation algorithms are introduced and in Section 3, the proposed ASA algorithm, which is intelligent sub-channel allocation algorithm for OFDMA system, is introduced. In Section 4, the performance of the proposed algorithm is validated with the given simulation environments and parameters. Finally, we conclude in Section 5.

## 2 Sub-channel Allocation Algorithm for Adaptive Transmission

### 2.1 BABS Algorithm

The BABS (Bandwidth Assignment Based on SNR) algorithm decides the minimum number of sub-channels that a user will be assigned to satisfy the required data rate of each user [5]. This means that, to consider fairness among users, sub-channels are not all allocated to the users with better channel condition.

First, we solve for the data rate for each user utilizing the given received SNR in the multi-user OFDM system [8]. Assuming that QAM modulation and ideal phase detection are used, the BER for the  $k^{th}$  user's  $m^{th}$  sub-channel signal is bounded by

$$BER \leq \frac{1}{5} \exp\left(\frac{-1.5\gamma_k(m)}{(2^{q_k(m)} - 1)}\right), \tag{1}$$

where  $q_k(m)$  is the number of bits in each data symbol which is allocated to the  $m^{th}$  sub-channel of the  $k^{th}$  user and  $\gamma_k(m)$  is the SINR of the  $m^{th}$  sub-channel of the  $k^{th}$  user. The BER bound is valid for  $q_k(m) \geq 2$  and  $0 \leq \gamma_k(m) \leq 30$  dB. For a given BER, re-arranging (1) yields the maximum number of bits in a symbol to be transmitted for the  $m^{th}$  sub-channel of the  $k^{th}$  user as

$$q_k(m) = \log_2\left(1 + \frac{\gamma_k(m)}{\Gamma}\right), \tag{2}$$

where  $\Gamma = -\ln(5BER)/1.5$ . So, in the multi-user OFDM system, the total data rate can be represented by

$$R_k = \sum_{m=1}^M \frac{q_k(m)}{T} = \frac{1}{T} \sum_{m=1}^M \log_2\left(1 + \frac{\gamma_k(m)}{\Gamma}\right), \tag{3}$$

where T is the OFDM symbol duration and M is the total number of users. The base station calculates a total data rate of each user across all sub-channels using the average feedback SINR.

Then the number of sub-channels to be allocated to each user is presented by

$$\overline{R}_k = \frac{R_k}{M}, \quad m_k = \left\lceil \frac{R_{k,req}}{\overline{R}_k} \right\rceil, \tag{4}$$

where  $R_k$  is the total data rate of each user within all sub-channels. That is,  $\overline{R}_k$  is the average data rate acquired by one sub-channel and  $R_{k,req}$  is the required data rate of each user. Therefore the number of sub-channels  $m_k$  to be allocated for each user is obtained by (4).

If the sum of all users'  $m_k$  exceeds the total number of sub-channels, sub-channels which are allocated to the user at minimum level (i.e., minimum number of sub-channels) are cancelled one by one until the sum of all users'  $m_k$  equals to the total number of sub-channels. However, if the sum of all users'  $m_k$  is smaller than the total number of sub-channels, sub-channels are additionally allocated to the user who requires minimum additional power for the allocation of the additional sub-channels until the sum of all users'  $m_k$  equals to the total number of sub-channels.



## 2.2 ACG Algorithm

The ACG (Amplitude Craving Greedy) algorithm is a technique which allocates sub-channels to each user as much as  $m_k$  which is already solved by BABS. The sub-channel is allocated to the user who has maximum channel gain in that sub-channel. During the allocating process, base station does not allocate sub-channels to users who already satisfy  $m_k$ .

In ACG algorithm, sub-channels are sequentially allocated from the 1st sub-channel to the last sub-channel. This is why the complexity of the ACG algorithm complexity is very low compared with that of other algorithms. However, system throughput is not very high, because sub-channel which has the highest channel gain is not preferentially allocated to the potential users.

## 3 Proposed Aggressive Sub-channel Allocation Algorithm

The proposed ASA (Aggressive Sub-channel Allocation) algorithm does not have the procedure to predetermine the number of sub-channels for the user by BABS algorithm. Instead following basic assumptions are being considered.

1. Every user reports channel state information of all the sub-channels to a base station using uplink feedback channel.
2. Each sub-channel is allocated to one user only, not to allow intra-cell interference.
3. There may be multiple number of sub-channels which has the same level of maximum channel gain due to the quantized feedback channel state information.

The proposed ASA algorithm can be implemented in the following four steps;

In the first step, the base station determines a user to be scheduled considering fairness as in (5)

$$k_{ASA} = \arg \min_{(k=0:K-1)} \frac{\sum_{m \in C_k} r_{k,m}}{R_{k,req}}, \quad (5)$$

where  $k_{ASA}$  is the selected user index by the ASA scheduler considering fairness,  $C_k$  is the group of the sub-channels allocated to the  $k^{th}$  user during a schedule process,  $r_{k,m}$  is the data rate of the  $m^{th}$  sub-channel which is allocated to the  $k^{th}$  user, and  $R_{k,req}$  is the required data rate of the  $k^{th}$  user. That is, fairness among users is guaranteed by preferential allocation of sub-channels to the user who has the minimum value of (Acquired Data Rate / Required Data Rate).

In the second step, a base station searches a sub-channel which allows maximum data rate to the selected user  $k_{ASA}$  among sub-channels which are not yet allocated to other users.

$$m^* = \arg \max_{m=0:M-1} (r_{k_{ASA},m}), \quad (6)$$

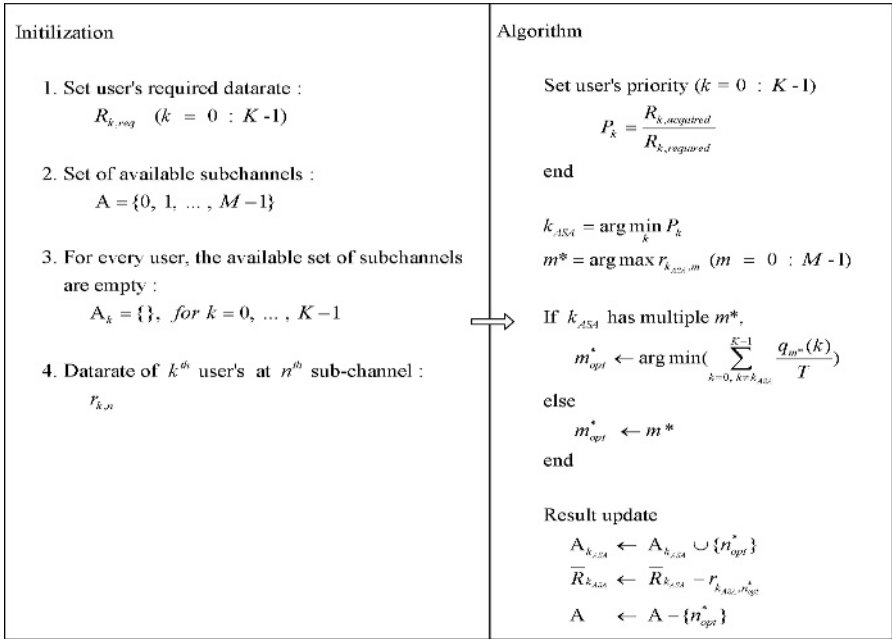
where  $r_{k_{ASA},m}$  is the data rate of a user  $k_{ASA}$  which is allowed in the  $m^{th}$  sub-channel. Here,  $m^*$  denotes the multiple sub-channel indexes which allow same maximum data rate to a user  $k_{ASA}$ .

In the third step, the base station evaluates data rates on sub-channel  $m^*$ , which might be allowed for other users, and then calculate the sum of data rates  $R_{m^*}$  of all users except user  $k_{ASA}$  on the sub-channel  $m^*$ .

$$R_{m^*} = \sum_{\substack{k=0 \\ k \neq k_{ASA}}}^{K-1} r_{k,m} = \sum_{\substack{k=0 \\ k \neq k_{ASA}}}^{K-1} \frac{q_{m^*}(k)}{T}, \tag{7}$$

In the last step, the base station determines the sub-channel index to be allocated to the selected user considering data rate  $R_{m^*}$  as following. If the sub-channel  $m_{opt}^*$  supports minimum data rates to other users (i.e.,  $m_{opt}^* = \arg \min(R_{m^*})$ ), the sub-channel  $m_{opt}^*$  is allocated to the selected user  $k_{ASA}$ .

Therefore, system throughput can be maximized by referring the channel state information of the selected user  $k_{ASA}$  and those of other users when there exist multiple number of the same maximum level of channel state information, and this procedure is applied to all the users in the system. The 1-dimensional search procedure of the ASA algorithm to find the optimum sub-channel index for a selected user is a lot more simpler than the 2-dimensional search procedure of the RCG (Rate Craving Greedy) algorithm [5]. Loss of system throughput by the ASA algorithm compared to the RCG algorithm is negligible. Overall procedure of the proposed ASA algorithm is illustrated in fig. 1.

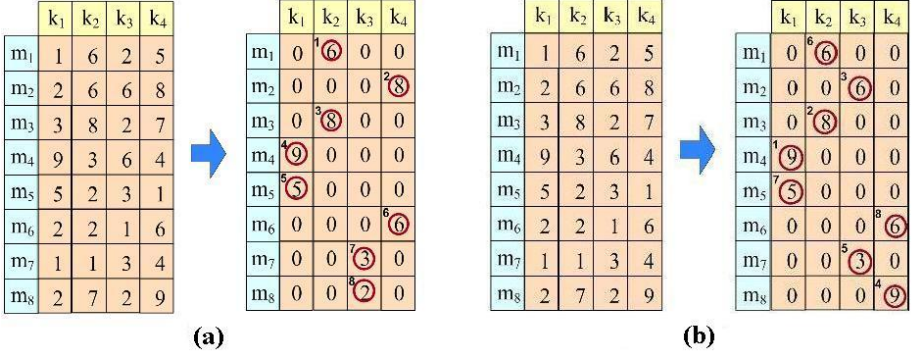


**Fig. 1.** Proposed Aggressive Sub-channel Allocation Algorithm

In fig. 2, example procedures for the sub-channel allocation are given for the cases of the ACG algorithm with BABS and the ASA algorithm. Both (a) and (b) assume that the total number of users is 4, the total number of sub-channels are 8, and all

users' required data rates are equal, where rows indicate sub-channels and columns indicate users. Data rate is given in the box for a user  $k_i$  and a sub-channel  $m_j$ .

In the first step, the base station determines a user to be scheduled considering fairness using (5). So a user is selected as  $k_{ASA} = k_1$  for scheduling.



**Fig. 2.** (a) Example procedure of the ACG algorithm with the BABS ( $m_k=2$ ). (b) Example procedure of the ASA algorithm.

Following the first step, the base station searches a sub-channel which allows maximum data rate to the selected user  $k_{ASA}$  among sub-channels which are not yet allocated to other users. In fig. 2, sub-channel  $m_4$  allows maximum data rate for  $k_1$ , so  $m_4$  is allocated to user  $k_1$ .  $m_3$  is allocated to  $k_2$ ,  $m_2$  is allocated to  $k_3$ , ..., and  $m_8$  is allocated to  $k_4$ , consecutively. Then, the base station gives a priority to the user  $k_3$  because he acquires the minimum data rate. In this case, both  $m_5$  and  $m_7$  support the same data rate of 3,

$$r_{k_3, m_5} = r_{k_3, m_7} = 3. \quad (8)$$

But the base station select  $m_7$  for user  $k_3$  based on (9).

$$R_{m_5} = \sum_{\substack{k=1 \\ k \neq k_3}}^4 \frac{q_{m_5}(k)}{T} = 8, \quad R_{m_7} = \sum_{\substack{k=1 \\ k \neq k_3}}^4 \frac{q_{m_7}(k)}{T} = 6, \quad (9)$$

where  $R_{m_5}$  and  $R_{m_7}$  are the sum of the data rates of all users except user  $k_3$  on the sub-channel  $m_5$  and  $m_7$ , respectively.

In the case of the ACG algorithm,  $m_5$  is allocated to  $k_3$  because scheduling is now being processed in the order of sub-channel number. Eventually, we get

$$R_{ACG} = \sum_{k=1}^4 \sum_{m=1}^8 \frac{q_m(k)}{T} = 47, \quad R_{ASA} = \sum_{k=1}^4 \sum_{m=1}^8 \frac{q_m(k)}{T} = 52, \quad (10)$$

where  $R_{ACG}$  and  $R_{ASA}$  is the total data rate by the ACG and the ASA algorithm, respectively. The proposed ASA algorithm guarantees better throughput than the ACG algorithm, while requires less complexity than RCG algorithm.

## 4 Performance Validation

In this section, the performance gain by the proposed ASA algorithm is validated under a baseline frame structure. We set the cell radius to 1 km. There are 32 sub-channels in the total frequency-domain resource structure and all the 32 sub-channels are used as FRF (Frequency Reuse Factor) 1. A sub-channel consists of 3 continuous bins in 1 symbol and a bin consists of 9 sub-carriers. Among the 9 sub-carriers, 1 sub-carrier is used as a pilot sub-carrier for the channel estimation, synchronization, and SINR measurement. As a channel model, the extended Hata path loss model is employed, and the standard deviation of log-normal shadowing is set to 10 dB. The multi-path channel model of the ITU-R Ped A is adopted as a delay profile model. We assume there are 9 modulation and coding levels, and the target packet error rate for the adaptive modulation and coding is 1%. The other fundamental OFDMA system and link level simulation parameters are given in Table 1 and 2 as in [9]-[11].

**Table 1.** Fundamental system level simulation parameters

Items	System Parameters	Values	Items	System Parameters	Values
BS Tx Side	BS Tx Power	43 dBm	Channel Model	Path Loss Model	Extended Hata Model
	BS Tx Antenna Gain	15 dBi		Shadowing	Log-Normal Distribution St. Dev. : 10 dB
	BS Cable Loss	3 dB		Fading Channel	ITU-R Ped A 3km/h
	BS Max EIRP	55 dBm	Cell Site	Nr. of Cells	19
MS Rx Side	BS Rx Antenna Gain	0.0 dBi		Cell Configuration	Hexagonal
	BS Thermal Noise	-174 dBm / Hz		Cell Radius	1 km
	MS Noise Figure	7 dB		User Position	Uniform

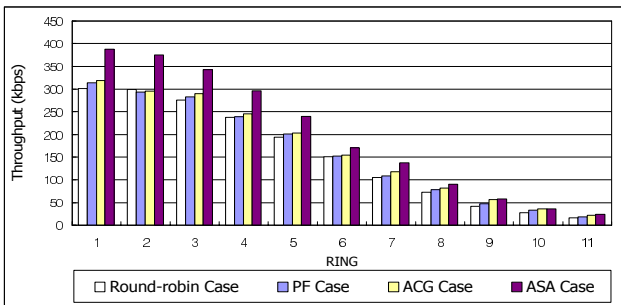
**Table 2.** Fundamental link level OFDMA system parameters

System Parameters	Values	System Parameters	Values
Carrier Frequency	2.3 GHz	Data Sub-carriers	768
Sampling Frequency	10 MHz	Pilot Sub-carriers	96
Effective Channel BW	8.75 MHz	FFT Size	1024

Four scheduling algorithms are considered as follows :

- Round-robin: Each sector utilizes all 32 sub-channels as FRF 1. This scenario performs adaptive modulation and coding per frame (5ms) to achieve a target PER (Packet Error Ratio) of 1%.
- PF (Proportional Fair) [3], ACG [5], ASA : Each sector utilizes 32 sub-channels as FRF 1. These scenarios perform adaptive modulation and coding per frame (5ms) and perform dynamic channel allocation per 4 frames (20ms) to achieve a target PER of 1%.

Fig. 3 shows the average downlink ring throughput per user for the four different scheduling algorithms with fully loaded adjacent cells. The  $n^{\text{th}}$  ring covers the area whose radius ranges  $(n-1)*100 \sim n*100$  meters. Simulation results show that the proposed ASA algorithm gets 24 %, 16 % and 14 % more ring throughput per user compared with the conventional Round Robin, PF and ACG algorithms, respectively. This benefit comes from the extensive use of the multi-user diversity gain considering all users' channel state information conditionally.



**Fig. 3.** Average downlink ring throughput per user for the four different scheduling algorithms with fully loaded adjacent cells

## 5 Conclusions

This paper proposes an ASA algorithm that system throughput can be maximized by referring the channel state information of the selected user and those of other users when there exist multiple number of the same maximum level of channel state information, and this procedure is applied to all the users in OFDMA system. Even though considering the additional downlink broadcasting overhead for the DCA procedure, the proposed ASA algorithm increases the downlink ring throughput by 24 %, 16 % and 14 % compared with the case of round-robin, PF and ACG, respectively. In the aspect of complexity, the 1-dimensional search procedure of the ASA algorithm to find the optimum sub-channel index for a selected user is a lot more simpler than the 2-dimensional search procedure of the RCG algorithm. Loss of system throughput by the ASA algorithm compared to the RCG algorithm is almost negligible.

## References

1. J. Chuang and N. Sollenberger, "Beyond 3G: wideband wireless data access based on OFDM and dynamic packet assignment," *IEEE Commun. Mag.*, vol. 38, pp. 78–87, July 2000.
2. G. Song and Y. Li, "Adaptive resource allocation based on utility optimization in OFDM," in *Proc. IEEE Globecom.*, San Francisco, Dec. 2003.
3. Wengerter. C., Ohlhorst. J., Elbwart. A. G. E., "Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA," in *Proc. IEEE VTC-Spring*, May. 2005, pp. 1903-1907.
4. C.Y. Wong, R.S. Cheng, K.B. Letaief and R.D. Murch, "Multiuser OFDM with adaptive subcarrier, bit and power allocation," *IEEE J. Select Areas Commun.*, vol. 17, pp. 1747-1758, Oct. 1999.
5. D. Kivanc, G. Li, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 1150-1158, Nov. 2003.
6. J. Li, H. Kim, Y. Lee and Y. Kim, "A novel broadband wireless OFDMA scheme for downlink in cellular communications," in *Proc. IEEE WCNC*, New Orleans, Mar. 2003.
7. G. Li and H. Liu, "Downlink dynamic resource allocation for multi-cell OFDMA system," in *Proc. IEEE VTC*, Orlando, Oct. 2003.
8. J. Jang and K.B. Lee, "Transmit power adaptation for multiuser OFDM system," *IEEE J. Select Area Commun.*, vol. 21, pp. 171-178, Feb. 2003.
9. I.S. Cha, "Effective adaptive transmission power allocation algorithm considering dynamic channel allocation in reuse partitioning based OFDMA system," *MS Thesis*, The Graduate School of Information Technology & Telecommunications, Inha Univ., Aug. 2005.
10. Recommendation ITU-R M.1225, Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000, 1997.
11. 3GPP R1-030042, Update of OFDM SI simulation methodology, Jan. 2003.

# A Novel Spectrum Sharing Scheme Using Relay Station with Intelligent Reception\*

Jaehwan Kim, Junkyu Lee, and Joonhyuk Kang

School of Engineering, Information and Communications University(ICU),  
119 Munjiro, Yuseong-gu, Daejeon, Korea, 305-732  
{mail, haibrid, jhkang}@icu.ac.kr

**Abstract.** In this paper, we propose a new spectrum sharing method using relay station with intelligent reception at the receiver. The relay station retransmits primary signal to destination which receives combined primary and secondary signals simultaneously. Then, the receiver detects the secondary signal by cancelling the primary signal out with the aid of the relayed signal. Simulation results demonstrate that the proposed method guarantees reliable communication for secondary service between some pairs of users, where there is another licensed primary service in the same frequency band.

## 1 Introduction

With the drastic growth of new wireless applications, the spectrum becomes more and more congested. On the other hand, it is commonly believed that the usage of spectrum is vastly underutilized all over the world. For example, a recent survey reports that only a little percent of allocated spectrum is actually in use not only in the United States but also in other countries [1].

In recent years, this inefficient utilization of spectrum motivates many communication engineers to pay attention to spectrum management such as resource allocation, coexistence, and spectrum sharing in order to promote more flexibility in spectrum usage. One way to increase the utilization of a precious natural resource is cognitive radio (CR) technology as a new paradigm in wireless communications that hold promise for new and better services to many markets. To date, there has been a lot of discussion and attention of CR from the wireless communications community.

Since the concept of CR has been firstly introduced by Mitola [2], it was shown that if the CR was profitably employed to communication systems, the unused or unlicensed spectrum could be exploited efficiently with adaptation to dynamical changes of their environment [1]. More specifically, the cognitive radio systems have cognitive cycle including Observe, Orient, Plan, Decide, Act, and Learn [2]. As a view point of the cognitive cycle, the communication system

---

\* This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

observes the outside world and determine the priority of the work in orient cycle. Then, the order of the tasks are planed and decided. Finally, the system performs the operations and recognizes the status of the circumstances. Several issues on CR have been well defined in [3] such as spectrum sensing, spectrum sharing, adaptive transmission, and power control.

Spectrum sharing methods have been considered through many literatures with active studies of cognitive radio lately. D. Willkomm et al. introduced negotiated spectrum sharing and opportunistic spectrum sharing method with the performance analysis for secondary usage system [4]. In case of the negotiated spectrum sharing method, primary user informs secondary user of the usage of the licensed spectrum. On the other hand, secondary user in opportunistic method monitors the frequency bands whether the primary user exists or not. Another spectrum sharing problem in unlicensed bands was raised by R. Etkin et al. [5]. It is divided into cooperative and non-cooperative situation by the properties of the systems. R. Etkin et al. analyzed the performance and concluded that Pareto efficient vector of rates could maintain constant power allocation in cooperative situation. For non-cooperative situation, game theory was used for analyzing the results, such as achievable rates and performance loss due to scarcity of cooperation. In addition, interference compensation for spectrum sharing was proposed in [6], where the users in network exchanged “price” signals corresponding to interference at the receiver. According to the “price” signals, the users select a channel which has the maximum of its benefit and decide the power level.

In this paper, we propose a new spectrum sharing scheme using relay station for unlicensed secondary users with intelligent reception. The existing relay station is, for example, used for cooperation with transmitter [7] in order to obtain diversity gain from virtual multi-antenna structure [8]. However, the relay station in our scheme is used for cancelling out the interference signal, the primary signal. The results show that the simple spectrum sharing method gives reliable link maintenance between secondary users and overall network capacity can be increased.

This paper is organized as follows. Section 2 describes the proposed spectrum sharing method. In section 3, we present simulation results of proposed method in AWGN and flat fading channel. Finally, we make the conclusion of this paper in section 4.

## 2 Spectrum Sharing with Intelligent Reception Using Relay Station

### 2.1 Proposed Method Using Relay Station

Consider a network composed of one base station, one relay station, and many users. Fig. 1 shows the network with 5 users and relay operation. The base station broadcasts to all users for primary service through licensed band. When the users in the network try to communicate with each others, for example,



user A with B and user C with D, the primary broadcasting signal affects to the receiver for secondary service as an interference. The received signal can be represented as,

$$y(n) = h_1(n)m(n) + h_2(n)s(n) + w(n) \tag{1}$$

where  $m(n), s(n)$  are primary and secondary signal, respectively, and  $h_1(n), h_2(n)$  are channel coefficients.  $n$  is time index. The primary signal,  $h_1(n)m(n)$ , disturbs communication between two users,  $h_2(n)s(n)$ .

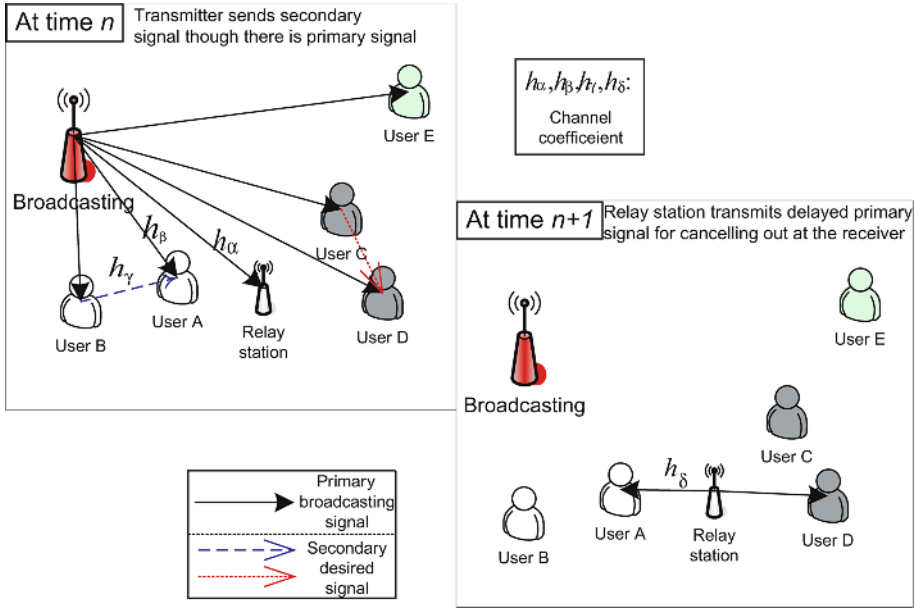


Fig. 1. Proposed spectrum sharing scheme

We consider two-user case for simplicity. At first, the base station sends signal for primary service and user B sends secondary signal to user A, simultaneously. After sending desired signal from user B to A, the relay station sends previous primary broadcasting data to user A, illustrated in Fig. 1. We assume that there is another frequency band for communication from relay station to users and the bandwidth is same to primary signal bandwidth. Finally, user A performs interference cancellation using both two signals which are from B and relay station. From now on, transmitter and receiver means user B and user A, respectively.

The received signals in Fig.1 can be represented by

$$y_a(n) = h_\beta(n)m(n) + h_\gamma(n)s(n) + w_1(n) \tag{2}$$

$$y_b(n) = h_\delta(n)\hat{m}(n - 1) + w_2(n) \tag{3}$$

where  $y_a(n)$  and  $y_b(n)$  are received signal from the base station and the transmitter simultaneously and the relay station, respectively.  $\hat{m}(n)$  denotes relayed

signal and  $w_i(n)$ s are complex Gaussian noise. In the relay station, the received signal is

$$y_r(n) = h_\alpha(n)m(n) + w_3(n) \quad (4)$$

Then, the relay station sends decoded data( $\hat{m}(n)$ ) to the receiver for error compensation.  $h_\alpha(n)$ ,  $h_\beta(n)$ ,  $h_\gamma(n)$ , and  $h_\delta(n)$  are channel coefficients for each link. Finally, the receiver applies received data,  $y_a(n)$  and  $y_b(n)$  for interference cancellation, it is assumed that the channel coefficients are known. Using these conditions, the receiver can acquire the desired signal.

We show the results for AWGN and flat fading channels. Brief mathematical representation is followed and simulation results will be shown in section 3. In AWGN channel, all the channel coefficients are one since there are no fading effect for signals. In this case, we can detect desired data only subtraction between  $y_a(n)$  and  $y_b(n)$ . It can be expressed as follows,

$$y_a(n) - y_b(n+1) = m(n) + s(n) - \hat{m}(n) + w_1(n) - w_2(n+1) \quad (5)$$

$$= s(n) + w'(n) \quad \text{where } m(n) \approx \hat{m}(n) \quad (6)$$

In fading channel, we can also avoid errors caused by primary service. The receiver eliminates the effect of the channel distortions by processing the received signal with the channel coefficients. It can be written as,

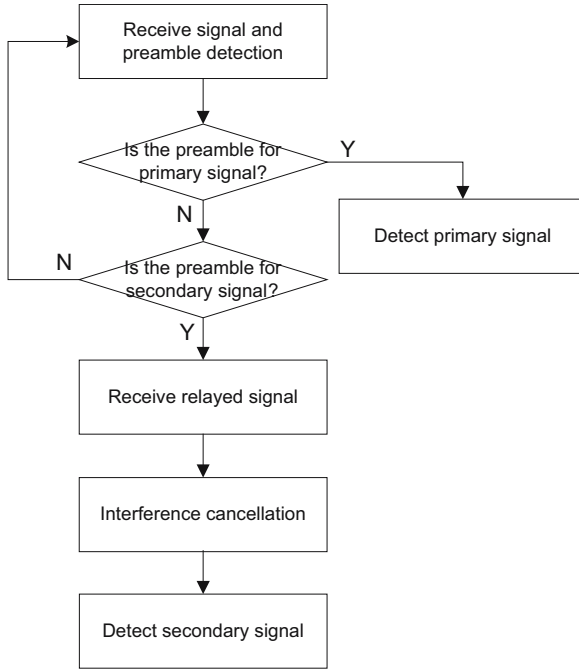
$$\left[ y_a(n) - \left\{ \frac{y_b(n+1)}{h_\delta(n+1)} \right\} \times h_\beta(n) \right] / h_\gamma(n) = s(n) + w''(n). \quad (7)$$

## 2.2 Intelligent Reception

The receiver must distinguish whether the unlicensed signal exists or not to decide reception of the relayed signal. The flow chart at the receiver is given in Fig. 2. First, the receiver receives signal and detects the preamble. If the preamble is primary's, the receiver detect primary signal. However, if the preamble is not primary's, it means that there is a secondary signal or preamble detection error. In this case, if it is secondary's, the receiver receives the delayed primary signal from the relay station, and if not, the receiver is initialized and waits for a new signal. We call this procedure as intelligent reception. After the intelligent reception when the secondary signal exists, the receiver cancels out the primary signal and obtains the secondary desired signal.

## 3 Simulation Results

We investigate the performance of the proposed scheme in AWGN and flat fading channels. We use QPSK modulation for data transmission, 20 symbols for a packet and Jakes' simulator with Rayleigh distribution for fading channel. The results show that the proposed method provides stable communication for secondary users and increases overall network capacity. The performance of secondary service, communication between user A and B in Fig. 1 is examined from packet error rate (PER) and spectral efficiency, and we assume that all the signals are combined with primary and secondary service.



**Fig. 2.** Flow chart of intelligent reception

### 3.1 AWGN Channel

Simulation results in AWGN are shown in Fig. 3 and Fig. 4. If we do not use a relay station, error floor occurs as in Fig. 3. It is from the intuition that there is no information to eliminate primary signal at all. However, through using relay station to retransmit primary signal at following time interval, user A can extract desired signal, secondary service, from the combined received signal. As a result, the error floor is eliminated; then the proposed method shows 3 dB difference compared to theoretical QPSK. It might be from the error occurrence at relay station before the retransmission. In addition, the proposed scheme achieves maximum one bit per symbol per channel when we use the relay station since there is another frequency band for the relay station as shown in Fig. 4. The bandwidth of the relay station is the same with the primary signal bandwidth so the spectral efficiency decreases by a half. However, if there is no relay station, communication between two users would be impossible.

### 3.2 Flat Fading Channel

Fig. 5 and Fig. 6 present the simulation results in fading channel. In flat fading channel, similar to AWGN, using relay station makes the error performance highly improve compared to that without the relay station. The result of PER has 8 dB difference at  $10^{-2}$  PER against to the theoretical result. The performance

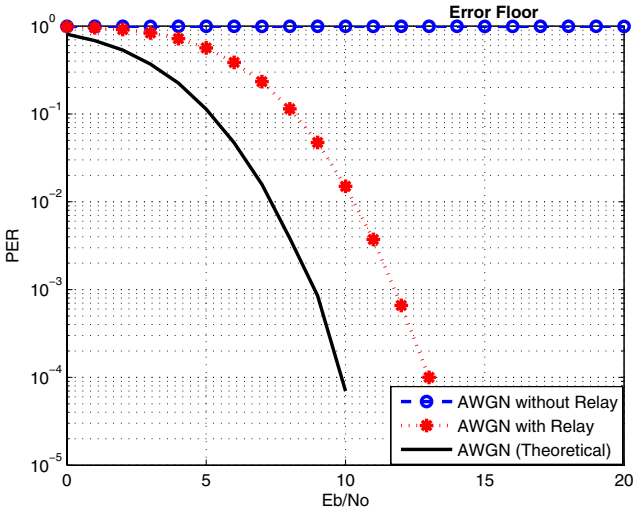


Fig. 3. Performance of proposed method in AWGN

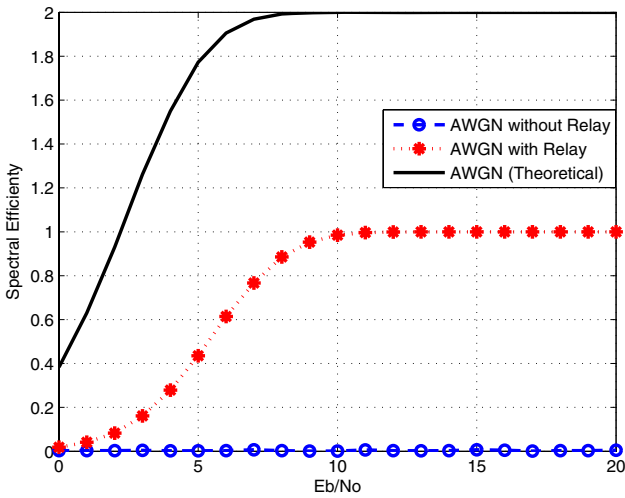


Fig. 4. Spectral efficiency of proposed method in AWGN

is worse than AWGN channel, because the signal suffers from channel distortion. In addition, the relay station maintains maximum spectral efficiency around 35 dB for  $E_b/N_o$  value.

Using the relay station gives 1 bps spectral efficiency to each pair of users. i.e, if there are many pairs of users in the network, the network capacity increases. For example, we can achieve network capacity by  $n$  bps when there are  $n$  pairs of users.

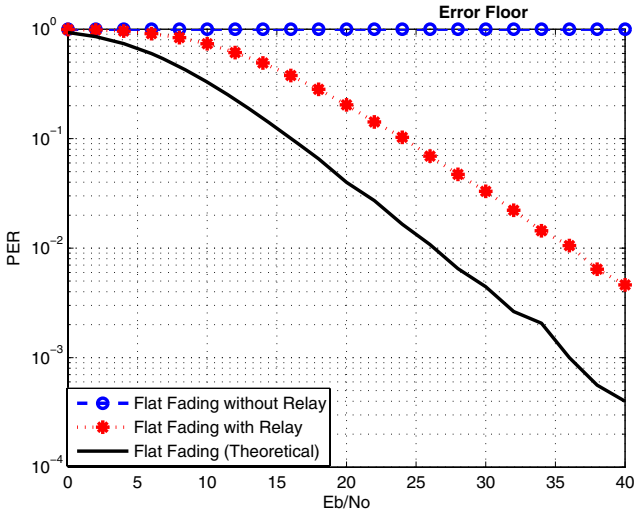


Fig. 5. Performance of proposed method in flat fading

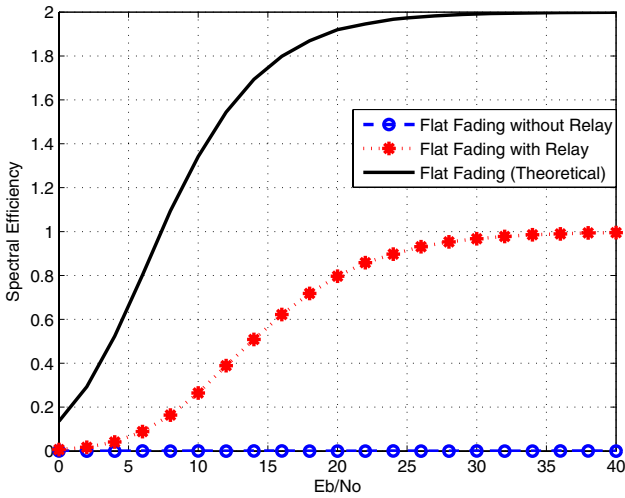


Fig. 6. Spectral efficiency of proposed method in flat fading

### 4 Conclusion

In this paper, we present a spectrum sharing scheme using relay station. To maintain reliable link for unlicensed users when there coexist licensed and unlicensed users, the relay station sends delayed version of the primary signal to the destination which performs the intelligent reception. Then, primary signal is cancelled out at the destination so that the secondary signal can be detected.

The simulation is performed in AWGN and flat fading channel. The results show that the retransmitted data from the relay station can be used for eliminating the packet error floor. Besides, a reliable communication for the secondary service within unlicensed band is provided. The proposed method is shown to give two considerable advantages. One is that we can make interference cancellation very simple. The other is that overall network capacity would be increased by increasing the number of users.

## References

1. V.D. Chakravarthy, A.K. Shaw, M.A. Ternple, J.P. Stephens, "Cognitive radio - an adaptive waveform with spectral sharing capability," in *Proc. IEEE WCNC*, vol. 2, pp. 724-729, New Orleans, USA, March 2005.
2. J. Mitola, III and G. Q. Maguire, Jr., "Cognitive Radio: Making Software Radios More Personal," *IEEE Personal Communications Magazine*, vol. 6, pp. 13-18, Aug. 1999.
3. S. Haykin, "Cognitive Radio: brain-empowered wireless communications," *IEEE J. Select. Areas Comm.*, vol. 23, no. 2, pp. 201-220, Feb. 2005.
4. Daniel Willkomm, James Gross, and Adam Wolisz, "Reliable Link Maintenance in Cognitive Radio Systems," in *Proc. IEEE DySPAN*, Baltimore, USA, Sep. 2005.
5. Raul Etkin, Abhay Parekj, and David Tse, "Spectrum Sharing for Unlicensed Bands," in *Proc. IEEE DySPAN*, Baltimore, USA, Sep. 2005.
6. Jianwei Huang, Randall A. Berry, Michael L. Honig, "Spectrum Sharing with Distributed Interference Compensation," in *Proc. IEEE DySPAN*, Baltimore, USA, Sep. 2005.
7. Aria Nosratinia, Todd E. Hunter, and Ahmadreza Hedayat, "Cooperative Communication in Wireless Networks", *IEEE Communicatios Magazine*, vol. 42, pp. 74-80, oct. 2004.
8. M. Dohler, E. Lefranc, and H. Aghvami, "Virtual Antenna Arrays for Future Wireless Mobile Communication Systems," in *Proc. IEEE ICT*, Beijing, China, 2002.

# Intelligent Beamforming for Personal Area Network Systems

Younggun Ji, Seokhyun Kim, Hongwon Lee, and Jaehak Chung\*

Inha University, Korea,  
jchung@inha.ac.kr

**Abstract.** The intelligent beamforming algorithm that transmits minimum power to the nodes in out-of main beam regions to prevent unwanted traffic in CSMA/CA (carrier sensing multiple access /collision avoidance) systems is presented, while main beams focus on specific nodes. To design the intelligent complementary beamforming, Gram-Schmidt orthogonalization is utilized, whose beam pattern exhibits perfect nulling performance at the main beam angles, and provides uniform power out of main beams. The proposed method can be utilized with any arbitrary beamforming algorithms and can control the power with a small value of  $\epsilon$  to satisfy detection criteria of communication standards.

## 1 Introduction

Array antennas can significantly increase the communication range and SNR (Signal-to-Noise Ratio) among nodes. Classical smart antenna system using array antennas is one of techniques to provide these advantages [GRO],[ZOO]. Therefore, in modern communication systems, many array antenna technologies have been developed and adopted as standards.

Among wireless communications, WPAN (Wireless Personal Area Network) is designed for supporting near areas, e.g., indoor environments, compared with outdoor cellular systems. Some of standards need to support high data rate such as IEEE802.15.3 [STD]. Since the WPAN utilizes distributed protocols rather than centralized systems, CSMA/CA (carrier sense multiple access/carrier avoidance) mechanism, i.e., listen-before-talk, is utilized for random multiple access. This technique is simple to implement and relatively efficient for small multiple access systems. In addition, wireless LAN (local area networks) also adopts this methodology.

In general, array antenna system focuses one-to-one communications without sending transmit power to others with nulling method. It reduces interferences to others. In turn, if other nodes want to make a new communication links, in CSMA/CA systems they should transmit data to acquire admission of link connection. However, they are not supposed to transmit data. This is because other nodes cannot listen the current nodes' communications. The unwanted

---

\* This research was supported by University IT Research Center (Inha UWB-ITRC), Korea.

trial transmission causes interference to others and consumes their own power. This scenario is inherently unavoidable if array antenna systems are utilized in CSMA/CA. Since it is similar to hidden-node-problem, we may call hidden beam problem [TAR].

To overcome this drawback, the authors in [TAR] designed complementary beamforming algorithms. In this paper, we develop an *intelligent* complementary beamforming method, which can be used for any arbitrary main beams. The main beams focus on specific users, while complementary beams generate beams to others with very small power to broadcast information that communication link is being occupied. First, a given main beam, we create complementary beams intelligently to let other node detecting the transmission. When array antenna systems consist of  $M$  transmit antennas and  $K$  nodes are connected at the same time, it is not easy to design arbitrary  $M - K$  complementary beams among  $M$  dimensional spaces. In this paper, Gram-Schmidt orthogonalization is utilized to design the intelligent complementary beams. The beams null the main beams effectively and just transmit some small power to others, which is borrowed from main beam power. Using this small cost from the main beams, the complementary beams can broadcast information of link occupation, and prevent other nodes' trials that cause interference and useless power consumption.

This paper is organized as follows: In Sect. 2, system model is described. Sect. 3 presents a new intelligent complementary beams. In Sect. 4, numerical results are exhibited. Finally, Sect. 5 concludes the paper.

## 2 System Model

For the array antenna systems, a receiver obtains signals from a central node with  $M$  transmit antennas. The received signal at the  $k$ th user can be written as

$$\mathbf{y}_k(m) = \sum_{l=0}^L \mathbf{h}_{k,l} s_k(m-l) + \mathbf{n}(m), \quad (1)$$

where  $\mathbf{y}_k(m)$  denotes the received signal vector at the  $m$ th sample for the user  $k$ ,  $\mathbf{h}_{k,l}$  denotes the channel spatial signature for the user  $k$  with the size  $M \times 1$  at the  $l$ th delay,  $s_k$  denotes the signal transmitted to user  $k$ , and  $\mathbf{n}(m)$  denotes white and independent identically distributed Gaussian noise with zero mean and variance 0.5 per dimension.

Since the channel  $\mathbf{h}_l$  in Eq. (1) contains spatial signatures, which can be presented as an angle  $\theta$  of DOA (direction of arrival), it can be written as

$$\mathbf{h}_l = g\mathbf{a}(\theta_l), \quad (2)$$

where  $g$  denotes a channel gain,  $\mathbf{a}$  denotes spatial signature, which can be written as

$$\mathbf{a}(\theta_l) = [1e^{j\pi \sin\theta_l} e^{j2\pi \sin\theta_l}, \dots, e^{j(M-1)\pi \sin\theta_l}], \quad (3)$$

where  $\mathbf{a}$  has DOA of  $\theta$ .



If we utilize the OFDM (orthogonal frequency division multiplexing) methodology, the delayed version of channel gain can be represented as frequency selected fading, and it can be written as a simple form as

$$\mathbf{Y} = \mathbf{H}\mathbf{S} + \mathbf{N}, \quad (4)$$

where  $\mathbf{Y}$ ,  $\mathbf{H}$ ,  $\mathbf{S}$ , and  $\mathbf{N}$  denote Fourier transform version of  $\mathbf{y}$ ,  $\mathbf{h}$ ,  $\mathbf{s}$ , and  $\mathbf{n}$ , respectively. For simplicity, in this paper, one ray between two nodes is used for connecting link and one main beam is applied for beam forming. Without loss of generality, we can present the link equation with matrix form.

For transmit beamforming, the weighting vector, which is calculated from many classical beamforming algorithms, is multiplied before transmission. Thus, Eq. (1) can be expressed as

$$y(m) = \mathbf{H}\mathbf{W}s(m) + n(m). \quad (5)$$

where  $y(m)$  denotes received signal at time  $m$ ,  $\mathbf{H} = [h_1, h_2, \dots, h_N]$ ,  $\mathbf{W} = [w_1, w_2, \dots, w_N]^T$ , and  $n$  denotes  $1 \times 1$  matrix.

In general, the weighting vector consists of conjugation of array response vector  $\mathbf{a}$ , and obtained as  $\mathbf{W} = \mathbf{a}^H$ , where  $H$  denotes Hermitian. Therefore, the received signal in Eq. (5) is  $y(m) = gs(m) + n(m)$ . The calculations of specific beamforming vectors are out of scope. It can found in [ZOO],[LIB]. If beamforming technique is applied at the receiver array antennas, we obtain array gain.

For conventional beamforming, the link is only establish between two nodes. Other nodes do not receive the transmit signals since they are in the shadowing region of main beamformings. Thus, if they want to access the central node, they try to transmit some signals, which causes interference and power consumption. This phenomenon is known as *hidden node problem*, some times called hidden beam problem [TAR] since the problem is originated from main beamforming. In the next section, to overcome this drawbacks, an intelligent complementary beamforming methodology is presented.

### 3 Intelligent Complementary Beamforming

As discussed in previous section, to avoid the hidden beam problem, we need to broadcast some signals to other nodes, whose power is small enough to be detected as a link connection. Since a node of WPAN can detect a signal of low level, the transmitter node needs to transmit signal with a small amount of power. The sum power of main beams and complementary beams is limited by regulation.

For reliable detection of complementary beams, we need to generate uniform level power and null the main beams. The design of this complementary beam is not trivial because of limitation of degree of freedom of antenna array. The requirements of the complementary beam is as follows: first, nulling the main beams, second, transmit uniform beams as possible. For the second condition, due to the limited number of transmit antennas, beam may be fluctuated. If we utilize a single antenna, it can be accomplished simply and easily. However, we do

not want to interfere the main beam region by complementary beams. To design the complementary beams that satisfy these conditions, we first assume the main beams is given by arbitrary beamforming techniques and create complementary beams which is orthogonal to the main beams.

The weighting vector  $W$  which includes the complementary beams can be written as

$$\mathbf{W} = \mathbf{W}_{main} + \mathbf{W}^c, \tag{6}$$

where  $\mathbf{W}_{main} \perp \mathbf{W}^c$  in spatial domain.

Thus, the transmitted signal  $\tilde{\mathbf{s}}$  can be presented as

$$\tilde{\mathbf{s}}(m) = \mathbf{W}_{main}\mathbf{s}(m) + \mathbf{W}^C\mathbf{s}^C(m). \tag{7}$$

To satisfy orthogonality to main beams  $\mathbf{W}_{main}$ , which is generated from array response vectors, we may utilize subspace method using singular value decomposition of channel matrix. However, various main beams may be generated with different techniques such as interference nulling and MVDR etc. [GRO],[GOD1],[GOD2]. In this paper, the complementary beam design method using Gram-Schmidt orthogonalization is presented. Therefore, the scheme can provides intelligent way that can generate orthogonal beams for any arbitrary main beams, which are obtained various beamforming vectors.

The procedure for obtaining complementary beams for a given beamforming is described as follows:

- **Step 1:** Let the given beamforming vectors, i.e., main beams, be fundamental beam vectors with normalization

$$\mathbf{w}_1^C = \mathbf{W}_{main}. \tag{8}$$

Then, we project the fundamental vector to other spaces and subtract it from the spaces.

- **Step 2:** For the  $i$ th complementary beam, project the calculated weighting vectors to the arbitrary space  $\mathbf{v}$  and subtract it from the space, then obtain the orthogonal vector as

$$\mathbf{w}_i^C = \mathbf{v}_i - \sum_{j=1}^{i-1} \frac{\mathbf{w}_j^{CH}\mathbf{v}_i}{\mathbf{w}_j^{CH}\mathbf{w}_j^C} \cdot \mathbf{v}_j \quad i = 2, 3, \dots, N. \tag{9}$$

The iteration is executed until obtaining  $\mathbf{w}_i$ ,  $i = N$ . Then, we acquire the complementary beams which is orthogonal to  $\mathbf{W}_{main}$ . We can also calculate orthogonal beams for multiple main beams. If we utilize  $K$  main beams, SDMA (Spatial Domain Multiple Access) technique can be implemented with the proposed method. By normalization, we obtain unit vector for the complementary beams by  $\mathbf{w}_i^c = \mathbf{w}_i/||\mathbf{w}_i||$  for  $i = 2, \dots, N$ .

Finally, if we have  $K$  main beams, we exhibit the beamforming vectors including main beams and complementary beams as

$$\mathbf{W} = \mathbf{W}_{main} + \mathbf{W}^C \tag{10}$$

$$= [\mathbf{w}_1, \dots, \mathbf{w}_k, \mathbf{w}_{k+1}^C, \dots, \mathbf{w}_N^C]. \tag{11}$$

In the conventional beamforming process, the power of the main beam utilizes the whole power to maximize SINR (Signal-to-Interference-Noise Ratio). However, in the proposed intelligent beamforming, we borrow the power of complementary beam from that of the main beam, which is capable to transmit minimum required power for link occupation information.

To keep the total power is the same at the transmitter, we should keep the total power of main beam and the complementary beam is the same. Thus, the following equation should be hold as

$$P_{tot} = P_{main} \mathbf{W}_{main} + P^C \mathbf{W}^C \quad (12)$$

$$= (1 - \varepsilon) P_{tot} \mathbf{W}_{main} + \varepsilon P_{tot} \mathbf{W}^C. \quad (13)$$

The  $\varepsilon$  value is very small to keep high SNR for the main beam, while the complementary beams transmit link occupation information. To keep high SNR of the main beam, we need to set this value as small as possible. In practice, SNR=-20dB is required for detecting the existence of the communications at the receiver. In this case, we let the  $\varepsilon = 0.01$ . Thus, at a cost of small power of the main beam, we can reduce unwanted traffics.

In addition, the main and complementary beam patterns are ringing because of finite number of antennas, we utilize Hamming window for the beamforming to mitigate the ringing. For the better result other specified windowing technique can be applied.

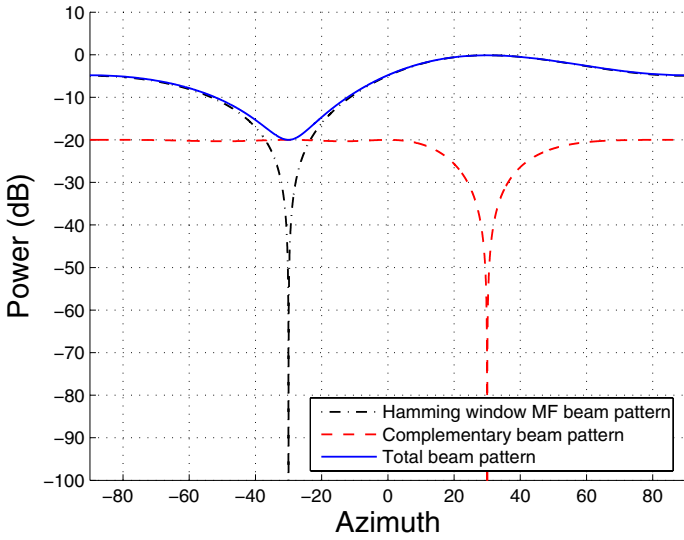
Finally, Because of inherent of the Gram-Schmidt orthogonalization, we can design arbitrary complementary beam which are orthogonal to the main beams.

## 4 Simulations

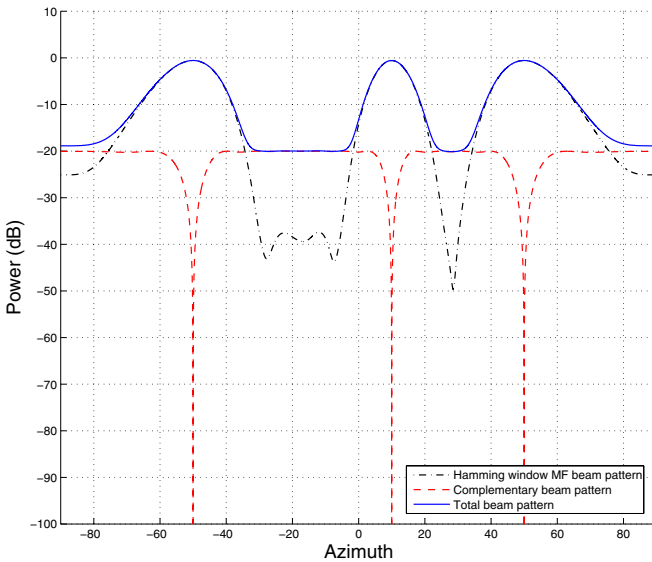
The simulation is performed by IEEE802.15.3 MBOA environments. For OFDM, we utilize 128 pt. sub-carriers and channel band 1, i.e., 3.1GHz ~ 3.628GHz.

To demonstrate the effect of designed complementary beamforming, we simulated two cases. The first simulation is for a single main beam with four transmit antennas, and the second simulation is executed for three main beams with 16 transmit antennas. To reduce the ringing of complementary beam pattern, as discussed in Sec. 3, Hamming window is used. Other windowing technique may be utilized for better results.

First, assuming that the number of transmit antennas is four and one main beam exist. For the power sharing, we let  $\varepsilon$  be 0.01. Fig. 1 displays the results of the first simulation. In Fig. 1, solid line presents the summation of the main beam and complementary beams, the dotted line denotes complementary beams, dash-dotted line denotes the main beams. As seen in Fig. 1, the main beam loses small amount of power, e.g., -20dB, between without complementary beam and with complementary beams. The gain of the main beam without complementary beam is 0 dB. Without complementary beam we observe shadowing region exist out of main beam region, i.e.,  $30^\circ$ . The complementary beams only null at the main beam angles. Thus, any user located out of main beam can detect complementary



**Fig. 1.** Beam pattern with  $N=4$ ,  $K=1$ ,  $DOA=30^\circ$



**Fig. 2.** Beam pattern with  $N=16$ ,  $K=3$ ,  $DOA=10^\circ, -50^\circ, 50^\circ$

beam power, does not try to proceed random access, and the main beam dose not have any interference.

In Fig 2, we increase the number of transmit antennas by 16 and the number of main beams is three, i.e.,  $DOA = -50^\circ, 10^\circ, \text{ and } 50^\circ$ . The  $\epsilon$  is also set 0.01 to keep -20dB, which is the same as the first scenario. The notation is the same

as in Fig. 1. Compare with Fig. 1, the main beam width of the main beam is narrower and all complementary beams need to provide three nulling angles. The proposed scheme with Hamming window demonstrated complete nulling at the main beams. For the multiple beam scenario, the proposed algorithm exhibits perfect nulling at the main beam and keeps -20dB complementary beam power out of main beams.

## 5 Conclusion

In this paper, the *intelligent* complementary beamforming method that transmits small power to out of main beam region users to prevent unwanted traffic in CSMA/CA (carrier sensing multiple access / collision avoidance) systems is presented when main beams focus on specific nodes. In order to design the intelligent complementary beamforming, Gram-Schmidt orthogonalization is utilized, whose beam pattern is always orthogonal to the main beams, and provides uniform power out of main beams with an aid of Hamming window. The simulations demonstrate the the proposed method nulls perfectly at the angle of main beams and transmit -20dB power level.

## References

- [ZOO] Zooghy, A.E.: Smart Antenna Engineering. Artech House, Boston, (2005)
- [GRO] Gross, F.: Smart Antenna for Wireless Communications. McGraw-Hill, New York, (2005)
- [STD] Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specification for High Rate Wireless Personal Area Networks (WPANs). IEEE Standard 802.15.3., (2003)
- [TAR] Tarokh, V., Choi, Y-S, Alamouti, S.: Complementary Beamforming. IEEE VTC 2003, **78** (2003) 3136–3140
- [GOD1] Godara, L.C.: Applications of antenna array to mobile communications-Part I: Performance improvement, feasibility, and considerations. *Proc. IEEE*, vol. 85, no. 7, (1997) 1031–1060
- [GOD2] Godara, L.C.: Applications of antenna array to mobile communications-Part II: BEamforming and directiona-of-arrival considerations. *Proc. IEEE*, vol. 85, no. 7, (1997) 1195–1245
- [LIB] Liberti, J.C., Rappaport, T.S.: Smart antennas for wireless communications. Englewood Cliffs, NJ:Prentice-Hall 1999

# Multiband Radio Resource Allocation for Cognitive Radio Systems

Yangsoo Kwon, Jungwon Suh, Jaehak Chung, and Joohee Kim\*

Inha University, Korea  
jchung@inha.ac.kr

**Abstract.** A multi-band radio resource allocation method in cognitive radio systems is presented. To maximize the cell throughput with minimizing outage rate, multi-band intelligent radio resource management method is described. The cell division and spectrum sharing methodologies are also applied. The numerical simulations exhibit the maximum capacity with the lowest outage rate point is located in  $r=700\text{m}$  with  $p = q$ .

## 1 Introduction

In the wireless communication systems, spectrum is getting important since many wireless applications require their own spectrums to communicate robustly. As the number of new communication standards and services emerges, more spectrums are needed. For future wireless communication systems, cognitive radio technology can search empty spectrum intelligently and increase spectrum usages. Therefore, this methodology may give a clue for the lack of spectrum in future wireless communications [MIT].

As the first step of exploring this technique, FCC (Federal Communication Committee) released NPRM (Notice of Proposal for Rule Making) by December 2003 [FCC1],[FCC2], and a standard of the cognitive radio started by November 2004 as IEEE802.22 [IEEE]. Hence, the cognitive radio technology is considered as more realistic systems for the future communication systems.

One of the advantages of the cognitive radio systems is searching the empty spectrum when the other users do not occupy their spectrums, and utilizing the multi-band spectrums if it needs. Inherently, the cognitive radio systems only use empty spectrum, and if incumbent users, who have rights to use the spectrums, start to use the spectrums, the cognitive radio systems should move their spectrums to other empty bands within a certain time. Fortunately, current smart spectrum sensing technology can detect the incumbent users quickly, and search and move to other empty spectrums.

Compared with the conventional communication systems, since cognitive radio systems utilize multi-bands if it is available. Thus, the amount of radio resource that we need to manage increases. If multiple radio channels are available, the resource management method is supported to increase trunk efficiency [LAG],[PAP].

---

\* This research was supported by Inha University Research grant, No. 32736, Korea.

In this paper, we present resource management technologies that manage the radio resource not only to increase capacity, but also reduce outage capacity in a cell using proposed criteria. To manage the resources, first we divide a cell into multiple circles and allocate multiple resources of channels to the divided region. Then, we manage using the multiple resources, e.g., capacity or SINR (Signal to Interference Noise Ratio) etc. For the specific purpose of the cell environment, we may choose one of proposed methods.

The organization of this paper is follows: in Sect. 2, radio management technologies for cognitive radios are presented. In Sect. 3, the proposed radio manage methods are described. In Sect. 4 numerical simulations are provided to present how we can apply the proposed algorithms, Finally, Sect. 5 concludes the paper.

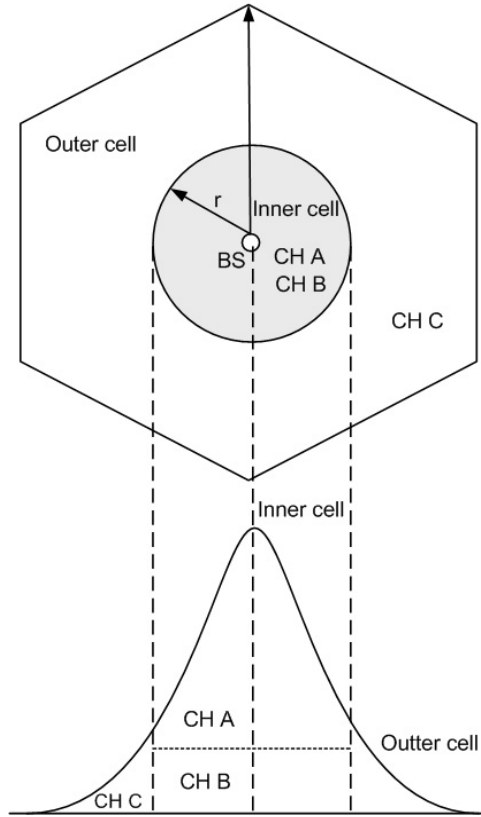
## 2 Multi-bands Radio Resource Management for Cognitive Radios

The multi-band applications, and various strategies have been developed for their own purposes. These management methods can be categorized as: Cell splitting, Reuse partitioning, hierarchical cell structure etc [LAG]. These algorithms have advantages and drawbacks. For example, if we allocate every spectrum for each cell, the SINR increases because of small intercell interference, but total throughput decreases due to small trunk efficiency. These two parameters are trade-off because of spectrum sharing problem.

In this paper, for multi-band cognitive radio systems, we assume following scenarios: Cell reuse factor is equal to one since we utilize OFDM (Orthogonal Frequency Division Multiplex), and multiple channels by reuse partitioning and hierarchical cell structure, which are effective to *hot spot* areas. Thus, the user distribution in a cell is unequal, and at the center of the cell more users are located. The requirement of radio resources for each user is fixed. The average transmit power is constant and the other BSs(Base Stations) transmit power, i.e., interference to observed BS, is the same as the observed BS. Because we have multi-bands, trunk efficiency can increases by spectrum sharing. Interference, however, may increase with the allocation of resources or transmit power among base stations. Therefore, we have to decide the best point of resource management rule. Unfortunately, the trade-off between SINR and trunk efficiency is nonlinear function.

To utilize this multi-band scenario, we need to manage the multi-band radio resources to maximize the cell throughput or reduce the outage probability in a cell. If we have a single spectrum band, it is relatively hard to handle it. However, using multi-bands we can control the resources SINR, which is closely related to capacity and outage ratio.

Fig. 1 exhibits the concept of the proposed resource allocation rule. For simplicity, in Fig. 1, we assume two multi-bands. The number of multi-bands can be increased for general scenarios. For the hot spot cell, the user distribution may be modeled as Gaussian, and the each user requires the same amount of resources in average sense. We divide the cell using inner cell region and outer cell region similar to reuse partitioning. This division provides interference control because of distance between two spectrum sharing BSs. Thus, inner cell is provided by



**Fig. 1.** Cell splitting with inner circle and allocation of multiple channels, e.g., CH A, CH B, CH C

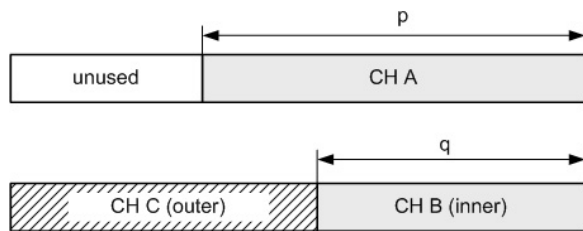
two channels, i.e., CH A, CH B, and outer region is supported by CH C. In general, for reuse partitioning, the resource are allocated from the inner cell and allocate outer cell. If inner cell user increases and requires more resources, we may allocate the resources using outer cell resources if it is available. In this paper, we assume the amount of resources is enough. In Fig. 2, each channel is shown and the allocated portion is presented as  $p$  and  $q$  for CH A, CH B, respectively. If the resource requirement is larger than the resources or SINR (Signal-to Interference noise ratio) is not satisfied with the requirements, we adjust the inner cell size, i.e.,  $r$ , to allocate the resources using multi-bands. In other words, resources can be allocated by adjusting cell size, which provides trunk efficiency.

To measure the throughput and outage, SINR needs to be calculated first, which is defined as

$$SINR_{BS_k} = \frac{P_{BS_k}}{\sum_{i=1, i \neq k}^N P_{BS_i} + noise_k}, \tag{1}$$

where  $k$  denotes the BS index and  $i$  denotes the interference cell index.





**Fig. 2.** Resource allocation with multiple channels with  $p, q$  for CH A, CH B, respectively

For simplicity, we assume the other cells utilize the same amount of resource allocation rule as in the current cell. If other cell resource allocation statistics are varying, it is harder to evaluate SINR. This, however, is reasonable assumption in the statistical point of view. Therefore, if we increase the amount of allocation to a CH A, the other cell also increases the amount of allocation which causes the interference to the current cell. In turn, decreasing the amount of allocation, it decreases interference to current cell. In this case, the SINR can be measured using Eq. (1). Since the relationship between increasing allocation and variation of SINR is nonlinear function, it is hard to analyze analytically. In this paper, we provide the resource allocation rules first, and using simulation, find the allocation parameters based on different criterions.

### 3 Some Allocation Criteria of the Proposed Algorithm

As mentioned in Sect. 2, we can set different rules to allocation radio resources. First, if we maximize the cell throughput, we can set the following equation as,

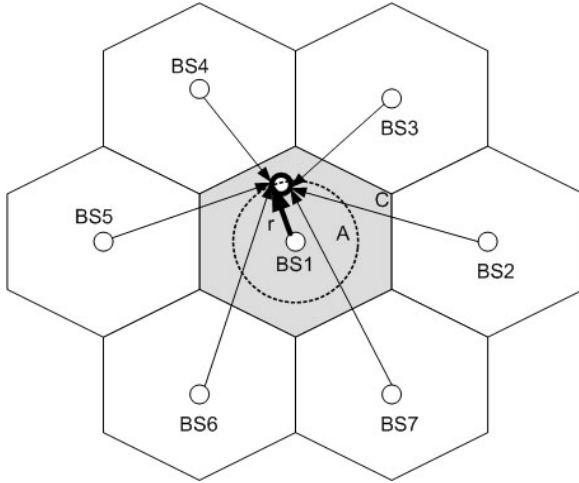
$$\begin{aligned}
 SINR_{opt}(p, q, r) &= \arg \max_{p, q, r} \sum_{i=1}^N SINR_i(p, q, r) & (2) \\
 \text{for } & 0 \leq p \leq 1, 0 \leq q \leq 1, 0 \leq p + q \leq 1,
 \end{aligned}$$

where  $N$  denotes the number of channels. We need to decide  $p, q$ , and  $r$  satisfying Eq. (2). For the calculation of the capacity from Eq. (2) can be obtained as

$$\begin{aligned}
 C_{opt}(p, q, r) &= \arg \max_{p, q, r} \sum_{i=1}^N C_i(p, q, r) & (3) \\
 \text{for } & 0 \leq p \leq 1, 0 \leq q \leq 1, 0 \leq p + q \leq 1,
 \end{aligned}$$

where  $\mathcal{C}$  is theoretically obtained from  $\mathcal{C} = \log_2(1 + SINR)$ . For more precise calculation for a specific systems, we need to utilize their MCS (Modulation Coding Scheme) tables. To calculate the total throughput of the cell is given by

$$C_{tot} = C_A + C_B + C_C. \tag{4}$$



**Fig. 3.** SINR calculation for multi-cell environment

As mentioned in Sect. 2, maximizing capacity is not the only optimum solution, but outage probability to keep QoS (Quality of Service) needs to be considered. Because of region division and allocation of channels separately, we may control the amount of interference, which may prevent the trunk efficiency. For the outage capacity, we need to apply different criteria as

$$P_{out} = P(C \leq C_{THR})(p, q, r) \tag{5}$$

Therefore, to maximize the capacity with minimum outage probability, we have to combine the criterion with Eq. (3) and Eq. (5). Then, we write the equation as

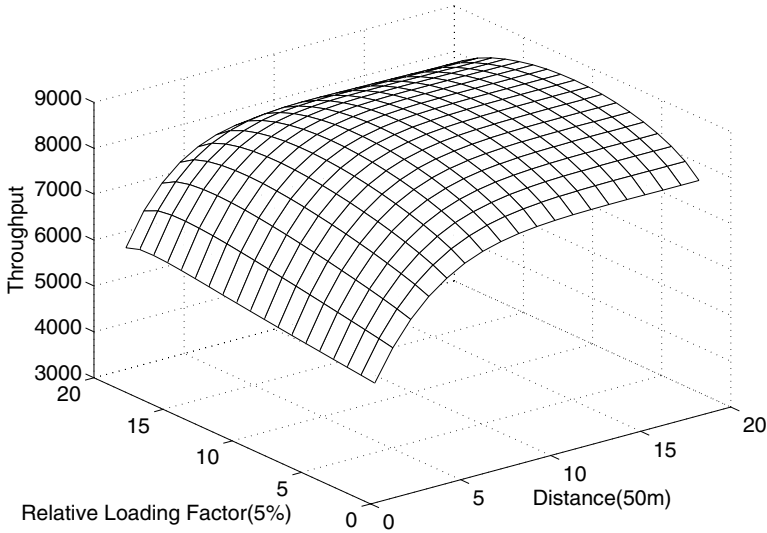
$$C_{opt}(p, q, r) = \arg \max_{p, q, r} \sum_{i=1}^N C_i(p, q, r) \tag{6}$$

$$\begin{aligned} & s.t. \quad \min P(C \leq C_{THR})(p, q, r) \\ & for \quad 0 \leq p \leq 1, 0 \leq q \leq 1, 0 \leq p + q \leq 1. \end{aligned} \tag{7}$$

To evaluate the above equations, we exhibit the numerical results in Sect. 4.

## 4 Simulations

To show the decision rule of Eq. (6), we assume that the user distribution is Gaussian, and each user requires the same amount of resource. For simplicity, OFDM is utilized for the simulation and one subcarrier is allocated to each user. The cell size is 1000 m, which is similar to WiMax, e.g., IEEE802.16d/e standard. The carrier frequency is 2.3GHz. The user is located in every 100m ring. For total resource loading, we allocated 80% of one OFDM symbol and the



**Fig. 4.** BER performance of the proposed HR-STBC and the conventional O-STBC with two transmit antennas

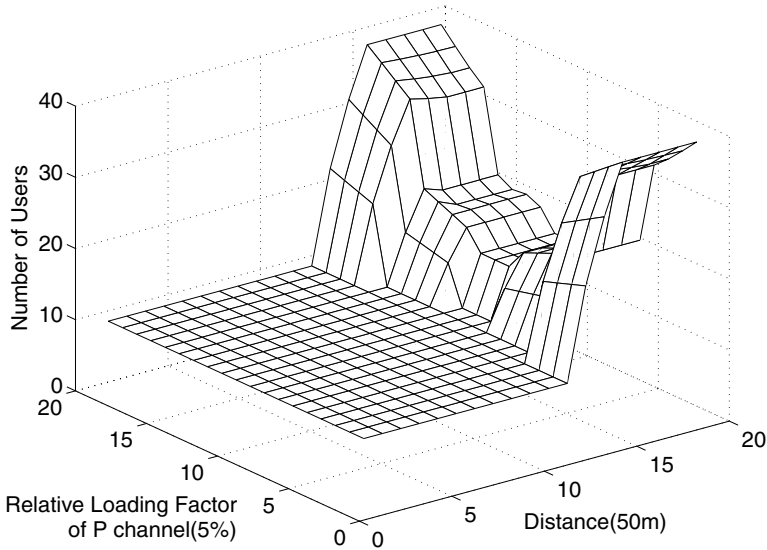
number of user is the same as the number. Two channels are available. In Fig. 1, we show the distribution of user. Assuming that shadowing effect is omitted for measuring SIR, e.g., only path loss is considered. For calculation of the path loss, the following channel is utilized as [ITU],

$$\begin{aligned}
 PL(r) = & 40 \times (1 - 4 \times 10^{-3} \Delta h_b) \times \log_{10}(r) - 18 \times \log_{10}(\Delta h_b) \\
 & + 21 \times \log_{10}(f_C) + 80,
 \end{aligned}
 \tag{8}$$

where  $D$  denotes a distance between BS and users (km),  $\Delta h_b$  denotes height of BS (m), and  $f$  denotes carrier frequency (MHz).

Fig. 4 presents the results of the cell capacity, e.g., throughput, using Eq. (3). In Fig. 4, the radius denotes distance of the  $r$  from the cell center, and the relative loading factors denotes loading factor of  $p$ . The capacity increases as the  $r$  increases. After a certain range, the capacity is increasing dramatically. The center of the loading factors,  $p$  and  $q$ , exhibits the maximum point at the given  $r$ . From this figure, trunk efficiency plays a great role in maximizing capacity. In other words, when the degree of allocation has maximum freedom, we obtain the maximum capacity. The maximum throughput is given at  $r=800m$ ,  $p=q$  since the throughput shape is concave.

In Fig. 5, the outage rate is shown when the required SINR is equal to 4.5 dB. The horizontal axes are the same as in Fig. 4. In Fig. 5, we observe that the smallest point of the outage rate is not matched with the value of the maximization of the capacity. The smallest number of the outage is located about  $r=700m$ . The vertical axis denotes the number of outage users. The minimum outage rate is obtained wide regions. However, using Eq. (6), the maximum capacity with



**Fig. 5.** Multi-band resource management

the smallest outage can be chosen at  $r=700\text{m}$ ,  $p=q$ . Since the division of the region and at a cost of trunk efficiency, we can reduce the outage rate.

## 5 Conclusion

Intelligent radio resource management method to maximize the cell throughput such that minimizing outage rate is present. To maximize the capacity with a QoS, we consider not only maximizing the capacity, but also minimizing outage capacity. In this paper, cell division is utilized and spectrum sharing method also applied. The numerical simulation exhibits the maximum capacity with the lowest outage rate point is located in  $r=700\text{m}$  with  $p = q$ .

## References

- [MIT] Mitola, J.: Cognitive radio for flexible multimedia communications. Proc. of IEEE workshop on Mobile Multimedia Comm. (1999) 3–10
- [FCC1] FCC. : Spectrum policy task force report. ET Docket No.02-155, Nov. (2002)
- [FCC2] FCC. : ET Docket No. 03-322: Notice of Rule Making and Order. (2003)
- [IEEE] Stevenson, C.:Functional requirements for the IEEE802.22 WRAN standard. IEEE802.22 draft (2005)
- [LAG] Lagrange, X. : Multitier Cell Design. IEEE Comm. Mag. (1997) 60–64
- [PAP] Papavassilou, S., Tassiulas, L.:Improving the capacity in wireless Networks through integrated channel base station and power assignment. IEEE tran. Veh. Tech. (1999) 417–427
- [ITU] Recommendation ITU-R.M. 1225 :Guidelines for Evaluation of Radio Transmission technologies for IMT-2000. ITU, (1997)

# Testing Online Navigation Recommendations in a Web Site

Juan D. Velásquez<sup>1</sup> and Vasile Palade<sup>2</sup>

<sup>1</sup> Department of Industrial Engineering,  
University of Chile, Chile  
jvelasqu@dii.uchile.cl

<sup>2</sup> Computing Laboratory, University of Oxford, UK  
Vasile.Palade@comlab.ox.ac.uk

**Abstract.** An online navigation recommendation system provides the prospective web site visitor with a set of pages that could be of his/her interest. Because the recommendations are given during the user session in the web site, it could be very damaging for the overall business of the company owning the web site, if the recommendations are erroneous. In this paper, we introduce an a priori method to estimate the success of an online navigation recommendation. The methodology was tested in a recommendation system that works with the data generated in a real web site, which proved the effectiveness of our approach.

## 1 Introduction

An efficient way to improve the relation between the web site and its users is by the personalization of the site structure and content, i.e., “*actions that tailor the web experience to a particular user, or a set of users*” [6,10].

The web site personalization can be implemented by recommendations done directly to the web site users, which can be web masters, web designers, anonymous web visitors; in short, any users of the web site [5,13].

Depending on the web user, the recommendations can be:

- Online - Provide the web site visitor, during his/her session in the site, with recommendations about interesting topics, for instance a web page, text contents, etc.
- Offline - Done mainly to the web master, web designers and, in general, to any person in charge of maintaining the structure and the content of the web site.

This paper is concerned with how to do online recommendations, especially those that are relative to the user navigation in the web site.

When a user visits a web site, the browsing behavior is related to what he/she is looking for. Identifying the user’s behavior usually becomes a true challenge, because the site doesn’t consider the particular needs of each user, e.g. if the user is an amateur in the web and needs assistance to find the desired information.

An online navigation recommendation system helps the user in searching for the desired information, through suggestions about web pages whose content could be of interest to that particular user [13].

Because these kind of systems give the recommendations during the user session in the web site, in some cases, it may happen that, if the user does not like the recommendations done by the system, he/she may decide to leave the site. This is a big problem, and our task of providing online recommendations could be a risky one for the business of the company owning the web site, if the recommendations are not done correctly.

In this paper, we introduce an a priori method to estimate the success of an online navigation recommendation. The paper is organized as follows. In Section 2, a short review about related research work is done. Section 3 shows the main characteristics of the online recommendation system that was used for testing the proposed methodology. This methodology is introduced in Section 4, and in Section 5 its application to a real-world case is presented. Finally, Section 6 presents the main conclusions and future work.

## 2 Background and Related Work

Web personalization approaches aim to personalize the user interaction with a web-based system, such as a web site, by understanding the user's explicit or implicit interests and desires.

### 2.1 Classifying the Personalization Approaches

A good classification of the web personalization approaches, considering functionality reasons, is presented in [8] and outlined below:

**Memorization.** This is the simplest expression of personalization, and it implies the storing of basic information about the user, such as name and visited pages. When the user visits again the web site, the system recognizes the user and shows the user's name and part of the last visit(s).

**Guidance.** It represents approaches that assist the user for finding what the user is looking for. In this sense, the personalization systems can recommend links to other pages and related contents.

**Task performance support.** It involves actions on behalf of the user, like sending e-mails, complete queries, or, in advanced systems, represent the user's interest, for example in a negotiation.

The common requirement in any web personalization system is to be provided with some mechanics to understand the user behavior in a web site. The web usage mining algorithms [3] are the current techniques used for this purpose.

### 2.2 Web Site Changes and Recommendations

In creating a web site, a complex and meticulous process is followed, in order to get the best look and feel. Consequently, off-line changes in its structure and

content are not frequent, even in web sites prone to changes, such as newspapers web sites. Although every day the newspaper's pages change, the structure and the main themes follow an editorial line.

If off-line changes could be a bit risky, the on-line changes are even more, because the visitors may lose the notion about "where they are" in the web site. In fact, some changes can violate the original web site links structure, i.e., there may be no physical link between the actual and the imposed page. When the visitor revisits the web site and wants to review the imposed page again, it may be hard for him/her to get it directly.

In order to avoid the above described problem, some authors [2] have proposed on-line recommendation systems that maintain the web site structure, i.e., there must be a physical link between the actual page and the recommended page. Those recommendations that break the structure should be considered for off-line changes.

### 2.3 Creating the Online Navigation Recommendations

Facing a high competition for catching or retaining customers in the digital market, an important way for companies to generate value is to understand, within the web site users sessions, what the users are looking for. In other words, "*users need to feel they have a unique personal relationship with the business*" [4].

A navigation personalization system assists the user by making recommendations about contents that could be of interest for him/her.

Independently of the technology used in the creation of the system, a common element is the representation of the knowledge about the user browsing behavior in the web site [1]. This knowledge can be extracted from usage data, like web log files [5,14,13], and from the experience of the web site users, for instance through questionnaires about the usability of the web site [7].

The online navigation recommendation is created by making a comparison between the current user session and the patterns and rules that represent the knowledge about the user browsing behavior in the system. Then, a set of web page links are showed to the user, who can choose to follow the recommendation(s) or continue with his/her own navigation.

Here, the question is: what happens if the recommendation is erroneous and the user is lost in the hyperspace? Naturally, a bad recommendation would get the relationship between the web site and its users worse, and it represents a high risk for the business.

### 2.4 Considerations for Testing an Online Recommendation

Usually, the offline structure and content changes in a web site can be reviewed through an "usability test" [7], where a group of selected users review the web site structure and fill a questionnaire with their impressions.

The same methodology could be used in the case of online recommendations using a group of simulated visitors, but it will not be a real situation, as the group of simulated visitors may usually have no incentive to search for some topics and show a real visitor behavior.

In [16], an association rule algorithm is applied for predicting the web log access. The effectiveness test is realized by comparing the prediction with the real navigation of a group of visitors, whose web log access is previously known. The problem with this method is that it needs the visitors' reaction for measuring the effectiveness of the prediction. A similar approach is applied in [17], where the test group is selected based on similar educational background and common interests. The web log registers are generated by giving to the users a search task about a specific topic. Then, the path prediction aims to help the users to find information related to that topic. The problem in this method is that a real test needs real users, i.e., persons with a real intention of searching a topic, without an external influence.

Independent of the navigation recommendation prediction method, the real test will always be when real users visit the web site, which could be very risky for the business. Therefore, an a priori effectiveness test will help improve the recommendation.

### 3 The Online Navigation Recommendation System

In order to test the proposed methodology, the online recommendation system introduced in a previous paper [12] is used. The main characteristics of this system are shortly detailed below.

#### 3.1 Modeling the User Browsing Behavior in a Web Site

Our user behavior model uses three variables: the sequence of visited pages, their text contents and the time spent on each page. The model is based on a  $n$ -dimensional visitor behavior vector which is defined as follows.

**Definition 1 (User Behavior Vector).** *It is defined as a vector  $v = [(p_1, t_1) \dots (p_n, t_n)]$ , where the pair  $(p_i, t_i)$  represents the  $i^{\text{th}}$  page visited ( $p_i$ ) and the percentage of time spent on it within a session ( $t_i$ ), respectively.*

#### 3.2 Comparing User Sessions

Let  $\alpha$  and  $\beta$  be two user behavior vectors of dimension  $C^\alpha$  and  $C^\beta$ , respectively. Let  $\Gamma(\cdot)$  be a function that returns the navigation sequence corresponding to a visitor vector. A similarity measure has been proposed elsewhere to compare visitor sessions, as follows [15]:

$$sm(\alpha, \beta) = dG(\Gamma(\alpha), \Gamma(\beta)) \frac{1}{\eta} \sum_{k=1}^{\eta} \tau_k * dp(p_{\alpha,k}, p_{\beta,k}) \quad (1)$$

where  $\eta = \min\{C^\alpha, C^\beta\}$ , and  $dp(p_{\alpha,k}, p_{\beta,k})$  is a similarity measure for comparing the free text inside two web pages [11], in this case between the  $k^{\text{th}}$  page of vector  $\alpha$  and the  $k^{\text{th}}$  page of vector  $\beta$ . The term  $\tau_k = \min\{\frac{t_{\alpha,k}}{t_{\beta,k}}, \frac{t_{\beta,k}}{t_{\alpha,k}}\}$  is an indicator of the visitor's interest in the visited pages. The term  $dG$  is the similarity between the sequences of pages visited by two visitors [9].



### 3.3 Extracting Knowledge from Web Logs

The knowledge extracted can be represented as patterns, and rules about how to use the patterns. Because the users' behavior can be grouped based on similar preferences, it looks convenient to apply a clustering technique in order to extract the navigation patterns.

For clustering the user sessions, a Self-Organizing Feature Map (SOFM) [14] was applied by using the similarity measure given in Eq. 1. The SOFM requires vectors of the same size. Let  $H$  be the dimension of the visitor behavior vector. If a visitor session has less than  $H$  elements, the missing components up to  $H$  are filled with zeroes. Otherwise, if the number of elements is greater than  $H$ , only the first  $H$  components are considered.

### 3.4 Creating the Online Navigation Recommendation

Let  $\alpha = [(p_1, t_1), \dots, (p_m, t_m)]$  be the user behavior vector that corresponds to the current user session and  $C_\alpha = [(p_1^\alpha, t_1^\alpha), \dots, (p_H^\alpha, t_H^\alpha)]$  the closest centroid, such as  $\max\{sm(\alpha, C_i)\}$ , with  $C_i$  the set of centroids discovered. The recommendations are created as a set of pages whose text content is related to  $p_{m+1}^\alpha$ .

Let  $R_{m+1}(\alpha)$  be the online navigation recommendation for the  $(m+1)^{th}$  page to be visited by the user  $\alpha$ , where  $\delta < m < H$  and  $\delta$  is the minimum number of visited pages necessary to prepare a suggestion. Then, we can write  $R_{m+1}(\alpha) = \{l_{m+1,0}^\alpha, \dots, l_{m+1,j}^\alpha, \dots, l_{m+1,k}^\alpha\}$ , with  $l_{m+1,j}^\alpha$  the  $j^{th}$  page link suggested for the  $(m+1)^{th}$  page to be visited by visitor  $\alpha$ , and  $k$  the maximum number of pages that can be suggested. In this notation,  $l_{i+1,0}^\alpha$  represents the "no suggestion" state.

## 4 The Proposed Methodology for Testing Online Recommendations

Using the discovered clusters as outlined above, we can classify the user browsing behavior into one of them, by comparing the cluster centroid with the current user session, using the similarity measure introduced in Eq. (1).

Here, we propose a method to test the effectiveness of the recommendations, based on the same kind of web data used in the pattern discovery stage [14]. More exactly, a percentage of all available web data is used to extract significant patterns about user behavior, and for these we define a set of rules. Then, we test the effectiveness of the recommendations using the remaining web data.

Let  $ws = \{p_1, \dots, p_n\}$  be the web site and pages that compound it. We can define some equivalence classes of pages, where the pages belonging to the same class contain similar information. The classes partition the web site in disjoint subsets of pages. Let  $Cl_x$  be the  $x^{th}$  equivalence class for the web site  $ws$ .  $\bigcup_{x=1}^w Cl_x = ws$ , where  $w$  is the number of equivalence classes.

Let  $\alpha = [(p_1, t_1), \dots, (p_H, t_H)]$  be a user behavior vector from the test set. Based on the first  $m$  pages actually visited, the proposed system recommends for the  $(m+1)$  page several possibilities, i.e., pages to be potentially visited.

We test the effectiveness of the suggestions made for the  $(m + 1)^{th}$  page to be visited by the visitor  $\alpha$  following this procedure. Let  $Cl_q$  be the equivalence class for  $p_{m+1}$ ; if  $\exists l_{m+1,j}^\alpha \in R_{m+1}(\alpha)$  and  $l_{m+1,j}^\alpha \in Cl_q$ ,  $j > 0$ , then we assume the suggestion was successful.

The number of recommended pages obtained during the construction of the recommendation could be high, and the user may get confused on which page to follow next. We set in  $k$  the maximum number of pages per recommendation. By using the page similarity measure (see Eq. (1)), we can extract from  $Cl_q$  the  $k$  closest pages to  $p_{m+1}$  in the recommendation, as follows:

$$E_{m+1}^k(\alpha) = \{l_{m+1,j}^\alpha \in sort_k(dp(p_{m+1}, l_{m+1,j}^\alpha))\}. \quad (2)$$

The “ $sort_k$ ” function sorts the result of similarity measure  $dp$  in descending order and extracts the “ $k$ ” link pages closest to the  $p_{m+1}$  page. A particular case is when  $E_{m+1}(\alpha) = \{l_{m+1,0}^\alpha\}$ , i.e., no suggestion is proposed.

The methodology proposed here allows to work with real user sessions and estimate the effectiveness of the online recommendations. It represents an alternative to testing the user answer in front of a navigation recommendation.

## 5 A Real World Application

We worked with a recommendation system that used data originated in the web site of the first Chilean virtual bank ([www.tbanc.cl](http://www.tbanc.cl)). The web data we used were collected between January and March 2003. Approximately eight millions of raw web log registers were collected.

After applying a cleaning and session reconstruction process, approximately 100,000 user behavior vectors were selected.

### 5.1 Applying Clustering Techniques

The SOFM used for clustering had 6 input neurons (that is,  $H = 6$  in the user behavior vector) and 32 X 32 output neurons with a toroidal topology in the feature map.

From the extracted user behavior vectors, we only considered those that contained six or more visited pages. With this restriction, around 65000 vectors were used in the experiment. A 75% of these vectors were applied to the SOFM to extract navigation patterns, and the remaining 25% formed the testing set.

The cluster identification is performed by using a visualization tool supported by a density cluster matrix, called winner matrix. It contains the number of times the output neurons win, during the training of the SOFM.

Table 1 contains the 4 discovered cluster centroids, represented by the sequence of visited pages and the the time spent on each centroid. The pages were previously labelled with a correlative number.

Analyzing the cluster “A”, we can see that the users are interested in the content of pages number 10 and 135. These pages contain, respectively, a general description of bank’s soft credit products (low interest rate) and information

**Table 1.** Visitor behavior clusters

Cluster	Pages Visited	Time spent in seconds
A	(1,5,7,10,135,191)	(30,61,160,110,175,31)
B	(162,157,172,114,105,2)	(3,71,112,110,32,3)
C	(72,87,154,188,140,85)	(8,57,31,3,71,91)
D	(110,104,128,126,31,60)	(25,73,42,65,98,15)

about a specific credit card. Then, we can infer that the users are looking for information about soft credits, i.e., small amount of money to borrow, short period of repayment and low interest rate.

## 5.2 Representing the Knowledge as Patterns and Rules

Table 2 shows an example about how the recommendation is created. Under the assumption that the cluster more similar to the current user session is  $C_A$ , the recommendation for the fourth page to be visited is prepared.

In this example, “S” is a stack implemented using a simple linked list that contains the pages to be suggested for the fourth page to be visited (navigation component), and the usage statistics associated with each of them (statistics component).

If a page that we are recommending does not exit in the current web site anymore (because of previous changes), the function **compare\_page** will compare the suggested page with all pages in the entire web site, in order to recommend a page with similar content. The function **Pop(S)** extracts the next element in “S”. The final set of pages is the list “L”. **Extracted\_Three\_Links** represents

**Table 2.** Extracting patterns and rules

```

 $C_A \rightarrow [(1,30),(5,61),(7,160),(10,110),(135,175),(191,31)]$ 
 $\alpha \rightarrow [(2, 10), (11, 50), (25, 120)]$  % current visitor
ws  $\rightarrow \{p_1, \dots, p_{217}\}$  % current web site pages
S.navigation  $\rightarrow \{p_{33}, p_{38}, p_{41}, p_{118}, p_{157}, p_{201}\}$ 
S.statistic  $\rightarrow \{1.2, 2.1, 1.8, 0.9, 0.8, 0.1\}$ 
Case  $C_A$  and SuggestionPage=4 :
  Prepare_suggestion( $p_0, 0, L$ ); % default “no suggestion”
% L: link page suggestion, 0: statistic associated
while S not null loop
  if (S.navigation not in ws) then
    S.navigation = compare_page(ws, S.navigation);
  else if ((S.navigation <>  $\alpha_{p_{1..3}}$ ) and (S.statistic >  $\gamma$ )) then
    Prepare_suggestion(S.navigation, S.statistic, L);
    Pop(S); % Next element in S
  end if;
end loop;
send(Extracted_Three_Links(L)); %  $L \rightarrow \{p_{38}, p_{41}, p_{33}\}$ 

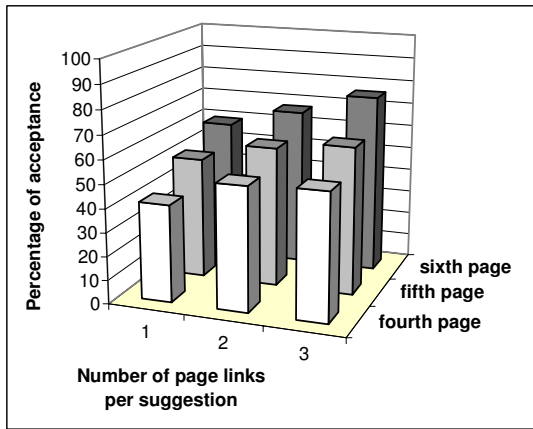
```

the expression in Eq. 2 and extracts a subset with a maximum of three links, based on the associated statistics. The default is the no suggestion state. **Send** is an instruction for sending the recommendation to the system who will prepare the final web object containing the recommendation for the user.

### 5.3 Results

In Fig. 1, the histogram shows the percentage of the accepted recommendations, using the proposed validation method.

If, by using the proposed methodology, just one page is suggested, slightly more than 50% of the users would accept it. This could be considered a very successful suggestion by the business expert, since we are dealing with a complex web site with many pages, many links between pages, and a high rate of visitors that leave the site after few clicks.



**Fig. 1.** Percentage of acceptance of online navigation recommendations

Furthermore, it should be mentioned that the percentage of acceptance would probably have been even higher if we actually had suggested the respective page during the real session. Since we compared past visits stored in log files, we could only analyze the behavior of visitors that did not actually receive any suggestion we proposed.

## 6 Conclusions

We introduced a methodology to a priori test the effectiveness of a system that provides online navigation recommendations for the prospective visitors of a web site.

The methodology used a percentage of all available web data to extract the navigation patterns, and the remaining data to test the effectiveness of the

recommendation. With this procedure, an estimation on the success of the recommendation can be calculated.

This methodology was tested in a recommendation system that used data originated in a real web site. There is good evidence to suggest that, if this approach is used, the users will follow the recommendation in a high percentage of them.

As future work, it is interesting to compare our estimation with the real situation, i.e., to apply the recommendation on real users and to analyze its acceptance or rejection, in order to see if the estimation was correct or not.

## References

1. P. Brusilovsky. Adaptive web-based system: Technologies and examples. Tutorial, IEEE Web Intelligence Int. Conference, Halifax, Canada, October 2003.
2. F. Coenen, G. Swinnen, K. Vanhoof, and G. Wets. A framework for self adaptive websites: tactical versus strategic changes. In *Procs. in 4th PAKDD Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 1–6, April 2000.
3. M. Eirinaki and M. Vazirgiannis. Web mining for web personalization. *ACM Trans. Inter. Tech.*, 3(1):1–27, 2003.
4. A. Kobsa, J. Koenemann, and W. Pohl. Personalised hypermedia presentation techniques for improving online customer relationships. *Knowledge Engineering Review*, 16(2):111–155, 2001.
5. B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.
6. B. Mobasher, T. Luo, Y. Sung, and J. Zhu. Integrating web usage and content mining for more effective personalization. In *Procs. of the Int. Conf. on E-Commerce and Web Technologies*, pages 165–176, September, Greenwich, UK, 2000.
7. J. Nielsen. User interface directions for the web. *Communications of ACM*, 42(1):65–72, 1999.
8. D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos. Web usage mining as a tool for personalization: A survey. *User Modeling and User-Adapted*, 13:311–372, 2003.
9. T. A. Runkler and J. Bezdek. Web mining with relational clustering. *International Journal of Approximate Reasoning*, 32(2-3):217–236, Feb 2003.
10. S. Sae-Tang and V. Esichaikul. Web personalization techniques for e-commerce. *Lecture Notes in Computer Science*, 2252(1):36–44, 2001.
11. G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM archive*, 18(11):613–620, November 1975.
12. J. D. Velásquez, P. Estévez, H. Yasuda, T. Aoki, and E. Vera. Intelligent web site: Understanding the visitor behavior. *Lecture Notes in Artificial Intelligence*, 3213(1):993–1003, 2005.
13. J. D. Velásquez, R. Weber, H. Yasuda, and T. Aoki. Acquisition and maintenance of knowledge for web site online navigation suggestions. *IEICE Transactions on Information and Systems*, E88-D(5):993–1003, May 2005.
14. J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. Acquiring knowledge about users's preferences in a web site. In *Procs. 1<sup>th</sup> IEEE Int. Conf. on Information Technology: Research and Education*, pages 375–379, Newark, New Jersey, USA, August 2003.

15. J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. A new similarity measure to understand visitor behavior in a web site. *IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization*, E87-D(2):389–396, February 2004.
16. H. Yang, S. Parthasarathy, and S. Reddy. On the use of constrained association rules for web mining. In *Procs. Int. Conf. WebKDD workshop on Knowledge Discovery in the Web*, pages 77–90, Edmonton, Alberta, Canada, 2002.
17. T. Zheng and R. Goebel. An ncm-based framework for navigation recommender systems. In *Procs. 19<sup>th</sup> Int. Conf. Artificial Intelligence workshop Multi-Agent Information Retrieval and Recommender Systems*, pages 80–86, Edinburgh, Scotland, UK, July 2005.

# An Overview of Agent Coordination and Cooperation

Angela Consoli<sup>1</sup>, Jeff Tweedale<sup>2</sup>, and Lakhmi Jain<sup>1</sup>

<sup>1</sup> School of Electrical and Information Engineering, Knowledge Based Intelligent Engineering Systems Centre  
University of South Australia, Mawson Lakes, SA 5095, Australia  
{Angela.Consoli, Lakhmi.Jain}@unisa.edu.au  
<http://www.kes.unisa.edu.au>

<sup>2</sup> Airborne Mission Systems, Defence Science and Technology Organisation  
[Jeffrey.Tweedale@dsto.defence.gov.au](mailto:Jeffrey.Tweedale@dsto.defence.gov.au)

**Abstract.** By their very nature, intelligent agents possess four important social abilities. These include the ability to communicate, cooperate, collaborate and the need to be coordinated. This paper presents an overview of two of these social abilities, that of being coordination and cooperation. The discussion develops the theory of each and derives the current definitions. The definitions will then be linked into a single multi-agent system (MAS) model, Agent Coordination and Cooperation Cycle Model. This shows a cognitive loop that replicates the link between coordination and cooperation in systems such as organizations, management and biological systems. This paper will also present the advantages, consequences and challenges associated with the implementation of Agent Coordination and Cooperation Cognitive Model (AC<sup>3</sup>M) within intelligent multi-agent systems.

## 1 Introduction

There are many applications where agents are designed or based on human behavior. Most have some form of HCI that communicate in a human-like fashion. The personification of agents is possible using a single agent [1] and is proving more successful in MASs [2].

Bratman (1990) introduces Practical Reasoning and discusses how personification can lead to an agent's ability to possess desires, values, cares and beliefs [6]. Rao and Georgeff (1995) further explain this definition based on an agent's actions [7] in order to achieve its objectives. The essential characteristic of the state of the environment is representative of the informative component and is seen as the *beliefs* of an agent, while priorities and rewards associated with the completion of tasks are seen as an agent's *desires* to complete a specific goal.

Finally, the current chosen course of action provides an agent with a deliberate component and can be seen as the *intentions*. Wooldridge (2002) illustrates how these characteristics combine into the Beliefs, Desires and Intentions (BDI) framework [8] called the Practical Reasoning System Architecture.

The introduction of the personification into agents has set the stage for the research in agent interaction [1, 2], and has been heavily focused in the social abilities of coordination and cooperation. The decentralised nature of personification enables the development and implementation of organisational structures and behavioural strategies using coordination and cooperation within MASs [2].

## 2 Multi-agent Interaction

Broadly speaking, interaction can be defined as the formation of a dynamic relationship of two or more agents through their influential actions [3]. Interaction between agents occurs either through direct or indirect contact in a mutual environment. Communication is an integral part of interaction but does not have to be direct. It can be indirect by means of a resulting action. Communication in MAS can be implemented either as message passing or data transactions to the agent or its environment [4]. Weiß (1999) also include competition and negotiation as important attributes of interaction [5]. This paper concentrates on cooperation and coordination in MAS.

## 3 Coordination and Cooperation Theory

### 3.1 Introduction to Coordination Theory

Ehlert and Rothkrantz (2001) discuss a simple way of managing coordination via task allocation methods [4]. They classify task allocations as: centralized, distributed or emergent in nature. Using centralized allocation, one central ‘leader’ conducts task distribution either by imposing tasks upon agents (hierarchical downwards through coordination or delegation) or by trading/brokering tasks (hierarchical upwards through cooperation or liaison). Using distributed task allocation, each agent attempts to obtain the services it requires from other agents either by sending requests to agents whom it knows have the required services or by sending requests to all agents and accepting the best offer. Alternatively emergent cooperation, is the characteristic of reactive systems, where each agent is designed to perform a specific task, therefore no negotiation is necessary. It is important to note that task allocation may not always be adequate in attaining a goal as some tasks may require additional coordination methods such as planning, synchronization and learning.

Prior to this Malone and Crowston (1990) focused on the notion of coordination theory by describing how the actors or objects work together harmoniously [9]. These key arguments include the need to subdivide goals and deciding how these actions can be assigned to one or more agent. They also identified resources usage as a critical factor for success [9] and defines coordination as: “... *the act of working together harmoniously ...*” By refining the definition of each keyword a simple, but powerful definition of coordination is established as:

*“The act of managing interdependencies between activities performed to achieve a goal” [8].*

Based on this definition, four distinct components of coordination must be examined. They include actors, activities, goals and management of interdependencies [9]. More importantly, three interaction processes need to be clarified: group-decisions, communication and the perception of common objects [9]. Malone and Crowston (1990) provide the taxonomy of these components in Table 1.



**Table 1.** Components of Coordination

<i>Process Level</i>	<i>Components</i>	<i>Examples of Generic Processes</i>
Coordination	Goals, activities, actors, resources, interdependencies.	Identifying goals, ordering activities to actors, allocating resources, synchronizing activities.
Group decision making	Goals, actors, alternatives, evaluations, choices.	Proposing alternatives, evaluating alternatives, making choices.
Communication	Senders, receivers, messages, languages.	Establishing common languages, selecting receiver, transporting message.
Perception of common objects	Actors, objects.	Seeing same physical objects, accessing shared database.

### 3.2 Coordination of Agents

Based on the working definition of coordination, agent coordination within a MAS is the act of managing interdependencies between agents' activities, which are performed to achieve a system goal<sup>1</sup>.

The advantages of coordination allow agents to specify and achieve a set of goals. It also provides a group of agents the ability to aspire to desired properties, such as coherency, and completion of plans and actions to achieve these goals [8, 10].

However, there have been problems. Nwana, Lee and Jennings (1996) provide some drawbacks relating to the lack of flexibility of coordination models in current applications due to erroneous assumptions about an agent's behaviour [11]. Furthermore, they state that current models do not take into consideration an agent's ability to conduct complex reasoning and have difficulty in validating the strategies used.

### 3.3 Introduction to Cooperative Theory

As with coordination, the term cooperation has multiple definitions. A universally accepted definition of cooperation is *acting together with a common purpose* [13]. In simple terms, cooperation is achieved when a number of persons enter a relationship with others for a common benefit or collective action in the pursuit of the common well-being. Cooperation requires an actor or an object belonging to a community to willingly share their knowledge. However, cooperation requires a group of actors or objects to make a voluntary association for a mutual benefit.

<sup>1</sup> Agent coordination is an important aspect of a MAS because, as with human society, agents need to be coordinated so they will act desirably. With the generic processes which Malone and Crowston (1990) have identified in their Coordination Theory, the three main reasons for agent coordination are completion of goals, plans and actions, coherency and distribution of resources [11, 12].

Tulken (2001) uses game-theory cooperation to describe cooperation in economics and describes it in terms of explicit influence from either a leader or referee and implicit influence of norms and values. This influence is from the norms and values that are common to the actors [14].

### 3.4 Cooperation of Agents

The advantage of autonomous agents is their ability to generate their own goals and to also decide when they wish to adopt the goals of others. When an autonomous agent enters a relationship with another agent voluntarily, they are said to be cooperating. Therefore the definition of agent cooperation is when an agent enters a relationship voluntarily and adopts the goals of an agent. There are two important aspects to this definition. First, the agent is autonomous, and by the nature of intelligent agents, this is assumed. Secondly is the goal acquisition that occurs during cooperation. An intelligent agent will acquire a goal of another if there is some positive motivational effect that will eventuate [15, 16]. Wooldridge and Jennings (1999) provide four important characteristics of agent cooperation. These include recognition, team formation, plan formation and team action [15].

Agent cooperation also relies on seven assumptions asserted by Wooldridge and Jennings (1999). A closer examination of their fifth assumption (agents initiate the social processes) shows that for an agent to effectively cooperate with interaction components, cooperation must take on either an external or internal perspective. An external perspective determines how an agent is to cooperate and the effectiveness of the cooperation. The internal perspective uses an agent's internal state to form the basis of cooperation [15].

Cooperation is said to take on an internal perspective. As with an external perspective, there may be difficulties in distinguishing between the coordination and collaboration of actions. However, cooperation should include both an external and internal perspective. Just relying on the agents internal states may not provide for effective cooperation. By allowing external perspective in cooperation, agents can perform cooperative actions, but also form a team and manage these actions [15].

## 4 AC<sup>3</sup>M – Agent Coordination and Cooperation Cognitive Model

The purpose of this model is to show the link between coordination and cooperation. Furthermore, it can show that coordination and cooperation do not simply co-exist, but is a cognitive loop that will lead into one another. There are two components in AC<sup>3</sup>M. They include the Cooperative Coordination and Coordinative Cooperation models, where each component combines the definitions of cooperation and coordination.

Both models are cyclic and viewed from a coordinative or cooperative perspective. They are designed to show that when a coordinative or cooperative event occurs, it will result in either cooperation or coordination respectively. Cooperative Coordination occurs when the agent has entered a voluntary relationship. However, once cooperation is achieved, the managing of the interdependencies between agents must occur. Hence, this gives rise to *Cooperative Coordination*. A draft of this

concept is shown in figure 1 and demonstrated during the invited session on Intelligent Agents and their applications at KES 2006. Further detail will be provided in future articles [17].

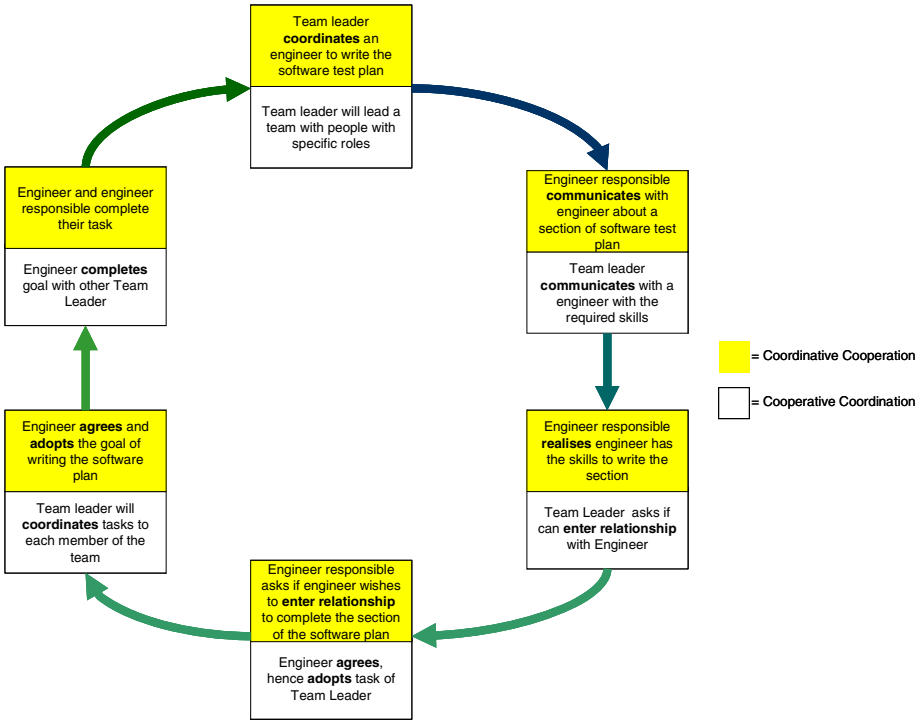


Fig. 1. Cooperative Coordination and Coordinative Cooperation Example

## 5 Conclusion

This paper provides an overview of the current research in relation to coordination and cooperation in a MAS. Agent coordination and cooperation are essential as it ensures agents behave desirably, do not waste efforts and squander resources in a system. It is also important to realize that if cooperation and coordination within a MAS are not achieved, contingency plans must be in place so the objectives are still met.

Another important assumption within an MAS is that agents must be the first to initiate the social processes within the system. This can be achieved with the proposed model by the use of Coordinative and Cooperative Events. This model can also provide some solutions to the current drawbacks in current models of coordination and cooperation. One solution is the possibility of reducing inflexibility. This can be achieved by simply using coordination and cooperation theories as well as the concepts of the personification of agents to emulate more realistic and human-like MAS.

## Acknowledgments

I wish to extend my gratitude to the KES group at the University of South Australia, especially to Nikhil Ichalkaranje, for their assistance, guidance and wisdom to all things concerning multi-agent systems.

## References

1. van Mulken, S., Andre, E. and Muller, J.P.: The Persona Effect: How Substantial Is It? In: Johnson, H., Nigay, L and Roast, C. (eds.): Proceedings of HCI'98 - Human Computer Interaction - People and Computers XIII, Springer-Verlag (1998) 53-66
2. Castelfranchi, C., de Rosis, F. and Falcone, R.: Social Attitudes and Personalities in Agents, In: Proceedings of AAAI Fall Symposium on Socially Intelligent Agents, AAAI Press, Cambridge, Massachusetts, United States of America (8-10 November 1997) 16-21
3. Ferber, J.: Multi-agent systems: an introduction to distributed artificial intelligence. New York: Addison Wesley Longman Inc (1999)
4. Ehlert, P. and Rothkrantz, L.: Intelligent Agents in an Adaptive Cockpit Environment, Delft University of Technology, Netherlands, Research Report DKE01-01, Version 0.2 (October 2001)
5. Weiß, G.: Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, The MIT Press (1999)
6. Bratman, M.E.: What is Intention?, In Intentions in Communications, MIT Press, Cambridge, MA (1990) 15-32
7. Rao, A. S. and Georgeff, M. P.: BDI Agents: from theory to practice, Proceedings of the First International Conference on Multi-Agent Systems, San Francisco, The MIT Press (1995) 384-389
8. Wooldridge, M.: An Introduction to Multiagent Systems, Chichester, John Wiley & Sons (2002).
9. Malone, T.W. and Crowston, K.: What Is Coordination Theory and How Can It Help Design Cooperative Work Systems? In: Proceedings of CSCW'90 - Conference on Computer-Supported Cooperative Work, Los Angeles, California, United States of America (1990) 357-370
10. Denti, E., Omicini, A. and Ricci, A.: Coordination Tools for the Development of Agent-Based Systems, In: Trapp, R. (ed): Proceedings of EMCSR 2002: 16th European Meeting on Cybernetics and Systems Research, , Austrian Society of Cybernetic Studies, Vienna, Austria (2-5 April 2002) 671-676
11. Nwana, H.S., Lee, L. and Jennings, N.R.: Coordination in Software Agent Systems, BT Technol J, 14(4), (October 1996) 79-89
12. McBurney, P. and Parsons, S.: Engineering Democracy in Open Agent Systems, In: Omicini, A., Petta, P. and Pitt, J (eds.): Proceedings of ESAW 2003 - 4th International Workshop on Engineering Societies in the Agents World IV, from Lecture Notes in Artificial Intelligence, , Springer-Verlag, London, UK, October 29-31, (2003) 66-80
13. Hua, Z.: Study of Multi-Agent Cooperation, In: Proceedings of Third International Conference on Machine Learning and Cybernetics, vol. 5, IEEE, Shanghai, China, August (2004) 3014-3017
14. Tulken, H.: Economic Theory and International Cooperation on Climate Change Issues, In: Proceedings of International Climate Policy after COP6 – Workshop on Frontiers in International Climate Policy Research, University of Hamburg, Germany, September 24-25 (2001)

15. Wooldridge, M.: The Cooperative Problem-Solving Process, *Journal of Logic and Computation*, vol. 9, no. 4, (1999) 563-592
16. D'Inverno, M., Luck, M.: *Understanding Agent Systems*, Springer-Verlag, Heidelberg (2004)
17. Consoli, A., Tweedale, J. and Jain, L.: The Link Between Agent Coordination and Cooperation, To appear in: KES'06 – 10th International Conference on Knowledge Based Intelligent Information and Engineering Systems, from *Lecture Notes in Computer Science*, Springer-Verlag, Bournemouth, England, October 9-11 (2006)

# Innovations in Intelligent Agents and Web

Gloria Phillips-Wren<sup>1</sup> and Lakhmi Jain<sup>2</sup>

<sup>1</sup> Loyola College in Maryland, 4501 N. Charles Street, Baltimore, MD 21210 USA  
gwren@loyola.edu

<sup>2</sup> University of South Australia, School of Electrical and Information Engineering, Adelaide,  
Mawson Lakes Campus, South Australia SA 5095  
Lakhmi.Jain@unisa.edu.au

**Abstract.** Intelligent agents are an integral and expanding part of practical systems. The ability of agents to generate goals and determine whether to accept the goals of others provides a powerful approach to autonomous computing, particularly in Web-based systems. This invited session of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems focuses on innovations in intelligent agents and Web-based agent systems.

## 1 Introduction

Humans have long dreamed of creating and owning robots or machines to perform mundane tasks such as household chores, while developments in the last fifty years have edged closer to that realization with advances in technology such as transistors and integrated circuits. Humans have also envisioned electronic personal assistants that provide human-like support for numerous tasks, such as purchasing, selling, handling email, managing files, navigating within computer space or while traveling, and scheduling jobs or meetings. Advances in computer science in the 1980s led to numerous applications in the 1990s with intelligent agents [1]. In general, intelligent agents are software programs that act on behalf of someone or something such as another agent [2, 3, 4]. A primary characteristic of agents is that they are autonomous [4]. The advantage of autonomous agents is their ability to generate their own goals and to decide when to adopt the goals of others. By allowing external perspective in cooperation, agents can perform cooperative actions, but also form a team and manage these actions.

The growth of the Internet and distributed systems has created fertile ground for research in mobile technologies such as intelligent agents [5, 6]. Teaming is possible in multi-agent systems with coordination and cooperation among agents. In fact, multi-agent systems are said to create an “artificial social system” [5] that involves agent architecture, cooperation among agents and human-agents, human-like learning and trust.

A number of successful agent-based systems are reported in the literature [7-10]. New applications of intelligent agents are emerging rapidly [10] such as Web-based businesses, Web-based education, Web publishing, Web-based healthcare, and so on.

## 2 Challenges in Multi-agent Systems and Web Architectures

The future presents research challenges in multi-agent systems. The challenge is to design robust:

- Human-agent teams
- Human-like agents
- Trust based agent models
- Emotion-based agents
- Agent communication and cooperation models
- Self-creating agent architectures

To this list, the Web revolution has added new challenges such as:

- Adaptive websites
- Evolution of websites
- Web intelligence
- Web inference engines
- Web mining
- Web agents
- Adaptive Web interfaces
- Web security and trust issues

This paper introduces an invited session of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems focused on innovations in intelligent agents and Web-based agent systems that takes a small step toward advancing these goals.

## 3 Session Papers

This session includes four papers.

The first paper by Consoli and Tweedale [11] is entitled “An Overview of Agent Coordination and Cooperation.” It presents an overview of two social abilities of agents, coordination and cooperation. The definitions and theory are derived and linked into a single multi-agent system model, called the Agent Coordination and Cooperation Cognitive Model, that replicates a linkage seen in organizational systems. These attributes are essential to ensure that agent systems efficiently and effectively utilize system resources.

An online navigation recommendation system is provided by Velasquez and Palade [12] in “Testing online navigation recommendations in a web site.” They introduce an *à priori* method to estimate the success of the recommendation system. It provides the web site visitor with a set of pages that could be of interest to him/her. The system is tested in a real web site to demonstrate its effectiveness.

A multi-agent system environment that is not constrained by time and space is demonstrated by Yang [13] in “From community models to system requirements: A cooperative multi-agents approach.” Middleware is presented along with a community computing meta-model to develop a framework for the system.

The final paper by Nguyen, Faulkner and Kolp [14] is entitled “A peer-to-peer agent architecture for information integration.” It presents a multi-agent system architecture based on organisational styles in designing a peer-to-peer information integration system. The system uses a multi-criteria search engine to increase efficiency of the peer-to-peer search using the Gnutella protocol.

## Acknowledgements

We appreciate the excellent contribution of the authors. The efforts of the reviewers greatly contributed to the quality of the papers and are gratefully acknowledged.

## References

1. Hyacinth, S., Nwana, D. and Ndumu, T.: A perspective on software agents research. *The Knowledge Engineering Review* **14** (1999) 1-18
2. Bradshaw, J. (ed.): *Software Agents*. The MIT Press: Cambridge, MA (1997)
3. Huhns, M. and Singh, M. (eds.): *Readings in Agents*. Morgan Kaufmann Publishers, Inc, San Francisco, CA (1998)
4. Jennings, N. and Woolridge, M. (eds.): *Agent Technology: Foundations, Applications and Markets*. Springer-Verlag, Berlin, Germany (1998)
5. Wooldridge, M.: *An Introduction to MultiAgent Systems*. John Wiley and Sons, LTD, West Sussex, England (2005)
6. Padgham, L., Winikoff, M.: *Developing Intelligent Agent Systems*. John Wiley and Sons, LTD, West Sussex, England (2004)
7. Phillips-Wren, G., Jain, L.C. (eds.): *Intelligent Decision Support Systems in Agent-Mediated Environments*, IOS Press, The Netherlands (2005)
8. Khosla, R., Ichalkaranje, N., Jain, L.C. (eds.): *Design of Intelligent Multi-Agent Systems*, Springer-Verlag, Germany (2005)
9. Jain, L.C., Chen, Z., Ichalkaranje, N. (eds.): *Intelligent Agents and Their Applications*, Springer-Verlag, Germany (2002)
10. Howlett, R.J., Ichalkaranje, N., Jain, L.C., Tonfoni, G. (eds.), *Internet-Based Intelligent Information Processing*, World Scientific Publishing Company Singapore (2002)
11. Consoli, A., Tweedale, J.: An overview of agent coordination and cooperation. In *Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Verlag-Springer, Berlin (2006) in press
12. Velasquez, J., Palade, V.: Testing online navigation recommendations in a web site. In *Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Verlag-Springer, Berlin (2006) in press
13. Yang, J-J.: From community models to system requirements: A cooperative multi-agents approach. In *Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Verlag-Springer, Berlin (2006) in press
14. Nguyen, T., Faulkner, S., Kolp, M.: A peer-to-peer agent architecture for information integration. In *Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Verlag-Springer, Berlin (2006) in press



# Robust Segmentation for Left Ventricle Based on Curve Evolution

Gang Yu, Yuxiang Yang, Peng Li, and Zhengzhong Bian

School of Life Science and Technology, Xi'an Jiaotong University  
Xi'an, 710049, China  
yugang@mailst.xjtu.edu.cn

**Abstract.** This paper presents a novel multi-resolution framework for the segmentation of left ventricle in echocardiographic images. This framework is based on curve evolution and nonlinear diffusion pyramid. At the low resolution, a statistical region-based model is applied to analyze the echocardiographic images and it is combined with a boundary-based model for the pre-segmentation. The pre-segmentation result is used to initialize the front for the high resolution. Meanwhile, a fast mathematical morphology-based method is used to pass the solution from low to high resolution. This method is competent to fast narrowband re-initialization. Furthermore, a local Snake model is used as an external constraint to optimize segmentation at the high resolution. Segmentation results of left ventricle images show that the multi-resolution segmentation method is accurate and robust.

**Keywords:** Level set; Echocardiographic Image; Mathematical morphology; Nonlinear diffusion; Gaussian distribution.

## 1 Introduction

Automatic segmentation of echocardiographic images provides a powerful tool for the extraction of lesion shape and quantitative image analyses that are useful for diagnosis of heart diseases. Early approaches for segmentation of left ventricle in echocardiographic images include some statistical methods. Many statistical models were introduced to improve the feature extraction [2]. However, the detection accuracy of these methods is to be validated. Other researchers developed segmentation algorithms with a priori knowledge to detect boundary in echocardiographic images [3]. However, the shape knowledge is usually difficult to learn. In most cases, the extensive training cost is necessary. Especially, a learned shape template can be only used to segment a specific class of images with a similar boundary shape.

“Snake” or active contour models have become effective tools for the extraction of region of interests (ROI), which were widely investigated for overcoming the limitations of traditional methods. Active contour models have been applied successfully to medical images including echocardiographic images in the last decade. Sethian *et al* firstly introduced the level set method into geometric active contour models for numerical implementation [1]. The level-set-based active contour methods are important for image segmentation, because its solution is steady and suitable for various topology changes. The selection of speed function is crucial to the

approaches. The boundary-based or edge-based speed functions may improve the segmentation results, but could not work well in low contrast or noise image, as the edge information is too weak there [4]. Region-based speed functions, such as geodesic active region model are more effective snake-based segmentation methods, because the prior knowledge about ROI is introduced into them[5]. However, the pure region-based approaches may bring expensive computational cost. The important problems are the design of region-based models and the combination with the snake energy minimization framework.

This paper presents an efficient multi-resolution segmentation framework for left ventricle in the echocardiographic images. The remainder of the paper is organized as follows. In Section 2, the basic methods are briefly introduced. In Section 3, the proposed multi-resolution framework is described in detail. In Section 4, experiments are presented; and finally, conclusions are reported.

## 2 Basic Methods

Though the intensity distribution of the original echocardiographic image is not Gaussian, the intensity distribution at a very high resolution in the Gaussian pyramid can be approximately models by Gaussian [6]. However, the approach in reference [6] only applied mean intensity inside and outside curve, not statistical models, to approximately analyze the echocardiographic images, which was difficult to segment complicated images such as ultrasound ones.

Christophe *et al* proposed a region snake approach, which is based on the calculation of the statistics of the inner and the outer regions [7]. It has thus been possible to develop optimal algorithms to the random fields, which describe the gray levels in the input image if the probability density function (PDF) is known. Martin *et al* analyzed level set implementation of this snake approach [8]. They showed that the approach can improve segmentation results in noisy images. Meanwhile, the speed function based on Gaussian distribution is defined as[8]:

$$F(x, y) = \frac{1}{2} [\log(\sigma_t^2) - \log(\sigma_b^2)] + \frac{(s(x, y) - m_t)^2}{2\sigma_t^2} - \frac{(s(x, y) - m_b)^2}{2\sigma_b^2} \quad (1)$$

Where  $t, b$  are the target to be extracted and background in the image respectively.  $m$  and  $\sigma$  are the mean and standard deviation of Gaussian distribution respectively. Obviously, this statistical model is based on region information.

The level set method embeds parameterized evolution equation to high dimensional level set function. Assume that the curve  $C$  is a level set of a function of  $u : [0, a] \times [0, b] \rightarrow R$ .  $u$  is therefore an implicit representation of the curve  $C$  [1]. This representation is parameter free, then intrinsic. If the curve  $C$  evolves according to

$$\frac{\partial C}{\partial t} = F \bar{N}$$

For a given speed function or “force”  $F$ , then the embedding function  $u$ , i.e. level set function, should deform according to

$$\frac{\partial u}{\partial t} = F|\nabla u|$$

Multi-resolution technique is based on a series of images with different scales. The noise in original image appears to be decreased with the increase of scale. It is well known that there is a great deal of speckle noise in the original ultrasound images. Therefore, multi-resolution method is more robust to ultrasound image segmentation than single-resolution method.

### 3 Multi-resolution Segmentation Framework

#### 3.1 Nonlinear Diffusion Pyramid

The Gaussian pyramid is a classical multi-resolution technique. Let original image  $I_0$  is denoted as level  $L_0$ . A 2D Gaussian kernel is used for smoothing the original image before sub-sampling the original image. The coarse image at Level  $L_1$  is obtained, which is smoother than the original image. Similarly, more levels can be obtained. Although the images at higher levels are getting blurred, the boundary shape remains similar at different pyramid levels. However, Gaussian pyramid may blur the edges and influence the segmentation result because the gradient information is weak here.

A scale-space consists of a family of image descriptions that vary from fine representations to coarse representations. Perona and Malik showed that a scale-space can be represented by a progression of images computed by the heat diffusion equation[10]. A Multiple Diffusion method(ADP-MD) for initializing anisotropic diffusion pyramids(ADP) were presented in Reference 9. Perona and Malik were the first to introduce non-linear diffusion within the image processing context. The Perona-Malik(P-M) diffusion equation is isotropic and nonlinear diffusion like Gaussian smoothing, but it reduce the diffusion coefficient in the edges, so it may protect the edges in the ultrasonic images while removing the noise. We then build an ADP-MD pyramid to analyze the echocardiographic image by P-M diffusion equation.

#### 3.2 Pre-segmentation Model

As mentioned above, the intensity distribution of the echocardiographic image at low resolution in the diffusion pyramid can be modeled by Gaussian. An efficient statistics-based speed function was presented in equation (1) if the PDF in the image is the Gaussian distribution.

In most cases, the region-based models, such as equation (1) may bring expensive computational cost. More importantly, the models may converge at a local minimum solution. Therefore, the combination with an edge-based model helps improve the performance of curve evolution. In this paper, we choose geodesic active contour as the boundary estimation term. It is presented as:

$$\frac{\partial C}{\partial t} = g(|\nabla I|)(c_1 + c_2 k) \cdot \vec{N} - (\nabla g(|\nabla I|) \cdot \vec{N}) \cdot \vec{N} \quad (2)$$

Integrate it with region-based model; the combinative model is given by:

$$\frac{\partial u}{\partial t} = \alpha \times F|\nabla u| + (1 - \alpha) \left\{ g(|\nabla I|)(c_1 + c_2 k) \cdot |\nabla u| - (\nabla g(|\nabla I|) \cdot \vec{N}) \cdot |\nabla u| \right\} \tag{3}$$

Equation (3) is the final curve evolution equation for pre-segmentation at low resolution, which can be implemented by level set method. The combinative model integrates the region- and edge-based model. The experiment will demonstrate the performance.

### 3.3 Passing Solution Between the Adjacent Levels

In order to improve the segmentation performance, the pre-segmentation solution at the low resolution should be passed to the high resolution and a new evolution is performed for desirable results. Conventional methods for passing solution are high computational cost because they interpolate the obtained contour at the high resolution. In this section, we present an efficient scheme based on mathematical morphology.

Step 1. Passing all the interior points to the high resolution

If  $\mu$  is a scaling factor from the low resolution  $L_{j+1}$  to the high resolution  $L_j$ , there are  $\mu * \mu$  points in Level  $L_j$ , which correspond to one point at level  $L_{j+1}$ . For example, if  $\mu$  is 2, for any point  $u(i, j)$  at  $L_{j+1}$ , the corresponding four points are  $u(2i, 2j), u(2j + 1, 2j), u(2i, 2j + 1), u(2i + 1, 2j + 1)$  at  $L_j$ . Therefore, we can obtain all interior point locations of the curve at level  $L_j$ .

Step 2. Extracting the Front

After Step 1, we obtain all interior point locations at high resolution level, seeing the left image of figure 1. We then apply morphological dilation to extract the boundary of the interior region, i.e. the front. The extraction operation is defined as:

$$\text{ImageB} = \text{ImageA} \otimes A - \text{ImageA} \tag{4}$$

The structure element  $A$  is 3\*3 strong template. ImageA is the mirror image at the high resolution, where let the all interior points be 1 and other points be 0. ImageB is the mirror image of the extracted front.

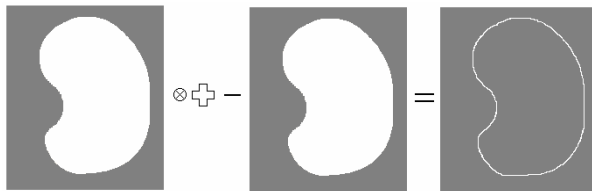


Fig. 1. Extracting the Front

Step 3. Dilation around the front  
The dilation operation is defined as:

$$\text{ImageC} = \text{ImageB} \otimes B \quad (5)$$

Dilation operation is performed around the extracted front, whose structure element  $B$  is a disc. In our method, usual computational cost, such as fast marching algorithm, is unnecessary. Meanwhile, the reconstruction method is self-adaptive, and the radius of the disc is also the narrowband width, which can be reset during the evolution. During the dilation operation, the distance between a point in the narrow band and the center of disc should be stored in the temporary memory, which is also the new distance function value.

After Step 3, a new narrow band at the high resolution is rebuilt. It is also the initial state of a new evolution. The scheme passing solution from the low resolution to high resolution is very rapid, because the mathematical morphology operators are more efficient than conventional interpolation computation.

### 3.4 Optimization Model at High Resolution

After extracting boundary at the low resolution, the region-based Gaussian model can no longer be used at the high-resolution image because the intensity distribution is not Gaussian here. Through passing solution from the low resolution to high resolution, the initial curve is obtained. Without the further constraints, the edge-based model such as geodesic active contour is easily to leak from weak boundary.

Yezzi proposed a fully global approach to image segmentation that is derived based on the global segmentation of an image [5]. This global approach is a region-based approach designed to optimally separate the certain image over a known number of region types. However, it may be invalid in echocardiographic images because the global intensity distribution is more difficult to estimate. More importantly, we have a pre-segmentation result at the low resolution level, so the global separation is unnecessary. Therefore, we then develop a local approach based on Yezzi's method. The local energy function is given by:

$$E = -\frac{1}{2}(u' - v')^2 \quad (6)$$

where

$$u' = \frac{1}{N} \sum_{R_i} I(x, y) H(\Phi(I(x, y)) + a) H(-\Phi(I(x, y)))$$

$$v' = \frac{1}{N} \sum_{R_i} I(x, y) H(\Phi(I(x, y)) - a) H(\Phi(I(x, y)))$$

$\Phi$  is the level set function, where the function value of interior points is smaller than 0, and that of exterior points is bigger than 0.  $I$  is the intensity value.  $H$  is a Heaviside function.  $a$  is a positive constant, which is often defined as the width of narrow band.  $u', v'$  are just the average intensity of the interior and exterior points whose distance function absolute values are smaller than the constant  $a$ . Therefore,

the evolution is within a small region, and finds more desirable boundary based on the solution at the coarse level. According to gradient descent method, the local speed function is given by:

$$\frac{dC}{dt} = (u' - v') \left( \frac{I - u'}{A_u'} + \frac{I - v'}{A_v'} \right) \vec{N} \tag{7}$$

Where  $A_{u,v}' = \iint_{R_{u,v}} dA$ , are the area of inner and outer curve respectively.

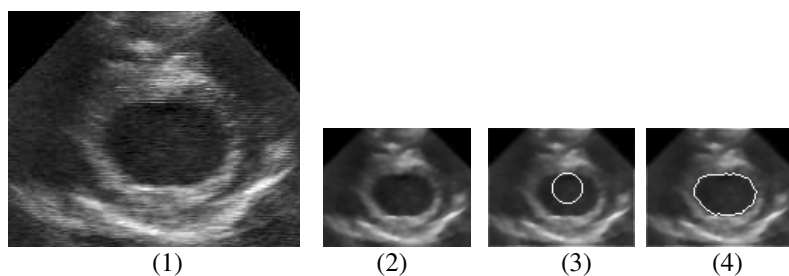
## 4 Experiment

We choose 4-Dimensional echocardiographic images as experimental datasets. Every dataset include 10~50 frame images, which describe a heart period. We segment every image in the dataset and reconstruct the heart. Before reconstruction, image segmentation must be done using the proposed method in this paper. Before the segmentation, the nonlinear diffusion pyramid is built. The parameters of P-M diffusion equation are chosen as:  $t = 0.2$ ,  $k = 0.1$ , which are suitable for most images. The default diffusion number is 10. After smoothing the original image, we sub-sample it to create the subsequent levels of the pyramid.

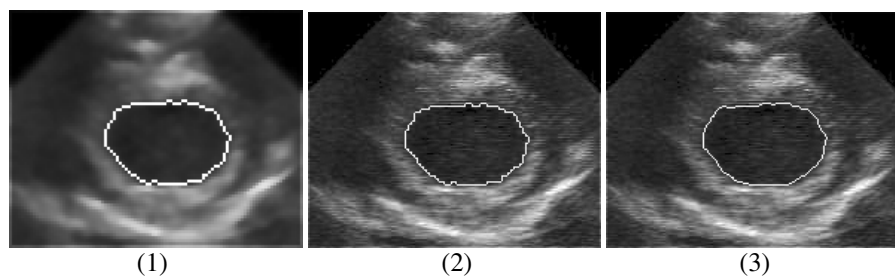
The parameters are chosen as follows. In equation 3,  $\alpha = 0.4$ ,  $c_1 = 1$ ,  $c_2 = 0.2$  work well in most images. The equation 7 is independent of any parameter. The original image is shown in figure 2.1. Figure 2.2 is a low resolution image which is obtained from diffusion pyramid. Figure 2.3 is the initial state of the front, which is described by a white curve. The curve is propagated under the equation 3, where the region-based and edge-based information are applied in the equation. Figure 2.4 is the pre-segmentation result at a low-resolution level. Although the edge-based model is easy to leak from weak boundary and region-based model may stop at local minimum location, the combinative model is robust.

Figure 3.1 magnifies the pre-segmentation result image, i.e. Figure 2.4, one times. Figure 3.2 is the initial state of the curve at high resolution level. The initial solution is passed from the low resolution image Figure 2.4 by our mathematical morphology-based scheme. The curve of Figure 3.1 and Figure 3.2 is almost equivalent to each other, which proves that the solution at the low resolution is transferred to high resolution level accurately. The result demonstrates our method based on mathematical morphology has an excellent performance in terms of accuracy. Figure 3.3 is the final result under the local constraint, i.e. the equation 7. Compared with the result at the coarse level, it is closer to real boundary.

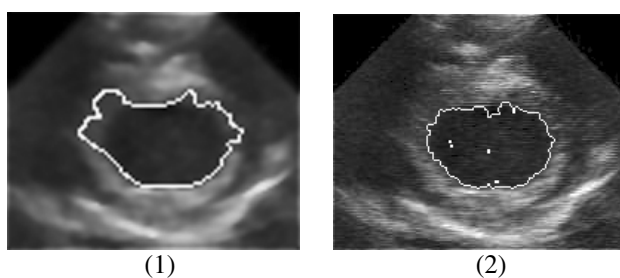
The segmentation results of two conventional snake model, Geodesic active contour and Gaussian statistical model, are given by Figure 4. In order to describe the difference, we magnify the coarse image one times. The curve of Geodesic model leaks from the boundary, because the gradient information is too weak here. Figure 4.2 shows the result of Gaussian statistical model presented by equation (1). The curve converges at a minimum location and four wrong regions are labeled, because the intensity of original image is not a Gaussian distribution.



**Fig. 2.** Original image and pre-segmentation at the low resolution level. (1) original image. (2) coarse image, (3) the initial state. (4) pre-segmentation result.



**Fig. 3.** The similarity of 3.1 and 3.2 shows the accuracy of our scheme for passing solutions between the adjacent levels. (1) is the magnified image of 2.4. (2) is the initial state at high resolution image. The optimized result is shown in (3), which demonstrate the performance of the proposed method.



**Fig. 4.** The segmentation result of Geodesic model and Gaussian statistical model

## 5 Conclusion

In this paper, we proposed a novel multi-resolution framework for left ventricle image segmentation. This framework is based on the curve evolution and nonlinear diffusion. A combinative model integrating region- and edge-based information function was designed to analyze the image at a coarse level. Meanwhile, a rapid scheme passing the solution from the low resolution to high resolution was developed.

The scheme is based on mathematical morphology and does not need interpolation computation, which makes it suitable for real-time applications. Furthermore, an efficient optimization method at the finer level was proposed, which propagates the curve towards the real boundary. The proposed optimization approach is local and rapid because the initial curve is close to desirable result. The proposed framework is implemented in a level set method and is suitable for various topologic changes. Moreover, it can be easily extended to 3D images. This segmentation framework was tested using real ultrasound images and experiments show that it is accurate and robust.

## Acknowledgement

This paper is supported by the national Natural Science Foundation of China under grant No. 60271022, 60271025.

## References

1. J.A.Sechian.: Level Set Methods and Fast Marching Methods. Cambridge University Press, New York(1999).
2. Xiao G., Brady M., Noble J., Zhang Y.: Segmentation of Ultrasound B-mode Images with Intensity in Homogeneity Correction. IEEE Transactions on Medical Imaging, Vol.21, No.1, 2002(48-57).
3. Chen Y., Thinuvenkadam S.: On the Incorporation of Shape Priors into Geometric Active Contours. IEEE Workshop on Variational and Level set Methods in Computer Vision (2001)45-152.
4. V. Caselles, R. Kimmel, G. Spairo: Geodesic Active Contours. International Journal of Computer Vision, Vol22,(1997)61-79.
5. Anthony Yezzi, Jr., Andy Tsai, Alan Willsky:A Fully Global Approach to Image Segmentation via Coupled Curve Evolution Equations. Journal of Visual Communication and Image Representation 13, (2002)195–216.
6. Ning L, Weichuan Y, James S.D.: Combinative Multi-scale Level Set Framework for Echocardiographic Image Segmentation. Medical Image Analysis, 7, (2004)529-537.
7. Christophe C, Philippe R, Vlady B.: Statistical Region Snake-based Segmentation Adapted to Different Physical Noise Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol21, No11, (1999) 1145-1157.
8. M Pascal, R Philippe and G Francois, G Prederic: Influence of the Noise Model on Level Set Active Contour Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.26, No.6, (2004).799-803.
9. S. T. Acton, A. C. Bovik, M.M Crawford: Anisotropic Diffusion Pyramids for Image Sgmentation . IEEE Conference on Image Processing, (1994)478-482.
10. P. Perona and J. Malik: Scale-space and Edge Detection using Anisotropic Diffusion. IEEE Transaction On Pattern Anal. and Mach. Intell., Vol. 12, No. 6,(1990)629-639.



# Segmentation of Ovarian Ultrasound Images Using Cellular Neural Networks Trained by Support Vector Machines

Boris Cigale, Mitja Lenič, and Damjan Zazula

University of Maribor, Smetanova 17, 2000 Maribor, Slovenia

**Abstract.** Various applications of cellular neural networks (CNNs) on complex image processing tasks raise questions about an appropriate selection of template elements that determine the CNN's behaviour. There are two possibilities: either to resort to the existing and published templates suitable for the problem under consideration or to construct the templates by one of well-known training methods, such as genetic algorithms, simulated annealing, etc. In this paper, a novel approach which utilizes the formalism of support vector machines (SVMs) is introduced. We found the CNN template optimisation done by this machine learning technique superior to other training methods. The learning time reduced from several hours to less than a minute. Testing our novel approach on ultrasound ovarian images, the obtained segmentation results and recognition rates for ovarian follicles were significantly better than with comparable solutions.

**Keywords:** support vector machines, cellular neural network, template optimisation, medical ultrasound image segmentation.

## 1 Introduction

A steady growth of computer performance and decrease of their costs have influenced all computer-assisted applications and services. One of the most rapidly developing fields encompasses medical imaging with a variety of imaging facilities and technologies [1]. Ultrasound, for example, is relatively inexpensive noninvasive method, being widely used in everyday's practice. Because of heavy speckle-noise corruption, the ultrasound image interpretation is not a straightforward and easy task.

Ultrasound images must first be segmented into smaller constituent components, such as edges and regions. The task is far from being plain even if the environment conditions are stable. Whenever the conditions change or the observed scene behaves with certain dynamics, the processing methods must resort to adaptive solutions [2]. The adaptability goes hand in hand with learning capabilities. Artificial neural networks (ANN) combine those properties in the way the human brains master the recognition and classification of visual information.

In 1988, Chua and Yang [3] introduced a special network structure which resembles the continuous-time Hopfield ANN, but it implements strictly local

connections. The proposal was influenced both by neural networks [4] and by cellular automata [5]. This new construct was called cellular neural networks (CNNs). An initial learning phase is also needed for CNNs, as it is needed for any other neural network. This adapts their properties to the identification problem to be resolved. In general, learning approaches for CNNs are based on genetic algorithms or simulated annealing. The implementation of the two methods has a rather unattractive characteristic: their convergency may be questionable, while the learning period always takes a lot of time (several hours on today's PCs with best performance). While training the CNNs means iterative adaptation to positive and negative examples, this procedure could be much shorter if the examples were preselected in two most discriminant groups. The idea coincides with the approach of support vector machines (SVM).

SVMs implement a general algorithm based on statistical learning theory [6], commonly applied in pattern recognition. They have been applied with success in different research fields for various learning problems with multi-dimensional data, such as text categorization, text mining, various classification tasks, face recognition, etc.

In this paper, we introduce an idea of how to use SVMs for obtaining the information needed in the process of the CNN template optimisation. A short overview of the CNN operation is given in Section 2, while the SVM application on linearly separable cases is explained in Section 3. Our novel CNN learning approach using SVMs is presented in Section 4. The algorithm is applied to the ovarian ultrasound image segmentation in Section 5. Section 6 concludes the paper.

## 2 Cellular Neural Networks

The fundamental building block of the CNNs is a cell designated by  $C(i, j)$ , where  $i$  and  $j$  stand for the coordinates of the cell's position in a 2D representation of the net. Each cell is coupled only to its neighbouring cells  $C(k, l)$  in the  $r$ -neighborhood  $N_r(i, j)$ . The cell's state can be described by the following equation:

$$s_t(i, j) = \sum_{(k, l) \in N_r(i, j)} A_t(i, j; k, l) o_t(k, l) + \sum_{(k, l) \in N_r(i, j)} B_t(i, j; k, l) u(k, l) + I_t \quad (1)$$

where  $s_t(i, j)$  is the inner state of cell  $C(i, j)$ ,  $o_t(k, l)$  the output, and  $u(k, l)$  the input of cell  $C(k, l)$ . Subscript  $t$  stands for the observed time instant.  $A_t(i, j; k, l)$  and  $B_t(i, j; k, l)$  are called feedback and control parameters, respectively. Parameter  $I_t$  is a bias which is added to each inner state of the cell. In this paper, parameters  $A_t(i, j; k, l)$ ,  $B_t(i, j; k, l)$ , and  $I_t$  are time variant, in contrast to the traditional definition where these parameters are time invariant. In the sequel, we assume the cell's output is defined by the following non-linear equation:

$$o_{t+1}(i, j) = f(s_t(i, j)) = \frac{1}{2}(|s_t(i, j) + 1| - |s_t(i, j) - 1|) \quad (2)$$

although other non-linear transformations (i.e. sigmoid function) can also be used.

Parameters  $A_t(i, j; k, l)$  are collected in template (matrix)  $A_t(i, j)$  and parameters  $B_t(i, j; k, l)$  in template  $B_t(i, j)$ . In this paper it is assumed that all the CNN cells have equal templates at specific time  $t$ . The spatial coordinates  $i$  and  $j$  can therefore be omitted, so  $A_t(i, j)$  becomes  $A_t$  and  $B_t(i, j)$  becomes  $B_t$ . Templates  $A_t$ ,  $B_t$ , and  $I_t$ , also called feedback template, control template, and the bias, respectively, completely define the behaviour of the network with given input and initial condition (the initial inner state of the cells).

### 3 Linear Classification Using Support Vector Machines

SVMs solve classification problems by determining a decision function which separates samples from different classes with highest sensitivity and specificity. A learning set must be given with positive and negative samples for each class, so the samples that belong to the observed class and those that belong to other classes, respectively. Assume a two-class case and a hyperplane which separates positive from negative samples,  $\mathbf{x}$ , defined as  $\mathbf{x} \cdot \mathbf{w} + b = 0$ , where  $\mathbf{w}$  is the hyperplane normal and  $b$  describes its distance from the origin. Each sample  $\mathbf{x}_i$  must be given a classification (target) value  $y_i \in \{-1, 1\}$  where the value of 1 designates positive and of -1 negative samples.

In linearly separable cases, support vector machines simply look for the hyperplane with the largest possible margin, i.e. the hyperplane whose distance to both the positive and negative samples is maximum. It can be formulated as:

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 \tag{3}$$

where  $\langle \cdot, \cdot \rangle$  denotes scalar product.

Considering inequality (3), there are two parallel planes which introduce the margin of width  $2/\|\mathbf{w}\|$ . There is no training data between the two planes. To reduce the risk of misclassification the margin between the planes must be maximum, which can be achieved by minimizing  $\|\mathbf{w}\|^2$ .

To solve the problem, nonnegative Lagrangian multipliers  $\alpha_i, i = 1, \dots, l$  are introduced, one for each inequality constraint (3). To maximize the margin, the following Lagrangian dual has to be solved [7]:

$$\max_{\alpha_i \geq 0} (\min_{\mathbf{w}, b} (L(\mathbf{w}, b, \boldsymbol{\alpha}))) \tag{4}$$

where

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_i \alpha_i y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + \sum_i \alpha_i \tag{5}$$

The optimal solution of Eq. (5) is obtained when the derivatives of  $L$  taken to  $\mathbf{w}$  and  $b$  equal 0. This requires the following conditions:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \tag{6}$$

$$\sum_i \alpha_i y_i = 0 \tag{7}$$

The final decision function which classifies an unknown vector  $\mathbf{x}_i$  is then defined as:

$$f(\mathbf{x}) = \text{sgn}(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \quad (8)$$

The majority of Lagrangian multipliers  $\alpha_i$  linked to the learning vectors which are not important for classification (they are too far away from the separation hyperplane) equal 0. Vectors  $\mathbf{x}_i$  whose  $\alpha_i$  differ from 0 are named support vectors and are the only important data for the separation of positive and negative samples.

## 4 The CNN Templates Optimisation Using SVMs

Firstly, we have to reformulate the problem description to determine the CNN templates in the manner the SVMs can operate on them. For this reason, the templates  $A_t$  and  $B_t$  from Eq. (2) have to be vectorized. All rows of each matrix are concatenated into a single vector. Consequently, two vectors  $\mathbf{a}_t$  and  $\mathbf{b}_t$  replace matrices  $A_t$  and  $B_t$ , respectively. Secondly, the same modification is applied to the output and input neighbouring cells giving vectors  $\mathbf{o}_t(i, j)$  and  $\mathbf{u}_t(i, j)$ , respectively. With this vectorization, Eq. (1) changes to:

$$s_t(i, j) = \langle \mathbf{a}_t, \mathbf{o}_t(i, j) \rangle + \langle \mathbf{b}_t, \mathbf{u}_t(i, j) \rangle + I_t \quad (9)$$

Suppose the non-linear function from Eq. (2) is the signum function ( $\text{sgn}$ ). When the proper value of the CNN output  $o_t(i, j)$  is known in advance, this knowledge may be employed in the learning process. Denote such value by  $r(i, j) \in \{-1, 1\}$ . All these assumptions together lead to the following relationship:

$$r(i, j) = \text{sgn}(\langle \mathbf{a}_t, \mathbf{o}_t(i, j) \rangle + \langle \mathbf{b}_t, \mathbf{u}_t(i, j) \rangle + I_t) \quad (10)$$

Eq. (10) warrants the following inequality holds true:

$$r(i, j)(\langle \mathbf{a}_t, \mathbf{o}_t(i, j) \rangle + \langle \mathbf{b}_t, \mathbf{u}_t(i, j) \rangle + I_t) > 0 \quad (11)$$

If a concatenation of vectors is defined as

$$\mathbf{x}|\mathbf{y} = \{x_1, \dots, x_n, y_1, \dots, y_m\} \quad (12)$$

where  $\mathbf{x} \in \mathbb{R}^n$  and  $\mathbf{y} \in \mathbb{R}^m$ , the resulting  $\mathbf{x}|\mathbf{y}$  is an element of  $\mathbb{R}^{n+m}$ .

Thus, inequality (11) may further be derived into:

$$r(i, j)(\langle \mathbf{a}_t|\mathbf{b}_t, \mathbf{o}_t(i, j)|\mathbf{u}_t(i, j) \rangle + I_t) > 0 \quad (13)$$

The more the value of inequality (13) differs from 0, the better the discrimination of classes done by the applied CNN. Recalling the condition (3), it is evident that inequality (13) may certainly be considered of the same form. This consequently means that the CNN templates  $\mathbf{a}_t$  and  $\mathbf{b}_t$ , and the bias  $I_t$ , can be estimated using the SVM procedure described in Section 3.

#### 4.1 Optimisation of the CNN Templates for Image Processing

The natural way of applying CNNs to image processing is to make individual CNN cells responsible for single image pixel neighbourhoods. Therefore, the size and geometry of the CNN is equal to the size of image, whereas pixel  $p(i, j)$  and its  $r$ -neighbourhood correspond to the cell  $C(i, j)$ . This cell produces an output value which is assigned to the corresponding pixel,  $o_t(i, j)$ , of an interim output image  $O_t$ . The steady states of the CNN cell outputs swing either to  $+1$  or to  $-1$ . Consequently, the resulting output image pixels would be of the same values.

On the other hand, the acquired pixel values in real images occupy integer values between 0 and  $2^n - 1$ , where  $n$  stands for the number of bits determining the image contrast resolution. The lowest value, 0, designates black colour, the highest,  $2^n - 1$ , white colour. To unify the coding of input and output images and to adapt it for processing by a CNN, the best way is to normalize the pixel values on the interval  $[0, 1]$ . After this normalization, 0 stands for black, 1 for white.

Suppose the image to be processed by CNN has been normalized. It can then be transformed into the CNN input image, so that

$$u(i, j) = 1 - 2p(i, j) \quad (14)$$

Additionally, the question of the initial CNN cell status,  $s_0(i, j)$ , remains open. For the time being we set  $s_0(i, j) = 0$  for all image pixels. Similarly, after the CNN operation reaches the steady state the resulting image,  $O_\infty$ , can be transformed into a normalized output image,  $R$ , accordingly:

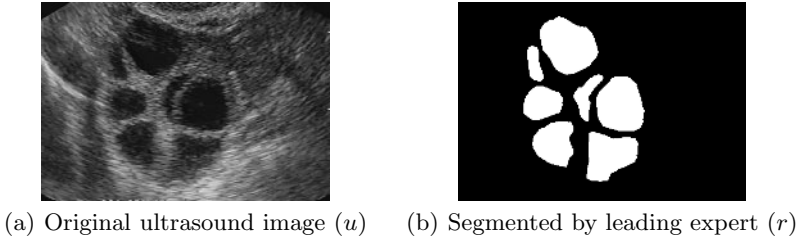
$$r(i, j) = \frac{1 - o_\infty(i, j)}{2} \quad (15)$$

The CNN learning sets of images must be formed in such a way that the objects looked for represent the positive samples, while the remaining background belongs to the opposite group. So, the annotation of the learning set images must correspond to the properties of the CNN output images—Eq. (15). The annotated values then apply directly to inequality (13), and are ready to be implemented in a SVM optimisation procedure.

## 5 Segmentation of Ovarian Ultrasound Images Using CNN Trained by SVM

A complete understanding of ovarian follicle dynamics is crucial for the field of in-vitro fertilisation. The main task is to succeed in determining the dominant follicles which have a potential to ovulate. For credible results, a doctor must examine patients every day during their entire menstrual cycle. Examination is usually done with ultrasonography. Because of the tedious and time-consuming nature of manual follicle segmentation, which is, on the other hand, also very demanding and inaccurate, an automated, computer-based method is desirable.

Segmentation of ultrasound images using the CNNs has been discussed in [8, 9]. The main drawback of the both methods is the time complexity of their learning procedures, completed by a genetic algorithm, for example. Typically, the learning times are in the range of hours.



**Fig. 1.** Image sample from learning set

The algorithm published in [8] consist of 5 successive steps where in the first step only a rough position of follicle is determined. A similar algorithm in 3 successive steps has been published in [9]. The first step of this algorithm is again a rough detection of follicles.

In order to test our optimisation method we tried to obtain the CNN templates for a rough detection of follicles. Our learning set consisted of 4 images, while the testing set consisted of 28 images, randomly selected from a database of 1500 ultrasound ovarian images. The selected images belong to 12 different patients. A leading expert manually annotated the positions of follicles and ovary in every image (an example of segmentation is depicted in Figure 1).

The real unfiltered ultrasound images, with no additional preprocessing, were used as an input to our method, both in learning and testing phase. Ultrasound images were sampled from the VHS tape using the MiroVideo DC30+ video card. A full resolution sampling to  $720 \times 540$  pixels and highest possible quality of JPEG movie (compression 2.66:1) were performed. However, it should be stressed that the VHS system is interlaced, where every image resolution is only  $352 \times 288$  pixels for PAL. After sampling all images were converted in 256 grey levels.

From each image in our experiment it is possible to generate 101376 learning samples for SVMs. Bigger learning sets can reduce errors in testing sets, but can also be quite time consuming, especially in the case of non-consistent samples. Therefore we decided to subsample the images according the CNN template size. Thus, the size of a single learning sample (number of attributes) varies with the template size.

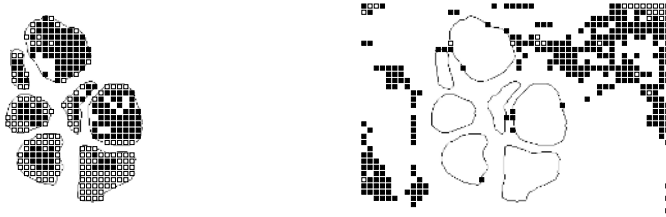
For this experiment we selected templates of size 15, which then generated 2185 learning samples with 225 attributes (pixels). Learning process took less then a minute on an average today's PC hardware. From the support vector distribution (Fig. 2), one can easily realize which regions of ovarian ultrasound images are harder to learn.

After the learning process completed, the CNN matrices  $A$ ,  $B$  and bias  $I$  were extracted from Lagrangian multipliers by applying Eqs. 6 and 12.

We verified the follicle recognition quality by using the so called ratios  $\rho^{(1)}$  and  $\rho^{(2)}$  [10]. This metric measures sensitivity and specificity of an image recognition algorithm. It calculates the intersections of the recognized and referential (annotated) image regions.  $\rho^{(1)}$  stands for the ratio between the areas of intersection and annotated follicle, while  $\rho^{(2)}$  for the ratio between the areas of intersection and recognized region. If the recognized regions entirely cover the annotated regions, then both ratios  $\rho^{(1)}$  and  $\rho^{(2)}$  are 1. In general, the closer the values of  $\rho^{(1)}$  and  $\rho^{(2)}$  to 1, the better is the matching of the regions being compared.

We validated our proposed learning and detection algorithm by observing only  $\rho^{(2)}$ . If the recognized, i.e. segmented, region of a follicle gives  $\rho^{(2)} > 0.5$ , we consider it a proper detection of the corresponding annotated follicle. This criterion warrants that more than a half of any recognized follicle region overlays a particular annotated follicle.

In 28 images of our testing set 168 follicles were annotated. The proposed detection algorithm recognized 113 regions of which 97 (86 %) belong to the follicles. Applying the abovementioned criterion for  $\rho^{(2)}$ , changed to  $\rho^{(2)} > 0$ , the number of recognized follicles rises to 99 %.



(a) Support vectors of area of interest      (b) Support vectors of background

**Fig. 2.** Support vector distribution after the first iteration. Black filled squares represent support vectors with proper classification, empty boxes represent support vectors that were misclassified.

If the results are compared to those published in [9], where on the same testing set only 81 regions were detected of which 63 (78 %) belong to the follicles, it is clear that we improved the recognition rate and reliability of the results. We have to emphasize again that the algorithm in [9] consists of 5 refining steps, from which we compare only the results after the first step.

Additionally, we evaluated the results of our newly proposed algorithm with the metric based on both ratios  $\rho^{(1)} > 0.5$  and  $\rho^{(2)} > 0.5$ . This more demanding condition lowered sensitivity to approximately 0.3, as only 52 follicles were recognized (46 % of all recognised regions). However, comparing this with the first step of algorithm in [8], the result is about three times superior.

Finally, it should be stressed that our novel method takes less than one minute to obtain the satisfactory CNN templates, in contrast to several hours needed when using genetic algorithms.

## 6 Conclusions

A novel approach based on SVM for the CNN template optimisation was presented in this paper. Because of different way of learning, the CNN templates obtained this way are time variant, in contrast to the conventional ones whose parameters are stationary.

We achieved two major advantages. Firstly, the learning times shrank from several hours to only a minute using an average today's PC hardware. Secondly, the CNN templates trained by SVM only on a small learning set are much more representative and robust than in the cases of genetic learning algorithms or simulated annealing. They give even three times higher sensitivity for detection of ovarian follicles in ultrasound images.

Our algorithm produces very efficient CNN templates in a very short time. In practice, separate medical examinations usually mean different settings on ultrasound machines, which produces different images. As suggested in [8], the best segmentation results on ultrasound images can therefore be obtained if the CNN templates are adapted to each examination separately. This was impossible with previous learning approaches, now our novel algorithms makes it possible.

## References

- [1] Bronzino, J.D.: The Biomedical Engineering Handbook. CRC Press, Boca Raton, USA (1995)
- [2] S. W. Perry, H.S.W., L, L.G.: Adaptive Image Processing: A Computational Intelligence Perspective. CRC Press, Boca Raton, USA (2002)
- [3] O., C.L., L., Y.: Cellular neural networks: Theory. IEEE Transactions on Circuits and systems **35**(10) (1988) 1257–1272
- [4] Hopfield, J.J.: Neural networks and physical systems with emergent computational abilities. Proc. Natl. Acad. Sci. **79**(8) (1982) 2554–2558
- [5] Hänggi, M., Moschzty, G.S.: Cellular Neural Network: Analysis, Design and Optimisation. Kluwer Academic Publishers, Boston, USA (2000)
- [6] Vapnik, V.N.: The nature of statistical learning theory. Springer, New York (1995)
- [7] Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery **2**(2) (1998) 121–167
- [8] Cigale, B., Zazula, D.: Segmentation of ovarian ultrasound images using cellular neural networks. IJPRAI **18**(4) (2004) 563–581
- [9] Zazula, D., Cigale, B.: Intelligent Segmentation of Ultrasound Images Using Cellular Neural Networks. Intelligent Processing Paradigms in Recognition and Classification of Astrophysical and Medical Images, In press (2006)
- [10] Potočnik, B., Zazula, D.: Automated analysis of a sequence of ovarian ultrasound images. part i. Imag. Vis. Comput. **20**(3) (2002) 217–225



# Bayesian Decision Tree Averaging for the Probabilistic Interpretation of Solar Flare Occurrences

Vitaly Schetinin<sup>1</sup>, Valentina Zharkova<sup>2</sup>, and Sergei Zharkov<sup>3</sup>

<sup>1</sup> Computing and Information Systems, University of Bedfordshire, Luton, Park Square, LU1 3JU, UK

Vitaly.Schetinin@luton.ac.uk

<sup>2</sup> Cybernetics Department, Bradford University, Bradford, BD7 1DP, UK

V.V.Zharkova@bradford.ac.uk\

<sup>3</sup> Department of Applied Mathematics, University of Sheffield, Sheffield, S3 7RH, UK  
S.Zharkov@sheffield.ac.uk

**Abstract.** Bayesian averaging over Decision Trees (DTs) allows the class posterior probabilities to be estimated, while the DT models are understandable for domain experts. The use of Markov Chain Monte Carlo (MCMC) technique of stochastic approximation makes the Bayesian DT averaging feasible. In this paper we describe a new Bayesian MCMC technique exploiting a sweeping strategy allowing the posterior distribution to be estimated accurately under a lack of prior information. In our experiments with the solar flares data, this technique has revealed a better performance than that obtained with the standard Bayesian DT technique.

**Keywords:** Machine learning, uncertainty, Bayesian averaging, decision tree, Markov Chain Monte Carlo, solar flare.

## 1 Introduction

Solar activity is characterized by patterns which can be represented by features of active regions, sunspots, solar flares, coronal holes, and/or filaments which are observable on full disk solar images taken from the ground and space-based instruments [1 - 3]. In the theory, these patterns have a complex dynamics associated with an 11 year solar activity cycle represented by the features dependent on the time and location on the solar disk [3 - 5]. The observed full-disk images are stored and then used in order to discover new knowledge about unknown phenomenon of solar activity. During the last decade many research have been done in order to discover phenomenon models in data by using the machine learning paradigms which potentially are capable of providing a high performance in terms of predictive accuracy [2 - 8].

In this paper we describe a new machine learning method developed for an automated classification of solar activity which is associated with solar flares. This method is based on the methodology of Bayesian model averaging which under some conditions is able to provide the best performance [9 - 12].

The Bayesian model averaging methodology has revealed promising results when the uncertainty in classification outcomes has to be estimated [13, 14]. The use of

Bayesian averaging over decision trees (DTs) allows domain experts to understand the nature of a phenomenon by treating the features as explanation variables involved in a model of probabilistic inference [14 - 18].

In practice, a prior information, which is required within the Bayesian methodology, can be distorted or unavailable in a full volume. For example, the nature of flare occurrences is not yet fully understood in terms of such features as brightness, magnetism, and topology of the observable regions of solar activity. Obviously, in such cases the domain experts cannot provide a high volume of a prior information, and therefore the Bayesian methodology cannot yield the optimal estimates [11, 13]. However, when the prior information is fully available, the Bayesian methodology provides the best performance, although this technique is still computationally expensive. Fortunately, this obstacle can be overcome by using Markov Chain Monte Carlo (MCMC) technique of stochastic approximation [13, 14].

In this paper we aim to explore the potential of the Bayesian DT MCMC technique on the benchmark solar flares data taken from the Machine Learning Repository [19]. The comparisons are made in terms of the predictive accuracy and uncertainty in classification outcomes estimated within the Uncertainty Envelope Technique described in [20].

Further in Section 2 we describe the bases of the MCMC sampling for the Bayesian averaging over DT models which can be easily interpreted by domain experts. In Section 3 we explore the conditions under which the Bayesian DT averaging allows the posterior distribution to be estimated accurately. The application of the Bayesian DT technique to the classification of solar flares is described in Sections 4 and 5, and finally Section 6 concludes the paper.

## 2 The Bayesian Decision Tree Technique

Bayesian model averaging methodology allows the uncertainty in classification outcomes to be evaluated. In this section first we consider how the Bayesian approach can be practically implemented on the base of the MCMC technique of stochastic approximation. Then we consider the use of DT models for probabilistic interpretation of the classification outcomes which gives domain experts useful information for understanding.

### 2.1 Bayesian Averaging over Decision Trees

The main idea of DT classification models is to recursively partition data points in an axis-parallel manner. Such models provide natural feature selection and uncover the features which make the important contribution to the classification. The resultant DT classification models can be easily interpretable by domain experts [11, 17].

By definition, DTs consist of splitting and terminal nodes, which are also known as tree leaves. DTs are binary if the splitting nodes ask a specific question and then divide the data points into two disjoint subsets, assigned to the left and the right branches [11, 17].

The number of the data points in each split should not be less than that predefined by a user, which has to properly specify this number; otherwise DT model can lose

the ability to generalise well. All data points fallen in a terminal node are assigned to a class of majority of the training data points residing in this terminal node. Within a Bayesian framework, the class posterior distribution is calculated for each terminal node [11 - 14].

The Bayesian MCMC methodology has revealed promising results in the applications to some real-world problems [13 - 16]. To deal with large DTs, Chipman *et al.* [13] and recently Denison *et al.* [14] have developed the MCMC techniques using the RJ extension suggested by Green [21]. These techniques make the moves such as *birth* and *death* in order to induce large DTs under the priors given on the shape or size of the DTs. In the theory, RJ MCMC technique exploring the posterior distribution has to keep the balance between the birth and death moves which is required to obtain the desired estimates of the posterior unbiased [13, 14, 21].

Within the existing RJ MCMC techniques the proposed moves are assigned unavailable when the number of data points, falling in one of splitting nodes, becomes less than the given number. In practice, a user can improperly set up an acceptable number of data points in splits as well as the priors on favourite shape of the DTs. In such cases, the resultant estimates of class posterior distributions become biased [12, 13].

Moreover, within the standard RJ MCMC technique suggested for Bayesian DT averaging, the desired balance between the birth and death moves practically cannot be achieved as shown in [16]. This observation is based on the facts that the RJ MCMC technique exploring DTs makes some moves unavailable because of an unacceptable number of data points in splits. As a result, such moves cause a disproportion in the given probabilities of moves. Next we describe the standard Bayesian RJ MCMC technique.

## 2.2 The Methodology of Bayesian Averaging

In general, the class posterior distribution we are interested in is written as an integral over parameters  $\theta$  of the classification model

$$p(y | \mathbf{x}, \mathbf{D}) = \int_{\theta} p(y | \mathbf{x}, \theta, \mathbf{D}) p(\theta | \mathbf{D}) d\theta, \quad (1)$$

where  $y$  is the predicted class (1, ...,  $C$ ),  $\mathbf{x} = (x_1, \dots, x_m)$  is the  $m$ -dimensional input vector, and  $\mathbf{D}$  denotes the given training data.

In practice, except some simple cases, the posterior density  $p(\theta | \mathbf{D})$  cannot be evaluated analytically. However, when values  $\theta^{(1)}, \dots, \theta^{(N)}$  are drawn from the posterior distribution  $p(\theta | \mathbf{D})$ , we can write:

$$p(y | \mathbf{x}, \mathbf{D}) \approx \sum_{i=1}^N p(y | \mathbf{x}, \theta^{(i)}, \mathbf{D}) p(\theta^{(i)} | \mathbf{D}) = \frac{1}{N} \sum_{i=1}^N p(y | \mathbf{x}, \theta^{(i)}, \mathbf{D}). \quad (2)$$

This is the basis of the MCMC technique for approximating integrals (1) [13, 14]. To perform such an approximation, we need to run a Markov Chain until it converges to a stationary distribution. After this we can draw  $N$  samples from the Markov Chain and calculate the class posterior density (2).

### 2.3 Reversible Jump MCMC

To sample models of a variable dimensionality, it has been suggested to extend the MCMC method by the Reversible Jumps (RJ) [21]. The RJ MCMC technique allows large DT models to be sampled from real data [13 - 15]. Within this technique the posterior probability is explored by using the following types of moves.

1. *Birth*. Randomly split the data points falling in one of the terminal nodes by a new splitting node with the variable and rule drawn from the corresponding priors.
2. *Death*. Randomly pick a splitting node with two terminal nodes and assign it to be one terminal with the united data points.
3. *Change-split*. Randomly pick a splitting node and assign it a new splitting variable and rule drawn from the corresponding priors.
4. *Change-rule*. Randomly pick a splitting node and assign it a new rule drawn from a given prior.

The first two moves, birth and death, are reversible and change the dimensionality of  $\theta$  as described in [13, 14, 21]. The remaining moves make jumps within the current dimensionality of  $\theta$ . Note that the change-split move is included to make “large” jumps which can increase the chance of sampling from a maximal posterior, whilst the change-rule moves do “local” jumps.

Because of a hierarchical structure of DTs, the changes at the nodes located at the upper levels can significantly change the location of data points at the lower levels. For this reason there is a very small probability of changing and then accepting a DT split located near a root node. As a result, the RJ MCMC algorithm cannot explore a full posterior distribution properly.

One way to extend the search space is to restrict DT sizes during a given number of the first burn-in samples as described in [14]. This strategy, however, requires setting up in an *ad hoc* manner the additional parameters such as the size of DTs and the number of the first burn-in samples.

Alternatively, the search space can be extended by using a restarting strategy described in [13]. Clearly, both strategies cannot guarantee that most of DTs will be sampled from a model space region with a maximal posterior. In the next section we describe our approach based on a sweeping strategy.

## 3 The Bayesian Averaging with a Sweeping Strategy

The main idea of using a sweeping strategy is to assign the prior probability of further splitting DT nodes to be dependent on the range of values within which the number of data points will be not less than a given number of points. Such a prior is explicit because at the current partition the range of such values is unknown.

Formally, the probability  $P_s(i, j)$  of further splitting at the  $i$ th partition and variable  $j$  can be written as:

$$P_s(i, j) = \frac{x_{\max}^{(i, j)} - x_{\min}^{(i, j)}}{x_{\max}^{(1, j)} - x_{\min}^{(1, j)}}, \quad (3)$$

where  $x_{\min}^{(i,j)}$  and  $x_{\max}^{(i,j)}$  are the minimal and maximal values of variable  $j$  at the  $i$ th partition level.

For all the partition  $i > 1$  we can see that  $x_{\max}^{(i,j)} \leq x_{\max}^{(1,j)}$  and  $x_{\min}^{(i,j)} \geq x_{\min}^{(1,j)}$ , and therefore there is partition  $k$  at which the number of data points becomes less than a given number  $p_{\min}$ . Therefore, probability  $P_s$  ranges between 0 and 1.0 for any variable  $j$ , and its value is dependent on the level  $i$  of partitioning a data set.

Form this point of view, prior (3) favors splitting the terminal nodes which contain a large number of data points. This allows accelerating the convergence of Markov chain and, therefore, the RJ MCMC technique can explore an area of a maximal posterior in more detail.

To make the birth and change moves within such a prior, the new splitting values  $s_i^{\text{rule,new}}$  for the  $i$ th node and variable  $j$  are drawn from a uniform distribution  $s_i^{\text{rule,new}} \sim U(x_{\min}^{1,j}, x_{\max}^{1,j})$ . Likewise, for the change-split moves, a new variable is assigned as follows  $s_i^{\text{var,new}} \sim U\{S_k\}$ , where  $S_k$  is the set of features except variable  $s_i^{\text{var}}$  currently used at the  $i$ th node.

For the change-rule moves, the value  $s_i^{\text{rule,new}}$  is drawn from a Gaussian with a given variance  $\sigma_j$ :

$$s_i^{\text{rule,new}} \sim N(s_i^{\text{rule}}, \sigma_j), \tag{4}$$

where  $j = s_i^{\text{var}}$  is the variable used at the  $i$ th splitting node.

For some moves, the number of data points becomes less than a predefined number  $p_{\min}$ . Within the existing Bayesian DT techniques such moves are assigned unavailable [13, 14].

Within our approach after making the birth or change move, three possible cases arise. In the first case, the number of data points in all the partitions is larger than  $p_{\min}$ . In the second case, the number of data points in one partition becomes larger than  $p_{\min}$ . In the third case, the number of data points in two or more partitions becomes larger than  $p_{\min}$ . These cases are processed as follows.

For the first case, the RJ MCMC algorithm runs as usual. For the second case, a node with an unacceptable number of data points in the split is removed from the current DT. If the move was of the birth type, the RJ MCMC just resamples the current DT; otherwise, this move is considered as the death move. For the last third case, the RJ MCMC algorithm simply resamples the DT.

Because the unacceptable nodes are removed from the DT, we named such a strategy *sweeping*. Next we describe the application of the Bayesian DT using such a strategy to the solar flares data.

## 4 Application to Solar Flares Data

The solar flares data were taken from the Machine Learning Repository [19]. These data contain three classes of the observations represented by the number of times of a certain type of solar flares occurred in a 24 hour period in various active regions allocated to different groups of sunspot complexities. Each observation represents captured features for one active region on the Sun. The total number of the observations is 1066, and each observation is presented by 10 features listed in Table 1.

**Table 1.** The solar flares data features

#	Features	Values
1	Code for class (modified Zurich class)	{A,B,C,D,E,F,H}
2	Code for largest spot size	{X,R,S,A,H,K}
3	Code for spot distribution	{X,O,I,C}
4	Activity	{1 = reduced, 2 = unchanged}
5	Evolution	{1 = decay, 2 = no growth, 3 = growth}
6	Previous 24 hour flare activity code	{1 = nothing as big as an M1, 2 = one M1, 3 = more activity than one M1}
7	Historically-complex	{ 1 = Yes, 2 = No}
8	Did region become historically complex on this pass across the sun's disk	{1 = yes, 2 = no}
9	Area	{1 = small, 2 = large}
10	Area of the largest spot	{1 = <=5, 2 = >5}

Three classes of flares are predicted as C, M, and X classes. However, in our first experiments this domain problem was considered as a 2-class problem of predicting either flare or non-flare outcome.

In these experiments, no prior information on the preferable DT shape and size was available. The minimal number of data point allowed being in the splits was set to 1. The proposal probabilities for the death, birth, change-split and change-rule moves were set to 0.1, 0.1, 0.1, and 0.7, respectively. The numbers of burn-in and post burn-in samples were set to 50000 and 5000, respectively.

All these parameters of the MCMC sampling were the same for the standard and proposed Bayesian DT techniques. The performance of these techniques was evaluated within 5 fold cross-validation and  $2\sigma$  intervals. The uncertainty in classification outcomes was evaluated within the Uncertainty Envelope technique providing the rate of sure correct classifications as described in [20]. Next we present the experimental results.

## 5 Experimental Results

Both Bayesian DT techniques with the standard (DBT1) and the suggested (BDT2) strategies have correctly recognized 82.1% and 82.5% of the test examples, respectively. The average number of DT nodes was 17.5 and 10.1, respectively. Table 2 shows the obtain results. This table shows also the rate of sure correct classifications which in accordance with the Uncertainty Envelope Technique is proportional to the classification confidence.

**Table 2.** The performance and size of the BDT1 and BDT2 on the Solar Flares Data

Strategy	Number of DT nodes	Perform, %	Sure correct, %
BDT1	17.5±1.5	82.1±4.5	67.4±3.3
BDT2	<b>10.1±1.6</b>	82.5±3.8	<b>70.2±4.2</b>

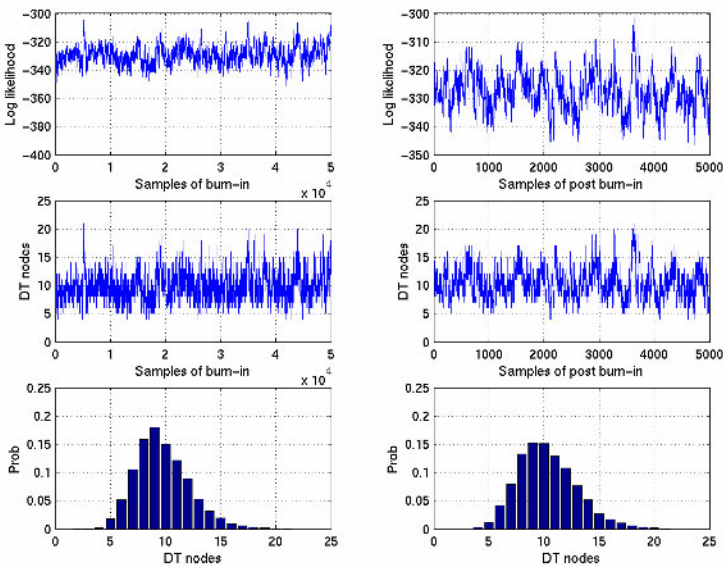
From Table 2, we can see that both strategies reveal the same performance on the test data. However, the number of DT nodes induced by the suggested BDT2 strategy is much less than that induced by the standard BDT1 strategy. Besides, the BDT2 strategy provides more reliable classifications than the BDT1 strategy: the rate of sure correct classification provided by the BDT2 is higher than that of the BDT1.

Fig. 1 depicts the samples of log likelihood and numbers of DT nodes as well as the densities of DT nodes collected during the burn-in and post burn-in phases for the suggested BDT2 strategy. From the top left plot of these figures we can see that the Markov chain very quickly converges to the stationary value of log likelihood near to  $-320$ . During the post burn-in phase the values of log likelihood slightly oscillate around this value that allows us to conclude that the samples of DTs are drawn from a stable Markov Chain.

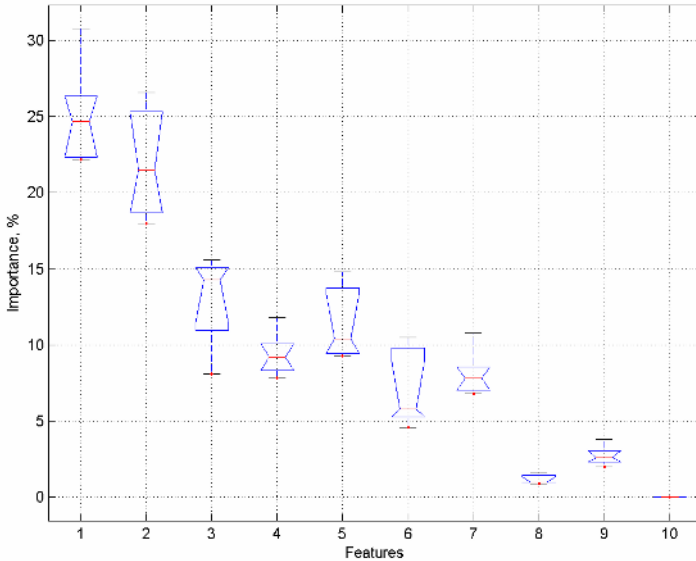
Fig. 2 depicts the contributions of the 10 features to the classification outcome. The feature importance is estimated in terms of the posterior weights with which the features were used in the DT models collected during the post burn-in phase.

From Fig. 2 we see that the most important contribution are made by features  $x_1$  (Code for class) and  $x_2$  (Code for largest spot size). Much less contribution are made by features  $x_8$  (the region complexity),  $x_9$  (Area) and  $x_{10}$  (Area of the largest spot).

Thus, based on these experimental results, we can conclude that the suggested BDT2 strategy using a sweeping strategy allows the classification uncertainty to be decreased without affecting the classification accuracy. At the same time, the suggested Bayesian strategy provides the shortest DTs which are easy-to-understand by domain experts.



**Fig. 1.** The BDT2 strategy. Samples of the burn-in and post burn-in



**Fig. 2.** The importance of the 10 features

As the Bayesian DT techniques require extensive calculations, the computational time required to run these techniques becomes important for real-world applications. Having compared the computational time in our experiments, we found that for the suggested BDT2 technique the computational time is less than that for the standard BDT1 technique on average on 20%. We can explain this by reducing the size of DTs induced by the suggested BDT2 strategy.

## 6 Conclusions

The classification technique we developed on the basis of Bayesian DT methodology has revealed promising results in our experiments on predicting the occurrences of solar flares. Both the standard and suggested Bayesian DT techniques provide a high performance in terms of predictive accuracy. However, the suggested technique using a sweeping strategy outperforms the standard Bayesian DT technique in terms of classification uncertainty. The suggested technique also provides shortest DTs which can be easily interpreted by domain experts.

The implementation of the Bayesian DT model averaging methodology is still computationally expensive and, therefore, further research should be done in this direction in order to reduce the computational expenses and make this methodology applicable to large-scale problems. Further research should be also done in order to verify the advantages of the proposed technique on new domain problems.

Overall, based on the obtained results, we believe that the Bayesian DT technique presented in this paper can be successfully applied to solar data. This technique is able to provide high performance, accurate estimates of uncertainty in classification outcomes, as well as the interpretability of models.



**Acknowledgments.** This work was partially (V. Zharkova) supported by the project European Grid of Solar Observations (EGSO), funded by the European Commission, Grant IST-2001-32409, and (V. Schetinina) by EPSRC, Grant GR/R24357/01. The authors also are thankful to the two anonymous reviewers for their constructive comments.

## References

1. Bentley R.D. *et al*: The European Grid of Solar Observations. In: Proc. Solar Cycle and Space Weather Euro Conference, Vico Equense, Italy (2001) 603
2. Turmon M., Pap J., Mukhtar S.: Automatically Finding Solar Active Regions Using SOHO/MDI Photograms and Magnetograms. In: Proc. Structure and Dynamics of the Interior of the Sun and Sun-like Stars, Boston (1998)
3. Zharkova V.V., Ipson S.S., Zharkov S.I., Benkhalil A., Abouharham J., Bentley R.D.: A Full Disk Image Standardization of the Synoptic Solar Observations at the Meudon Observatory. *Solar Physics* 214/1 (2003) 89
4. Zharkova, V.V., Ipson. S.S., Zharkov, Abouharham, J., Benkhalil, A.K., Fuller, N.: Solar Feature Catalogues in EGSO. *Solar Physics*, 228/1 (2005) 139-150
5. Zharkova V.V., Ipson S. S., Qahwaji R., Zharkov S., Benkhalil A.: An Automated Detection of Magnetic Line Inversion and Its Correlation with Filaments Elongation in Solar Images. In: Proc. SMMSP-2003, Barcelona, Spain (2003) 115-121
6. Bader D.A., Jaja J., Harwood D., Davis L.S: Parallel Algorithms for Image Enhancement and Segmentation by Region Growing with Experimental Study. In: Proc. IEEE IPSP'96 (1996) 414
7. Gao J., Zhou M., Wang H.: A Threshold and Region Growing Method for Filament Disappearance Area Detection in Solar Images. In: Proc. Information Science and Systems, Johns Hopkins University (2001)
8. Turmon M., Mukhtar S., Pap J.: Bayesian Inference for Identifying Solar Active Regions. In: Proc. Knowledge Discovery and Data Mining (1997)
9. Koppurapu S., Desai U.: Bayesian Approach to Image Interpretation. Kluwer (2002)
10. Duda, R.O., Hart, P.E.: Pattern Classification. Wiley Interscience (2001)
11. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Belmont, Wadsworth (1984)
12. Buntine, W.: Learning Classification Trees. *Statistics and Computing* 2 (1992) 63-73
13. Chipman, H., George, E., McCulloch, R.: Bayesian CART Model Search. *J. American Statistics* 93 (1998) 935-960
14. Denison, D., Holmes, C., Mallick, B., Smith, A.: Bayesian Methods for Nonlinear Classification and Regression. Wiley (2002)
15. Schetinina, V., Partridge, D., Krzanowski, W.J., Everson, R.M., Fieldsend, J.E., Bailey, T.C., Hernandez, A.: Experimental Comparison of Classification Uncertainty for Randomized and Bayesian Decision Tree Ensembles. *J. Math. Modeling and Algorithms* 4 (2006) forthcoming
16. Schetinina V., Fieldsend J.E., Partridge D., Krzanowski W.J., Everson R.M., Bailey T.C., Hernandez A.: The Bayesian Decision Tree Technique with a Sweeping Strategy. In: Advances in Intelligent Systems - Theory and Applications, In cooperation with the IEEE Computer Society, Luxembourg (2004)
17. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
18. Kuncheva, A.: Combining Pattern Classifiers: Methods and Algorithms. Wiley (2004)

19. Blake, C.L., Merz, C.J.: UCI Repository of Machine Learning Datasets. [www.ics.uci.edu/~mlearn/MLRepository](http://www.ics.uci.edu/~mlearn/MLRepository). Irvine, University of California (1998)
20. Fieldsend J.E., Bailey T.C., Everson R.M., Krzanowski W.J., Partridge D., Schetinin V.: Bayesian Inductively Learned Modules for Safety Critical Systems. In: Proceedings of the 35th Symposium on the Interface, Computing Science and Statistics, US, Salt Lake City (2003)
21. Green, P.: Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* 82 (1995) 711-732

# An Efficient 3-D Positioning Method from Satellite Synthetic Aperture Radar Images

Yeong-Sun Song, Hong-Gyoo Sohn, and Choung-Hwan Park

School of Civil Eng., Yonsei University, 134, Seodaemooon-Gu, Seoul, Korea  
{point196, Sohn1, c142520}@yonsei.ac.kr

**Abstract.** This paper describes an efficient 3-D positioning method of SAR image only using ephemeris data and a single control point. The proposed method is applied to RADARSAT-1 SGF images of which ephemeris information is not accurate enough to directly use for the dynamic geometric modeling. The RMSE of the 3-D model by using a single CP (Control Point) is 43.86m, while that of traditional method using 10 CPs is 45.79m in comparison with reference 3-D model generated from 1:5,000 digital topographic maps of the study area. The test results show that even without the use of many CPs, it is possible to generate similar accuracy of 3D-model comparable to traditional method.

## 1 Introduction

Accurate orbit parameters determination of satellite is essential process for 3-D positioning from satellite imagery. Provided that accurate state vector of ephemeris data is available, together with precise SAR (Synthetic Aperture Radar) processing parameters, orbit parameters can be precisely determined and CPs (Control Points) are not required. That is the case of ERS SAR imagery, which has a geocoding accuracy of 10m [3]. However, most satellites have a low accurate state vector on its orbit data due to the precession, the nutation, or the polar motion etc., then the orbit refinement process should be done by a number of well-distributed CPs. In the geometric processing, acquisition of a sufficient number of well distributed CPs in SAR images is often very difficult and time consuming and is proved to be a problem in using SAR data for various applications. The quality of the provided CPs is not always reliable due to speckle, poor illumination of features appearing in SAR images, mapping accuracy, and scale. Also, the quality of the measurements depends on the experience of the human operator.

A number of studies about 3-D positioning by radargrammetry have been proposed [6-9]. However, there were few studies that made an effort to reduce a number of CPs for efficient 3-D positioning. Chen and Dowman [2] addressed the particular problem of sensor position errors in the along-track direction in the context of control point acquisition for stereo measurements. They developed a weighted least-squares algorithm that made the radargrammetric model more robust to azimuth errors and showed that their method can significantly reduce the number of necessary ground control points and the requirements on their spatial distribution. Smith [5] described a near real-time geocoding method to effectively remove orbit errors from computed azimuth and slant-range coordinates using the error function so that

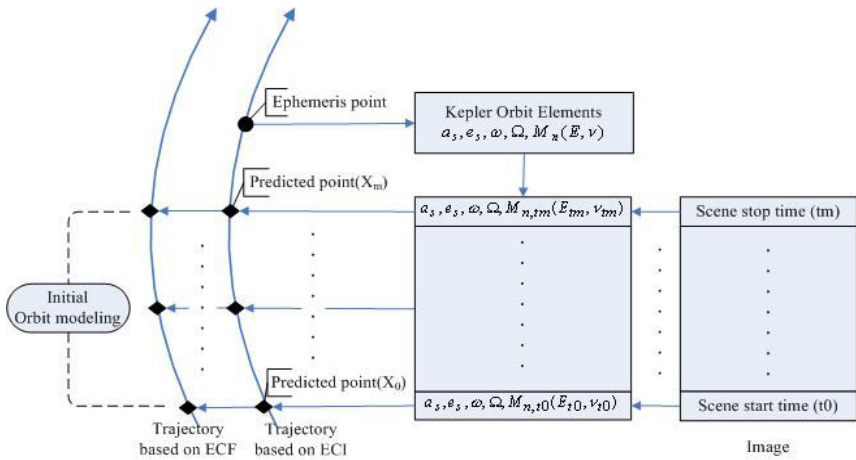
same geocoding accuracy can be obtained as when using precise orbits. The coefficients of the error functions was estimated from number of at least two ground control points in the SAR image and a digital land map displayed in two viewers. Sohn and others [4] proposed a method that determined the initial orbit from an ephemeris data, then calculated the systematic errors from a single GCP (Ground Control Point). They moved the geometrically corrected SAR image by a mount of systematic errors to reduce GCP requirements. This study investigated that the initial orbit determined by the ephemeris data has the systematic relationship between the object space and the image space. The proposed methods by Smith [5] and Sohn and others [4], however, could not estimate the accurate satellite orbit, 3-D positioning process could be problematic.

This paper proposes the concept of Pseudo Control Points (PCPs) that were defined by assuming the systematic relationship between the object and the image space if only ephemeris information is used. PCPs can be used instead of real CPs to refine the initial orbit parameters that are calculated by only ephemeris data. The seven RADARSAT-1 processing level 1 SGF images were used for the test. For the accuracy assessment of orbit parameters determined by a single CP, 3-D positioning was performed. The results from our method were compared with that of traditional method using not pseudo but actual 10 CPs.

## 2 Satellite Orbit Parameters Determination

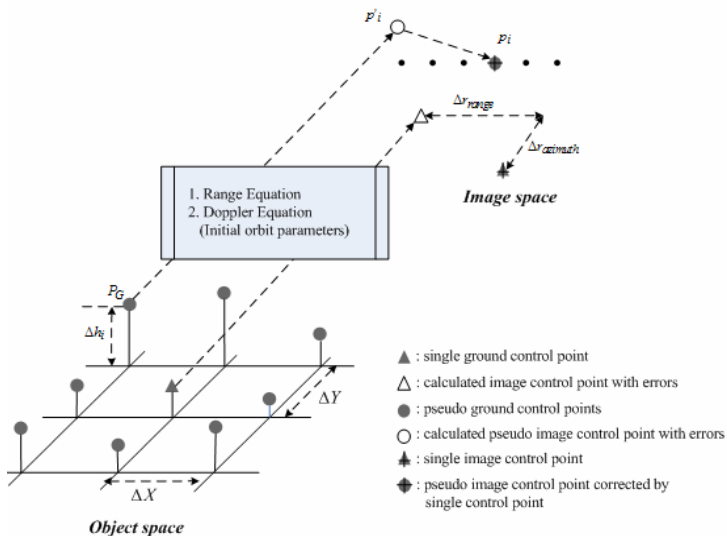
In order to uniquely define the position and velocity of satellite at any instant time, six parameters should be required to estimate. They are the three position components and three velocity components. In addition to these parameters, there are other forms of orbit representation which have more geometrical significance. One such representation is Kepler elements. The six components of Kepler elements are the semi-major axis ( $a$ ), the eccentricity ( $e$ ), the inclination ( $i$ ), the argument of perigee ( $\omega$ ), the right ascension of ascending node ( $\Omega$ ) and the mean anomaly ( $M$ ). These elements uniquely decide the position and velocity at a given time commonly known as an epoch time. The main advantage of using the orbital elements is that they are related to the equations of motion of a satellite. This advantage allows us to extrapolate a single state vector to determine its position and velocity at next time interval. All of the elements except for the anomaly are fixed to a purely Keplerian orbit. The only varying element as time to an orbit is the anomaly, whether it is true, eccentric or mean.

Satellite's position and velocity vector at any time can be calculated from Kepler's law and ephemeris data in header information. The image acquisition time is listed in header information, so the position and the velocity with respect to all image lines can be calculated. If satellite position and velocity are 2<sup>nd</sup> polynomial equations, the equations can be solved by more than three satellite positions and velocity vectors. Fig. 1 shows the algorithm to determine initial orbit parameters from only ephemeris data.



**Fig. 1.** Algorithm to extrapolate an orbit from a single state vector and determination of initial orbit parameters

The initial orbit parameters calculated by erroneous ephemeris data must be refined by well defined CPs and least square adjustment. For appropriate corrections, a sufficient number of CPs should be required to consider the degree of freedom and should be distributed well over the image. The required number of CPs can be substantially reduced by which satellite orbit parameters determined by ephemeris data has a systematic shift between the ground coordinates and the image coordinates [1][2]. Using the systematic shift, a single CP can produce a necessary amount of PCPs (Fig. 2).



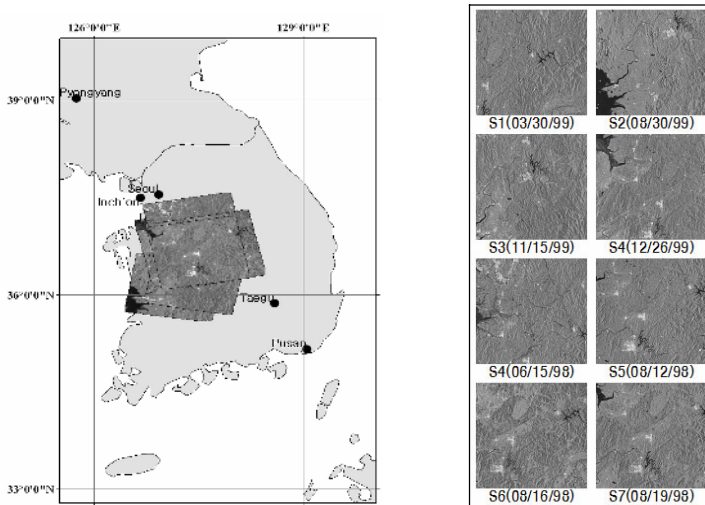
**Fig. 2.** Generation of PCPs from a single CP

When the real ground coordinate ( $\blacktriangle, P_G$ ) are transformed into the image coordinate ( $\triangle, p_i'$ ) using initial orbit parameters and Doppler and range equation, the systematic shift ( $\Delta r_{range}, \Delta r_{azimuth}$ ) is caused. The shift simply means the distance difference between real image coordinate and image coordinate transformed from the real ground coordinate. PCPs are composed of virtual grid in ground space and transformed image points that are corrected by the shift. The PCPs can be used for least square adjustment of initial orbit parameters instead of real CPs, so we need only one CP to determine the accurate shift error vector.

### 3 Experiments and Results

#### 3.1 Study Area and Datasets

The methodology proposed in this study was applied to eight RADARSAT-1 SAR images acquired over the middle parts of Korea peninsula from different look angles and orbits. The area is characterized by a rugged topography where the elevation range from 29m to 580m with up to  $70^\circ$  slopes. Fig. 3 represents an overview of the target site and RADARSAT-1 SAR images of the study area. Table 1 summarizes the general characteristics of the images used for the test. The images are in ground range representation (ellipsoid projection without relief correction), orbit oriented and coded in 16 bits without any radiometric processing.



**Fig. 3.** Overview of target site and SAR images for test

CPs and check points were obtained from 1:5,000 scale digital topographic maps compiled photogrammetrically from 1:20,000 aerial photographs. The overall accuracy of digital topographic maps is about 2m horizontally and vertically. To extract the accurate GCPs were chosen mainly intersections of streams and road, bridges center crossing rivers. The customized GUI application program was developed for

the study using MFC 6.0 (Microsoft Foundation Class 6.0) to process the RADARSAT-1 image header file analysis, the initial satellite orbit determination, and the 3-D positioning.

**Table 1.** Description of the RADARSAT images

Scene no.	Beam mode	Acquisition date	Orbit	Full-scene size (azimuth, range)	Nominal resolution (range $\times$ azimuth)	Pixel spacing (m)
1	Standard 1	08/16/98	Asc.	(8971, 9158)	(25.0 $\times$ 25.0)	12.5
2	Standard 2	12/26/99	Asc.	(8547, 9109)	(24.3 $\times$ 27.3)	12.5
3	Standard 3	06/15/99	Asc.	(7816, 8903)	(27.3 $\times$ 27.3)	12.5
4	Standard 4	03/30/99	Asc.	(8078, 9036)	(25.5 $\times$ 27.2)	12.5
5	Standard 4	11/15/99	Des.	(8019, 9014)	(25.4 $\times$ 27.2)	12.5
6	Standard 5	08/19/98	Asc.	(8441, 8979)	(25.0 $\times$ 25.0)	12.5
7	Standard 6	08/12/98	Asc.	(8630, 8724)	(25.0 $\times$ 25.0)	12.5
8	Standard 7	08/30/99	Des.	(7918, 8958)	(19.9 $\times$ 27.1)	12.5

### 3.2 Satellite Orbit Parameters Determination

For initial orbital parameters, only the nearest state vector to the scene centre was selected and used to calculate Kepler elements of the corresponding images. Since RADARSAT-1 ephemeris data was not accurate enough to directly determine orbit parameters, Kepler elements from each image showed slight discrepancies when compared with the actual CPs.

2<sup>nd</sup> polynomial equation was selected as the satellite position and velocity equations, so we need more than three satellite positions and velocities to estimate the initial satellite orbit parameters. As for the three satellite positions, the first row, the centre row, and the last row, were calculated, then eighteen initial satellite orbit parameters were estimated for each image.

It is extremely important to have a highly accurate CPs using estimated parameters for generating accurate PCPs. A single CP can be visually and easily identified on whole SAR images and digital maps and derived the systematic shifts.

PCPs from a single CP were used to process the least square adjustment for the correction of the initial orbit parameters. For PCPs, ground uniform grid with 20km $\times$ 20km was generated and elevations range of each grid points were confined to a single CP $\pm$ 500m. The least square adjustment set the critical value to 0.0001 and stably converged within 5 iterations for all images.

### 3.3 3-D Positioning and Accuracy Assessment

For assessing the performance of 3-D positioning from a single CP, the orbit parameters were determined by two methods. Hereafter we refer that method I is a proposed method using single CP and method II is traditional method using 10 CPs.

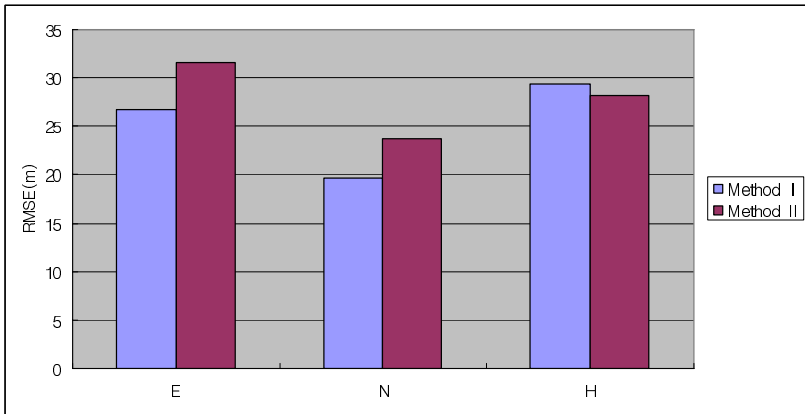
The 3-D stereo intersection was performed using the previously computed orbit parameters by method I, II to convert the image coordinates in both images to 3-D

data. The 8 sets of stereo image were selected from images in Table 1 based on covered areas and intersection angle. Table 2 shows the stereo sets for 3-D positioning and the number of check points for accuracy assessment.

**Table 2.** Stereo sets for 3-D positioning

	1	2	3	4	5	6	7	8
Used scene	2 and 7	2 and 6	1 and 7	3 and 7	3 and 6	5 and 7	5 and 8	1 and 4
No. of check points	10	12	11	14	10	13	12	11

We converted the above 10 check points for one stereo pair to terrain height by computing the rigorous stereo solution, that is, the intersection of the two range/Doppler circles which are defined by corresponding points in the two stereo images. Fig. 4 shows RMSE of check points by both methods.

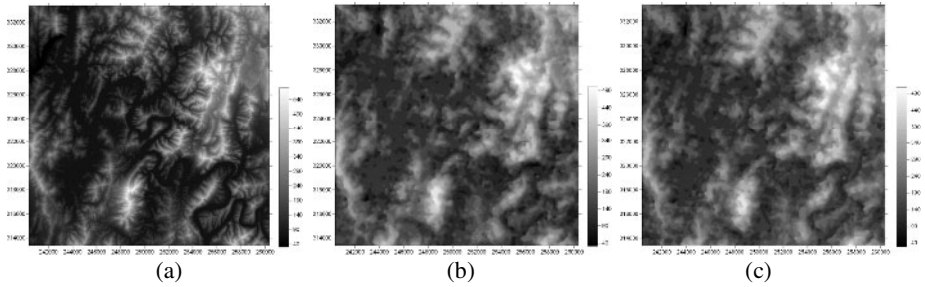


**Fig. 4.** Accuracy of 3-D positioning

From various stereo sets, Standard 1 and Standard 6, has selected by which its radiometric distortion is minimal and it is geometrically stable as intersection angle 21, and generated 3-D models. For the matching of noisy SAR image, a hierarchical correlation-based algorithm is applied in order to automatically determine the parallaxes between the stereo pairs. The images were filtered using Gamma Map filter in the initial processing stage to remove speckle noise as well as to smooth the images.

Using the method I and method II, 20m grid 3-D model was generated from the stereo RADARSAT-1 SAR images (Fig. 5). To assess the quality of the resulting 3-D models, they were compared with those of 3-D model generated from the contour of a 1:5000 scale digital topographic map. The 3-D models generated cover approximately 20km×20km and have 20m spatial resolution with Bessel TM 127°.





**Fig. 5.** 3-D models from (a) digital topographic maps (reference model), (b) method I (using a single CP, proposed method), and (c) method II (using ten CPs, traditional method)

The results of the comparison are summarized in Table 3. As shown in Table 3, RMSE of 3-D models generated by two methods are 43.86m and 45.79m, respectively. The reason that RMSE is larger than the pixel resolution is because the study area is composed of rugged terrain with high relief.

**Table 3.** Statistical comparisons of the two models (unit:m)

	Reference model	Method I	Method II
Mean height	174.52m	201.10m	200.576m
Maximum height	580.68m	514.33m	520.34m
Minimum height	29.54m	23.42m	31.58m
Mean height error	-	36.39m	37.39m
Height RMSE	-	43.86m	45.79m

## 4 Conclusion

In this paper we developed the determination method of orbital parameters using only a single CP for efficient 3-D modeling by reducing required CPs and applied the method to RADARSAT-1 SAR images that have relatively inaccurate ephemeris data. For the accuracy assessment of orbit parameters from a single CP, 3-D positioning was performed using orbit parameters determined by a single CP, and the results from our method and from 10 control points were compared.

For 3-D modeling results, 8 stereo images were tested by check points, and their average RMSE from a single CP were 26.8m in east-west direction, 19.6m in north-south direction, and 29.3m in elevation. From a stereo image from 10 control points, its RMSE was 31.6m, 23.7m, and 28.2m, respectively. Among 8 stereo sets, a selected set constructed 3-D model and assessed accuracy. A 3-D model from a single CP has 43.86m RMSE in elevation and the average error 36.39m, whilst a 3-D model from 10 CPs showed 45.79m RMSE in elevation and 37.39m in average errors.

Given a high accurate single CP and ephemeris data, the orbital parameters of satellite can be adjusted. Therefore the method proposed in this paper can be applied to the areas where it is difficult to collect well defined and good quality of CPs. The fact

that many CPs are not required for 3-D positioning translates to very significant cost and time savings for the user.

## Acknowledgements

The work was supported by grant No. NIDP-PR-2005-1-1 from practical use of National Institute for Disaster Prevention.

## References

1. Chen, P. H., Dowman, I. J.: Space Intersection from ERS-1 Synthetic Aperture Radar Images, *Photogrammetric Record*, 15(88), (1996) 561-573.
2. Chen, P. H., Dowman, I. J.: A Weighted Least Squares Solution for Space Intersection of Spaceborne Stereo SAR Data, *IEEE Trans. Geosci. Remote Sensing*, 39(2), (2001)233-2401
3. Mohr, J. J., Madsen, S. N.: Geometric correction of ERS Satellite SAR images, *IEEE Trans. Geosci. Remote Sensing*, 39(4), (2001)842-850.
4. Sohn, H. G, Song, Y. S, Kim, G. H., Bang, S. N.: A Rigorous Geometric Rectification of RADARSAT SAR Imagery Using a Single Control Point, *J. of The Korean Society of Civil Engineers*, 24(1D), (2004)107-115.
5. Smith, A. J. E.: Near Real-Time Geocoding of SAR Imagery with Orbit Error Removal, *Int. J. Remote Sensing*, 24(24), (2003)5219-5228.
6. Toutin, T., Opposite Side ERS-1 SAR stereo Mapping over Rolling Topography, *IEEE Trans. Geosci. Remote Sensing*, 34(2), (1996)543-549.
7. Toutin, T., Error Tracking of Radargrammetric DEM from RADARSAT Images, *IEEE Trans. Geosci. Remote Sensing*, 37(5), (1999)792-789.
8. Toutin, T., Evaluation of Radargrammetric DEM from RADARSAT Images in High Relief Areas, *IEEE Trans. Geosci. Remote Sensing*, 38(2), (2000)792-789.
9. Toutin, T., L. Gray, State-of-Art of Elevation Extraction from Satellite SAR Data, *ISPRS Journal of Photogrammetry & Remote Sensing*, 55, (2000)13-33.

# Generalized Logit Model of Demand Systems for Energy Forecasting

Hong Sok Kim<sup>1</sup>, Hoon Chang<sup>2</sup>, and Young- Kyun Lee<sup>3</sup>

<sup>1</sup> Dept. of Urban Planning and Eng., Yonsei Univ., Korea  
hskim66@yonsei.ac.kr

<sup>2</sup> Dept. of Urban Planning and Eng., Yonsei Univ., Korea  
hchang@yonsei.ac.kr

<sup>3</sup> Director ITS Team, Ministry of Construction & Transportation, Korea  
ykleefiu@hanafos.com

**Abstract.** A number of different versions of the Generalized Logit (GL) model have been applied in the literature, and the primary objective of the paper is to determine which one is the best. Using annual data for energy demand in the USA at the state level, the final model selected is similar to a simple form that was originally proposed by Considine and Mount (1984). A second objective of the paper is to demonstrate that the estimated elasticities are sensitive to the units specified for prices, and to show how price scales should be estimated as part of the model. The GL models of demand systems for energy and other factors have been shown to work well in comparison with other popular models, such as the Almost Ideal Demand System and the TransLog model. The main reason is that the derived price elasticities are robust when expenditure shares are small.

## 1 Introduction

Generalized Logit (GL) models have been used in a number of different applications to estimate demand systems for energy. Rothman et. al. [1] have shown that a GL model of consumer demand performed much better than the popular Almost Ideal Demand System (AIDS), proposed by Deaton and Muellbauer [2], or the TransLog (TL) model, proposed by Christensen et. al. [3] Although all three models gave similar estimates of price elasticities at the mean of the sample, the economic consistency of the AIDS and TL models tended to breakdown when expenditure shares differed from the mean values.

Reasons for using a GL model are not limited to judging its relative performance with other models. The structure of the GL model also enhances the types of analysis that can be conducted by making it possible to consider extreme situations which are not observed directly in the sample. For example, Dumagan and Mount [4] show how a GL model of a demand system, which includes electricity, natural gas and oil, can be used to represent an all-electric customer who is not affected by changes in the prices of natural gas or oil. In this case, the issue is how price elasticities behave when some expenditure shares are zero.

The basic tradeoff between using an AIDS model (or a TL model) and a GL model is that the structure of the GL model is more difficult to estimate. Hong [5] has shown

that a Generalized Barnett model, which was much harder to estimate than a GL model, also performs better than the AIDS and TL models using the United Nation’s data from 53 different countries.

The main complication of estimating a GL model compared to the AIDS model, for example, is that weighting functions for prices must be specified to approximate the symmetry restrictions. A variety of different parameterizations of the weighting function have been specified for GL models in the literature. The primary objective of this paper is to specify a general form of weighting function, and to determine which specific parameterization is supported best by the data. A secondary objective of the paper is to introduce a new issue concerning how to scale the price variables. This issue is shown to affect the economic properties of the estimated elasticities in a significant way.

## 2 Model Specification of the Generalized Logit Model

The general form of a demand system for  $n$  input factors can be written as a series of  $(n-1)$  linear regression equations:

$$\begin{aligned}
 y_i &= \alpha_{i0} + \alpha_{i1}x_{i1} + \dots + \alpha_{i(n-1)}x_{i(n-1)} + \beta_{i1}z_{i1} + \dots + \beta_{im}z_{im} + e_i \\
 &= \alpha_{i0} + \sum_{j=1}^{n-1} \alpha_{ij}x_{ij} + \sum_{k=1}^m \beta_{ik}z_{ik} + e_i \qquad i = 1,2,\dots,n-1
 \end{aligned}
 \tag{1}$$

where  $y_i$  is the dependent variable,  $x_{ij}$  a price variable,  $z_{ik}$  is a non-price variable, and  $e_i$  is a residual.

The GL model is a simple modification of the standard regression Eq. (1):

$$y_i = \alpha_{i0} + \sum_{j=1, j \neq i}^n \alpha_{ij}x_{ij} - \sum_{j=1}^{n-1} \alpha_{nj}x_{nj} + \sum_{k=1}^m \beta_{ik}z_{ik} + e_i \tag{2}$$

where

$$\begin{aligned}
 y_i &= \log(w_i / w_n) & i = 1,2,\dots,n-1, \\
 x_{ij} &= \theta_{ij} \log(p_j / p_i) & \text{for all } j \neq i
 \end{aligned}$$

where  $\theta_{ij}$  is a known function of  $w_i$  and  $w_j$ . In the GL models, the restrictions  $\alpha_{ij} = \alpha_{ji}$  are implied by economic theory. Even though the form of the regression equations in the GL model is complicated, the expressions for the Hicksian price elasticities for the GL are simple.

Cross-price

$$E_{ij} = \alpha_{ij}\theta_{ij} + w_j, \qquad \text{for all } i \neq j \tag{3a}$$

Own-price

$$E_{ii} = - \sum_{k=1, k \neq i}^n \alpha_{ik}\theta_{ik} + w_i - 1 \tag{3b}$$

The GL elasticities are not sensitive to small expenditure shares ( $w \rightarrow 0$ ) if the form of  $\theta_{ij}$  is specified appropriately.

### 3 Functional Form of the Cross-Price Weight in the GL Model

The functional form of  $\alpha_{ik}$  in a GL model Eq. (2) is critical in determining the properties of the price elasticities. All elasticity expressions in Eq. (3) have been derived conditionally on the value of  $\theta_{ij}$ . Using this simplification, any form of  $\theta_{ij}$  must satisfy the property  $w_i \theta_{ij} = w_j \theta_{ji}$  (for  $i \neq j$ ) to ensure that the symmetry conditions are met.

When  $w_j \rightarrow 0$ , it is desirable to have the cross-price elasticity  $E_{ij} \rightarrow 0$  because it implies that the demand for commodity  $i$  is unresponsive to changes in prices for commodities that are not purchased. The following forms of weighting scheme have been used in previous studies:

- (i)  $\theta_{ij} = w_j$ .
- (ii)  $\theta_{ij} = \frac{w_j^{1-\gamma}}{w_i^\gamma}$  where  $0 \leq \gamma \leq 1$  is a parameter.
- (iii)  $\theta_{ij} = w_i^{-\gamma} w_j^{1-\gamma}$  and  $\gamma \leq 0$ .
- (iv)  $\theta_{ij} = w_i^{-\gamma} w_j^{1-\gamma} (1 - w_i - w_j)$  and  $\gamma \leq 0$ .
- (v)  $\theta_{ij} = \frac{w_j}{(w_i + \delta)^\gamma (w_j + \delta)^\gamma}$  and  $\delta > 0$ .

For standard data sets which are characterized by substitution among commodities, forms (i), (ii) and (iii) are possible choices for  $\theta_{ij}$ , and all three can be approximated by (v) when  $\delta \rightarrow 0$ . All three cases exhibit the desirable property  $E_{ij} \rightarrow 0$  when  $w_j \rightarrow 0$ . However, the behavior of  $E_{ij}$  when  $w_i \rightarrow 0$  is determined by the sign of  $\gamma$ , and this has implications for the economic logic of the model. One would expect that the elasticity for changing price  $j$  (for a given  $w_j$ ) would be larger if  $w_i$  was small. Consequently, GL models with  $\gamma > 0$  should be preferred.

### 4 Estimation and Price Scaling

GL models with and without price scaling are estimated by using a range of values of  $\gamma$  and a specified value of  $\delta = .005$ . By varying  $\gamma$  from -1 to 1, the goodness of fit and the economic validity of the model can be determined and the effects of scaling prices assessed. First, any set of estimated price elasticities should be logical and consistent with economic theory. For estimation, the best value of  $\gamma$  is selected by finding the smallest determinant of the variance-covariance matrix of residuals across equations, which corresponds to the best fit of the model. A finer grid of  $\gamma$  values is used close to the best fit ( $\gamma = 1, 0.5, 0.1, 0.05, 0.01, 0, -0.01, -0.05, -0.1, -0.5, -1$ ).

The data used for estimation are a pooled cross-section of 48 states and an annual time-series from 1980 to 2001 (Residential) and 1988 to 2001 (Commercial & Industrial) using data from the Energy Information Administration and the Bureau of Economic Analysis.

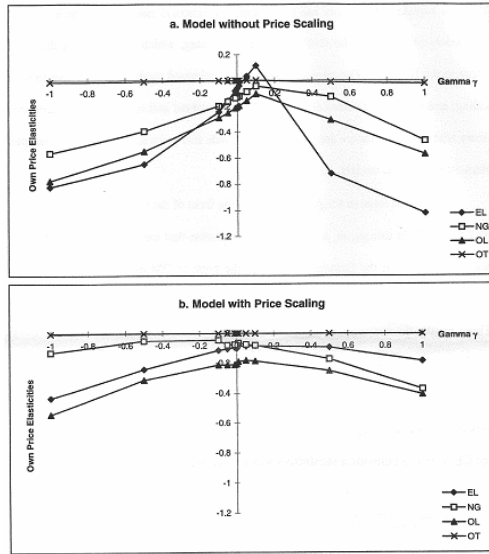


Fig. 1. The Estimated elasticities for different values of gamma

Fig. 1 shows a summary of the estimated own price elasticities in the Residential sector for different values of  $\gamma$ . The first observation is that the price elasticities, particularly for electricity in the model without price scaling, are sensitive to the value of  $\gamma$ . Demand is generally more price responsive at the extreme values of  $\gamma$  and less responsive for values close to zero. The second observation is that price scaling matters. In the model without price scaling, two of the cases violate economic logic and give price elasticities for electricity that are positive. In addition, the value of gamma with the best fit ( $\gamma = 0.01$ ) is close to the invalid models ( $\gamma = .05$  and  $.1$ ). The model with price scaling is consistent with economic theory for all values of  $\gamma$ , and the corresponding price elasticities are much more robust to different values of  $\gamma$ . Thus, the model with price scaling is preferred, and this conclusion is also reached in the Commercial and Industrial sectors.

For  $\delta = .005$ , the best fit is obtained at  $\gamma = .01, .075$  and  $.05$  for the Residential, Commercial, Industrial sectors, respectively. In order to get an economically valid model for the Industrial sector, weak separability among fossil fuels was imposed. This type of simplification is easy to impose using the restrictions  $\alpha_{ik} = \alpha_{jk}$  for all  $i, j$  that belong to the fossil-fuel group and all  $k$  that do not belong to that group.

## 5 Numerical Results

The matrices of estimated elasticities for the three sectors using the GL model are presented in Tables 1-3. Since the models include a dynamic adjustment process, estimates of both the short-run and long-run elasticities are given. The reported elasticities use the data for New York State in 2000 as the base point.

**Table 1.** The Estimated Demand Elasticities for Residential Sector

<b><u>Short Run Marshallian Income &amp; Price Elasticities</u></b>					
	Electricity	N gas	Oil	Other	Income
Electricity	-0.09767	0.03979	-0.00071	-0.84035	0.89894
N gas	0.07112	-0.07349	-0.00784	-0.80765	0.81787
Oil	0.00429	-0.00923	-0.19312	-0.13134	0.32939
Other	-0.01060	-0.00630	-0.00322	-0.98490	1.00502
<b><u>Short Run Hicksian Elasticities</u></b>					
	Electricity	N gas	Oil	Other	
Electricity	-0.08780	0.04539	0.00287	0.03954	
N gas	0.08010	-0.06840	-0.00459	-0.00711	
Oil	0.00791	-0.00718	-0.19181	0.19107	
Other	0.00044	-0.00005	0.00078	-0.00118	
<b><u>Long Run Marshallian Income &amp; Price Elasticities</u></b>					
	Electricity	N gas	Oil	Other	Income
Electricity	-0.47167	0.17652	0.01189	-0.28263	0.56586
N gas	0.39810	-0.35114	-0.02373	-0.27262	0.24940
Oil	0.16085	0.06583	-0.89820	2.81973	-2.14868
Other	-0.00912	-0.00638	-0.00040	-1.00666	1.02245
<b><u>Long Run Hicksian Elasticities</u></b>					
	Electricity	N gas	Oil	Other	
Electricity	-0.46066	0.18275	0.01587	0.69618	
N gas	0.40908	-0.34490	-0.01975	0.70619	
Oil	0.17184	0.07205	-0.89414	3.79854	
Other	0.00187	-0.00015	0.00358	-0.02775	

**Table 2.** The Estimated Demand Elasticities for Commercial Sector

<b><u>Short Run Price Elasticities</u></b>					
	Electricity	N gas	Oil & Coal	Capital	Labor
Electricity	-0.04744	0.02598	0.01556	0.03275	-0.02685
N gas	0.13502	-0.18649	0.05598	0.12371	-0.12821
Oil&Coal	0.08633	0.05974	-0.19266	0.38053	-0.33394
Capital	0.00244	0.00177	0.00510	-0.34246	0.33315
Labor	-0.00060	-0.00055	-0.00135	0.10054	-0.09803
<b><u>Long Run Price Elasticities</u></b>					
	Electricity	N gas	Oil & Coal	Capital	Labor
Electricity	-0.04682	0.05601	0.02884	0.03356	-0.07164
N gas	0.40422	-0.41604	0.13057	0.24739	-0.36625
Oil&Coal	0.33236	0.18585	-0.40085	0.93375	-1.05142
Capital	-0.01774	-0.00083	0.00713	-0.83853	0.84996
Labor	0.00331	0.00004	-0.00174	0.24746	-0.24906

In general, own price elasticities for all sources of energy in all three sectors are price inelastic in the short-run and in the long-run. Cross-price elasticities between

sources of energy are very small ( $|E_{ij}| < 0.1$ ) in the Residential and Industrial sectors, but generally exhibit strong substitutability ( $E_{ij} > 0.1$ ) in the Commercial sector. Complementary relationships between energy and non-energy exist in the Residential sector. Strong substitutability between energy factors and capital exists in both the Commercial and Industrial sectors. In contrast, all but one of the relationships between energy and labor are complementary. Labor and capital are strong substitutes in both the Commercial and Industrial sectors. One surprise in the Residential sector is that the long run income elasticity for oil is highly negative. This may reflect a general movement away from using oil for heating homes during the eighties.

**Table 3.** The Estimated Demand Elasticities for Industrial Sector

<b>Short Run Price Elasticities</b>						
	Electricity	N gas	Oil	Coal	Capital	Labor
Electricity	-0.19054	-0.00010	-0.00018	-0.00003	0.10498	0.89894
N gas	-0.00038	-0.11256	0.00217	0.00211	0.18996	-0.08132
Oil	-0.00038	0.00124	-0.10676	-0.00276	0.18996	-0.08132
Coal	-0.00038	0.00731	-0.01665	-0.09894	0.18996	-0.08132
Capital	0.01047	0.00509	0.00890	0.00147	-0.59988	0.57395
Labor	0.00367	-0.00094	-0.00163	-0.00027	0.24613	-0.24697
<b>Long Run Price Elasticities</b>						
	Electricity	N gas	Oil	Coal	Capital	Labor
Electricity	-0.35704	-0.01227	-0.00974	-0.00234	0.35870	0.02261
N gas	0.00225	-0.35350	-0.00039	-0.00332	1.07140	-0.71659
Oil	0.00806	0.08300	-0.26201	-0.01121	0.09881	0.08325
Coal	0.00431	0.06738	-0.01212	-0.24444	0.44219	-0.25744
Capital	0.02815	0.02207	0.00655	0.00357	-1.44409	1.38371
Labor	0.00300	-0.00676	0.00291	-0.00035	0.58817	-0.58694

One issue about the form of the GL models deserves further elaboration, and this relates to the relatively large number of negative estimates of the cross price coefficients ( $\alpha_{ij}$ ). Given the chosen form of the cross price weight ( $\theta_{ij}$ ), the cross price elasticities for the GL model in Eq. (3) can be written:

$$E_{ij} = w_j \left[ \alpha_{ij} / \left( (w_i + \delta)^\gamma (w_j + \delta)^\gamma \right) + 1 \right] \tag{4}$$

Consequently, the sign of  $\alpha_{ij}$  determines whether  $E_{ij}$  increases more than ( $\alpha_{ij} > 0$ ) or less than ( $\alpha_{ij} < 0$ ) proportionally with  $w_j$ . If  $\alpha_{ij}$  is sufficiently negative, then  $E_{ij}$  is also negative and the relationship is complementary.

Since the effect of  $\alpha_{ij}$  in Eq. (4) is largest, for any given  $w_j$  and  $\gamma > 0$ , when  $w_i = 0$ , the discussion will focus on how  $E_{ij}$  changes as  $w_j$  increases from 0 to 1 holding  $w_i = 0$ . If  $\alpha_{ij} > -\delta^{2\gamma}$ ,  $E_{ij}$  is always positive (substitute), and if  $\alpha_{ij} < -(1 + \delta)^\gamma \delta^\gamma$ ,  $E_{ij}$  is always negative (complement). For  $-(1 + \delta)^\gamma \delta^\gamma < \alpha_{ij} < -\delta^{2\gamma}$ ,  $E_{ij} < 0$  for small  $w_j$  and  $E_{ij} > 0$ , as economic theory requires, if  $w_j \rightarrow 1$ . Given these desirable properties, one could consider using  $-\delta^{2\gamma}$  as a lower bound for  $\alpha_{ij}$ , but the problem with this restric-



tion is that the magnitude of  $|E_{ij}|$  is too small to capture strong complementary relationships. Hence, the presence of  $\alpha_{ij} < -(1 + \delta)^\gamma \delta^\gamma$  must be accepted as a possibility, and the economic logic of the model would only hold for a limited range of  $w_i$  and  $w_j$ .

It is interesting to note that capital and labor are found to be substitutes in Table 2 and 3. Since these two factors generally account for almost 90% of total expenditures, finding complementary relationships would have posed a potential problem. If complementaries are important in a particular application, then modifying form (v) of  $\theta_{ij}$  could be considered using the same rationale adopted to convert form (iii) to form (iv) in Section 3.

## 6 Conclusion

This paper has focused on two practical issues related to estimating GL models of demand. The first issue is price scaling. The results show that the estimated models are sensitive to price scaling, and that the estimated elasticities are more robust and consistent with economic theory when price scales are estimated. We conclude that price scaling should be adopted when estimating GL models.

The second issue in the paper considers the form of the cross price weights ( $\theta_{ij}$ ) in the GL model. In this paper, a general form is chosen that can approximate a range of models discussed in the literature. The key parameter ( $\gamma$ ) that determines the form of  $\theta_{ij}$  is estimated using a grid search over the range -1 to 1. The estimated values are positive and close to zero in all three sectors. These results do not support the form of elasticity derived from a AIDS or TL model ( $\gamma = 1$ ), and are closest to the GL model proposed by Considine and Mount [6] ( $\gamma = 0$ ). They are consistent with the economic expectation of how price elasticities should change when expenditure shares change ( $\gamma > 0$ ). In this respect, the results provide more evidence that the GL model can provide a satisfactory way to represent demand systems for energy.

## References

1. Rothman, D.S., J.H. Hong and T.D. Mount (1994). "Estimating Consumer Energy Demand Using International Data: Theoretical and Policy Implications." *The Energy Journal* 15(2): 67-88.
2. Deaton, A. and J. Muellbauer (1980). "An Almost Ideal Demand System." *American Economic Review* 70(3): 312-326.
3. Christensen, L.R., Jorgenson, D.W. and Lau, L. (1975). "Transcendental Logarithmic Utility Functions." *American Economic Review* 65(3): 367-383.
4. Dumagan, J.C. and T.D. Mount (1996). "Global Properties of Well-Behaved Demand Systems: A Generalized Logit Model Specification." *Economic Modelling* 13: 235-256.
5. Hong, J-H. (1994). "The Performance of Alternative Flexible Functional Forms in Model Demand Systems: Theory and Application to Consumer Energy Demand." Ph.D. Dissertation, Agricultural, Resource and Managerial Economics, Cornell University.
6. Considine, T. and T.D. Mount (1984). "The Use of Linear Logit Models for Dynamic Input Demand Systems." *The Review of Economics and Statistics* 66(3): 434-444.

# Secure Authentication Protocol in Mobile IPv6 Networks

Jung Doo Koo<sup>1</sup>, Jung Sook Koo<sup>2</sup>, and Dong Chun Lee<sup>3</sup>

<sup>1</sup>Dept. of Computer Science and Eng., Hanyang Univ., Korea  
jdkoo@cse.hanyang.ac.kr

<sup>2</sup>Dept. of Information Security, Kyonggi Univ., Korea

<sup>3</sup>School of Computer Science, Howon Univ., Korea  
ldch@sunny.howon.ac.kr

**Abstract.** We propose new binding update protocol in Mobile IPv6. In our protocol, the home agent (HA) of each node creates Diffie-Hellman session key using middle key and safely transmits the middle key, cookie, etc via secure IPsec tunnel. These are used to generate secret keys in a mobile node (MN) and a correspondent node (CN). We present that the proposed protocol is satisfied with security requirements of key agreement protocol and shows the appropriate defense in existed attack scenarios.

## 1 Introduction

In Mobile IPv6 (MIPv6) [1], a MN is identified by its home address, which is within the home link of MN. If moving away from its home link, the MN is also associated with a care of address, which provided information such as router about its current location. The care of address must be registered with its HA, which transparently routes IPv6 packets sent to the home address to the care of address. If the MN moves foreign links, the MN is cut the connection with current communication nodes. Therefore, it must be binding update with existing communication nodes. But it is able to disclosure from many attack danger if binding update message transmits after is don't encrypt [5]. For easing this attack, various protocols is proposed [2-4]. But this protocol doesn't satisfy partially authentication requirements and have to process many computation quantities such as public key operations or exponential operations (i.e., Diffie-Hellman key establishment protocol).

This paper proposes the secure binding update protocol satisfying the upper items, which is divided with key agreement step, binding update step, and re-keying step. In the key agreement step, the MN sends the parameters using key agreement to its HA. In this step, we use the cookie to protect necessary parameters in key agreement and prevent the DoS attack and redirect attack. Also, it offers the user authentication through digital signature.

## 2 Related Work

To be safely the binding update, the existing protocols is divided largely two patterns.

A one type is the pattern that performs the binding update using the Return Routability (RR) scheme. The Binding Authentication Key Establishment (BAKE) protocol

uses the RR procedure for binding update. To create the session key in the BAKE, the CN transmits  $K_h$  (i.e., the result that applies Message Authentication Code (MAC) using MN's home address and nonce) and  $r_c$  (i.e., the result applying MAC that uses the care-of address and nonce of MN) to MN. After the CN combines with these values, it creates session key using hash function. But the CN in this protocol transmits the messages that don't encrypt  $K_h$  and  $r_c$ . The attacker is able to intercept these values because of not encoding this message. Therefore, there are dangers which are able to create session key.

A different type is the method that uses to combine Diffie-Hellman key exchange form with the identifiers associated with public key. A representative example is the very Purpose Built Keys (PBK) [4] and Extended Certificate Based Binding Update (ECBU)[2]. In the PBK, the MN sends many times a Purpose Built ID (PBID) applying hash function into public key. If the MN then comes about hand off, it transmits the signed binding update using a public key and PBID to its CN. The CN which receives this message verifies the signature using public key which covers the PBID. This protocol is advantages that don't require the special security infrastructure. But it is disadvantages that are weak the man-in-the-middle attack and that give the loads to the MN and its CN which may be the limit in battery and computation ability. The ECBU uses also the public key scheme. The HA at home links receives the public key authentication from the Certificate Authority (CA). It is similar to the PBK method. But it doesn't perform the public key computation and the exponential computation to MN. Only, these computations deal with the HA and the CN which is the plentiful battery, high bandwidth, and computation ability. Therefore, it is advantage that doesn't give computation loads. But, the CN may be the mobile device such as mobile phone, PDA limiting in the battery and computation quantities. In this case, the ECBU protocol may fit the ubiquitous environment which is able to use low power devices.

### 3 The Protocol

The proposed protocol is divided largely with middle key agreement step, binding update step, and re-keying step. First, the middle key generation courses safely exchange necessary parameters for creating the middle key into the HAs abounding with the battery and computation abilities. Then, it creates the key. Also, to transfer securely between the MN and its CN with messages, it generates the encryption key. In the binding update step, the MN and its CN create the binding update key and the binding acknowledgement key by using the stored middle key in advance.

Each HA in the protocol assumes that it is the wire and wireless devices which doesn't limit the battery and is full of the computation quantities such as the serve or workstation.

#### 3.1 Notations

- *MN/CN*: a mobile node and correspondent node.
- *HA*: a home agent.
- *MH/CH*: a mobile node's home agent and correspondent node's home agent.

- $HoA / CoA$ : a mobile node's home address and care of address.
- $MH_{ADDR} / CH_{ADDR}$ : a mobile node's home agent address and correspondent node's home agent address.
- $CN_{ADDR}$ : a correspondent node's address.
- $SIG_A$ : a digital signature on node A.
- $HMAC(K, M)$ : a keyed hash of message M with key K for authenticating message M.
- $H(M)$ : a one-way hash function such as SHA or MD5 and M is the message.
- $( )_K$ : a encoding of message with the secret key K.
- $+ K_A / -K_A$ : a node A's public key and private key.
- $K_{DH}$ : a Diffie-Hellman session key ( $K_{DH} = (g^{xy}) \bmod p$  ( $Z_p^* = \langle g \rangle$ ),  $p$  is large prime number and  $g$  is generator).
- $K_{BU} / K_{BA}$ : a key which uses to authenticate the binding update message and binding acknowledgement message.
- $K_{EN}$ : a encryption key using to encode messages between MN and CN.
- $K_{MN-MH} / K_{CN-CH}$ : a shared key between MN and MH and between CN and CH.
- $cookie_{A-B}^i$ : a  $i$ th cookie between node A and node B[6].
- $N_A^i$ : a  $i$ th nonce of node A.
- $NAI_A^i$ : a  $i$ th network access identifier[7,8] of node A.
- $LT/T_A$ : a lifetime and timestamp of node A.
- $SN$ : a sequence number.
- $AllB$ : a bitwise concatenation of message A and B.

### 3.2 Protocol Description

The purpose of this step is thing which exchanges safely necessary parameters for creating the middle key. Also it is purpose to decrease load for exponential operation in the MN and CN. Each node's HA performs these operations instead of MN and CN which have the low battery or computation abilities.

The binding update step is carried as figure 1. The MN first creates BU key using the middle key. Then, the CN received this message creates the binding acknowledgement key with using BU, Cookie, and Nonce received in the MN. The MN verifies the message after created the BA key by using BU and parameters received in the CN. The explanation of specific protocol is as follows.

#### 3.2.1 Middle Key Agreement Step

Message ① and ③ are a message which transmits necessary parameters in key agreement to HA. The  $N_{MN}^i$  is nonce of MN, which is used to generate a session key. Also, it is used to prevent replay attack. The  $NAI_{MN}^i$  is a network access identifier [7, 8] of MN, which is used to authenticate the MN in HA. The  $K_{MN-MH}$  is the shared symmetric key between MN and MH. The message ③ has the same form with message ①. Also, it is similar in the parameters of message.

$$\textcircled{1}: MN \rightarrow MH : \{ HoA, MH_{ADDR}, NAI_{MN}^1, SN_{MN}, (N_{MN}^1, CN_{ADDR})_{K_{MN-MH}} \}$$

$$\textcircled{3}: MN \rightarrow MH : \{ CN_{ADDR}, CH_{ADDR}, NAI_{CN}^1, SN_{CN}, (N_{CN}^1, HoA)_{K_{CN-CH}} \}$$

The ② message is created by MH and is sent to generate the middle key to CH. The MH verifies on the validity of  $NAI^1_{MN}$ , which is used to discriminate MN. It then generates a  $cookie^1_{MH-CH}$  to prevent the redirect attack or Denial-of Service (DoS) attack. Also, it is used to generate the session key. MH generates the  $g^x$ , which is used to create the middle key. Finally, it then sends the following message to CH along with signature value to verify a user or an entity.

$$\textcircled{2}: MH \rightarrow CH : \{HoA, CN_{ADDR}, cookie^1_{MH-CH}, N^1_{MN}, N^1_{MH}, g^x, SN_{MH}, T, SIG_{MH}\}$$

$$SIG_{MH} = sig(-K_{MH}, h(N^1_{MN} \parallel N^1_{MH} \parallel g^x \parallel cookie^1_{MH-CH} \parallel SN_{MH} \parallel T))$$

The CH stores securely important parameters in its cache. It then transmits  $cookie^2_{MH-CH}$  to MH. The MH then confirms the signature and other parameters. The nonce and sequence number is used to prevent replay attack and man-in-the-middle (MITM) attack. The MH and CH generate a Diffie-Hellman session key  $K_{DH}=(g^{xy} \text{ mod } p, p \text{ is a large prime number})$ . The purpose of creating the middle key adds to the safety of key establishment. And it also reduces the load for CN's exponential operations.

$$CH \rightarrow MH : \{CN_{ADDR}, HoA, cookie^1_{MH-CH}, cookie^2_{MH-CH}, N^1_{CN}, N^1_{CH}, N^1_{MH}, g^y,$$

$$\textcircled{4}: SN_{MH}, T, SIG_{CH}\}$$

$$SIG_{CH} = sig(-K_{CH}, h(CN_{ADDR} \parallel HoA \parallel cookie^1_{MH-CH} \parallel cookie^2_{MH-CH} \parallel N^1_{CN} \parallel N^1_{CH} \parallel N^1_{MH} \parallel g^y \parallel SN_{MH} \parallel T))$$

Both message ⑤ and message ⑥ are the messages of transmitting the middle key via the secure channel. Each HA adds the nonce into message. Then, both MN and CN verify the on the validity of nonce.

$$MH \rightarrow MN : \{CN_{ADDR}, HoA, SN_{MN}, (N^1_{MN}, N_t, cookie^1_{MH-CH},$$

$$\textcircled{5}: cookie^2_{MH-CH}, K_{DH})K_{MN-MH}\}$$

$$N_t = h(N^1_{MN} \parallel N^1_{CN} \parallel N^1_{MH} \parallel N^1_{CH})$$

$$CH \rightarrow CN : \{HoA, CN_{ADDR}, SN_{CN}, (N^1_{CN}, N_t, cookie^1_{MH-CH},$$

$$\textcircled{6}: cookie^2_{MH-CH}, K_{DH})K_{CN-CH}\}$$

$$N_t = h(N^1_{MN} \parallel N^1_{CN} \parallel N^1_{MH} \parallel N^1_{CH})$$

The MN and CN create the encryption key  $K_{EN}$  by using the middle key and cookie. This key is used to encrypt delivered messages between nodes.

$$K_{EN} = hmac - sha1_{128}(K_{DH}, h(cookie^1_{MH-CH} \parallel cookie^2_{MH-CH}))$$

### 3.2.2 Binding Update Step

#### • Binding update between MN and MH

The binding update between MN and MH uses the existing generative secret session key. This course is same to fig 2.

$$\begin{aligned}
 & MN \rightarrow MH : \{CoA, MH_{ADDR}, HoA, cookie^1_{MN-MH}, NAI^2_{MN}, N^2_{MN}, B_{LT}, \\
 BU: & T, SN_{MN-MH}, MAC_{BU} \} \\
 & MAC_{BU} = hmac - sha^1_{128}(K_{MN-MH}, h(CoA \parallel cookie^1_{MN-MH} \parallel HoA \parallel NAI^2_{MN} \\
 & \parallel N^2_{MN} \parallel SN \parallel B_{LT} \parallel T))
 \end{aligned}$$

The binding update between MN and MH uses the shared secret session key  $K_{MN-MH}$  in advance. The  $cookie^1_{MN-MH}$  is used to prevent the DoS attack or redirect attack and is used to update the secret key between MN and MH. After transmitted the binding acknowledgement message, MH stores the binding information into cache.

$$\begin{aligned}
 & MH \rightarrow MN : \{MH_{ADDR}, CoA, HoA, cookie^1_{MN-MH}, cookie^2_{MN-MH}, N^2_{MN}, N^2_{MH}, \\
 BA: & T, SN, MAC_{BA} \} \\
 & MAC_{BA} = hmac - sha^1_{128}(K_{MN-MH}, h(CoA \parallel HoA \parallel cookie^1_{MN-MH} \parallel cookie^2_{MN-MH} \parallel \\
 & T \parallel SN \parallel N^2_{MN} \parallel N^3_{MH}))
 \end{aligned}$$

• **Binding update between MN and CN**

Before MN sends the following BU message to CN, it creates a  $BU(K_{BU})$  key. This key is used to authenticate binding update message. It is generated as follows.

$$K_{BU} = hmac - sha^1_{128}(K_{DH}, h(cookie^1_{MH-CN} \parallel cookie^2_{MH-CN} \parallel N_t))$$

The MN authenticates the message to use securely exchanged parameters. The  $cookie^1_{MN-CN}$  is used to prevent the DoS and redirect attack. Also, it is used to create  $BA(K_{BA})$  key.

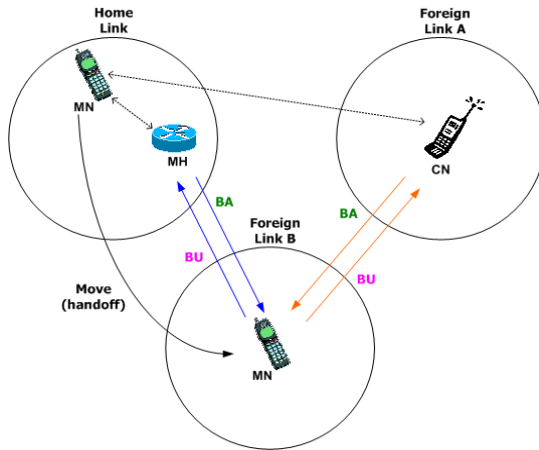


Fig. 1. Binding Update Protocol

$$\begin{aligned}
 BU: & MN \rightarrow CN : \{CoA, CN_{ADDR}, HoA, cookie^1_{MN-CN}, B_{LT}, T, MAC_{BU} \} \\
 & MAC_{BU} = hmac - sha^1_{128}(K_{BU}, h(CoA \parallel HoA \parallel cookie^1_{MN-CN} \parallel N^3_{MN} \parallel B_{LT} \parallel T))
 \end{aligned}$$

After received the binding update message, it verifies on the validity of the message. To create the BA key, which is used to authenticate the binding acknowledgement, it generates  $cookie^2_{MN-CN}$ . The BA key then is generated as follows.

$$K_{BA} = hmac - sha_{128}(K_{BU}, h(cookie^1_{MN-CN} \parallel cookie^2_{MN-CN} \parallel N^3_{MN}))$$

The CN sends the following binding acknowledgement message to MN.

$$CN \rightarrow MN : \{CN_{ADDR}, CoA, HoA, cookie^1_{MN-CN}, cookie^2_{MN-CN}, N^3_{MN}, N^2_{CN},$$

$$BA: T, SN, MAC_{BA}\}$$

$$MAC_{BA} = hmac - sha_{128}(K_{BA}, h(CoA \parallel HoA \parallel cookie^1_{MN-CN} \parallel cookie^2_{MN-CN} \parallel$$

$$T \parallel SN \parallel N^3_{MN} \parallel N^2_{CN}))$$

On receiving the binding update message, the MN verifies on the validity of the nonce  $N^3_{MN}$ . It then confirms  $cookie^2_{MN-CN}$ . The MN generates the BA key in the same way with CN. It then verifies the binding acknowledgement message by using this key.

### 3.2.3 Re-keying Step

When expired the lifetime of key, each node is able to regenerate the following secret key with exchanging parameters.

$$- K_{BU}/K_{BA}, K_{EN}$$

The key  $K_{BU}$  can be freshly generated by using master key  $K_{DH}$  as follows. Therefore, no matter when previous key  $K_{BU}$  is exposed, the attacker can't forge new binding update and acknowledgement messages. And, the  $K_{BA}$  will be generated by CN after it received the binding update message, which is hashed by  $K_{BU}$ .

$$K_{BU}^{new} = hmac - sha_{128}(K_{DH}, h(cookie^1_{MN-CN} \parallel cookie^2_{MN-CN} \parallel N^3_{MN} \parallel N^2_{CN} \parallel 0))$$

The encryption key  $K_{EN}$  is created as follows.

$$K_{EN}^{new} = hmac - sha_{128}(K_{DH}, h(N^3_{MN} \parallel N^2_{CN} \parallel 1))$$

It is used to encrypt a transferring message between nodes. Through these courses, the proposed protocol is able to save the consumption of nodes, which have the limitation of computational cost and power.

## 4 Analysis

In this section, we present the security, efficiency, and key agreement requirements of proposed binding update protocol.

### 4.1 Efficiency Analysis

In table 1, we show that our protocol is more efficient than existing proposed protocols. The following table presents comparison analysis with existing binding update protocols. Like the table analysis results, our scheme has the low battery consumption and computation quantity. On the other hand, the BAKE and PBK protocol is high.

Therefore, our protocol is very efficient and is able to fit the mobile network environment which uses the low power mobile devices.

**Table 1.** Efficiency Analysis

	Efficiency analysis ( $MN$ , $CN$ )	
	Battery consumption	Computation quantity
BAKE method	Low( $MN$ , $CN$ )	Low ( $MN$ , $CN$ )
PBK method	High( $MN$ , $CN$ )	High( $MN$ , $CN$ )
ECBU method	High ( $CN$ )	High ( $CN$ )
Proposed method	Low ( $MN$ , $CN$ )	Low ( $MN$ , $CN$ )

## 4.2 Security Analysis

We present the security analysis of using existed attack forms. The analysis is same as table 2. Our binding update protocol is more secure than BAKE and PBK protocol. The detailed results are same to the following table 2.

- **Known-key security:** In our protocol, security parameters such as middle key, nonce, NAI, and cookie are used to generate the binding update key. The  $K_{BU}$ ,  $K_{BA}$ , and  $K_{EN}$  are always freshly generated by transferring binding update message and binding acknowledgement message. As a result, our protocol is secure.
- **DoS attack and redirect attack:** To alleviate these attacks, we use cookie and NAI. Also, we are able to use client puzzle, which is used to authenticate client.
- **Replay attack and man-in-the-middle attack:** These attacks can be prevented by using nonce and timestamp.

**Table 2.** Security Analysis

	Secrecy analysis			
	DoS attack	Man-in-the-middle attack	Replay attack	Redirect attack
BAKE method	weak	weak	weak	weak
PBK method	weak	weak	weak	weak
ECBU method	Good	Good	Good	Good
Proposed method	Good	Good	Good	Good

## 5 Conclusions

In this paper, the proposed protocol is able to resolve the attacks which are able to exist the Internet and computation loads in the MN, CN which have the limited battery quantities and computation ability by using the following methods.



First, we are able to ease attacks by using the middle key creation, Cookie, *NAI*, other security parameters, and encryption algorithm. Second, in operations of plentiful computation quantities such as public key operations, exponential operations case, these operations are dealt with each HA. The other operation such as hashing function is computed in the MN and CN. Therefore, our protocol is good to not only efficiency side but security side more than previous methods.

## Acknowledgements

This work was supported by grand from NSRI Support Project, 2006.

## References

1. D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," RFC 3775, Jun. 2004.
2. Y. Qiu, J. Zhou, and F. Bao, "Protecting All Traffic Channels in Mobile IPv6 Network," Conf. on the Wireless Communications and Networking, Mar. 2004.
3. P. Nikander and C. Perkins, "Binding Authentication Key Establishment Protocol for Mobile IPv6," IETF Internet Draft, Jul. 2001.
4. S. Bradner, A. Mankin, and J.I. Schiller, "A Framework for Purpose Built Keys (PBK)," IETF Internet Draft, Jun. 2003.
5. A. Mankin, B. Patil, D. Harkins, E. Nordmak, P. Nikander, P. Roberts, and T. Narten, "Threat Models introduced by Mobile IPv6 and Requirements for Security in Mobile IPv6," IETF Internet Draft, May 2001.
6. C. Kaufman, "Internet Key Exchange (IKEv2) Protocol," IETF Internet Draft, Sep. 2004.
7. A. Patel, K. Leng, H. Akhtar, and M. Khalil, "Network Access Identifier Option for Mobile IPv6," IETF Internet Draft, Jul. 2004.
8. P. Calhoun and C. Perkins, "Mobile IP Network Access Identifier Extension for IPv4," RFC 2290, Mar. 2000.
9. H. Krawczyk, M. Bellare, and R. Canetti, "HMAC: Keyed-Hashing for Message Authentication," RFC 2104, Feb. 1997.
10. P. Kern and W. Simson, "Photuris: Extended Schemes and Attributes," RFC 2523, Mar. 1999.
11. Wide Kame Project, <http://www.kame.net>.

# Experiments and Experiences on the Relationship Between the Probe Vehicle Size and the Travel Time Collection Reliability

Chungwon Lee<sup>1</sup>, Seungjae Lee<sup>2</sup>, Taehee Kim<sup>3</sup>, and Jeong Hyun Kim<sup>4</sup>

<sup>1</sup>Dept. of Transportation Engineering, The University of Seoul, Korea  
chungwon@uos.ac.kr

<sup>2</sup>Dept. of Transportation Engineering, The University of Seoul, Korea  
sjlee@uos.ac.kr

<sup>3</sup>POSCO E&C, Seoul, Korea  
theekim@poscoenc.com

<sup>4</sup>Associate Research Fellow, Land & Urban Institute, Korea  
jeonghkim@iklc.co.kr

**Abstract.** It is obvious that more probe vehicles guarantee better reliability of travel time data collection. Because of cost involved, increasing the probe vehicles is not always possible. Thus, the mechanism of the relation between the probe vehicle size and the travel time data collection reliability needs to be investigated for optimizing the ATIS implementation. On the relationship between the probe vehicle size and the travel time collection reliability, this study contains two different kinds of data in two test networks. One came from simulation based experiments and the other came from field experiences of a real time travel time collection system. There are a bit comprehensive differences between the results of experiments and experiences. The similar trend has been observed in both networks. In spite of many possible sources of the differences, it can be learned that in selecting probe vehicles, the operational characteristics of vehicles are very important to construct reliable information systems to guarantee to meet network coverage requirements. Furthermore, the effectiveness of probe size increases marginally decreases.

## 1 Introduction

Because of difficulties of field trials, experts in transportation engineering frequently encounter the situation that an experimental study with the knowledge of traffic flow theories and driver behaviors are a sole alternative to apply. In designing or deploying the Advanced Traveler Information Systems (ATIS), one important issue is how many probe vehicles need to be equipped with on-board travel time collection devices. Furthermore, what would be the marginal effect of the probe vehicle size change (generally increase). Simulation based experimental studies have been commonly adopted to resolve this issue. However, it is hard to find any paper with field validation and investigation. A probe-based ATIS has been tested in Seoul Metropolitan in Korea and it is under operation. With these field data, some simulation experiments have been performed to investigate the relation between experiences and experiments.

Previous studies provided various estimates on the optimal number of probes required in obtaining reliable measures of travel times in various networks. Boyce et al. (1991) have estimated that about 5,000 probe vehicles were required to cover 60 percent of the northwest suburbs of Chicago for a measurement period of 10min during the morning peak period. Turner et al. (1995) and Gorys et al. (1999) performed similar studies on probe vehicles in Houston and Toronto, respectively. The estimated probe numbers in these studies are quite different because they may depend on the network characteristics including link capacities and flow levels. All studies did not contain any comprehensive field validation and hence no general procedure has been established for the optimal probe vehicle size decision problem yet.

This study tried to compare simulation based experimental results with field experiences in two test networks in Seoul, Korea. One network is from new CBD area with some residential area and the other is from old CBD area. The former is relatively smaller than the latter. For the simulation, INTRGRATION (Van Aerde, 1985) was selected to utilize its stochastic and dynamic nature of the route choice component. This study is an enhanced version of the author's previous work, Lee and Park (2001), which dealt with the former test network of the two. This study confirmed Lee and Park's results in the different network and includes more investigation as well as some suggestion to construct a new ATIS system and/or to improve existing systems.

## 2 Methodology

### 2.1 Notation and Sampling Probe Vehicles

First, the following notations are used for link  $l$  and measurement unit time  $t$ .

$\mu_{lt}$  : the true mean of link travel time experienced by vehicles passing link  $l$  during time interval  $t$

$\sigma_{lt}$  : the true variance of  $\mu_{lt}$

$n_{lt}$  : the smallest number of probe vehicles required

$\varepsilon_{max}$  : the maximum permitted relative error

$r$  : the percentage of time that the absolute value of relative error is less than  $\varepsilon_{max}$

$\Phi(x)$ : the cumulative distribution function of travel time evaluated at  $x$

$\Phi^{-1}(x)$ : the reverse of  $\Phi(x)$

It is common to adopt a statistical sampling methodology to determine the minimum required number of probe vehicles that would provide reliable link travel time estimates. The assumption that vehicle travel time on a particular link is an identically and independently distributed random variable is adopted in spite of its excessiveness due to no practical alternative. Then, the number of probe vehicles required could be calculated from Eq.(1) in the following:

$$n_{lt} = \left[ \frac{\Phi^{-1}\left(\frac{1+r}{2}\right)\left(\frac{\sigma_{lt}}{\mu_{lt}}\right)}{\varepsilon_{max}} \right]^2 \quad (1)$$

Assuming that  $\Phi$  is normally distributed and  $r=90\%$ ,  $\epsilon_{max}=10\%$ , Eq.(1) is rewritten as

$$n_{it} = \left[ 16.5 / \frac{\sigma_{it}}{\mu_{it}} \right]^2 \quad (2)$$

Turner et al.(1995) employed this statistical sampling method to obtain the minimum number of probe vehicles corresponding to a pre-specified permitted relative error( $\epsilon_{max}$ ) and confidence level( $r$ ). Srinivasan et al.(1996) and Lee and Park(2001) used the similar method as well.

## 2.2 Coverage for Travel Time Data Collection System

An important aspect for determining the number of probe vehicles required is the adequate area coverage, which is frequently misunderstood even within ITS experts. To provide reliable real time traffic information, the probe vehicles should traverse the sufficient proportion of links in the network during the measurement period. Area coverage can be defined as the proportion of links in the network *reliably* covered by probe vehicles during a measurement unit time interval. To investigate this area coverage, the distribution pattern of individual vehicle is important because the same number of probe vehicles may report different area coverage depending on network flow loading pattern. In other words, how effectively probe vehicles are distributed affects on the reliable travel time data collection. If the probe vehicles drive only on some specific highways, simple probe size increases will not improve the coverage of the data collection system.

Boyce et al. (1991) and Srinivasan et al. (1996) studied this using a static user-optimal route choice model for the peak period and not static route choice model, respectively. These studies were totally simulation based and hence needs to be validated with field data to ensure the results.

## 2.3 Estimating Optimal Number of Probe Vehicles

To determine the number of probe vehicle required for both reliability and adequacy, the following procedure is used as shown in Fig. 1, similarly in Lee and Park (2001).

The procedure can be explained sequentially as follows;

First, determine the number of probe reporting required for satisfying the reliability criteria on each link  $n_{it}$  using Eq(1). Assuming that the value of  $\sigma_{it} / \mu_{it}$  is constant across all links and for all the time period and referring that the coefficient of variation  $\sigma_{it} / \mu_{it}$  is in the range of 0.08 to 0.17(May, 1984), this study adopted the range and  $n_{it}$  becomes 2 to 8 when  $r=90\%$ ,  $\epsilon_{max}=10\%$  as like as Lee and Park(2001). Second, sample  $N\%$  probe vehicle trips from the pool of all vehicle trips in the network in proportion to each OD amount. Third, assign these vehicle trips using a simulation model, INTRGRATION with capability of the stochastic user-optimal route choice. Finally, determine the proportion of links covered reliably by probes,  $p_r$ . The area coverage defined as the averages of  $p_r$ ,  $p$ . This procedure was applied repetitively until the coverage requirement  $p_o$  is satisfied.

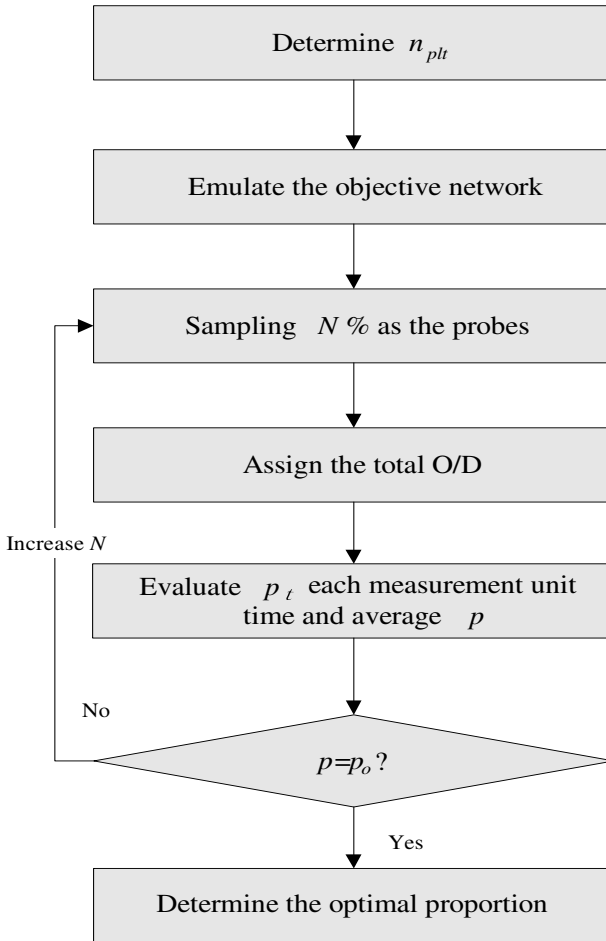


Fig. 1. Estimating the number of probe vehicles

### 3 Experiments

#### 3.1 Simulation Networks

The two test networks have been chosen. Network A is the same network form Lee and Park (2001) and Network B is relatively bigger network. Both are signalized major arterials in Seoul, Korea. The dynamic OD matrix is estimated and calibrated with various historical data sources to emulate the real flow pattern. Table 1 summarizes the characteristics of the two networks. Both are CBD area and suffer from heavy congestion.

**Table 1.** Two test networks

	Network A	Network B
Number of Nodes	11	20
Number of Links	23	31
Total Links	32Km	42Km
Network shape	Grid network	Grid network
Area Characteristics	New CBD area including some residential area	Old CBD area including little residential area

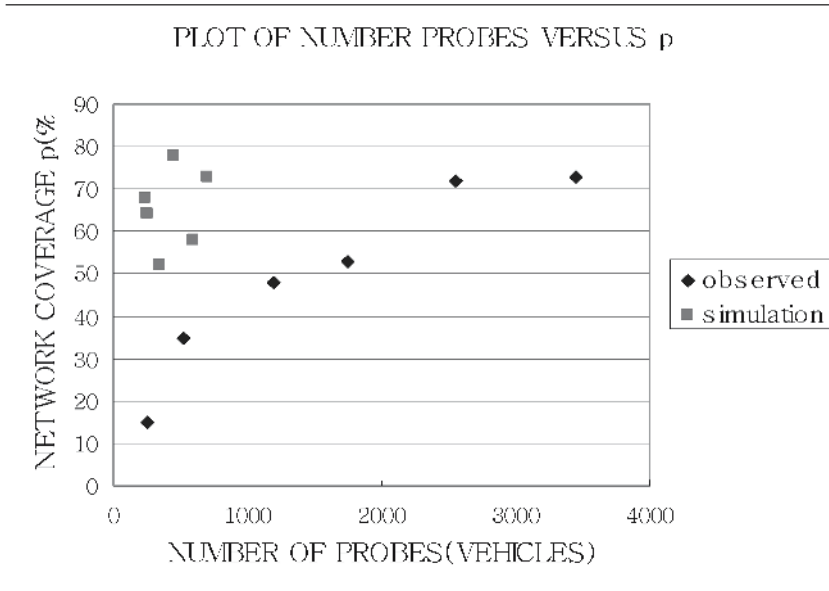
### 3.2 Probe Data Experiences and Simulation Experiments

Field observed data from probe vehicles from private ATIS were collected during the morning peak hours, 7Am-10Am, Oct. 17-19, 2000 and Dec. 19-21, 2000 for Network A and B. The total probe vehicles ranged from 300 to 800 during the periods. The INTEGRATION model was utilized to perform the Figure 1 procedure. Traffic assignment is based on the stochastic user equilibrium route choice rule. The simulation includes the repetitive trials with various proportions of probe trips. The values of 1%, 2.5%, 5.0%, 7.5%, 10% and 12.5% are considered. This simulation output provides vehicle probe records chronicling link-by-link.

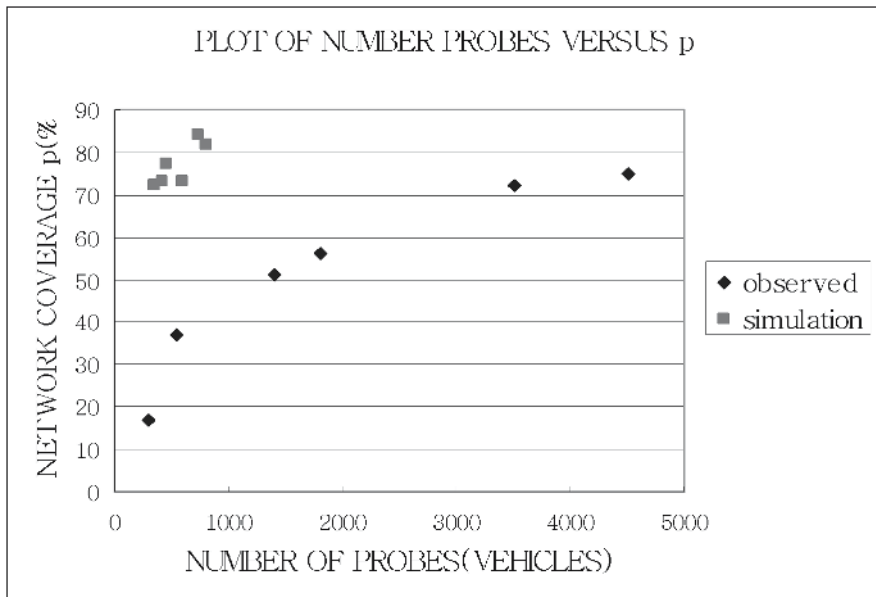
### 3.3 Experiment Results

Figure 2 and Figure 3 show the network coverage of the real probes and the simulated probes for the test networks during one peak hour, respectively. The simulation underestimates the coverage than the actual operations by quite a big amount in both networks. In both test networks, 300-800 real probe vehicles did show much more than 50% network coverage. On the other hand, these probe vehicles could cover at most 40% network area in simulation experiments. This confirms the previous work, Lee and Park (2001). Relatively higher coverage of the Network B might be caused by the fact that Network B does not contain residential area and real probes may travel to various destinations in general. Network A does contain resident areas a bit more, and hence real probes might have less various destinations and less dispersed in the network.

The differences between experiments and field experiences may be caused by many sources and it can be argued that simulation may not so informative in designing or deciding probe vehicle size decision. However, it should be noticed that what probe vehicles need to be selected. Most of real probe vehicles in our travel time data collection system were taxis. In terms of network-wide dispersion of probes, taxi would be relatively effective compared with passenger cars because of its continuous roaming. Thus, the probe vehicle selection process looks outweighing the simple increase of probe vehicle sizes in some cases. As can be seen in simulation results, the effectiveness improvement diminishes marginally as the probe vehicle size increase, which means simple sampling could easily fail to improve the reliability of travel data



**Fig. 2.** Comparison between simulation results and observed data in Network A



**Fig. 3.** Comparison between simulation results and observed data in Network B

collection system. If a system exceeds some level of coverage, say 75% in our sample networks, and has a chance to increase probe vehicles, two alternatives need to be considered instead of simple sampling. First one is to search effective category of probe vehicles and to investigate their operational characteristics individually before including as a probe vehicle. Second is to replace redundant vehicles among existing probes with more effective probe vehicles. The second alternative will save operational costs and computational loads on the ATIS center as well. At this stage, a canonical procedure to decide the probe vehicle size needs to be further explored with more field data.

## 4 Conclusion

For ATIS application, utilizing probe vehicles is commonly used. The question that how many probes will guarantee what level of reliability has not been fully investigated. This paper contains the comparison of experimental results and real experiences of the question in two test networks in Seoul, Korea. In spite of the relative big differences between those two results, the comparison is pretty informative. The result confirms the conjecture that simple sampling is hard to guarantee to meet the required reliability. For guarantee of satisfying the reliability requirement, much more probes need to be involved in the system resulting in high costs. Probe vehicles' operational characteristics are important factor of the reliability matter. For improving reliability of existing systems, the same argue applies. Simulation results indicate that the marginal effect of simple probe vehicle size increase decreases and may not guarantee to meet required reliability criteria although the tendency may be dependent on network characteristics. Further intensive studies for different and larger networks will be able to set up the systematic probe vehicle size decision.

## References

1. Aerde, M.V.(1985). Modeling of Traffic flows, Assignment and Queuing in Integrated Freeway/Traffic Signal Networks, PhD dissertation, University of Waterloo.
2. Boyce, D., Hicks, J. and Sen, A.(1991). "In-vehicle Navigation System Requirements for Monitoring Link Travel Time in a Dynamic Route Guidance System", Presented at the 70<sup>th</sup> annual meeting of the Transportation Research Board, Washington, DC.
3. Chen, M. and Chien, S.(2000). "Determining the Number of Probe Vehicles for Freeway Travel Time Estimation Using Microscopic Simulation", the 79<sup>th</sup> Annual Meeting of the Transportation Research Board, Washington, D.C.
4. Gorys, J. and Keen, S.(1999). "Measuring Congestion Through the use of Probe Vehicles in Toronto", the 78<sup>th</sup> Annual Meeting of the Transportation Research Board, Washington, D.C.
5. Hellinga, B. and Fu, L.(1999). "Assessing Expected Accuracy of Probe Vehicle Travel Time Reports", Journal of Transportation engineering, pp. 524-530.
6. May, A.D. (1984). Traffic Flow Fundamentals, Englewood Cliffs.
7. Lee, C. and Park, J.(2001) "Determining the Optimal Number of Probe Vehicles for ATIS Applications in Urban Networks", Proceedings of the 9th World Conference on Transport Research, Seoul.



8. Sen, A., Thakuriah, P., Zhu, X. and Karr, A.(1997). "Frequency of Probe Reports and Variance of Travel time Estimates", *Journal of Transportation Engineering*, pp. 290-297.
9. Srinivasan, K. and Jovanis P.(1996). "Determination of Number of Probe Vehicles Required for Reliable Travel Time Measurement in Urban Betwork", *Transportation Research Record 1537*, TRB, National Research Council, Washington D.C., pp. 15-22.
10. Turner, S. and Holdener, D.(1995). "Probe Vehicles Sample Sizes for Real-time Information: The Houston experience", *Proceedings of the Vehicle Navigation & Information Systems conference*.

# Conversion Scheme for Reducing Security Vulnerability in IPv4/ IPv6 Networks

Do Hyeon Lee<sup>1</sup> and Jeom Goo Kim<sup>2</sup>

<sup>1</sup>Division of Electrical and Computer Eng., Hanyang Univ., Korea  
dohyeon@mnlab.hanyang.ac.kr

<sup>2</sup>Dept. Computer Science, Namseoul Univ., Korea  
jgoo@nsu.ac.kr

**Abstract.** There are various ways exist for converting between IPv4 address and IPv6 address. A header conversion is the method that converts IP Header of the IPv4 and IPv6 network address throughout converters. A header conversion method has the merit that conversion procedures were simple and easy for implementation. But, there is the problem that can contain security vulnerability, which is existed in IPv4 address. We propose an encrypted header conversion method by applying ESP of IPsec. to a header conversion method. The proposed method can improve end to end network layer security.

## 1 Introduction

Explosive increase of the internet application and an available IP address of Internet are absolutely insufficient. Therefore, in order to solve these problems, IPv4 (64 bits address system) will be replaced by IPv6 (i.e., 128 bits address system) in the near future. However, there are no automatic conversions between IPv6 and IPv4. Therefore, IPv6 and IPv4 are must co-exist for considerable times.

Dual stack method, Tunneling method, and Header conversion method are proposed for changing IPv4 address system to IPv6 address system [2]. Header conversion method converts IPv4 packet header to IPv6 packet header and vice versa. Header conversion methods expected apply plentifully, because conversion process is transparent and implementation is easy. However, this method cannot provide end-to-end network layer security [5].

In this paper we propose an encrypted header conversion method which is applied ESP of IPsec to a header conversion method in order to improve a disadvantage of the original header conversion method. A conversion rule of header changes in the ways that proposed at Stateless IP/ICMP Translation (SIIT). In here, IP header field are changed depend on IP version. In the case of convert from IPv4 to IPv6, attach additionally ESP expansion header to a basic IPv6 header. In reverse case, convert from IPv6 to IPv4, remove an ESP expansion header and convert it to IPv4 packet header according to SIIT ways. In this way, encryption of a packet unit is continuously performed during conversion process, can improve security vulnerability between end-to-end communication.

## 2 Related Work

One of a representative way of a header conversion method is a NAT-PT method. A NAT-PT method uses address pool (pool of IPv4) to assign to IPv6 nodes of dynamic bases to generate a session in IPv4 and IPv6 boundaries. Terminal nodes of each network do not need to change, and IP packet routing is completely done transparency in terminal node. However, the datagram in the same session must pass throughout the same NAT-PT router [7]. NAT-PT has two functions. First, NAT function it owns an address pool to assign an IPv4 address to a dynamic IPv6 node when session was created. NAT function is located in a boundary router of networks, and carrying out address conversion. Second, Protocol Translation (PT) carries out address translation at host based on RFC 2765 SIIT. For performing a dynamic address allocation, in some cases, additional requirement is needed by the applications included an IP address or port information at payload scope.

**Security lack between end-to-end terminations:** NAT-PT cannot provide a security of network layer between end-to-end terminations. Also, transmission layer and application layer securities are impossible if they use an IP address at application layer. These are basic limitation of a NAT function. These leakages based on the session were transmitted by NAT-PT.

**Topology restriction:** All requests and answers participating in a session must be transmitted through the same NAT-PT router. For providing this scheme, implementing a NAT-PT in the unique boundary router at stub-domain. This scheme is not applied about the packet which is generated at the dual-stack node which does not require packet conversion. IPv4 address can be distinguished at IPv6 addresses of a PREFIX: x.y.z.w form. The dual-stack router routing can do a packet without state information coordination between IPv4 nodes and dual-stack nodes. This way does not affect communication between IPv6 nodes, in fact, this conversion only used without communication ways.

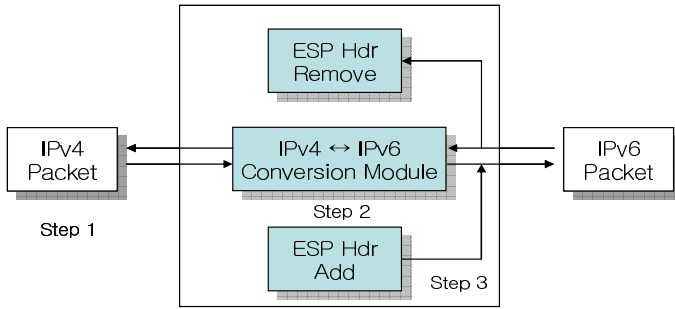
**Protocol conversion restriction:** A lot of fields in IPv4 was changed its' meaning in IPv6, when the conversion occurred, this meaning is not delivered exactly in IPv6

**Influence of address conversion:** An application communicating an IP address is not operated well at upper layer as NAT-PT carries out address conversion, ALG is essential

## 3 Design of Encrypted IPv4/IPv6 Header Conversion System

Encrypted IPv4/IPv6 header conversion system has three phases. Each phase is illustrated in Fig. 1. First phase transmits IPv4 packet and second phase converts IPv4 packet into IPv6 basic header and fragmentation expansion header. And finally third phase generates a perfect IPv6 header by appending ESP expansion header.

**IPv4 packet transmission phase:** In IPv4 packet transmission phase, IPv4 packet is same as in general IPv4 transmission environment. It is composed to IP header, TCP header, and data.



**Fig. 1.** IPv4/IPv6 header conversion system architecture

**Compose basic IPv6 header and fragmentation expansion phase:** In this phase, convert IPv4 header into IPv6 header. First, compose an IPv6 basic header and next compose a fragmentation expansion header. A configuration of basic IPv6 header passes the value to header structure of IPv6 to field value of IPv4 header recalculated or newly carry out a by generation or abolition. Address conversion between IPv4 and IPv6 uses an "IPv4-maped IPv6 address" way for the IPv4 nodes that do not support IPv6. This address is an IPv4 address embedded IPv6 address. A conversion method about each field value is same as Table 1 and Table 2. A router supporting IPv6 cannot divide an IP packet. Therefore, in this paper, IP packet size is smaller than 1,280 bytes, which is a MTU (Maximum Transmission Unit) size in IPv6.

**Table 1.** IPv6 basic header conversion

Field	Value
Version	IP version. 6
Traffic Class	Correspond to IPv4 Type of Service value, exact same copy of IPv4, generally '0'
Flow Label	Not defined. '0'
Payload Length	Payload length = IPv4 Total Length value - IPv4 IHL(IP Header Length) value If Fragmented, Payload length = Total Length value - IHL value + Fragment Header Length value
Next Header	Define expanded header type after IPv6 basic Header Fragmentation expansion Header : 44
Hop Limit	Copy of IPv4 Header TTL value
Source Address	Upper 96bits : IPv6-maped Prefix (::fff:0:0/96), Lower 32bits : IPv4 Header Source Address
Destination Address	Upper 96bits : IPv6-maped Prefix (::fff:0:0/96), Lower 32bits : IPv4 Header Destination Address

**Table 2.** IPv6 Fragment expansion header conversion

Field	Value
Next Header	Decide Next expansion Header. ESP Expansion Header : 50
Fragmentation Offset	Copy of IPv4 Fragment Off field value
M Flag	In IPv4, MF flag set = true, 1, DF flag set = true, 0
Identification	Copy of IPv4 Identification value

**Table 3.** Expansion header conversion

Field	Value
Security Parameter Index(SPI)	Random s2 bits value, Necessary parameter, Denotes SA, Decided by destination
Sequence Number	Necessary parameter for prevention Reply Attack, Never re-use
Initial Vector	Initialized by DES-CBC mode
Data	Encrypted Data
Pad	Block Encryption Data, Decide Block size decide (0 ~ 255byte)
Pad length	Byte number of Pad
Next Header	Decide upper layer Header Number: TCP : 6, UDP : 17, ICMP : 1
Authentication Data	Variable Number decided by authentication function

**Appending ESP expansion header phase:** In this phase, append ESP expansion header into basic IPv6 header and fragment expansion header according to Table 3. SPI value, Sequence Number value, and Initial Vector value are generated with SA at ESP Header before communicating through IKE communication. Another field value was referenced and filled by security related value in SA.

### 3.1 Encrypt Conversion Process from IPv6 to IPv4

For processing received IPv6 packet is achieved by opposite step mentioned before. Received packet must be confirmed about its authorization using security related information saved in SA. If this packet is valid, remove ESP header and decrypt the information. Basic IPv6 header and fragment expansion header are transformed into IPv4 header. If IPv6 header has Hop-by-hop expansion header, destination expansion header, and routing expansion header, this information was disregarded. Cause IPv4 network cannot resolve this expansion header information. During conversion process from IPv6/ IPv4, check-sum value must recalculated depend on pre-defined routines.

## 4 Implementation

For boundary router between IPv4 network and IPv6 network, this boundary router has encrypted conversion module for conversion between IPv4/IPv6 packets. One side, which connected with IPv4, only send/receive IPv4 packets, and the other side (connected with IPv6), only processing IPv6 packets. A basic protocol conversion method depends on RFC 2765 SIIT methods. Also, apply ESP Header of IPSec for encryption of a packet unit.

```

Struct ip6_hdr {
    union {
        struct ip6_hdrctl {
            uint32_t ip6_un1_flow ; /* 24 bits of flow-ID */
            uint16_t ip6_un1_plen ; /* payload length */
            uint8_t ip6_un1_nxt ; /*next header*/
            uint8_t ip6_un1_hlim ; /*hop limit*/
        } ip6_un1 ;
        uint8_t ip6_un2_vfc ; /*4 bits version, 4 bits priority*/
    } ip6_ctlun ;
}

```

**Fig. 2.** IPv4/IPv6 conversion module procedure

Using Fig. 2., perform type casting to struct ip6\_frag to the prepared buffer and convert it among struct ip member value. Then, assemble a value returned from buffer for upper layer and ESP expansion header. And transmit it to the LAN card eth1. IPv4/IPv6 conversion module is Fig. 3.

**A system is composed to four modules:** IPv4-IPv6 conversion module: receives IPv4 packets and converts it to IPv6 packets. Encryption module: encrypt upper layer data and return to it into IPv4/IPv6 conversion module. IPv6-IPv4 conversion module: receives IPv6 packets and converts it to IPv4 packets. Decryption module: perform integrity check and authentication compare SA value in IKE, and decrypt.

**IPv4/IPv6 conversion module:** This module assign buffer for IPv4, buffer for IPv6, and buffer for upper layer through LAN card connected to an IPv4 network after receiving an IPv4 packet. Buffer for IPv4 is 20 bytes. Buffer for IPv6 is designed the size adding basic header and fragment expansion header. Upper layer buffer size is decided by adding 1,260 bytes and ESP expansion header size. 1,260 bytes are came from subtract IPv4 header size from maximum packet size (1,280). Separate IP header and upper layer header using these buffers. After that, perform type casting to struct ip4 as read to buffer prepared IP Header as settle, and analyze field value of IPv4 Header. And assign buffer for an IPv6 packet, and perform type casting to struct

ip6\_hdr like this Fig. 2. Fill field value corresponding to the structure member Using Fig. 2, perform type casting to struct ip6\_frag to the prepared buffer and convert it among struct IP member value. Then, assemble a value returned from buffer for upper layer and ESP expansion header. And transmit it to the LAN card eth1. IPv4/IPv6 conversion module is shown in Fig. 3.

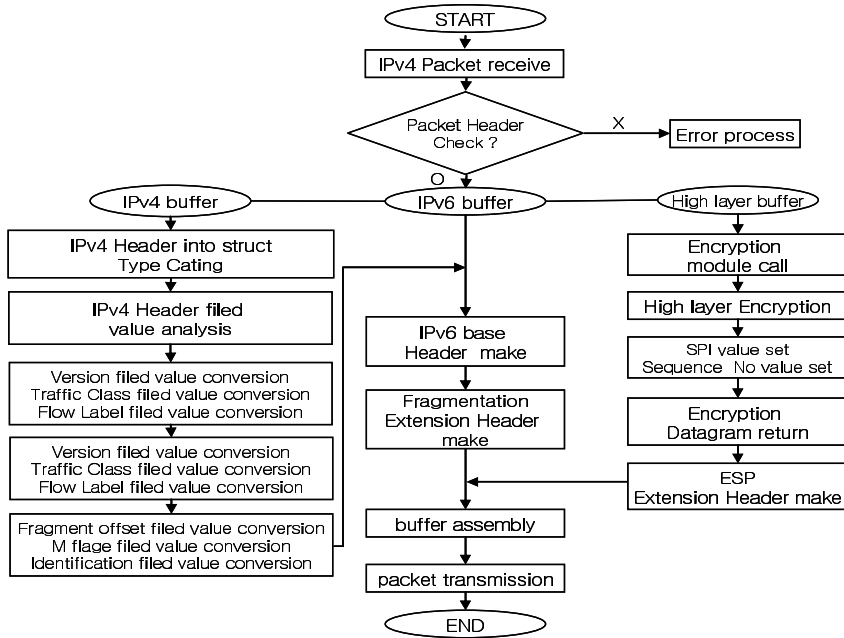


Fig. 3. IPv4/IPv6 conversion module process

**Encryption module:** IPv4/IPv6 conversion module invokes an encryption module using upper layer buffer pointer as a parameter. Encrypt upper layer data using encryption methods designated in SA. For composing an ESP expansion header, assign a struct esp as a member value refer SA. Encrypted upper layer data and ESP expansion header must returned into IPv4/IPv6 conversion module.

**IPv6/IPv4 conversion module:** IPv6/IPv4 conversion module check received IPv6 packet is valid or not. In this procedure, packet authentication procedure performed based on SA value. Analyze each member value by performing type casting to struct IPv6\_hdr and struct ip6\_flag. After that, compose an IPv4 header and calculate an ip\_sum for integrity check of composed IP header. Buffer size for IPv4 is 20 bytes same as IPv4/IPv6 conversion module. Buffer size for IPv6 is calculated by Basic header size and fragment expansion header size.

**Decryption module:** Inspect integrity of a received IPv6 packet as refer to SA with ESP field value. If packet is valid, perform a decryption process. After decryption

process, attach IPv4 header into the decrypted upper layer data. If IPv6 expansion header attached, eliminate that packet and generates an error message. IPv4 network cannot handle the expansion messages.

### 5 Performance Analysis

In this paper, we propose an encrypted header conversion method, in order to improve the security vulnerability between end-to-end communications. Proposed method applies encryption during conversion from IPv4 packet to IPv6 packet. Encryption method was based on ESP header format which was used in IPv6 network for packet unit encryption.

**Table 4.** System Comparison

	Comparison elements	NAT-PT Method	Proposed Method
IPv4/IPv6 conversion	Is conversion possible to IPv6 packets to IPv4 packet?	○	○
Confidentiality	Even if a packet transmitted becomes leakage, can try to know by the third party?	×	○
Packet change	A packet flowed out can modify contents as intention of a third party?	×	○
Data integrity	Can confirm data changed by third party? During transmission?	×	○
Original authentication	Is there exist method for confirming origin authentication?	×	○
Conversion Data number	A number of the Header field that must do a process in case of conversion	15	21

Try to compare two IPv4/IPv6 conversion methods mentioned before. The common procedure is exchange header of IPv4 and IPv6 header. The difference is applying an encryption procedure into the NAT-PT method. With the proposed method, we can perform packet unit encryption. Finally we can provide a packet unit security mechanism. Also, we can provide an integrity check for ESP header and payload using Integrity Check Value (ICV) in ESP header. Furthermore, can be easily checked alteration between transmissions and can provide a session origin authentication using IKE. However, processing time was increased because, encryption and decryption process were appended. In Table 4, NAT-PT and encrypted header conversion method, both methods can provide a packet conversion between IPv4 packet and IPv6 packet. When a packet was flowed out, a third party cannot inquire the contents of the packet in the encrypted header conversion method. Also the third party cannot change the contents of the packet without decryption key.



## 6 Conclusion

This paper proposed an encrypted header conversion method for a safe header conversion method among conversion methods of IPv4/IPv6. Encrypted header conversion method can convert one header into other version of header through one-on-one matching and recalculation. After that procedure, append an IPsec ESP header for encryption of a packet unit.

Proposed method can provide data integrity, original authentication, and data confidentiality. Because new method using format of IPsec ESP during encryption process, in order to provide packet unit encryption. And finally improve security mechanism better than provided by NAT-PT method.

## Acknowledgements

This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA.

## References

1. IPv6 Forum Korea "Next generation Internet Protocol (IPv6) Introduction", <http://www.ipv6.or.kr/members/materials/TM2000-001.pdf>
2. ETRI IT Information Center "IPv6 Trend" <http://new.itfind.or.kr/KIC/etlars/industry/jugidong/1018/101801.htm>
3. SinSunghuiSa "IPv6 Internet Protocol Version 6", MIC Graw Hill, 2001
4. S.M Lee, S.W Min "The IPv4 the Research Regarding the NAT-PT for the Gearing and an Interchange of and IPv6", Korea Information Science Society 2000, 2000.
5. Communication Protocol Engineering LAB, Kwangwoon Univ., "NAT-PT/SIIT" <http://cpe.gwu.ac.kr/data/cupid/Seminar/2000/nat-pt.pdf>
6. Communication Protocol Engineering Lab., Kwangwoon Univ., "SIIT" <http://cpe.gwu.ac.kr/data/cupid/T-report/2000/siit.pdf>
7. IPv6 Forum Korea "Linux base user mode NAT-PT" <http://mail.ipv6.or.kr/TM/TM2001-008.pdf>
8. S.C Park, J.Y Lee "IPv4/IPv6 plan and implementation of conversion protocol", Korea Information Processing Society Review C, Vol. 08, No. 06, 2001.
9. IPv6 Forum Korea "Next Generation IP IPv6" Dasung Inc., 2002
10. E.Nordmark, "Stateless IP/ICMP Translation Algorithm (SIIT)" RFC 2765 2000. 2.
11. S. Kent, R. Atkinson "Security Architecture for the Internet Protocol" RFC 2401 1998
12. Kent, S. R. Atkinson, "IP Encapsulating Security Payload (ESP)", RFC 2406, 1998. 11.
13. Kent, S. R. Atkinson, "IP Authentication Header", RFC 2402, 1998.
14. T.Larder "Transition Scenarios and Solution" Internet-Draft, IETF 1999.
15. ETRI krv6 Project "Linux base the Address Translation which Improves and Protocol Conversion Mechanism Implementation" <http://www.krv6.net/natpt-kr.htm>

# Improved Location Management for Reducing Traffic Cost in 3G Mobile Networks

Jae Young Koh

Principal Member of Eng. Staff, National Security Research Institute, Korea  
Jykoh@etri.re.kr

**Abstract.** For reducing the load of Home Location Register (HLR) in 3G mobile networks, we propose a mobility scheme to efficiently provide Intelligent Networks Service (INS). Once INS service profile is downloaded from Service Control Point (SCP) into Visitor Location Register (VLR), the INS is managed by only VLR not to connect SCP, so the service management cost is hardly affected by the call arrival rate as INS profile is provided by VLR where Mobile Host (MH) locates, not SCP. The proposed scheme is better performance than previous methods in 3G mobile networks.

## 1 Introduction

For raising the use of the wireless networks dramatically, these have been much research in 3G mobile networks (i.e., IMT-2000 and UMTS). It is the latest trend to divide the traffic load of the HLR occurring in the location registration and search the VLR to increase network efficiency. Also, the traffic of the SCP of the IN is overloaded on supporting INS. If we distribute the INS profile of the SCP into VLR, the INS provision in a Mobile Network (MN) will be very efficient.

Many researches [3-8, 12] have carried out the study of the requirements for Intelligent Network (IN) structure to support the MN, and the interest in integrating 3G mobile networks and IN has increased. As the integration model provides 3G mobile communications services which support IMT-2000 networks and Universal Personal Telecommunications (UPT), the model makes it possible to efficiently implement various services as well as ensure mobility. With regard to IMT-2000 networks, the services provision should be independent of the physical devices and provide of the same quality as a fixed network. These requirements could be satisfied with IN.

SSP in IN recognizes the IN call and sends SCP the query to request the INS processing. Then SCP processes the INS profile and returns the call to SSP. Because an SCP is connected to many SSPs, SCP is overloaded. A MN uses VLR and HLR to manage the location of the MH and the subscriber profiles of the users. In MN architecture, HLR is also bottlenecked. As both SCP of IN and HLR of a MN have the function of managing the subscriber's data, several studies [4-8,12] have proposed that the location management of a MN through IN is efficient. We propose location scheme providing INS in the IMT-2000 networks.

## 2 IMT-2000 Networks

IMT-2000 includes both the mobile network structure and the major functional entities of IN. Namely, IN CS-2 and IMT-2000 include similar functional entities. So the integration of those entities might be possible and expand the characteristics of IN by introducing and efficiently processing the service to IMT-2000 networks.

Fig. 1 shows the location registration in IMT-2000 [11, 12] and Fig. 2 shows the general procedure of call setup in the IMT-2000 model. They take charge of location management for MHs and INS profile management for INS that are performed by old HLR and old SCP, respectively. The detailed operations of each entity are as follows:

*SCP/HLR* is to copy VLR of not only subscriber profiles from HLR but also INS profiles of the user from SCP, when MH performs the location registration. These could be processed at a time without any special procedure under *SCP/HLR* integration. *VLR* is to receive the subscriber location profiles and INS from *SCP/HLR* at location registration procedure and responds to INS request from the corresponding MH at call setup procedure. *MSC/SSP* is to recognize the intelligent network calls by MH and send the INS query to VLR instead of SCP. Also, the location registration algorithm and the call setup algorithm in the IMT-2000 [11, 12] are in given Fig.1 and Fig.2.

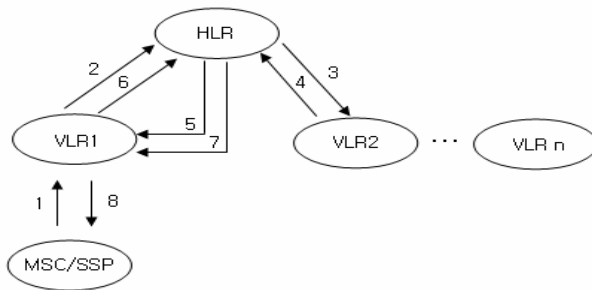


Fig. 1. The location registration in IMT-2000 models

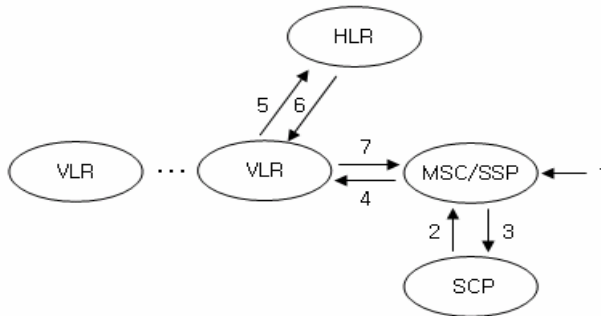


Fig. 2. Call setup steps in IMT-2000 model

The location registration algorithm in the IMT-2000 is following steps:

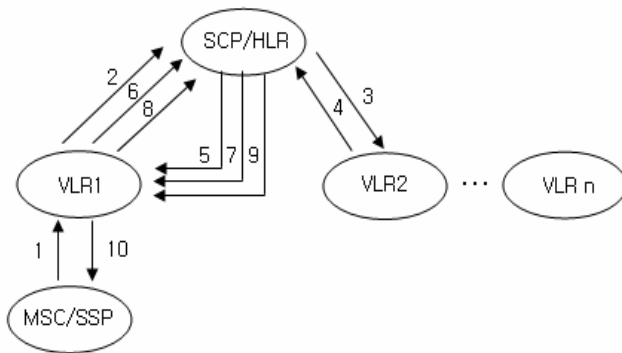
- Step1 and Step 2:** Location Update Request;
- Step 3 and Step 4:** User Profile Removal Request/Response;
- Step 5 and Step 6:** User Profile Copy Request/Response;
- Step 7 and Step 8:** Location Update Response

The call setup algorithm in the IMT-2000 [11, 12] is following steps:

- Step 1:** Call setup Request;
- Step 2 and Step 3:** INS Query Request/Response;
- Step 4 and Step 5:** Routing Query Request;
- Step6 and Step 7:** Routing Query Response.

### 3 Proposed Model

Fig. 3 show the location registration algorithm includes the INS profile distribution procedure in the proposed model.



**Fig. 3.** The location registration steps in the proposed model

The location registration algorithm in the proposed model is following steps:

- Step 1 and Step 2:** Location Update Request;
- Step 3 and Step 4:** User Profile Removal Request/Response;
- Step 5 and Step 6:** User Profile Copy Request/Response;
- Step 7 and Step 8:** INS Profile Copy Request/Response;
- Step 9 and Step 10:** Location Update Response.

There are two possibilities about INS process in VLR. First, various profiles are stored in an auxiliary device (ex, SDP) accessed by VLR. If many VLRs try to access this device, it might cause another network bottleneck. Second, VLR itself already stores INS profiles and copies just data necessary for the subscriber from SCP/HLR. Fig. 4 shows the INS call setup algorithm in the architecture. In the proposed model, we propose two architectures of signaling links. First, the architecture has the direct signaling link between MSC/SSP and SCP/HLR. Direct signal between MSC/SSP and

SCP/HLR is needed when SSP sends the IN query to SCP and MSC requires the location information of MH to HLR. In the proposed method, since MSC/SSP request the INS to VLR, a direct link between SSP and SCP is not required. All messages consist of request and response.

Next, there is no direct signaling link between MSC/SSP and SCP/HLR in the architecture. In this architecture, MSC/SSP sends all of the calls to VLR. If a call is the mobile service call, VLR requests the location of MT to HLR. If it is an IN service call, it processes the call using its own service profile and requests the location information to HLR like a mobile call. And then it returns the result to MSC/SSP.

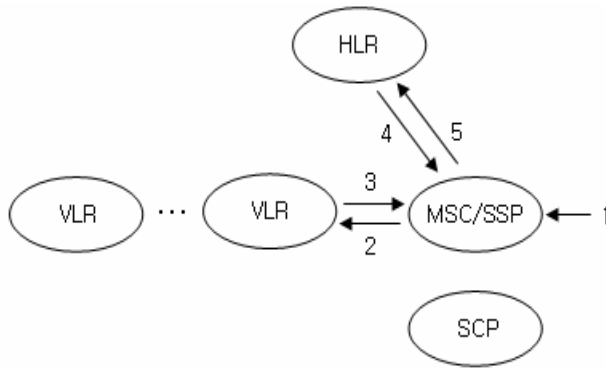


Fig. 4. IN Call setup in the proposed model

The IN call setup algorithm in the proposed model is following steps:

- Step 1:** Call setup Request;
- Step 2 and Step 3:** INS and Routing Query Request/Response;
- Step 4 and Step 5:** Routing Query Request/Response;

## 4 Performance Analysis

### 4.1 Performance Model

In performance model, we have some assumptions. There are  $n$ VLSs,  $n$ MSC/SSPs and an SCP/HLR in the model. A VLR is assumed to serve exactly one RA and to be connected to the corresponding MSC/SSP. All VLRs are connected to one SCP/HLR. We assume that there is one HLR/SCP and  $n$  VLRs within one signaling network boundary and considers only single network boundary. VLR and MSC/SSP are simply modeled by an M/M/1 queuing system. Though HLR and SCP are integrated, they are composed of separate modules and each also has a server with exponential distribution service time and a queue with infinite length. The average arriving call rate of MSC/SSP is  $\lambda_0$  with Poission distribution. Of the call, probability of IN call is  $P_{IN}$ . The average location registration rate of an RA is  $\lambda_r$  with Poission distribution. The probability of a user subscribing various network elements as follows:  $AC_V$  is average

cost to process a query or response by VLR;  $AC_H$  is average cost to process a query or response by HLR;  $AC_S$  is average cost to process a query or response SCP;  $S1$  is signaling link cost between VLR and SCP/HLR;  $S2$  is signaling link cost between VLR and MSC/SSP;  $S3$  is signaling link cost between SCP/HLR and MSC/SSP.

The processing cost may constitute the sum of queuing and service delay, and signaling link cost may represent transmission delay. To simplify the analysis, we assume that  $S1, S2, S3,$  and  $l$  are same.

To evaluate the performance, we define INS call processing cost as the delay between INS call processing cost as the delay between INS call arrival at MSC/SSP and service completion. So the INS management cost  $IN_{COST}$  is  $(D_{IN} \times R_{DN}) + (P \times R_{IN})$ .

### 4.2 Numerical Results

We calculate access rates of SCP and VLR for the proposed model in Fig. 3 and Fig.4, and we compare them with the IMT-2000 model. In the proposed model, SCP is accessed only for location registration. So the access rate of SCP is  $\lambda_r \times n \times P_{IS}$ , for INS profile copy form  $n$  VLRs. In IMT-2000 model, since INS is processed during call connection, the SCP access rate is  $\lambda_o \times n \times P_{IN}$ .

In VLR of proposed model, accesses of VLR comprise  $\lambda_r$ , registration requests from MSC/SSP,  $\lambda_r$ , user profile copy responses from HLR,  $\lambda_r \times P_{IS}$ , INS profile copy request from SCP,  $\lambda_r$ , registration responses from HLR, during location registration and  $\lambda_o \times P_{IN}$ , IN service queries from MSC during call connection. The access rate of VLR is  $3\lambda_r + \lambda_r P_{IS} + \lambda_o P_{IN}$ . In the case of MSC/SSP in proposed model, since routing requests to HLR and INS queries to VLR are not separated, the access rate of MSC/SSP is reduced. Table 1 shows access rates of VLR and SCP in proposed model and IMT-2000 model.

**Table 1.** Network signaling link cost

	Proposed model	IMT-2000 model
Call setup	$2+2P_{IN}$	$4+2P_{IN}$
Location management	$8+2P_{IS}$	8

As  $0 < P_{IS} < 1$  and  $0 < P_{IN} < 1$  in the worst case of  $P_{IS} = 1$  and  $P_{IN} = 1$ , the access rates of VLR in proposed model is  $4\lambda + 2\lambda_o$ . Therefore the increments of VLR access rate is  $\lambda_r$ . If  $P_{IS}$  is 0 and  $P_{IN}$  is 0, the access rates of VLR in proposed model is  $3\lambda_r + 2\lambda_o$ .

The SCP access ratio of the proposed model to IMT-2000 is  $\lambda_r P_{IS} / \lambda_o P_{IN}$ . If the ratio is greater than one, the SCP access rate of the proposed model is greater than the IMT-2000 model. INS processing delay and network load, we can get the service processing delay at each node,  $W_v, W_H,$  and  $W_S$  by Little’s result. Let Cost  $u$  of querying from MSC/SSP to VLR and receiving a response is  $2l + AC_v$ ; Cost  $v$  of querying from MSC/SSP to HLR and receiving a response is  $2l + AC_H$ ; Cost  $w$  of querying from MSC/SSP to SCP and receiving a response is  $2l + AC_S$ ; Cost  $x$  of querying from VLR to HLR and receiving a response is  $2l + AC_H$ ; Cost  $y$  of querying from

HLR to VLR and receiving a response is  $2l + AC_V$ ; Cost  $z$  of querying from VLR to SCP and receiving a response is  $2l + AC_S$ .

From the above,  $u, y, v,$  and  $x$  are same and  $w$  and  $z$  are same. The INS delay of proposed model and IMT-2000 model,  $P_a$  and  $P_b$  are  $4l + AC_V + AC_H$ , and  $P_{IMT}$  is  $6l + AC_V + AC_H + AC_S$ .

To simplify, if we ignore the signal link cost, as  $l = 0$ , INS delay in IMT-2000 model takes more delaying of  $w_s$  than the proposed model.

**Table 2.** Comparison of access rates of VLR and SCP

Node	Proposed model	IMT-2000 model
VLR	55.68	97.53
SCP	86.68	154.43

Let consider signaling network traffic load. In proposed model, call setup is  $(2l + W_V) P_{IN} + 2l + W_H$ . In IMT-2000 model, call setup cost is  $4l + AC_V + AC_H + (2l + AC_S) P_{IN}$ . If we assume only signaling network traffic load and  $l$  is 1, Table 1 shows the average network signaling link cost due to a call setup and a location management. When  $P_{IN}$  is 1, the maximum call setup cost of proposed model is 4. Even if  $P_{IN}$  is 0, the minimum call setup cost of IMT-2000 model is 4.

Also, For INS management cost, we should consider that there are more delays in location management due to INS profile download in the proposed model. In the IMT-2000 model, location registration cost  $D$  is  $3u + 2v$ . For INS,  $D_{IN} = w$  is added. So the location registration cost  $D'$  of an IN subscriber is  $D + D_{IN}$ . For INS management cost, we can calculate the INS management cost  $INS_{cost}$  is  $\left( P \times \frac{\lambda_0 P_{IN}}{\lambda_0 + \lambda_r} \right) +$

$$\left( D_{IN} \times \frac{\lambda_r P_{IS}}{\lambda_0 + \lambda_r} \right).$$

In IMT-2000 model, during only call connection, INS is managed. So the IN management cost  $INS_{IMT}$  is  $\frac{\lambda_0 P_{IN}}{\lambda_0 + \lambda_r} P_{IMT}$ .

We define the used parameters obtained from IMT-2000 recommendation [10]: Number  $n$  of VLRs per RA is 9; Traffic per terminal  $T_{MT}$  is 0.0417 (Erlang/MT); Average call holding time  $T$  is 100 (s); Average speed of users  $V$  is 2 (Km/h); Density of users  $\rho$  is 36,000(MT/km<sup>2</sup>); Single RA  $L_{RA}$  is 2286(km<sup>2</sup>); One side-length  $L$  of RA is 1.512 (km); Probability  $P_{IS}$  of subscribing INS is 1; Probability  $P_{IN}$  of arriving INS call is 0.5. Using these parameters, we can get the call rate  $\lambda_0$  arriving at MSC is  $P T_{MT} (1/T) L_{RA}$  and the location registration rate  $\lambda_r$  is  $\rho V L / \square$ . Fig. 5 shows the total network messages including the call setup and location management procedure per unit time according to the movement rate of MT. In the figure, the point where the X-axis is 1.75 is the movement rate threshold of MT in proposed model if the movement rate of MT is greater than the threshold, network messages of proposed model exceed the ones of IMT-2000 model.

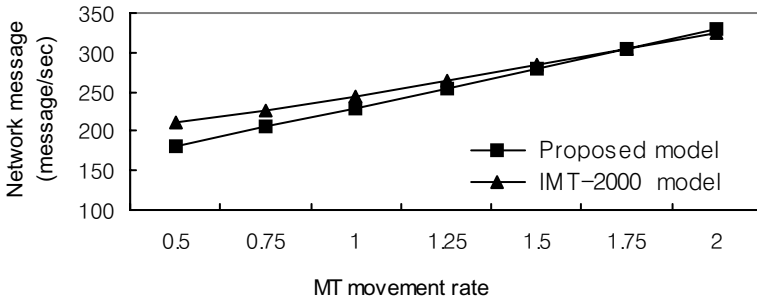


Fig. 5. Network load comparison

Fig. 6 shows the INS management cost according to the movement rate of MH and the call arriving rate. The units of the X-axis are normalized with respect to  $\lambda_r$  and  $\lambda_o$ . We could know that the INS management costs of proposed model increase slowly as the movement rate of MH increases, and the movement rate threshold of MH is also 1.75.

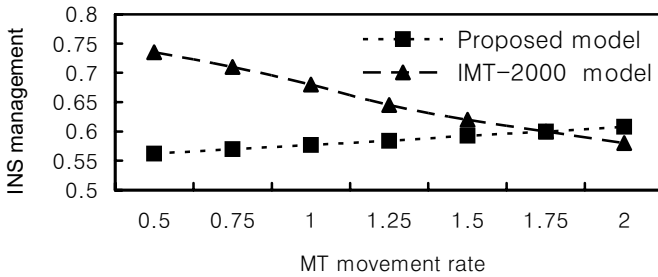


Fig. 6. INS management cost according to the call arrival rate

## 5 Conclusions

This paper described improved location management in 3G mobile networks and evaluated the performance of the proposed model and the IMT-2000 model using SCP and VLR load and network signaling link load. In IMT-2000 networks, INS provision is irrelevant to the location registration and as the call arrival rate from MH in RA increases, the INS management cost also increased rapidly. In the proposed method, the load of SCP is distributed to VLRs. It is suitable for the recent studies [9] which insist that location management data of HLR might be distributed to VLR. VLR manages the INS profiles and the location information about the user in its area. This proposed method is to download INS profile originally provided by SCP into VLR located by MH during location registration. It also provides INS through VLR when the IN service call arrives from the MH.



## References

1. J. Homa, S Harris, Intelligent Network Requirements for Personal Communications. Services, IEEE Commun. Mag. 1 (2) (1992) 70-76.
2. G.B. Choi, K.R. Kim, T.I. Kim, B.N. Yoon, Intelligent Network, KICS, Seoul, 1997.
3. Y.W. Chung, M.y. Chung, Y.H. Kwon, D.K. Sung, S. Kim, Performance Evaluation of Advanced Intelligent Networks Supporting Mobile Communications, Proc. of IEEE GLOBECOM'97, 1997
4. N. Faggion, T. Hua, Personal Communications Services Through the Evolution of Fixed and Mobile Communications and the Intelligent Network Concept, IEEE Network Mag. July/August, 1998.
5. N. Faggion, Intelligent Networks: a Key to Provide Personal Communications, Proc. of IEEE GLOBECOM'98, November 1998.
6. M. Laitinen, J. Rantala, Integration of Intelligent Network Services into Future GSM Network, IEEE Commun. Mag., 1995.
7. K.S. Meier-Hellstern, D. Alonso, The Use of SS7 and GSM to Support High Density Personal Communications, in: ICC'92, Chicago, USA, June 1992,
8. CCIR Recommendation 687-1, Future Public Land Mobile Telecommunication systems, 1992.
9. ITU-T, Draft Recommendation Q.FIF. Geneva, Switzerland, September 1997.
10. ITU-T Draft Recommendation Q.FNA, USA, January 1998.
11. X. Qui, V.O.K. Li, Performance Analysis of PCS Mobility Management Database System, Proc. of USC/IEEE Int. Conf. On Computer Communications and Networks, September 1995.
12. Young Lee, J. S Song, IN Service Provision Using VLR in IMT-2000 Network, Computer Communication, 2003

# A Semantic Web Portal for Semantic Annotation and Search

Norberto Fernández-García, José M. Blázquez-del-Toro,  
Jesús Arias Fisteus, and Luis Sánchez-Fernández

Telematic Engineering Department, Carlos III University of Madrid  
{berto, jmb, jaf, luiss}@it.uc3m.es

**Abstract.** The semantic annotation of the contents of Web resources is a required step in order to allow the Semantic Web vision to become a reality. In this paper we describe an approach to manual semantic annotation which tries to integrate both the semantic annotation task and the information retrieval task. Our approach exploits the information provided by Wikipedia pages and takes the form of a semantic Web portal, which allows a community of users to easily define and share annotations on Web resources.

## 1 Introduction

The Semantic Web [1] vision promises to expand the capabilities of current Web search engines by providing more powerful search mechanisms which exploit not only the data to be searched, but the real meaning, the semantics, of such data.

In order to make this vision become a reality the semantics of the data needs to be described in a formal, computer understandable manner. This process is known in the literature as semantic annotation. In this paper we describe an approach to manual semantic annotation which tries to take advantage of the effort of users looking for information on the Web, by integrating both the semantic annotation and the information retrieval tasks. Basically, our approach requires that the users define keyword-based queries and annotate them, associating them a concept or set of concepts taken from a semantic source. Once annotated, the keywords in the query are sent to a Web search engine and the results shown to the user. By providing relevance feedback over the results of the query, a user claims that a certain Web resource is relevant for his query. By doing so, the system associates the concepts previously used to annotate the query to the Web resource, generating an annotation of such resource.

Following the ideas that we introduced in [2], our approach uses Wikipedia [3] as semantic source. But, as a difference with our previous work, our current system adopts the architecture of a semantic Web portal, instead of that of a peer to peer network. This makes easier the process of annotation sharing and the implementation of services exploiting such annotations, as semantic search services. As a proof of concept, we have implemented a prototype of this portal. It allows a community of users to define and share annotations. It also allows external applications to access to such annotations using a Web service

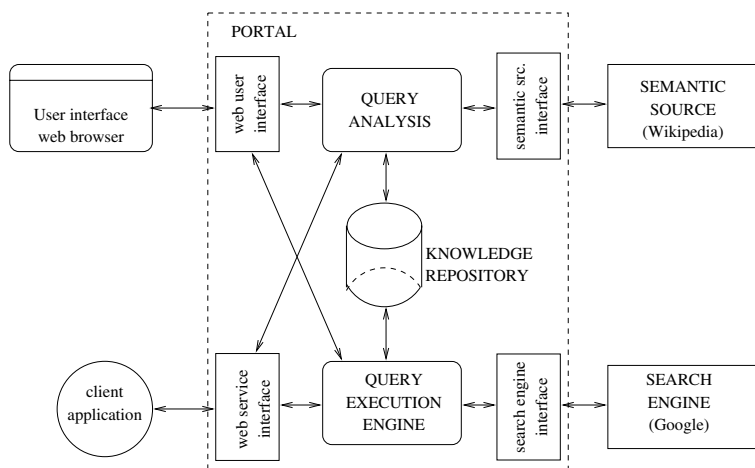


Fig. 1. System Architecture

interface. This prototype is publicly available at [4]. The rest of this paper is devoted to the description of such prototype. It is organized as follows: section 2 describes the details of our portal, including its architecture and working model. Implementation issues are covered in section 3. Section 4 briefly describes some related work in the state of the art. Concluding remarks and future lines, in section 5, complete this paper.

## 2 The Portal in Depth

As we have said in the introductory section, we will start by describing the main building blocks of the architecture of our semantic annotation portal and the main steps in the portal working model.

### 2.1 Architecture

Figure 1 shows the architecture of our system, including its internal modules, external information providers, client applications and Web users.

The *client interface module* is the interface between the clients and the server. Human users interact with the server from their Web browser, using a standard Web interface. In addition, the system allows client applications to interact with the system via a Web service interface. This Web service facilitates the development of third party semantic Web applications that use the information inside our system. For example, an *intelligent agent* could use the Web service interface and suggest resources related to the one the user is reading or alert the user when new resources are annotated for a given concept in which (s)he is interested. It can also be used to federate the portal with other ones.

The *semantic source module* is the interface of the server with the semantic source, i.e. the provider of the concepts used by the portal for annotating

queries and resources. The semantic source provides at least, for each concept, a label, a human readable description and a unique identifier. Although other semantic sources, like WordNet [5], were considered in the past, we are currently using Wikipedia. The reason is that it is much more dynamic and open to public creation and modification of entries, making easier the process of knowledge maintenance. In addition, it provides us with the information required by our model (label, definition and identifier), it covers a broad range of topics and it has thousands of users collaborating in maintaining the information. Furthermore, several external analyses have shown that its contents seem to have a reasonable quality [6]. On the negative side, one drawback of Wikipedia is that the semantics of the information is not very formal. It can be seen as a semantic network of linked topics which, for instance, makes difficult the usage of reasoning mechanisms. However, works dealing with making Wikipedia a more formal source can be found in the literature [7]. Our work may benefit from that.

The *search engine module* is the interface between our server and external search engines. The external engine to which our portal is currently connected is Google, but other search engines might be also easily integrated.

The *query analysis module* annotates keyword-based queries introduced by the user. Its objective is to associate terms in the query with concepts obtained from the semantic source. A given concept can be associated with one or more terms (for example, *semantic web* is a unique concept named by two terms). For a given query, the module provides a list of possible concepts that match the terms in the query and the user selects the one(s) (s)he is referring to.

The *knowledge repository* stores all the knowledge acquired by the system (basically resource annotations but also information about users, concepts and documents being annotated). All this information is represented in RDF [8] format according to the RDFS [9] model available at [4]. This model provides basic annotation characterization based on Annotea schema [10], but also takes advantage of well-known metadata vocabularies such as Dublin Core [11], SKOS [12] or FOAF [13] to describe the other entities involved in the system process (concepts, documents, users).

The *query execution module* performs the actual search of relevant resources for a given annotated query. The search is performed both in the internal knowledge repository and in the external search engine. Resources are ranked by the system giving more weight to those which are annotated by the concepts of the query in the knowledge repository. The user interface allows users to annotate any result by specifying that it is relevant to the concepts of the query.

## 2.2 System Working Model

Once we have introduced in the previous section the main building blocks of our system, we will describe here how these blocks interact to allow the system operation. Basically the system operation consist of the following steps:

1. *Authentication*. In order to gain access to the system the user needs to be authenticated using a login/password scheme. The system exploits the user

information in order to associate the annotations with their authors, but also in providing personalized search services. If the user has not an account, (s)he can create a new one by providing an e-mail address and optionally the URL of his/her FOAF profile. All this information about the user is stored inside the knowledge repository.

2. *Query annotation.* Once the user has been authenticated, (s)he can access the system and start querying and annotating. When a user types in a query it is sent to the query analysis component. A two step process occurs here. The first step tries to match all the terms in the query with a single concept (simple query definition). If this does not fit user intentions, a complex query definition step tries to match the query with several concepts by looking for different concepts for each term or section between quotes of the query. We have taken the decision of providing both steps by observing that most popular queries in Google, though sometimes contain more than one term, usually refer to a single concept [14]. So, we expect that a significative percentage of queries will be annotated as simple queries. As the simple query annotation process is easier than the complex query one, we expect that our decision will make the system more usable.

In order to annotate the query, the query analysis component uses the information in the semantic source, which in our current system is based on Wikipedia. Basically to provide to the user a set of candidate concepts to annotate a full query or a query term, the system performs a query on Google using the restriction *site:wikipedia.org*. The Google results are sorted giving more weight to concepts used in the past by the same user.

Every time a user expresses his/her interest in a certain Wikipedia page representing a concept, the system looks for the URL of the Wikipedia page into the knowledge repository to know if it is already here. If not, the Wikipedia page is processed to obtain the concept label and definition in a certain language and the links to Wikipedia pages in other languages talking about the same topic. All this information is stored inside the knowledge repository.

3. *Query execution.* Taking as input a query, this step shows to the user a set of relevant Web resources for such query. If the query is not annotated, the system just shows Google results. No semantic annotations can be generated, as the query has no semantics. If the query is partially annotated (part of the terms have not an associated concept), the system queries Google using the keywords in the query (we are not using query expansion in our current prototype) and sorts the results giving a better ranking to Web resources annotated with concepts in the query (the more concepts, the better the ranking). If the query is fully annotated, the system also queries Google and sorts the results, but also queries the knowledge repository for resources annotated with all the concepts in the query. These knowledge repository results are sorted by the number of users who have annotated them. The more users, the better. Then, results from Google and most relevant results from local repository are combined, using the following rules: resources from local repository are given better ranking than results from Google, but

if a certain resource both appears in Google and knowledge repository results, it is given more ranking. It has to be noted that the semantic query, as is based in concepts not in terms, can provide results in languages different than that of the original query. All the results shown to the user have a button. By clicking on that button, the user can say that a certain resource is relevant for his query. By doing so, the annotation generation process starts. As can be seen, the user explicitly gives relevance feedback in order to generate an annotation. So (s)he decides at any moment which information is shared with the system and its users, minimizing privacy problems.

4. *Annotation generation.* This process just associates an annotated query with a Web resource, generating an RDF representation of the annotation and storing it inside the knowledge repository. Every annotation includes information about its author, a creation timestamp, a link to the resource being annotated and the terms and concepts of the query which has been used to generate the annotation. During this process it is also checked if information about the Web resource is available at the knowledge repository. If not, some RDF triples containing information about the URL of the Web resource, its title and its snippet are inserted into the repository.

### 3 Implementation

Our portal implementation takes the form of a Java based Web application. The data is managed with Jena [15], and stored in a MySQL 5.0 [16] database. The



Fig. 2. Results for the query 'owl' without any annotations of the query (left) and after annotating it (right)

Web service interface of the system has been created with Axis [17]. Because of the limitations of space, we refer the reader to [4] in order to test the system.

Let us see an example. As explained before, once logged into the system, the user has the possibility of sending a query. Let us suppose that the user sends the query `owl`. First, the system shows a list of candidate concepts for this term. In the example, there are concepts about the bird, the Web Ontology Language and persons with this name. If the user wants results about the ontology language, (s)he selects this concept by clicking on a **More about this** button. Then, the system returns the results for the query over Google. Fig. 2 (left side) shows part of the results, assuming that there are no resources in the knowledge repository annotated for this concept. As can be seen, the results that correspond to the ontology language are not the first ones. But, if the user annotates one of them (clicking on its **Annotate** button), future queries asking for the same concept will give more relevance to this annotated result, as shown in Fig. 2 (right side). As can be seen, the resources have also a **View Annotations** button which allows to the users to see the concepts used to annotate such resources.

## 4 Related Work

Due to its critical importance in order to build the Semantic Web, semantic annotation has been and still is an active area of research for the Semantic Web community. In the state of the art in semantic annotation, we can find tens of proposals, ranging from completely manual approaches to automatic ones. For instance, from the manual annotation perspective, we found works such as Annotea [18], the SHOE Knowledge Annotator [19], SMORE [20], CREAM [21] or SemanticWord [22]. These proposals allow human annotators to define annotations on Web resources using the knowledge provided by one or several ontologies while these annotators browse the resources or while they are creating them. On the other hand, from the (semi)automatic annotation world, we found works such as, for instance, AeroDAML [23], SemTag [24], S-CREAM [25], PANKOW [26], C-PANKOW [27], KIM [28] or MnM [29]. Basically these systems exploit natural language processing techniques in order to extract the references in text to certain concepts in ontologies. Usually these systems require as input language dependant seed patterns or document corpus for training purposes. Of course there are also works in between of both worlds like [30] where the authors propose a system where a Web site manager can annotate manually SQL queries to a database used to generate dynamic Web pages. Then, the pages generated by the system are automatically annotated using the annotations on SQL. But, as far as we know, none of the systems described here explores the possibility of integrating the semantic annotation task with the keyword-based information retrieval task. Additionally, exploiting the information maintained collaboratively by Wikipedia users as knowledge source for the annotation process is also not suggested by none of these works.

## 5 Conclusions and Future Lines

In this paper we have described an approach to manual semantic annotation of Web resources that could exploit the effort of the millions of users who every day look for information on the Web by integrating the semantic annotation task with the keyword-based information retrieval task. From our point of view, our approach has the advantage of combining the easy knowledge maintenance of Wikipedia, with the low effort manual approach based on exploiting user queries and the powerful annotation sharing and accessing capabilities provided by the semantic Web portal. An additional advantage is that it can be used not only in Web scenarios, but in any other where keyword-based information retrieval is currently in use.

On the negative side, one of the main limitations of our system comes from the dependence on user collaboration. But, from our point of view, integrating the annotation activities with habitual user actions, as Web search, should be a point in favor.

A Web portal showing the functionalities of the system described in the paper is currently implemented and being tested. It is publicly accessible from [4]. Future lines of development of this prototype could include the integration of some trustness approach to determine the degree of trust of a certain annotation. This would allow us to deal with users inserting into the system wrong annotations both by error or malice. We also plan to perform usability analysis in order to test the validity of our approach.

## Acknowledgements

This work has been partially funded by the *Ministerio de Educación y Ciencia de España*, as part of the Inflex Project, TIC2003-07208.

## References

1. T. Berners-Lee, J. Hendler, O. Lassila, "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities". *Scientific American*, May 2001.
2. Fernández-García, N.; Sánchez-Fernández, L.; Blázquez-del-Toro, J.; Luque, V.; Exploiting User Queries and Web Communities in Semantic Annotation. In 5th International Workshop in Knowledge Markup and Semantic Annotation, located at the 4th International Semantic Web Conference, ISWC 2005, November 2005.
3. Wikipedia: The Free Encyclopedia. <http://www.wikipedia.org/>
4. SQAPS Homepage. <http://www.it.uc3m.es/berto/SQAPS/>
5. WordNet Homepage. <http://wordnet.princeton.edu/>
6. Wikipedia evaluations. <http://en.wikipedia.org/wiki/Wikipedia#Evaluations>
7. Krtzsch M, Vrandecic D, Vlkcl M (2005) Wikipedia and the Semantic Web - The Missing Links. Proceedings of Wikimania 2005: The First International Wikimedia Conference, Germany, August 2005.
8. Resource Description Framework (RDF). <http://www.w3.org/RDF/>



9. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004. <http://www.w3.org/TR/rdf-schema/>
10. Annotea annotation schema. <http://www.w3.org/2000/10/annotation-ns>
11. Dublin Core Metadata Initiative (DCMI). <http://dublincore.org/>
12. Simple Knowledge Organization System. <http://www.w3.org/2004/02/skos/>
13. The Friend Of A Friend (FOAF) Project. <http://www.foaf-project.org/>
14. Google ZeitGeist. <http://www.google.com/press/zeitgeist.html>
15. Jena Semantic Web Framework. <http://jena.sourceforge.net/>
16. MySQL AB. <http://www.mysql.com/>
17. Apache Axis. <http://ws.apache.org/axis/>
18. Kahan, J.; Koivunen, M.-R.; Prud'Hommeaux, E.; Swick, R.R.; Annotea: An Open RDF Infrastructure for Shared Web Annotations. In WWW10 Conference, Hong Kong, May 1-5 2001.
19. SHOE Knowledge Annotator. <http://www.cs.umd.edu/projects/plus/SHOE/KnowledgeAnnotator.html>
20. Kalyanpur, A.; Hendler, J.; Parsia, B.; Golbeck, J.; SMORE - Semantic Markup, Ontology, and RDF Editor. <http://www.mindswap.org/papers/SMORE.pdf>
21. Handschuh, S.; Staab, S.; Authoring and Annotation of Web Pages in CREAM. Proceedings of the 11th International World Wide Web Conference, WWW 2002, Honolulu, Hawaii, May 7-11, 2002. ACM Press.
22. Tallis, M.; Semantic Word Processing for Content Authors. Second International Conference on Knowledge Capture (K-CAP 2003), Sanibel, Florida, USA, 2003.
23. Kogut, P.; Holmes, W.; AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages. First International Conference on Knowledge Capture (K-CAP 2001). Workshop on Knowledge Markup and Semantic Annotation, Victoria, B.C. October 21, 2001.
24. Dill, S.; Eiron, N.; Gibson, D.; Gruhl, D.; Guha, R.; Jhingran, A.; Kanungo, T.; Rajagopalan, S.; Tomkins, A.; Tomlin, J.A.; Zien, J.Y.; SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. WWW2003 Conference, Budapest, Hungary, May 2003.
25. Handschuh, S.; Staab, S.; Ciravegna, F.; S-CREAM: Semi-automatic CREAtion of Metadata Proceedings of the European Conference on Knowledge Acquisition and Management - EKAW-2002. Madrid, Spain, October 1-4, 2002.
26. Cimiano, P.; Handschuh, S.; Staab, S.; Towards the Self-annotating Web. In the 13th International World Wide Web Conference, WWW 2004, pp. 462-471, New York, USA, May 17-22, 2004.
27. Cimiano, P.; Ladwig, G.; Staab, S.; Gimme' The Context: Context-driven Automatic Semantic Annotation with C-PANKOW. In the 14th International World Wide Web Conference, WWW 2005, Chiba, Japan, May 10-14, 2005.
28. Popov, B.; Kiryakov, A.; Kirilov, A.; Manov, D.; Ognyanoff, D.; Goranov, M.; KIM, Semantic Annotation Platform. Proceedings of Second International Semantic Web Conference, ISWC 2003, LNCS 2870, pp. 835-849.
29. Vargas-Vera, M.; Motta, E.; Domingue, J.; Lanzoni, M.; Stutt, F.; Ciravegna, F.; MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup. In 13th International Conference on Knowledge Engineering and Management (EKAW 2002), Springer Verlag, 2002.
30. Handschuh, S.; Staab, S.; Volz, R.; On Deep Annotation. In 12th International World Wide Web Conference, WWW2003, Budapest, Hungary, May 2003.

# A Semantic Portal for Fund Finding in the EU: Semantic Upgrade, Integration and Publication of Heterogeneous Legacy Data

Jesús Barrasa Rodríguez, Oscar Corcho<sup>1</sup>, and Asunción Gómez-Pérez

Ontology Engineering Group, Departamento de Inteligencia Artificial,  
Facultad de Informática, Universidad Politécnica de Madrid, Spain  
jbarrasa@eui.upm.es, ocorcho@fi.upm.es, asun@fi.upm.es

**Abstract.** FundFinder is a Semantic Web portal that allows searching for and navigating through information about funding opportunities. This application has been created following a set of techniques and using a set of tools for the upgrade of legacy content to the Semantic Web, including databases and semi-structured documents. This process consists in extracting and populating knowledge from heterogeneous information sources and making it available on the Web.

## 1 Introduction

Nowadays there are several Web portals that contain information related to funding opportunities for different types of organisations and individuals. Examples of such portals in the context of the European Union are CORDIS<sup>2</sup> or the EU's Grants and Loans site<sup>3</sup>. These types of portals are also available at national, regional or local levels in the different EU member states.

One example of such portal, at the regional level, is the public Website of CIDEM<sup>4</sup>, which is a non-profit Catalan organisation that aims at improving the region's industrial community networks and at increasing their competitiveness. This Website contains information about funding opportunities gathered, manually and on a daily basis, by CIDEM's staff members from different sources (mainly official publications). Access to this content is provided by standard form-based web pages that allow users to specify some basic search criteria such as the productive sector (Agriculture, Industry, Services, Tourism, Non-profit Organizations, etc.) to which the funding applies, the funding's objective (Technical and Financial Consultancy, Business cooperation, Culture, Energy, Tax incentives, Environment, R&D, Training, etc.), the date of the last update (to get the newest ones), etc. A traditional full text search engine is also provided to ease the search for funding opportunities.

Search interfaces like this one are helpful for basic information retrieval, with questions like *"Give me all funding opportunities in the agriculture sector"* or *"Give*

---

<sup>1</sup> Currently at the University of Manchester (Oscar.Corcho@manchester.ac.uk).

<sup>2</sup> Community Research and Development Information Service (<http://ica.cordis.lu/search/>).

<sup>3</sup> [http://europa.eu.int/grants/index\\_en.htm](http://europa.eu.int/grants/index_en.htm)

<sup>4</sup> Centre for Innovation and Business Development (<http://www.cidem.com>).

*me all funding opportunities containing the words ‘sustainable development’*”. However, they fall short for dealing with complex queries involving relations between concepts, such as “*Give me all the funding opportunities that can provide a supplement to those aiming at company creation*” or “*Give me all the funding opportunities that are incompatible with funding 651*”. The reason for this is that answering these types of questions requires understanding the meaning of the relations “*provide a supplement*” and “*be incompatible with*”.

Ontologies can provide a shared understanding of such relations and, in general, of most of the terms used in such queries. When ontologies are integrated in Web portals we normally talk, indistinctly, about the terms knowledge portals, semantic portals, community Web portals or Semantic Web portals [7].

In this paper we describe how we have created the Fund Finder application, whose objective is to allow semantic access to the content available in the current CIDEM portal, integrated with content from other heterogeneous sources. In other words, we describe the process of upgrading the current CIDEM portal to the Semantic Web, for which we have used some of the approaches, techniques and tools developed in the context of the project Esperonto. These are:

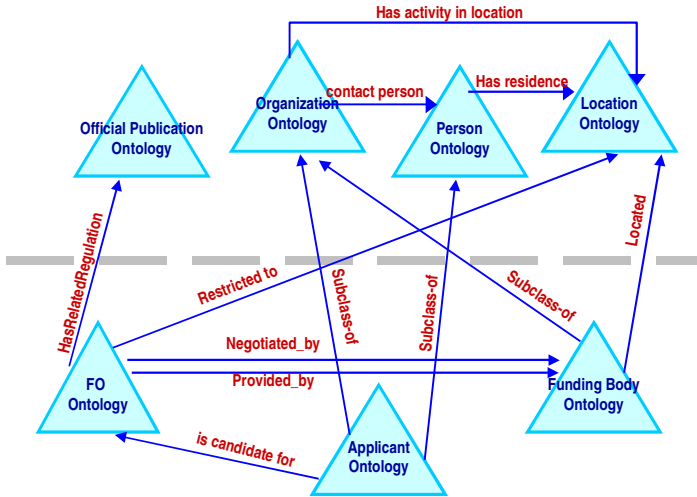
- A set of domain ontologies (covering the funding domain and other domains related to it.)
- An automatic processor called ODEMapster, capable of transforming information from databases into knowledge bases, according to a declarative mapping description document previously specified.
- An automatic processor called Knowledge Parser, capable of extracting information from semi-structured documents and populating it into a knowledge base, according to a configuration previously specified.
- A publication tool called ODESeW [5], capable of deploying Semantic Web portals with semantic navigation and querying functionalities.

In Section 2 we briefly describe the set of ontologies developed to formalize the Funding domain. Section 3 details how the generation of semantic content from heterogeneous information sources is carried out, providing an overview of the approach followed and of some of the tools involved in the process: the R<sub>2</sub>O mapping description language and the ODEMapster processor, on the one hand, and Knowledge Parser, on the other hand. Section 4 focuses on presentation, describing how the semantic content is presented and how it can be queried by final users. Section 5 concludes the paper and describes how to replicate this approach in another domain. A final appendix on the R<sub>2</sub>O mapping description language and the ODEMapster processor provides further details about them.

## 2 Ontologies in the Funding Domain

Ontologies are defined as formal, explicit specifications of a shared conceptualization [8]. In the context of our Fund Finder application, ontologies represent the domain of funding opportunities, funding bodies, applicants, organisations, persons, locations, publications, etc. These ontologies can be divided into two layers: the higher one

contains general-purpose highly reusable ontologies (Person, Location, Organization, Official Publication), while the lower one contains specific ontologies specifically related to funding (Funding Opportunity, Funding Body, Applicant). Figure 1 presents these two layers and an inter-ontology relation diagram that summarizes the main relations between these ontologies.



**Fig. 1.** Inter-ontology relationships between the different Fund Finder ontologies

It is not the purpose of this paper to explain in detail the knowledge formalised in these ontologies. We will just say that they have been developed by experts in the domain of funding in the European Union, following the Methontology methodology [6] and using the WebODE ontology engineering workbench [1].

### 3 Automatic Semantic Content Generation

One of the biggest barriers to large-scale deployment of Semantic Web applications is the availability of semantic content [3]. This content can be created by annotating existing information sources, by using different types of annotation techniques and tools, with different degrees of human supervision and annotation accuracy.

In the case of the Fund Finder application, existing content is stored in a relational database, owned by CIDEM and updated on a weekly basis, and in PDF and HTML documents available from several official journal Web sites (Catalan, Spanish, European, etc.). In the following sections we briefly describe how we extract knowledge from the different types of information sources, using the R<sub>2</sub>O language and ODE-Mapster for databases and Knowledge Parser for the PDF and HTML documents, and how we populate the funding opportunity ontologies integrating knowledge coming from these different sources.

### 3.1 R<sub>2</sub>O and ODEMapster: Database-to-Ontology Mapping

As aforementioned, some of the content to be upgraded to the Semantic Web resides in a legacy database that belongs to CIDEM. This database was developed several years ago with the purpose of being used as a backend for the Web application provided by this organisation. The database is updated manually, on a weekly basis, using different types of information sources as official journals, internal documents, faxes, etc.

Our objective is to be able to access the contents of the database as if they consisted of instances of the domain ontologies defined for our application. However, the process is not straightforward, since although the database schema and the ontologies cover overlapping parts of the domain, the models are usually different (databases are modelled with the objective of being used as data backends for applications while ontologies are modelled with the objective of representing the domain).

The R<sub>2</sub>O language [2] has been developed with the purpose of allowing the declarative specification of mappings between a database and a set of ontologies, so that these mappings can be later processed by the ODEMapster processor [2] in order to transform the content of the database into instances of an ontology implemented in a Semantic Web language like RDF Schema or OWL, as shown in Figure 2. The transformation can be done in a batch mode (all the RDF statements the result from applying the mappings are generated and stored somewhere in the application) or on-demand (only the mappings that are relevant for a query are executed when a query is sent to the system).

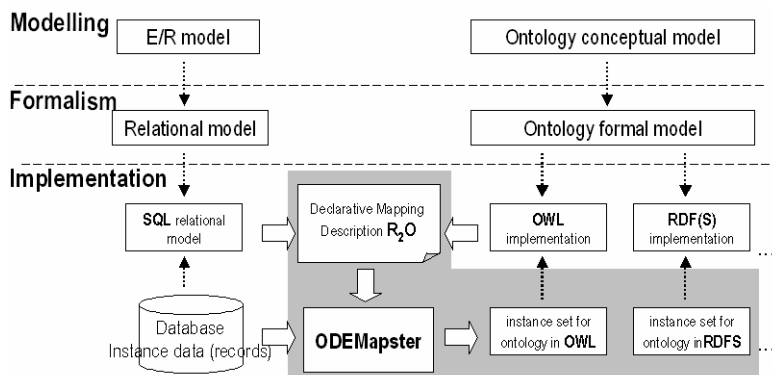


Fig. 2. R<sub>2</sub>O mapping architecture

R<sub>2</sub>O is intended to be expressive enough to describe the semantics of a great range of mappings between relational databases and ontologies. It is independent of the RDBMS, working with any DB implementing the SQL standard.

Because the domains covered by the ontology and the database do not always coincide and because the design modeling criteria used for building the DB are different from those used for ontology creation, the correspondences between their elements will be sometimes straightforward and sometimes difficult. R<sub>2</sub>O distinguishes the

following cases in concept transformation: 1) one DB table or view maps one concept in the ontology, 2) one DB table or view is used to instantiate more than one concept in the ontology, but only one instance per concept, 3) one DB table or view is used to instantiate more than one concept in the ontology, but multiple instances of the ontology can be generated.

Furthermore, before generating ontology instances, some standard relational algebraic operations (projection, selection, etc.) usually need to be executed, such as: Direct Mapping, Join/Union, Projection, Selection, or any combination of them.

Finally, the values of the attributes and relations can be filled in directly from the values of the fields in a DB record or after the application of a transformation function, which can affect more than one data field.

Although SQL relational algebra operations cover many cases, there are situations in which some additional transformations might be needed. Examples are more complex operations like natural language processing techniques over text data fields, regular expression matching for dates, URL or email extractions, etc. The R<sub>2</sub>O language provides means for specifying declaratively such selections and transformations.

### 3.2 Knowledge Parser: Knowledge Extraction from Documents from the Official Journal Web Sites

One of the objectives of the Fund Finder application is to allow content integration about funding opportunities coming from different legacy sources. This is particularly useful in the domain of fund finding because sometimes the information related to a specific funding opportunity is not complete or is spread over several Web sites. For instance, the CIDEM's database does not contain information about the documentation that a candidate needs to provide to apply for a specific funding opportunity. However, it contains the number and date of the official publication from which the information about the funding opportunity was taken, so that it is easy to locate it in the corresponding journal Web site by building automatically the URL of the on-line version of the journal or performing a search over the search facilities of the on-line journal.

Once the relevant document and the specific piece of text describing the funding opportunity are found, the relevant information has to be extracted. In these documents this information is usually available in natural language or in a semi-structured form (normally in the form of a bullet list where the types of necessary documentation are listed). R<sub>2</sub>O and ODEMapster cannot be used for this purpose, and hence we used iSOCO's<sup>5</sup> Knowledge Parser® [4], an automatic annotation system able to parse unstructured or semi-structured content, extract knowledge from it and populate it in an ontology.

### 3.3 Semantic Content Integration

Figure 3 shows how the task of integrating the information coming from the different sources is performed. We can see how two mappings (represented by arrows named Mapping1 and Mapping2) have been defined between columns in the database and properties in the ontology in an R<sub>2</sub>O mapping description document. The

<sup>5</sup> Intelligent Software Components (<http://www.isoco.com/>).

ODEMapster processor generates automatically instances for the ontology according to these mapping definitions with the information that it extracts from the database.

The official publication column in the database needs more complex processing. This record is filled in with semi-structured natural language text. This column contains all the elements needed to build the URL of the on-line version of the publication. These elements are extracted using regular expressions, which are also specified in the mapping description document. The generated URL is provided as an input to Knowledge Parser®, which retrieves this on-line resource from the Web and performs the information extraction process to generate instances with complementary information to that obtained from the database. Both instances (named Instance 85 in the diagram) are finally assembled in the resulting knowledge base.

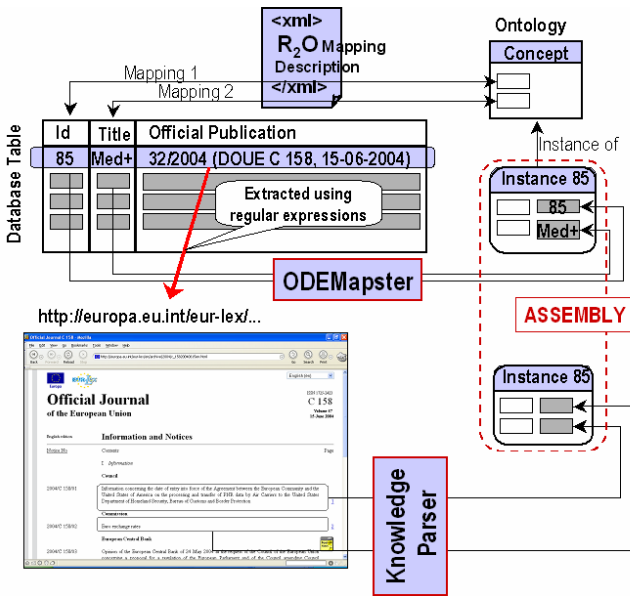


Fig. 3. Information integration in the Fund Finder application

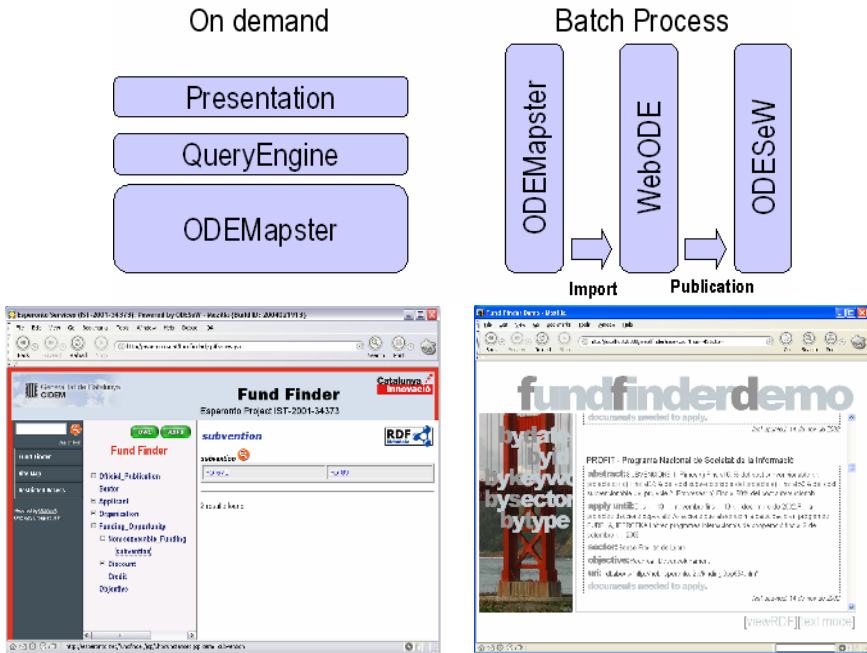
## 4 Semantic Publishing and Navigation

We explored two approaches for making the Fund Finder application publicly available, that is, for publishing the Semantic Web portal: batch and on-demand approaches. The selection of a specific approach has an important influence on the way that the knowledge extraction tools (ODEMapster and Knowledge Parser) are used. Figure 4 shows a schematic view of these two approaches, together with some screenshots of their user interfaces.

The batch approach is intended for massive batch semantic content generation and is especially useful when data does not change too often (as it is the case for this application). It is based on a three-step process. First, the content is extracted from the database by the ODEMapster processor and from the official journal documents by

the Knowledge Parser, and is represented in RDF. Then this content is imported into the WebODE workbench using WebODE import services. Finally the content is presented to the user using the ODESeW portal [5], which provides functionalities for semantic-based navigation and querying, different access control rights for different users, personalisation, etc.

The on-demand approach is focused on query processing and is more adequate when data changes frequently. It provides a lightweight presentation layer on top of a simple semantic query engine. The transformations are made on demand, based on the mapping description documents and configuration files needed by the ODEMapster processor and by Knowledge Parser.



**Fig. 4.** Two alternatives for deploying semantic portals. From left to right, the one generated by ODESeW and the web interfaces that use the semantic query engine.

## 5 Conclusions and Future Work

In this paper we have presented the Fund Finder application, which shows how a set of legacy databases and documents can be upgraded to the Semantic Web with some of the tools developed in the context of the project Esperanto, providing added value by integrating information from different heterogeneous sources, by allowing to perform additional types of queries that cannot be performed with the current application in place, and by allowing another type of navigation that was not foreseen with the current state of affairs.



This application is currently in the evaluation phase inside CIDEM and will be launched in the following year at their Web site, complementing the current application. Both portals are being evaluated and it seems that the early results confirm that the batch mode will be preferred in this application, given the fact that the information sources change only at regular weekly intervals.

Since all the technologies used in the construction of the Fund Finder Semantic Portal are domain independent, they can be easily reused in other domains. The description of another similar application can be found at [4], and other commercial applications are being also developed with this toolset at the time of writing this paper. By providing a toolset for the upgrade of legacy content to the Semantic Web and some hints on how to exploit the upgraded knowledge we strongly believe that we will allow others to implement other similar applications as well, hence fostering the vision of the Semantic Web.

## Acknowledgements

This work has been funded by the European Commission in the context of the project Esperanto Services IST-2001-34373 (<http://www.esperanto.net/>). We would like to thank Raúl Blanco and Carles Gómara for providing the application requirements and the CIDEM database, as well as Richard Benjamins, Jesús Contreras and Robert Salla for creating the official publication wrappers with Knowledge Parser.

## References

1. Arpírez, JC.; Corcho, O.; Fernández-López, M.; Gómez-Pérez, A. *WebODE in a nutshell*. AI Magazine 24(3):37-48. Fall 2003
2. Barrasa J, Corcho O, Gómez-Pérez A *R<sub>2</sub>O, an extensible and semantically based database-to-ontology mapping language*. Proceedings of the Second International Workshop on Semantic Web and Databases. Co-located with VLDB 2004 Toronto, Canada, 29-30 August 2004
3. Benjamins VR, Fensel D, Decker S, Gómez-Pérez A. *(KA)<sup>2</sup>: Building ontologies for the internet: a mid term report*. International Journal of Human-Computer Studies, 51(3):687-712, 1999.
4. Contreras J et al. *A Semantic Portal for the International Affairs Sector*. 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW'04). Springer-Verlag. Lecture Notes in Computer Science (LNCS) 3257:203-215. October 2004.
5. Corcho, O.; Gómez-Pérez, A.; López-Cima, A.; López-García, V.; Suárez-Figueroa, MC. *ODESeW. Automatic Generation of Knowledge Portals for Intranets and Extranets*. Lecture Notes in Computer Science Vol 2870. The Semantic Web - ISWC 2003. Springer-Verlag. pp:802-817. October 2003.
6. Fernández-López, M.; Gómez-Pérez, A.; Pazos-Sierra, A.; Pazos-Sierra, J. 1999. *Building a Chemical Ontology Using METHONTOLOGY and the Ontology Design Environment*. IEEE Intelligent Systems & their applications. January/February PP. 37-46.
7. Staab S, Angele J (2000) AI for the Web - Ontology-based Community Web Portals. 17th National Conference on Artificial Intelligence and 12th Innovative Applications of Artificial Intelligence Conference (AAAI 2000/IAAI 2000), Menlo Park/CA, Cambridge/MA, AAAI Press/MIT Press.
8. Studer R, Benjamins VR, Fensel D (1998). Knowledge engineering: Principles and methods. IEEE Transactions on Data and Knowledge Engineering, 25:161-197.

## Appendix: The R<sub>2</sub>O language and the ODEMapster Processor

### The R<sub>2</sub>O language

R<sub>2</sub>O is a declarative, XML-based language that allows the description of arbitrarily complex mapping expressions between ontology elements (concepts, attributes and relations) and relational elements (relations and attributes). The strength of the R<sub>2</sub>O language lies in its expressivity and in its DBMS independence. The elements of the language providing such qualities are **conditions & operations** and the **rule-style mapping definition for attributes**.<sup>6</sup>

### Conditions and Operations

Conditions and operations allow the description of *"under which circumstances a database individual (a relational tuple, a database record) can be upgraded to a Semantic Web individual (an instance of the target ontology)"* and *"what kind of transformations are needed to create a Semantic Web individual from a database individual"* respectively. Both are defined in terms of an extendable set of primitives and are identified by their names and the set of named parameters they accept. The values of such parameters can be constant values (**has-value**), variables referring record fields from the database (**has-column**), or the result of the execution of other operations (**has-transform**).

The first R<sub>2</sub>O excerpts describe a condition based on the *"match-regexp"* primitive. The condition is verified if the content of column *salaryRange* of table *jobs* matches the regular expression.

```
condition "match-regexp"
  arg-restriction
    on-param "string"
    has-column jobs.salaryRange
  arg-restriction
    on-param "regexp"
    has-value ([:digit:]*)-[:digit:]*
```

The second fragment describes an operation based on the *"concat"* primitive. The operation concatenates two constant strings with the content of column *id* of table *jobs*.

```
operation "concat"
  arg-restriction
    on-param "string1"
    has-value "http://net.testing.r2o/job-"
  arg-restriction
    on-param "string2"
    has-transform
      operation "concat"
        arg-restriction
          on-param "string1"
          has-column jobs.id
        arg-restriction
          on-param "string2"
          has-column jobtypes.code
```

---

<sup>6</sup> A complete description of the R<sub>2</sub>O Language is available in [2].

Other primitives defined in the first version of the language are: *plus*, *minus*, *multiply*, *divide*, *apply-regexp*, *in-keyword*, *hi-tan*, *lo-tan*, *equals*, *hieq-tan*, *loeq-tan*, etc.

### Attribute Mapping Definitions

Mapping definitions for attributes are defined as sets of *if-then* rules that allow the conditional generation of attribute values as well as multivaluation. The structure of an attribute mapping definition is described by the following example. The value of the ontology attribute *type* is calculated based on the application of the set of rules (**selector**): If the condition part (**applies-if**) is verified, then the action part (**aftertransform**) is executed to generate a value.

```

attributemap-def "http://net.testing.r2o/jobs#type"
selector
  applies-if
    condition [...condition desc 1...]
  aftertransform
    operation [...transformation desc 1...]
selector
  applies-if
  aftertransform ...

```

### The ODEMapster Processor

The ODEMapster processor generates Semantic Web instances from relational instances based on the mapping description expressed in an R<sub>2</sub>O document. ODEMapster offers two modes of execution (see figure 5): **Query driven upgrade** (on-the-fly query translation) and **massive upgrade batch process** that generates all possible Semantic Web individuals from the data repository.

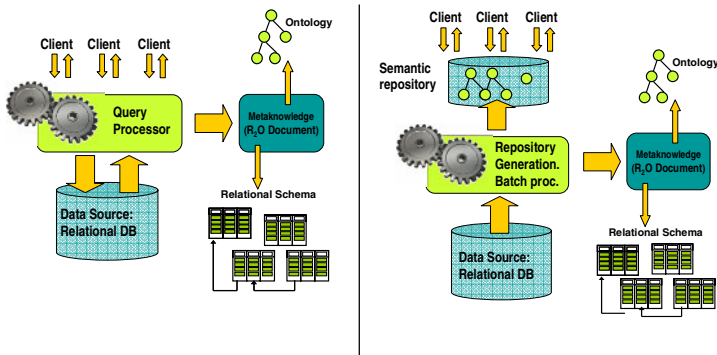


Fig. 5. ODEMapster execution modes

The operations of ODEMapster are not limited by the expressivity of the DBMS. The set of primitives can be extended with delegable or non delegable primitive conditions and operations. The processor will delegate the execution of certain actions to the DBMS and execute the rest itself (post processing). The main steps of its executions are: Query & R<sub>2</sub>O parsing, SQL generation, SGBD execution result grouping and finally post-processing.

# A Heuristic Approach to Semantic Web Services Classification

Miguel Ángel Corella and Pablo Castells

Universidad Autónoma de Madrid, Escuela Politécnica Superior  
Campus de Cantoblanco, 28049 Madrid, Spain  
{miguel.corella, pablo.castells}@uam.es

**Abstract.** Web service technologies, and the vision of semantically-enhanced services, aim to be key enablers to further exploit the potential of the Web as a platform where units of functionality can be deployed, shared, and assembled in a much more flexible way than it is possible today. Due to the expectable growth of the number of services offered in the WWW, the need for service repositories and mechanisms for their (semi)automatic organization and discovery of services is becoming increasingly important. In this paper, we propose a heuristic-based mechanism that enables service publishers to (semi)automatically classify their services in a service taxonomy managed by a service repository.

## 1 Introduction

Since the emergence of the semantic web [3], many research efforts have been aiming to use semantics to endow web services with a much higher potential for automation. These efforts have resulted in a new research trend called semantic web services [16]. The basis of this trend is to attach some semantic information to current WSDL – based web services descriptions [6] in order to enable their analysis and manipulation by software programs. This manipulation would be useful to enact powerful capabilities such as service automatic selection, invocation, composition or discovery.

Nowadays, UDDI [12] is the most widely accepted and used protocol for publishing and searching services over the web. These actions are usually performed using UDDI registries, which can be seen as service repositories easily accessed through a URL. In these registries, the published services are classified using some taxonomy (e.g. UNSPSC – United Nations Standard Products and Service Code, NAICS – North American Industry Classification System, etc.). Nevertheless, this classification is performed manually by a human publisher. Due to the huge quantity of classes in service taxonomies, the classification process is usually complex and costly. Furthermore, taxonomies are subject to evolution, or even complete replacement by new ones, making even heavier the maintenance effort load on repository administrators.

The central point of our approach is to provide automatic mechanisms to help service publishers in the classification task. We propose a heuristic that provides a ranked list of service categories in which the new published service fits best.

The paper is organized as follows: Section 2 introduces some previous work in the domain of web service classification and other areas related to the work presented here. Section 3 introduces the problem of web service classification and shows the scenario in which our work takes place. Section 4 presents some ideas demonstrating why web service semantics are needed in our classification approach. The complete presentation and explanation of our heuristic classification approach is described in Section 5. Finally, Section 6 presents some conclusions and outlines future work.

## 2 Related Work

Existent web service classification proposals may be divided into heuristic (e.g. [13]) and non-heuristic approaches (e.g. [5] and [9]). We briefly discuss next the most relevant initiatives in this scope related to our research.

In [9], two different approaches to web service classification are presented: a) selecting the category based on the service information (using Natural Language Processing, machine learning and text classification techniques) and b) creating categories based on service information (using clustering techniques). In this proposal the classification is based on extracting relevant words from service descriptions, and using them to build term vectors for classification mechanisms (e.g. Naïve Bayes), in such a way that the service classification problem is solved by a text classification approach.

In [5], the classification process follows the same steps than in [9], but Support Vector Machines are used as the classification method. In addition, service publishers are provided with some information (a concept lattice extracted using Formal Concept Analysis over service descriptions) about how the words used in their descriptions contribute to the selection of a specific category.

In [13], a framework to (semi)automate the semantic annotation of web services (i.e. the attachment of semantic information to web service descriptions, such as parameter description based on ontology concepts, service classification, etc.) is presented. A matching algorithm between web service data types and ontology concepts is defined (based on matching element schemas) in order to obtain a degree of similarity between services and domain ontologies. Having as many domain ontologies as service categories, they classify a service by finding the ontology that yields a higher similarity value when compared with the service.

Another related area to our work is that of service matchmaking (e.g. [10] and [14]). It is somehow related to web service classification but differs in the final objectives. Our research aims to find similarity degrees between services in order to assign them to a common category, admitting some degree of fuzziness in the matching. In contrast, work on service matchmaking follows a strictly boolean approach, aiming to find a service that matches some capabilities in order to e.g. invoke it or compose it into a more complex process.

## 3 The Service Classification Problem

Service categorization is commonly used to facilitate service retrieval, be it by manually browsing service repositories, or by automatic discovery mechanisms. Classification

taxonomies can be extremely large, comprising thousands of categories, within multiple hierarchical levels. Additionally, the number of services in a repository can grow quite large. Furthermore, the placement of a service under a proper category requires a considerable amount of knowledge of the taxonomy, the service characteristics, the application domain, the overall organization of the repository, implicit guidelines, etc., in order to make good classification decisions. As a consequence, the task often becomes overwhelming for repository administrators. Our work aims at alleviating the administrator's work by automatically providing her/him with a smaller set of likely appropriate taxonomy categories (ranked by likelihood of appropriateness), when a new service has to be registered in the repository.

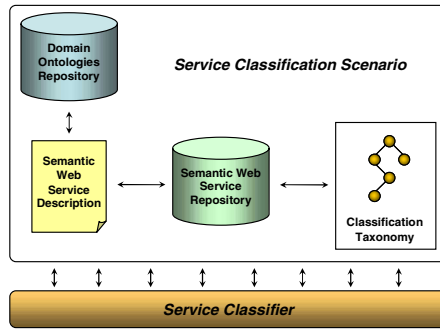
We approach the service classification task as follows. Given a set of service descriptions, already classified under some taxonomy, and a new service description, we propose a heuristic for automated service classification, based on the comparison of the unclassified service with the set of already classified services, whereby a measure of the likelihood that the service should be assigned a certain category is computed.

In our approach we make the following assumptions:

- A service taxonomy or classification is available, i.e. a set of different service types (e.g. currency exchange, flight information retrieval, etc.) arranged in some logical way (e.g. a list, a tree with parent-child relations, etc.). These could be standards commonly used for this purpose, such as UNSPSC or NAICS.
- A service registry is in place, where services are stored and classified in advance (manually or under human supervision). This could be an UDDI registry which seems to be the most widely accepted standard. The repository should be populated with classified services, which will serve as a basis for new ones classification.

## 4 The Need for Service Semantics

WSDL descriptions provide us with details about the operations a service provides, as well as the input and output information involved. While this enables comparisons already, the comparison would be greatly enhanced if the descriptions were enriched with further semantic information, as we discuss next. Consider this example: take a currency conversion service and a service that computes the expected time a trip between two cities would take. Since the WSDL description of a service only includes functional information (i.e. the description of its interface), at this level we can only compare service elements such as operations, messages, data types, etc. Let us say the currency conversion service has only one operation, which takes as inputs an amount of money (of type double) and two currency codes (of type string), and returns as output the converted amount (a double). On the other hand, the trip time calculator has also one operation taking as inputs an average speed (a double) and two city names (strings), and returns as output the estimated trip time (a double). It is easy to see that using only the WSDL descriptions (without using NLP mechanisms) we would get a very high similarity value between these services, since their interface is (syntactically) exactly the same. Since the similarity measure between services is a critical point in our approach, cases like this one would lead to misclassification.



**Fig. 1.** Overview of the main elements involved in our approach to web service classification

Therefore, more information is needed in the descriptions, semantic information enabling to make the difference between e.g. a currency code and a city name. Thus we advocate here for using semantic web services, which raises two further issues: the choice of a semantic description language and the availability of domain ontologies where the concepts involved in a service are defined.

Any of the currently available semantic service description languages, such as OWL-S [11], WSMO [15], WSDL-S [1] and SWSO [2], could be used in our approach, since our heuristic is language agnostic. We only assume a set of domain ontologies is available as a common vocabulary for all the classified services.

## 5 The Service Classification Heuristic

As has already been introduced earlier, our approach to service classification is based on the comparison of a new unclassified service with a set of already classified services. Our heuristic can be divided into three different levels of granularity, each one corresponding to the comparison between two elements involved in the classification process: service to category similarity measure, service to service similarity measure, and, finally, concept to concept similarity measure. These are explained in detail next.

### 5.1 Service to Category Similarity

The comparison between a service and a category should provide evidence that a service should belong to a taxonomy category. The proposed similarity measure works as follows. Let  $\mathcal{S}$  be the set of all web services in a repository, and let  $\mathcal{C}$  be the taxonomy used to classify the services in the repository. If we allow a service to be classified under several categories of the taxonomy, we may define the classification by  $\mathcal{C}$  as a mapping  $\tau : \mathcal{S} \rightarrow 2^{\mathcal{C}}$ . Given a new service  $s$  to be added to  $\mathcal{S}$ , we want to find the categories in  $\mathcal{C}$  that best suit  $s$ . Given  $c \in \mathcal{C}$ , let  $P(s:c)$  be the probability that  $c$  is an appropriate classification for  $s$ . We define an estimate for this probability by comparison of  $s$  with all the services classified under  $c$ . With this aim, if we take  $P(s:c) \sim 0$  if

$\{x \in \mathcal{S} \mid c \in \tau(x)\} = \emptyset$  (i.e.  $c$  is disregarded as a potential category for  $s$  if there is no previous service  $s \in \mathcal{S}$  classified under  $c$ ), we can write:

$$P(s : c) \sim P\left(s : c \wedge \left(\bigvee_{x \in \mathcal{S}} c \in \tau(x)\right)\right)$$

By rewriting the right hand-side, it can be seen that:

$$P(s : c) \sim \sum_{A \subset \mathcal{S}} (-1)^{|A|+1} \prod_{x \in A} P(c \in \tau(x)) \cdot P(s : c \mid c \in \tau(x))$$

provided that  $s:c \wedge c \in \tau(x)$  are pairwise independent for all  $x \in \mathcal{S}$ . Since  $c \in \tau(x)$  is true iff  $x \in \{x \in \mathcal{S} \mid c \in \tau(x)\}$ , and assuming a crisp service classification (i.e.  $c \in \tau(x)$  is either true or false, as opposed to fuzzy classification where  $P(c \in \tau(x)) \in [0,1]$ ), we have: Now we shall estimate  $P(s:c \mid c \in \tau(x))$  by a measure of similarity  $\text{sim}(s, x)$ , that is:

$$P(s : c) \sim \sum_{A \subset \tau^{-1}(c)} (-1)^{|A|+1} \prod_{x \in A} \text{sim}(s, x)$$

whereby the appropriateness of a category for a service is computed in terms of the similarity between the service and the services classified under that category. The measure of the similarity between two services will be defined in the next subsection.

Finally,  $P(s:c)$  is taken as a score value for ranking the categories according to their predicted appropriateness for  $s$ . Note that  $P(s:c) \in [0,1]$  provided that  $\text{sim}(s, x) \in [0,1]$  and increases monotonically with respect to  $\text{sim}(s, x)$ .

### 5.2 Service to Service Similarity

The similarity between two services is measured in terms of the similarity of their operations and parameters. Let  $\mathcal{P}$  be the set of all the parameters of the services in  $\mathcal{S}$ , and  $\mathcal{OP}$  the set of service operations. If we denote by  $P_s \subset \mathcal{P}$  the set of the parameters of service  $s$ , and by  $OP_s \subset \mathcal{OP}$  the set of its operations, the similarity is defined as:

$$\text{sim}(s, s') = f(\text{sim}(P_s, P_{s'}), \text{sim}(OP_s, OP_{s'}))$$

Developing a measure for comparing the complete set of parameters (despite the operation they belong to) between services is still work in progress in our research at the time of this writing, so in the meantime we are working with  $f(x, y) = y$ .

The similarity between the sets of operations ( $OP$  and  $OP'$ ) of two services is computed as the average of the best possible pairwise similarities obtained by an optimal pairing of the elements from the two sets. We define the similarity between two operations as:

$$\text{sim}(op, op') = \text{sim}(I_{op}, I_{op'}) \cdot \text{sim}(O_{op}, O_{op'})$$



where  $I_{op}$ ,  $I_{op'}$ ,  $O_{op}$  and  $O_{op'}$  are the set of input and output parameters of the operations  $op$  and  $op'$  respectively. The similarity between two parameter sets is computed in turn as the average of the best possible pairwise similarities obtained by an optimal pairing of the elements from the two sets. The similarity between two parameters is defined as the similarity of their types, which are assumed to be classes in a domain ontology. The similarity between concepts in an ontology has been widely studied (see e.g. [4], [7], [8]). In the next section we describe our own proposal, suited to the purpose of concept comparison in our context.

Note that the service comparison defined here returns values in  $[0,1]$ , provided that the similarity between ontology concepts is also within that range.

### 5.3 Concept to Concept Similarity

The similarity measure between concepts describing service parameters is key to our heuristic, since the previously defined comparisons are layered on top of this measure. We define it as follows.

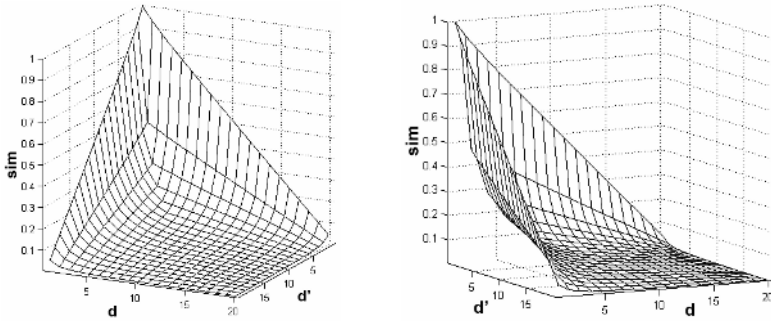
Let  $\mathcal{T}$  denote the set of all concepts in the domain ontology. The similarity between two concepts is measured in terms of their distance in the ontology class hierarchy. Given two concepts  $t \in \mathcal{T}$  and  $t' \in \mathcal{T}$ , let  $t_0$  be the lower common ancestor to  $t$  and  $t'$  in  $\mathcal{T}$ , and let  $d = \text{dist}(t, t_0) + 1$ ,  $d' = \text{dist}(t', t_0) + 1$  be the number of levels (plus 1) between  $t$ ,  $t'$  and  $t_0$  in the concept hierarchy. We define the similarity between  $t$  and  $t'$  as:

$$\text{sim}(t, t') = \left( 1 - \frac{\alpha}{h(\mathcal{T})} \cdot \frac{|d - d'|}{d + d'} \right) \cdot \frac{1}{\min(d, d')} \cdot \left( 1 - \frac{\max(d, d') - 1}{h(\mathcal{T})} \right)$$

where:

- $h(\mathcal{T})$  is the total height of the concept hierarchy, which is introduced to measure the distance between concepts as a proportion of the total depth of the ontology.
- The term  $\frac{|d - d'|}{d + d'}$  increases (that is, the similarity decreases) with the difference in the depth level between  $t$  and  $t'$ . Note that the similarity would seem to increase with the length  $d + d'$  of the shortest hierarchical path between them, but this is compensated by dividing by  $\min(d, d')$ , in a way that the similarity is essentially not sensitive to the depth of the concepts.
- $\alpha \in [0,1]$  is a parameter that ensures a minimum non-zero similarity value, even for the most dissimilar concepts, in a way that the similarity ranges in some interval  $[\text{min}, 1]$  above 0, in order to relax the influence of the measure in the heuristic.
- The factor  $1 - \frac{\max(d, d') - 1}{h(\mathcal{T})}$  is introduced to reinforce the decrease of the similarity when the concepts are in the same branch of the ontology hierarchy.

Figure 2 shows how the similarity function  $\text{sim}(t,t')$  depends on the distance  $d$  and  $d'$  of the  $t$  and  $t'$  to their lowest common ancestor.



**Fig. 2.** Concept to concept similarity measure graph (displayed from two angles) showing the behavior of the similarity measure with respect to  $d$  and  $d'$  in a twenty depth levels ontology

## 6 Conclusions and Further Work

We have developed a set of similarity measures, which assembled together result in an effective heuristic for the (semi)automatic classification of semantic web services. This heuristic offers an alternative mechanism to the ones explored in previous research, such as the usage of NLP techniques [5] [9], and the classification based on WSDL descriptions [13], to which our proposed approach is complementary. Our recursive definition of the similarity measures for service classification by successive levels has the advantage of modularity, in the sense that the measures can be studied and optimized at each level with a reasonable degree of independence. The main limitation of our approach is that it relies on the existence of a critical mass of semantic web service descriptions, which do not yet abound nowadays. Nevertheless, as the semantic web is gaining momentum, we rely on the growth and spread of such corpora, as a hypothesis for our research.

At the time of this writing, we are testing and refining all the measures defined here in order to achieve the best classification success rate. So far, we have performed several experiments over the concept to concept similarity measure obtaining good results. The next steps in our research, apart from finishing complete heuristic test and evaluating the results, include the creation of a sufficient repository of semantic web service descriptions, in order to obtain some realistic performance results of our approach and test its scalability.

## Acknowledgements

This research was supported by the Spanish Ministry of Science and Education (TIN2005-0685) and the Ministry of Industry, Tourism and Commerce (CDTI05-0436).

## References

1. Akkiraju, R., Farrel, J., Miller, J., Nagarajan, M., Schmidt, M., Sheth, A., Verma, K.: Web Service Semantics – WSDL-S, Technical Note, Version 1.0, 2005.
2. Battle, S., Bernstein, A., Boley, H., Grosz, B., Gruninger, M., Hull, R., Kifer, M., Martin, D., McIlraith, S., McGuinness, D., Su, J., Tabet, S.: Semantic Web Service Ontology (SWSO), Version 1.0, 2005.
3. Berners – Lee, T., Hendler, J., Lassila, O.: The semantic web. In *Scientific America*, 2001.
4. Bernstein, A., Kaufmann, E., Bürki, C., Klein, M.: How similar is it? Towards personalized similarity measures in ontologies. In the 7. Internationale Tagung Wirtschaftsinformatik. February 2005, pp. 1347 – 1366.
5. Bruno, M., Canfora, G., Di Penta, M., Scognamiglio, R.: An approach to support web service classification and annotation. In *Proceedings of the IEEE International Conference on e – Technology, e – Commerce and e – Services (EEE 2005)*, Honk – Kong, 2005.
6. Christensen, E., Curbera, F., Meredith, G., Weerawarana, S.: Web Service Description Language (WSDL), v1.1
7. Culmore, R., Rossi, G., Merelli, E.: An ontology similarity algorithm for BioAgent. In *NETTAB 02 Agents in Bioinformatics*, Bologna, 2002.
8. Ehrig, M., Haase, P., Stojanovic, N.: Similarity for ontologies – a comprehensive framework. In *Workshop Enterprise Modelling and Ontology: Ingredients for Interoperability*, at PAKM 2004. December 2004.
9. Heß, A., Kushmerick, N.: Automatically attaching semantic metadata to Web Services. In *Workshop on Information Integration on the Web (IIWeb2003)*, Acapulco, Mexico, 2003.
10. Li, L., Horrocks, I.: A software framework for matchmaking based on semantic web technology. In *the International Journal of Electronic Commerce*, 8(4):39 – 60. 2004.
11. Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., Paolucci, M., Parsia, B., Payne, T., Sirin, E., Srinivasan, N., Scycara, K. OWL-S: Semantic markup for web services, v1.1, 2004.
12. OASIS: UDDI: The UDDI technical white paper, 2004.
13. Oldham, N., Thomas, C., Sheth, A., Verma, K.: METEOR-S Web Service Annotation Framework with Machine Learning Classification. In *Proceedings of the 1<sup>st</sup> International Workshop on Semantic Web Services and Web Process Composition (SWSWPC'04)*, California, July 2004.
14. Paolucci, M., Kawamura, T., Payne, T., Sycara, K.: Semantic Matching of Web Service Capabilities. In *Proceedings of the First International Semantic Web Conference*, 2002.
15. Roman, D., Lausen, H., Keller, U., de Bruijn, J., Bussler, C., Domingue, J., Fensel, D., Hepp, M., Kifer, M., König-Ries, B., Kopecky, J., Lara, R., Oren, E., Polleres, A., Scicluna, J., Stollberg, M.: Web Service Modeling Ontology (WSMO), 2005.
16. Terziyan, V. Y., Kononenko, O.: Semantic web enabled web services: State-of-the-art and industrial challenges. In *Proc. International Conference on Web Services (ICWS)*, 2003.

# A RDF-Based Framework for User Profile Creation and Management

Ignazio Palmisano, Domenico Redavid, Luigi Iannone, Giovanni Semeraro,  
Marco Degemmis, Pasquale Lops, and Oriana Licchelli

Dipartimento di Informatica, Università degli Studi di Bari  
Campus, Via Orabona 4, 70125 Bari, Italy  
{palmisano, redavid, iannone, semeraro, degemmis, lops,  
licchelli}@di.uniba.it

**Abstract.** The semantic evolution of the Web has an heavy impact on traditional systems, as the ability to use a formal interoperable language simplifies information exchange between different systems. In order to foster information exchange and to easily connect new functionalities to semantic knowledge bases, in order to be able to use and reuse the valuable knowledge embedded in the existing systems, we designed a plugin-based framework, and used it to connect together different tools and systems developed in the LACAM laboratory. Our pilot project includes user profiling abilities coming from two components, namely Profile Extractor (PE) and Item Recommender (ITR), and storage capabilities implemented by a repository tool called RDFCore.

## 1 Introduction

One of the main points for the Semantic Web to be useful is interoperability; Semantic Web applications should be able to exchange information with (almost) no human intervention needed. By exchanging information, we mean exchanging meaningful information, i.e. two applications A and B should be able to share not only the bare data (which is already doable in a number of ways, one of which is through standards like XML), but the associate meaning, in a reliable way. For this to be possible, the process must be described in an unambiguous way; the easiest solution is then to express the knowledge that the applications want to share in a formal language, with well defined and logically based semantics. With this aim, currently two main languages have been defined by W3C: RDF (Resource Description Framework)<sup>1</sup> and OWL (Web Ontology Language).<sup>2</sup>

The use of the Semantic Web languages (OWL Full/DL/Lite) enables applications to decouple knowledge from application machinery, thus enabling other applications to share the meaning, provided that they can understand the same logic language.

---

<sup>1</sup> <http://www.w3.org/RDF/>

<sup>2</sup> <http://www.w3.org/2004/OWL/>

With this aim, we designed a framework to ease the realization of semantic applications. The framework is based on a set of interfaces that abstract some common functionalities, built around the concept of *flow* of information.

## 2 The Flow Metaphor

By *flow* of information, we mean the transmission of information (expressed as ontological information in SW languages, with DL semantics as background) from a *source* to a consumer for that information (*sink*).

Along the path, the information flow can be modified in many ways; we identify two main approaches for information change: *enrichment* and *transformation*.

Enriching an information flow means adding new information to this flow; an example could be the use of inference and deduction rules in order to explicit some implicit knowledge or to add new information coming from background knowledge.

Transforming an information flow involves the rewriting of the information; an example could be the change of background ontology for some data, or the merge of two information flows into a single one, involving the use of ontology alignment techniques.

Another kind of flow modification is the *store* of an information flow for later use; this is the typical job of a persistent storage component.

A sketch of the resulting conceptual architecture is in Figure 1.

The kind of components that implement these interfaces can then be summarized as:

- Source plugins: this kind of plugin creates new information (e.g., wrapping an external source of information, such as a database)
- Store plugins: this kind of plugin stores information, enabling both persistent storage and retrieval
- Transformer plugins: this kind of plugin modifies the information it is fed with (e.g., changing class or property definitions)
- Enricher plugins: this kind of plugin differs from the Transformer because it does not modify existing informations, but adds new information (e.g., an external reasoner could be wrapped in this kind of plugin)
- Sink plugin: this kind of plugin does not produce or modify information in the framework (e.g., a visualization plugin or an external application that needs to get information from the framework)

## 3 Information Flow Language

As already said, the Semantic Web languages for ontology expression are RDF and OWL, with the RDFSchema<sup>3</sup> language as an intermediate level of expressiveness (and of computational complexity).

<sup>3</sup> <http://www.w3.org/TR/rdf-schema/>

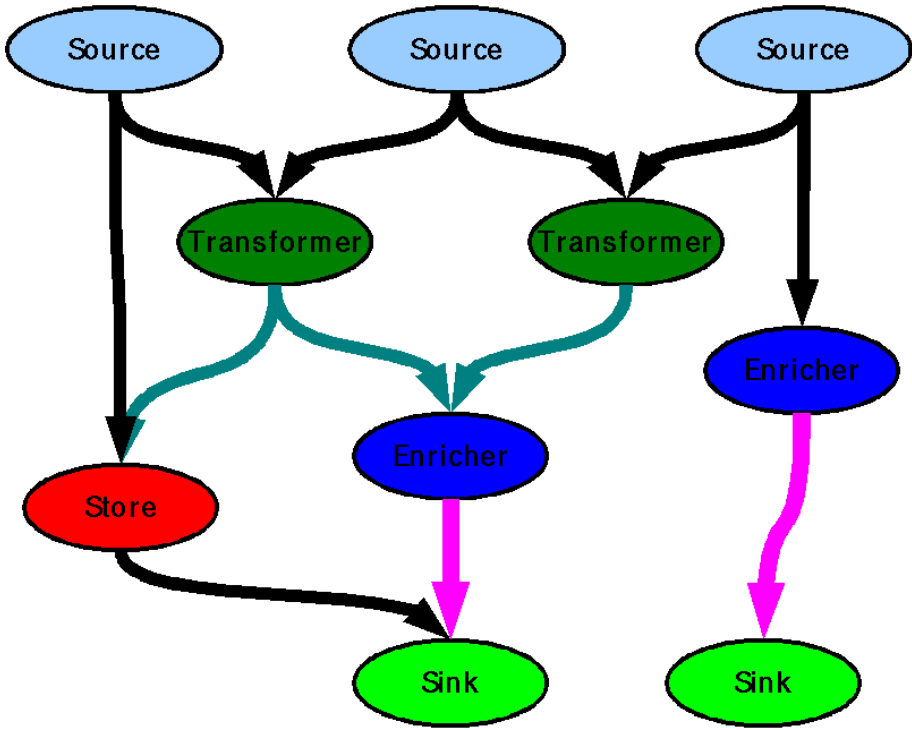


Fig. 1. Architecture Sketch

RDF is primarily focused on the concepts of resource and property: a resource is an identifiable entity, e.g. a human being, a web site, or a building, while a property is a relation between two resources or between a resource and a literal value (e.g. a human being is related to his name). A set of triples (*Subject, Predicate, Object*) is a RDF Model (or Description).

The RDF language is the base for the use of languages with a richer semantic, such as RDFSchema and OWL. OWL includes RDFSchema, in order to reuse the concepts already described there, and is divided into three sublanguages (Lite, DL, Full).

While in RDFS the main relation is inheritance, i.e. the definition of subclass/superclass relations between resources and subproperty/ superproperty relations for properties, OWL introduces a more complex semantic, e.g. restrictions on properties (it is possible to define cardinalities and data ranges for properties); the main advantage of this language, however, is the well defined semantic of the defined relations; this enables the construction of automatic reasoners that are not limited to a particular domain or to a particular implementation. Since OWL ontologies are expressed in RDF, there is no need for a separate storage layer for OWL data; and, since RDF is an abstract specification that can have different representations (see RDF/XML, Notation3, N-Triples, Turtle), it

is possible to exchange RDF data between application without enforcing an a priori representation. As a consequence, the language for the information flow in our framework is RDF.

## 4 Framework Test Case

In order to verify the framework in a practical scenario, we used the defined interfaces to wrap up other components developed in the LACAM lab. The components we included so far are:

- *RDFCore*: a component for RDF storage, wrapped as a *store* plugin
- *Profile Extractor*: a component for supervised learning of user classification rules, wrapped up as an *enrichment* plugin
- *ITR (ITem Recommender)*: a component for content based classification, based on naïve Bayes classifiers; from this component, which originally was a Java Web Application, many plugins have been created:
  - an *enrichment* plugin, that encloses the learning abilities of the system, in a way similar to Profile Extractor
  - a *source* plugin, that encapsulates the part of the web application that gains data from users and domain experts
  - a *sink* plugin, that contains the result display part of the system

The resulting instance of the framework is biased towards the user modeling domain, as is easy to see from the description of the systems that follows; other work in this area has been done, for example UUCM (Unified User Context Model) [4], which is based on an extensible representation for models. UUCM provides a simple schema to describe different dimensions of user models; each dimension can be described through values that can be either simple types (such as strings, numbers, dates) or be typed. In this last case, the type of the value is expressed as classes defined in OWL language.

The components we integrated are not tied to a particular ontology, but can be used with any OWL ontology like UUCM.

### 4.1 RDFCore

The RDFCore component, presented in [2], is a component used for RDF descriptions storage and retrieval, including multiuser support and extensible support for query languages.

The main modules of RDFCore are *DescriptionManager* and *TripleManager*. The first one gives access to Creation, Retrieval, Updating and Deletion (CRUD) operations on RDF models seen as a whole, while the second component enables the same operations at the single assertion level. Both modules use the Jena Semantic Web Toolkit[1] API to work with RDF models.

The component also offers multiuser support; users can choose whether some of the models they own should be private, publicly readable or writable, and can restrict access to single users or groups of users. This support is useful when

designing cooperative applications, thus enabling geographically dispersed teams to work together easily.

RDFCore has been adopted in the VIKEF Project as the basic component for RDF metadata storage in the VIKE (Virtual Information and Knowledge Environment) Framework, where its SOAP<sup>4</sup>-exposed services have been wrapped as a Web Service<sup>5</sup> for metadata storage, retrieval and querying.

In Figure 2 there is a small sketch of the system architecture.

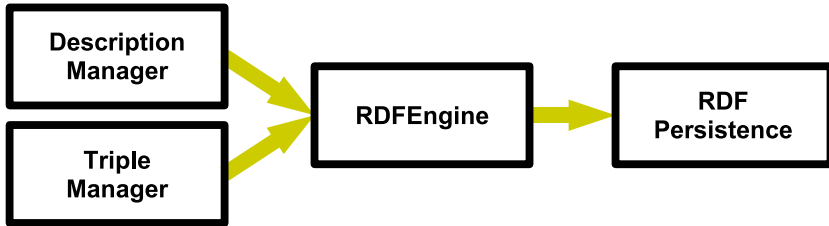


Fig. 2. Architecture of the RDFCore system

## 4.2 Profile Extractor

The Profile Extractor (PE) [7] is a module that classifies users by using supervised learning techniques. It can be used to discover user preferences by analyzing data related to user interaction or other data that are gathered from different data sources, such as data warehouse or transactions, in order to infer rules describing the user behavior. More in detail, these data are represented in RDF and refer to a simple ontology designed to be used as UUCM value type. The ontology is actually limited in its scope, since the PE component is limited to the use of zero order data (vectors of attribute/value pairs), and cannot exploit relational knowledge available in the input data.

To build profiles, the PE component uses decision rules induced from training data, through the use of well-known Machine Learning techniques, such as partition trees. In order for the rules to be inferred in an efficient way, and to maximize the predictive power of the inferred rules, it is necessary to establish what features and attributes, in the available data, are useful to accomplish the learning task, and what data, on the other hand, would not increase the predictive power or could waste computation time. The other main problem concerns the definition of meaningful classes to learn, which are to be defined before the learning task starts. The problem of learning user preferences can be considered as the problem of inducing general concepts from examples labeled as members (or non-members) of the concepts. In this context, given for example a finite set of categories of interest  $C = \{c_1, c_2, \dots, c_n\}$ , the task may consist in learning the target concept  $T_i$  “users interested in the category  $c_i$ ”. In the training phase, the users are positive examples for the categories they like/are interested, and

<sup>4</sup> <http://www.w3.org/2000/xp/Group/>

<sup>5</sup> <http://www.w3.org/2002/ws/>



negative examples for the categories they don't like/have interest. We chose an operational description of the target concept  $T_i$ , using a collection of rules that match against the features describing a user in order to decide if he/she is a member of  $T_i$ . Hence, the problem is reduced to the combination of a number of binary classifiers, in this specific context. For particular classes, where the expected value is not binary (like/dislike), but has more possible values (likes much/enough/little/nothing), the solution is still valid, but the classifier will not be binary; this could result in a small increase in the required computational time.

### 4.3 ITeM Recommender

ITR (ITeM Recommender) implements a probabilistic learning algorithm, the naïve Bayes classifier, relying on a content-based approach. The prototype is able to classify documents as interesting or uninteresting for a particular user, on the ground of the textual content of the documents. This approach is analogue to the relevance feedback in Information Retrieval [5], which adapts the query vector by iteratively absorbing users relevance judgments on newly returned documents. In the Information Filtering paradigm, the tuned query vector is actually a profile model, specifying both keywords and their informative power. Based on the constructed user profile, a new item relevance is measured by computing a similarity measure between the query vector and the item's feature vector. Learning a user profile generally involves the application of Machine Learning techniques to generate a predictive model based on information that has been previously labeled by the user. To learn user profiles, ITR casts the problem as a Text Categorization (TC) problem. The techniques used are those that are well-suited for text categorization [6].

We consider the problem of learning user profiles as a binary TC task: each document has to be classified as interesting or not w.r.t. the user preferences. Therefore, the set of categories is restricted to  $c_+$ , that represents the positive class (user-likes), and  $c_-$  the negative one (user-dislikes). ITR representation is based on *bag of concepts* (BOC) [8]. In this approach each feature corresponds to a single concept (that can be recognized from different words by using a lexical database) found in the training set. The final outcome of the learning process is a probabilistic model used to classify a new instance in the class  $c_+$  or  $c_-$ . The model can be used to build a personal profile that includes those concepts that turn out to be most indicative of the user preferences.

### 4.4 Other Plugins

Other algorithms developed in the LACAM lab have been ported in the framework or are migrating at the time of writing; in particular, the REDD algorithm [3] has been wrapped in a Transformer plugin, enabling the applications that use the framework to apply the redundancy detection algorithm on any RDF model they use. REDD is based on blank node semantics, and is able to detect redundancies in RDF (and OWL) models, where, for example, multiple blank nodes

with no distinguishing features are present in the same model. This is the case, for example, of a remote store that gets updates from other applications; it is possible that one or more applications send the same information more than once, and, while this is not a problem with RDF ground statements (i.e. statements with no blank nodes), since RDF models are defined as triple sets, blank nodes inserted during the life of the model are not recognized as already present; this increases the size of the models without reason, and could also be regarded as an error. Another possible application, which is under experimentation at the time of writing, is the use of REDD to detect redundant definitions in ontologies; the ongoing project aims at using the framework in the building of a Protégé<sup>6</sup> plugin.

## 5 Semantic Evolution

On the side of algorithm evolution, in the ITR component the update to OWL formalism is strictly related to the switching from keyword-based representation of the user profile to user profiles based on concepts (bags-of-concepts, BOC, instead of bag-of-words, BOW). While this shift of representation is natural when considering the new environment, we already demonstrated in [8] that the traditional TF-IDF heuristic gains some percents both in precision and recall from the new representation. Moreover, we are currently doing empirical measures on an evolution of TF-IDF that takes into account the hierarchical relations between concepts, that, informally, redefines the classical definition of TF-IDF, which is based on sheer concept occurrence number, taking into account that a more specific term is also an instance of a more general term, and as consequence, so to speak, each occurrence of the more specific term counts also as an occurrence of the more general term.

## Acknowledgments

This research was partially funded by the European Commission under the 6<sup>th</sup> Framework Programme IST Integrated Project VIKEF - Virtual Information and Knowledge Environment Framework (Contract no. 507173, Priority 2.3.1.7 Semantic-based Knowledge Systems; more information at <http://www.vikef.net>).

## References

1. B. McBride. JENA: A Semantic Web toolkit. *IEEE Internet Computing*, 6:55–59, Nov-Dec 2002.
2. F. Esposito, L. Iannone, I. Palmisano, and G. Semeraro. RDF Core: a Component for Effective Management of RDF Models. In Isabel F. Cruz, Vipul Kashyap, Stefan Decker, and Rainer Eckstein, editors, *Proceedings of SWDB'03, The first International Workshop on Semantic Web and Databases, Co-located with VLDB 2003, Humboldt-Universität, Berlin, Germany, September 7-8, 2003*, 2003.

<sup>6</sup> <http://protege.stanford.edu>

3. Floriana Esposito, Luigi Iannone, Ignazio Palmisano, Domenico Redavid, and Giovanni Semeraro. REDD: An Algorithm for Redundancy Detection in RDF Models. In Asunción Gómez-Pérez and Jérôme Euzenat, editors, *The Semantic Web: Research and Applications, Second European Semantic Web Conference*, volume 3532 of *Lecture Notes in Computer Science*, pages 138–152. Springer, 2005.
4. C. Niederée, A. Stewart, B. Mehta, and M. Hemmje. A Multi-Dimensional, Unified User Model for Cross-System Personalization. In Liliana Ardissono and Giovanni Semeraro, editors, *Proceedings of the AVI 2004 Workshop On Environments For Personalized Information Access*, pages 34–54, 2004.
5. G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.
6. F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 2002.
7. G. Semeraro, F. Abbattista, M. Degemmis, O. Licchelli, P. Lops, and F. Zambetta. Agents, Personalisation and Intelligent Applications. In R. Corchuelo, A. Ruiz Cortés, and R. Wrembel, editors, *Technologies Supporting Business Solutions, Part IV: Data Analysis and Knowledge Discovery, Chapter 7*, pages 141–160. Nova Sciences Books and Journals, 2003.
8. Giovanni Semeraro, Marco Degemmis, Pasquale Lops, and Ignazio Palmisano. WordNet-based User Profiles for Semantic Personalization. In P. Brusilovsky, C. Callaway, and A. Nurnberger, editors, *Proceedings of the Workshop on New Technologies for Personalized Information Access (PIA 2005), part of the 10th Int. Conf. on User Modeling (UM'05), Edinburgh, UK, 2005.*, pages 74–83, 2005.

# Integrated Document Browsing and Data Acquisition for Building Large Ontologies

Felix Weigel, Klaus U. Schulz, Levin Brunner, and Eduardo Torres-Schumann

Centre for Information and Language Processing (CIS)  
University of Munich (LMU), Germany  
{weigel, schulz, brunner, torres}@cis.uni-muenchen.de

**Abstract.** Named entities (e.g., “Kofi Annan”, “Coca-Cola”, “Second World War”) are ubiquitous in web pages and other types of document and often provide a simplified picture of the document’s content. We present an ontology currently containing 31,000 named entities in different languages from various domains such as history, geography, politics, sports, arts, etc., which is being developed at the University of Munich (LMU). The underlying graph data model is simple and yet extremely versatile in different application scenarios. We demonstrate a prototype of a graphical interface to both the ontology and to documents on the web or in a local document repository, with a tight interaction in both directions. Occurrences of concepts from the ontology are highlighted and hyperlinked in the documents. Unrecognized entities could be added to the database and related to other concepts in a semiautomatic process. The entity database can also be used for extending full-text queries on the web or the repository to semantically close documents, and for indexing different kinds of named entities in the document repository. Similar to a programming IDE, the system illustrates how integrated browsing, search and update functionality contributes to the construction of high-quality ontologies, fundamental to the vision of a truly “semantic” web.

## 1 Introduction

The Semantic Web is meant to enable reasoning not only on the contents of static web pages, but further and foremost on the underlying data sources such as databases, web services, document repositories, digital libraries, on-line encyclopaedias, etc. (sometimes called the “deep web”). A fundamental building block for the description, integration/mediation and inference on these heterogeneous data sources are *ontologies*, i.e., formal schemata of the concepts and relationships in one or more, possibly overlapping domains. Faced with a growing number of different knowledge representation formalisms [1,2] and tools for creating ontologies [3,4], we claim that in order to make *large* amounts of *existing* data accessible to the Semantic Web, formalisms and tools are needed which (1) make ontology development, maintenance and usage easy even for non-experts in knowledge representation, (2) strive for a feasible amount of inference rather than too much of fancy (and expensive) “magic”, and (3) offer a data-driven approach for integrating ontologies and existing data sources.

We would like to comment on these controversial issues. First, experts are most apt at giving some thematic domain a precise structure, but non-expert users may not be able to query this complex structure, falling back to less sophisticated queries or simply browsing. By contrast, collaborative projects such as Wikipedia [5] clearly show the benefits of both expert and non-expert contributions for gathering encyclopaedic knowledge. Second, even in the presence of emerging knowledge representation standards such as OWL [1], it seems natural to study simple formalisms by experimenting with prototypes of a limited functionality, before coming up with more sophisticated applications in a second step. Third, we believe that a fine-grained, up-to-date model of common-sense knowledge on a web scale is infeasible. Fully automated knowledge extraction may complement manual and semiautomatic acquisition techniques, but cannot replace them. We thus advocate a mainly data-driven process where ontologies are updated and enlarged while searching and browsing the actual contents.

**Contributions.** In this work we describe the *EFGT Net*, an ontology containing currently 31,000 named entities from various domains, which is being developed at the University of Munich. The underlying graph data model is deliberately simple and specifically designed for building scalable models of common-purpose knowledge. The ontology features multilingual concept descriptions, which we consider indispensable in a web context. Target applications include simple semantic search and inference on web contents or documents from a local repository or intranet. We also outline a new integrated tool – similar to an Integrated Development Environment (IDE) – for searching, exploring and updating the ontology while browsing the documents. An extensive sample session illustrates typical usage patterns with a prototype, emphasizing the benefit of a tight ontology/corpus integration for creating and using large-scale, high-quality ontologies.

The next section describes in a nutshell knowledge representation and inference with the EFGT Net. Section 3 presents our prototype using a real-world web example. The system architecture is sketched in Section 4. Finally, Section 5 concludes with a brief outlook on future work.

## 2 Knowledge Representation and Inference with EFGT

This section introduces the EFGT Net and its basic principles on an informal basis. For a more formal definition of the data model, see [6]. The EFGT Net is a directed acyclic graph (DAG) whose nodes represent concepts and whose edges represent directed binary relations between the concepts. Each concept has a natural-language definition in at least one of the supported languages, as well as optional semantic and syntactic information. In the current version of the EFGT Net there are about 31,000 nodes and 637,000 edges, with a language coverage of 100% German, 70% English, 30% French, 30% Polish and 10% Bulgarian. Every node has a unique identifier, its *ID string*, which determines the position of the corresponding concept in the structure of the EFGT Net and all its relations to other concepts. Thus from the complete set of identifiers, all the edges can be inferred by means of a couple of formal deduction mechanisms [7].

The EFGT Net is built and enlarged by generating ID strings according to a set of formal rules. Syntactically, ID strings are defined by the grammar in Fig. 1. The seven alternatives in the second production can be arranged into the four main types *E*, *F*, *G* and *T* (from which the acronym EFGT is derived). Uppercase letters denote sets and lowercase letters singleton elements:

$$\begin{aligned} \mathbb{I} &:= () \mid (\mathbb{X} \mathbb{I} \mid \mathbb{N}) \mid (\mathbb{I} \&\mathbb{I}) \\ \mathbb{X} &:= (\mathbf{e} \mid \mathbf{E} \mid \mathbf{F} \mid \mathbf{g} \mid \mathbf{G} \mid \mathbf{t} \mid \mathbf{T}) \\ \mathbb{N} &:= \mathbb{D} (0 \mid \mathbb{D})^* \\ \mathbb{D} &:= (1 \mid \dots \mid 9) \end{aligned}$$

Fig. 1. ID string syntax

- E, e* Type *E* denotes a set of *Entities* like *composers*, whereas type *e* denotes a singleton entity like *J. S. Bach*.
- F* Type *F* denotes a thematic *Field* (topic) like *politics*. Since every thematic field can be regarded as a set of subfields, there is no type *f*.
- G, g* Type *G* denotes *Geographical* sets like *rivers*, whereas type *g* stands for singleton geographic sites like *the Alps*.
- T, t* Finally, type *T* denotes a *Temporal* period like *epochs in art*, whereas type *t* denotes an individual time interval or point like *September 11<sup>th</sup>*.

As shown by the first production in Fig. 1, there are two ways to create a new ID string based on existing ID strings like the initial *top node* (). First one can refine a single existing ID string by a *local introduction* with the dot notation shown in Fig. 1. For instance, (F().1) defines the first concept below the top node, and is assigned to the topic “politics” in our ontology. When creating a new concept by local introduction, the only directly connected other concept is the existing concept being refined. The resulting edge can represent different relations, e.g., the definition of a subconcept – “foreign policy” (F(F().1).2) –, or subset – “political problems” (F(F().1).20) –, or the selection of a single member of a set (“Presidents of the United States” – “Bill Clinton”). This vagueness is intended to facilitate the modelling of real-world knowledge, mirroring the ambiguity of recursive constructs in natural language (“members of the Beatles” vs. “albums of the Beatles”). The second way to create a new ID string is to combine two existing ID strings with the & operator, which can either denote the intersection between two given sets, (X&Y), or the creation of a new set (X&y) by placing the original set X in the context of a singleton concept y. Note that (X&y) need not be a subset of X. As an example combining local introduction and intersection, consider the concept “politics” (F().1) and its subfield “defense politics” (F(F().1).14). The intersection with “persons” (E().1) yields the concept “defense politicians” ((F(F().1).14)&(E().1)). For convenience, we henceforth drop occurrences of the top node, writing (F.1) for (F().1), etc.

### Inference with the EFGT Net

Figure 3 illustrates how the new concept “Cities in Bavaria” is represented in the ontology (for simplicity, ID strings are symbolized by their descriptions). Assume there exist concepts “Germany”, “Bavaria” (a region in Germany), both of type *g*, and concepts “Cities”, “Cities in Germany” of type *E*. “Bavaria” is derived from “Germany” by local introduction, whereas “Cities in Germany”

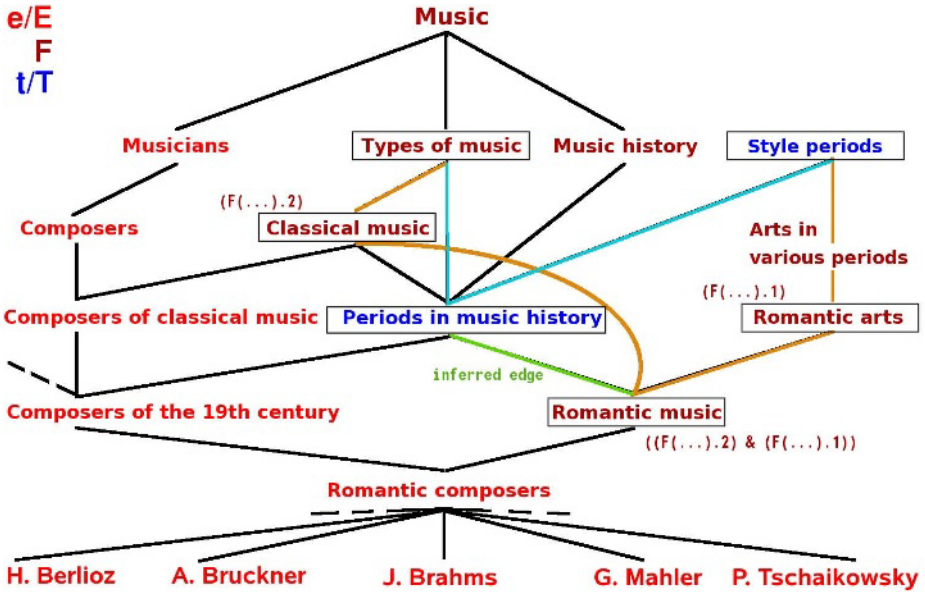


Fig. 2. Neighbourhood of the concept “Romantic composers” in the EFGT Net

results from combining “Cities” and “Germany” with the & operator. Analogously, the new concept “Cities in Bavaria” is defined simply by applying & to “Cities” and “Bavaria”, as indicated by the dashed arrows. However, the edges actually inferred (green solid lines) relate the new concept directly to “Cities in Germany” (since Bavaria is a part of Germany), which implies an indirect link to “Cities”.

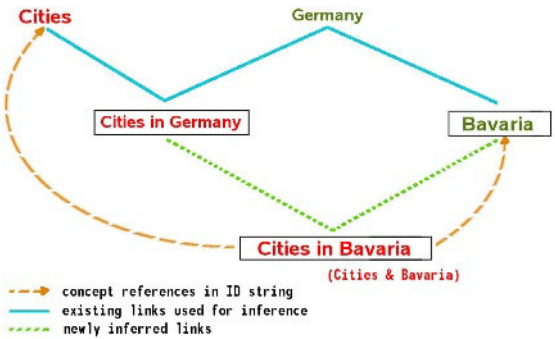


Fig. 3. Link inference in the EFGT Net

Typically many relations are inferred from a small number of references to existing concepts explicitly stated in the ID string. Figure 3 shows a similar case in a more complex subgraph. When inserting “Romantic music” into the graph, only “Classical music” and “Romantic arts” are explicitly mentioned in the ID string. The link stating that romantic music is a period in music history is inferred, based on the fact that the latter concept shares all higher-level ancestors (“types of music”, “style periods”) with the newly defined node. The exact inference algorithm is given in [7].

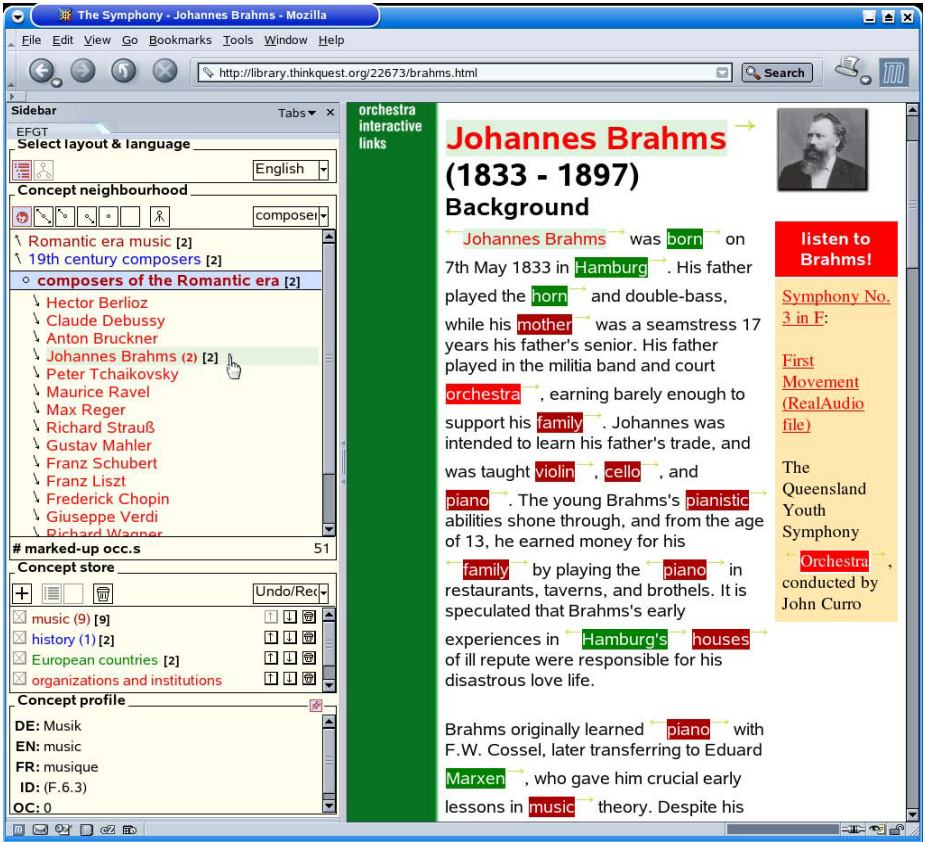










Fig. 4. Screenshot of the EFGT Net browser prototype

### 3 Using the EFGT Net for Semantic Web Browsing

This section describes a typical session with our prototypical ontology browser for the EFGT Net (see Figure 4). The system is explained in detail in Section 4. To achieve tight interaction of the documents and the ontology as claimed above, we integrated a GUI to the EFGT Net (left-hand side in Figure 4) into an ordinary web browser accessing documents in local repositories, intranets or the WWW (right-hand side). The EFGT GUI displays any concept along with its *neighbourhood*, i.e., the more general (specific) concepts one level above (below). In Figure 4, the concept “romantic composers” from Figure 2 is selected. Its ancestor (descendants) are labelled  $\nwarrow$  ( $\searrow$ ). Colours indicate the EFGT type of all concepts as in Figure 2. Clicking on any item in the list shifts the focus to the neighbourhood of that concept. History and bookmark functions (above/below the list) and a backlink to the top node ( $\boxtimes$ ) facilitate browsing the EFGT Net.



As outlined in Figure 5, the EFGT GUI interacts with the web page by highlighting selected occurrences of concepts in the document. The six highlighting modes (see the buttons on top of the neighbourhood) are: all EFGT concepts () , both more general and more specific () , more general/specific only (/) , current concept only () , or no highlighting at all (). In Figure 4, the  mode is active. A click on the  button removes the highlighting of all occurrences visible in the screenshot, except “Johannes Brahms” and “music”. Counters behind list items indicate how many occurrences of a concept itself (parentheses) or any of its descendants (square brackets) are highlighted in the document. The inference on ID strings for highlighting and counting ancestors or descendants is done on the fly with simple and fast string matching [6]. Occurrences of the same concept in the web page are chained through hyperlinks (light green arrows) for easy location of all paragraphs where a concept of interest is mentioned. Each occurrence is also backed by a hyperlink into the EFGT GUI, with the same functionality as links inside the GUI (focus shift, see above). Besides, by hovering an occurrence, the user obtains a *concept profile* (bottom left in Fig. 4) including, e.g., the life time of a person or the English description of the concept, which is useful when browsing documents in foreign languages.

Currently concepts in the GUI can be reached alternatively by (1) browsing the hierarchy down from the top node, (2) selecting a concept from the *concept store* containing bookmarks stored persistently in previous sessions, (3) selecting a concept visited earlier from the persistent history, or (4) clicking on an occurrence in the document. A simple string search on the natural-language concept definitions is being integrated into the prototype. Further extensions of the functionality, including ontology maintenance, are listed in Section 5.

## 4 Architecture of the Prototype

The ontology browser presented in Section 3 is implemented as a client-server application using Java Server Pages (JSP) and Java Servlets, as sketched in Figure 6. The EFGT GUI is a JSP displayed in the sidebar of the web browser (Mozilla 1.7.12). Each focus shift (i.e., request for the neighbourhood of a specific concept) triggers the dynamic generation of a new web page containing the entire GUI shown in Figure 4. This includes (partly hidden) concept and widget labels in all supported languages, such that no reload is necessary when the user selects a different language from the drop-down list on top. The concept information needed to build the neighbourhood is obtained from the EFGT Net, which resides in a RDBS backend (PostgreSQL 8.0). Requests for a specific neighbourhood are answered through a Web Service interface from tables containing all nodes and edges in the graph. The underlying inference relies on

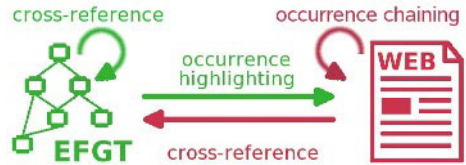


Fig. 5. Ontology interaction

simple and efficient string matching of ID strings. In order to reduce the response time even further, a server-side concept cache retains the most recently requested neighbourhoods.

The preparation of EFGT-enriched documents is a little more involved, as shown on the right-hand side of Fig. 6. The web browser directs each request for a new page (through a link or the address bar) to the servlet engine (Tomcat 5.0) running on a proxy server. The servlet receiving the request first fetches all data from the document repository or the web. Before the data is sent back to the client, all occurrences of EFGT concepts in the requested document must be located and

hyperlinked using the ID strings from the EFGT Net, which are needed for the interaction with the EFGT GUI. We found that when implemented naïvely, preprocessing the documents thus on the fly easily entails unacceptable response times. Therefore we compile the required EFGT data into a set of rewrite lexica to be used by an extremely efficient and scalable string transducer [8]. Each lexicon contains the ID strings of all nodes in the EFGT Net, along with the concept definitions in a specific language. Thus the size of the lexica varies with the coverage for the corresponding languages. For instance, the German lexicon comprising all 31,000 nodes in our EFGT Net takes up 15 MB on disk. Note that the lexica are created off-line once (in a few seconds) and then used repeatedly when requests come in. Annotating all concept occurrences in a given document with EFGT data (and HTML markup for links) is a matter of milliseconds.

There are a couple of caveats related to the document preprocessing. First, to determine which lexicon to use for a given web page we examine the HTTP header *Content-Language*, using a simple fall-back heuristic based on the top-level domain of the requested URL if the header is missing. More sophisticated techniques could infer the language from the page contents. Second, markup must be temporarily stripped off the document before the transduction, otherwise HTML or embedded script code might get corrupted. Finally, to recognize inflected occurrences of EFGT concepts and to avoid needless annotation, some linguistic normalization (stemming, stop-word removal) is in order—both on-line in the document and off-line when creating the lexica. While this works fine for English content (using Porter stemming), a simple adaptation of the Porter algorithm to German resulted in many needless ambiguities. Hence we automatically generated all noun inflections of the German concept descriptions, added them to the lexicon (included in 15 MB), and disabled stemming for German.

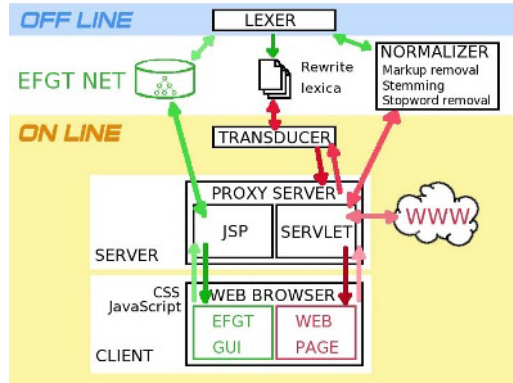


Fig. 6. Architecture of the prototype

## 5 Conclusions

In this paper we presented an integrated tool for document browsing with ontology support, and showed how the simple and versatile EFGT formalism provides efficient string-based inference on multilingual corpora such as local document repositories or the web. The underlying data model addresses common-sense knowledge and can be universally applied and manipulated by non-experts. Efficient reasoning is achieved by comparing language expressions directly, which ensures the scalability of our approach. While the prototype presented here exploits inference only for simple subsumption tests, we envisage a more sophisticated semantic search and classification of the documents, using ID string manipulation either directly or with graphic metaphors. In the EFGT Net, semantic querying can be realized in an intuitive way without imposing to the user the burden of a complex formal query language. Further query refinement is achieved through interleaved ontology navigation and document browsing as illustrated above. The browsable GUI to the multilingual ontology distinguishes our system from other tools for automatic document annotation and hyperlinking, such as Magpie [9] and COHSE [10]. We plan to extend the prototype to support the user in ontology maintenance: e.g., new concepts occurring in the documents could easily be added on the fly to the EFGT Net using a drag-and-drop interface. By virtue of its simple recursive structure, the EFGT formalism seems particularly promising for such combined knowledge management and acquisition techniques. As the ontology grows, the automatic linking of entities and documents becomes more difficult since ambiguities arise. Apart from user-driven disambiguation (showing alternatives in context menus), we intend to develop detection methods for new entities and linguistic variants (maybe using named-entity recognizers such as GATE [11]), as well as intelligent indexing with on-line disambiguation. These steps should allow us to minimize human effort during ontology development.

## References

1. Dean, M., Schreiber, G.: OWL Web Ontology Language Ref. (2005) W3C Rec.
2. Klyne, G., Carroll, J.J.: Resource Description Framework (2005) W3C Rec.
3. Sure, Y., Erdmann, M. et al.: OntoEdit: Collaborative Ontology Engineering for the Semantic Web. In: Proc. 1st Int. Semantic Web Conf. (2002) 221–235
4. Noy, N.F., Sintek, M. et al.: Creating Semantic Web Contents with Protege-2000. *IEEE Intelligent Systems* **16** (2001) 60–71
5. Wikipedia: The Free Encyclopedia. ([www.wikipedia.org](http://www.wikipedia.org))
6. Schulz, K.U., Weigel, F.: Systematics and Architecture for a Resource Representing Knowledge about Named Entities. In: Proc. Workshop on Principles and Practice of Semantic Web Reasoning. (2003) 189–207
7. Brunner, L., Schulz, K.U., Weigel, F.: Organizing Thematic, Geographic and Temporal Knowledge in a Well-founded Navigation Space: Logical and Algorithmic Foundations for EFGT Nets. *J. Web Serv. Research, Spec. Issue “Semantically Augmented Metadata for Services, Grids, and Software Engin.”* (2006) (in press).

8. Mihov, S., Schulz, K.U.: Efficient Dictionary-Based Text Rewriting using Subsequential Transducers. *Journal of Natural Language Engineering* (2005)
9. Dzbor, M., Domingue, J., Motta, E.: Magpie: Towards a Semantic Web Browser. In: *Proc. 2nd Int. Semantic Web Conf.* (2003) 690–705
10. Carr, L., Hall, W., Bechhofer, S., Goble, C.: Conceptual Linking: Ontology-based Open Hypermedia. In: *Proc. 10th Int. World Wide Web Conf.* (2001) 334–342
11. Cunningham, H., Humphreys, K. et al.: GATE – a General Architecture for Text Engineering. In: *Proc. 5th Applied Natural Lang. Processing Conf.* (1997) 29–30

# A Workflow Modeling Framework Enhanced with Problem-Solving Knowledge

Juan C. Vidal, Manuel Lama, and Alberto Bugarín

Department of Electronics and Computer Science  
Campus Universitario Sur s/n, University of Santiago de Compostela  
15782 Santiago de Compostela, A Coruña, Spain  
{jvidal, lama, alberto}@dec.usc.es

**Abstract.** This paper outlines a framework that extends the Unified Problem-solving Method description Language in order to enhance workflows modelling with knowledge. The framework defines the knowledge components needed to represent and reuse both the static and dynamic knowledge used to describe a business process. For this purpose, workflows are represented by means of an ontology based on High-level Petri Nets. This ontology provides the structure and axioms needed to validate and reason about the graph structure, the state and the execution of the workflow. The framework architecture is based Workflow Reference Model architecture defined by the Workflow Management Coalition and is built around the knowledge components that defines the structure and execution semantics of workflows.

**Keywords:** workflows; ontologies; problem-solving methods.

## 1 Introduction

Business Process Management (BPM) is a technology which gives support to the whole business process development and management [1]. BPM allows the explicit representation of the business process logic in a process-oriented view. In this context, a workflow consists of an ordered sequence of activities and task executors. Modeling techniques for these processes have risen since the first appearance of commercial Workflow Management Systems. Nevertheless, most of these techniques lack in theoretical foundation. At present there is no standard language for workflow specification, but Petri Nets [2] and Pi calculus [3] have proved to be a good starting point for a solid theoretical foundation. The main advantage of using Petri nets instead of Pi calculus is that Petri nets provide a solid graphical semantics and analysis methods (Pi calculus is based on text descriptions) which makes the labor of designers easier [4].

However, workflow modeling techniques are focused on the description of the coordination between the tasks to be executed in a process. These techniques do not incorporate a modeling of the problem-solving knowledge (how the tasks are executed and what relation exists between those tasks) neither the modeling of the static knowledge of the domain used to execute the process tasks. In fact, the workflows

themselves can also be considered as pieces of knowledge that indicates how the tasks of a problem-solving method can be coordinated to achieve a goal.

In this paper we describe a framework for workflow management and execution that integrates static and problem-solving knowledge. For this purpose, workflows are defined in an ontological framework based on the Unified Problem-solving Method description Language (UPML) [5]. The framework extends UPML with two new components: *resources* and *workflows*. On one hand, resource components define the organization model of workflows by means of the agents involved in the workflow execution. On the other hand, workflow components define the operational description of the methods used to solve a task. Workflow components are defined by means of an ontology stack that describes semantically all the elements of a workflow. Following this framework, a software architecture has been designed and implemented. This architecture is based on the workflow reference model [6] architecture defined by the Workflow Management Coalition and provides a workflow engine and a set of interfaces to support interaction of the system with agents (human or software). This architecture has been implemented and applied to the price estimation task of wood-based furniture.

This paper is structured as follows: section 2 outlines the framework solution that integrates workflow and knowledge technologies. In this section, we pay special attention to the ontology (section 2.1) that supports the definition and execution of workflows and to the framework architecture (section 2.2). Section 3 describes an example of application of this framework in the wood furniture industry. Finally, section 4 presents work contributions and its future developments.

## 2 Knowledge-Based Workflow Framework

We have developed a new framework that enhances the workflow modeling (such as is presented in [1]) with a new dimension which deals with knowledge modeling. More specifically, this dimension deals with the definition and modeling of the static knowledge (through ontologies) and the problem-solving knowledge (through PSM).

**Workflow Framework.** The workflow framework is composed of two dimensions: the *resource dimension*, which describes the organizational elements (humans and software) that take part in the workflow specification (such as organizational units and roles); and the *process dimension*, which deals with the definition and execution of the process structure, specifying the activities (also called tasks) to be executed, its order, the conditions that must verify the activities to enable its execution, and the specific case (or domain) in which the workflow will be executed. In this paper, we will focus on the integration of knowledge into the process dimension.

The workflow framework proposes to model the process dimension through High-Level Petri Nets (HLPN). A HLPN is a directed, connected, and bipartite graph with two kinds of nodes: places and transitions. Each place is associated with a color (*datatype*) and may contain a multiset of values of such color (*tokens*). Transitions are the active components of the net and they model activities whose execution (*transition firing*) changes the state of the net. Transitions and places are connected among them through arcs, which are edges in the graph. An arc can be annotated with an

expression that contains variables, values and operations. When this expression is evaluated, a multiset of values from the color of the adjacent place is generated.

Applying the HLPN formalism for modeling workflows, activities are modeled by means of transitions, conditions by means of places, and cases by means of colored tokens. Furthermore, the execution of the workflow follows the semantics defined to execute HLPN.

**Knowledge Framework.** UPML is the standard *de facto* that models the knowledge components of a PSM: tasks, methods, and domain models. A *task* describes the operation to be solved during the execution of a method, specifying the required input and output parameters and the pre- and post-conditions. This description is independent of the method used to solve the task. A *method* details the control of the reasoning process to achieve a task. For composite methods, they also describe both the decomposition of the general task into subtasks and the coordination of those subtasks (flow control) to achieve the required result. An *adapter* specifies mappings among the UPML knowledge components, adapting a task to a method and refining tasks and methods for generating more specific components. Finally, a *domain model* describes the domain knowledge of an application needed to define the tasks in the application domain and to carry out the inference steps to execute the methods.

To integrate these two frameworks with the aim of developing a knowledge-enhanced workflow, it is needed to establish a set of mappings between the components of each framework. Figure 1 depicts these mappings with an example: (i) a HLPN transition of the process dimension is associated with a UPML task (*Propose Workplan*), which will be solved by a given method (GA method); and (ii) the operational description of a composite method is associated to the process dimension of a

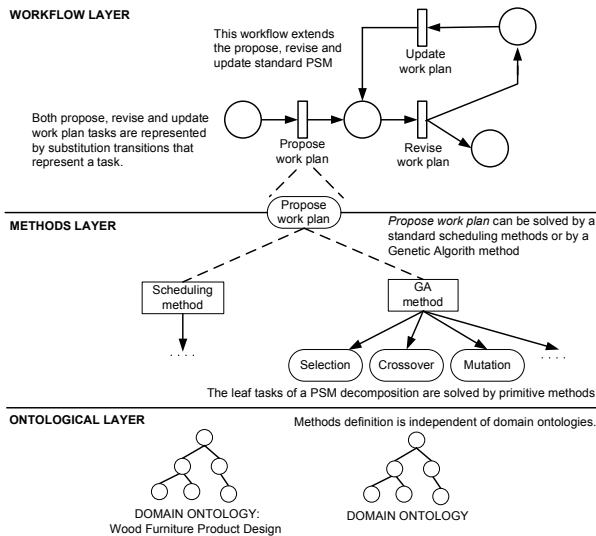


Fig. 1. Example of integration between workflows and problem-solving methods

workflow whose HLPN transitions are linked to the subtask (Selection, Cross-over, Mutation) in which such composite method is decomposed of. Therefore, from the process structure perspective, the result of the integration between the process and knowledge dimensions is a two layer model that delegates the execution of HLPN transitions to the UPML methods and models the control flow of those methods (its operational descriptions) through workflows.

### 2.1 Workflow Modeling Ontology

As the UPML framework is formalized by means of an ontology that describes semantically all the knowledge components, it is necessary to create a workflow ontology for connecting the workflow components (transitions, places, etc.) with the UPML elements (tasks, methods, etc.), and vice versa. Therefore, the core of the knowledge-enhanced workflow framework relies on a stack of ontologies that define all the components of a workflow based on HLPN. These ontologies are the following (Fig. 2(a)):

- *Data Types Ontology (DTO)* contains the data types associated to the concept attributes of the domain ontology used to annotate the arcs and places of the workflow. This ontology is based on the XML Schema Data Types specification [7].
- *Knowledge Representation Ontology (KRO)* is developed on the top of the DTO and describes the representation primitives (such as concepts or relations) used to specify the domain ontology managed by the HLPN. The components of the domain ontology, therefore, will be instances of the KRO and will be used to define the colors of the HLPN. The KRO selected has been the Frame Logic knowledge model [8], which provides the primitives for describing classes, attributes, formulas, and axioms.

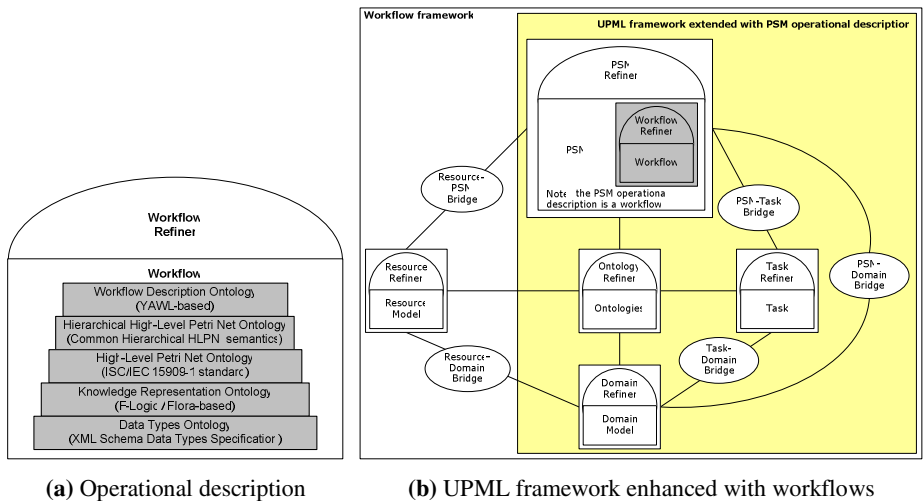


Fig. 2. The workflow ontology extends the UPML framework



- *High-level Petri Nets Ontology (HLPNO)* describes semantically all the static components of a HLPN (arcs, nodes, terms, etc.) and the dynamic behavior of the net when it is executed. The HLPNO is the core ontology of the stack, because of it defines a set of axioms that guarantee the correctness of both the static representation of the net (and, therefore, of the process structure) and the execution of the net. The HLPNO is based on the international standard ISO/IEC 15909-1 [9] that formulate the semantics of HLPN in a mathematical notation. The ontology is a translation of this notation into a logic formulation (such as Frame Logic).
- *Hierarchical High-level Petri Nets Ontology (HHLPNO)* is based on the HLPNO and approaches hierarchical representation and model composition of HLPN. This ontology extends the token-based folding defined by HLPN with some composition and structuring mechanisms [10]. This ontology has been introduced in the stack to avoid the rapid model growth of flat HLPN, and thus define the way several HLPN can be combined to obtain a more complex net.
- *Workflow Description Ontology (WDO)* contains the elements used in the definition of workflows and is based on the vocabulary defined in YAWL (Yet Another Workflow Language) [11]. This ontology extends the semantics of HHLPNO from which workflows are defined.

Once the workflow ontology has been developed, the integration with the UPML ontology can be carried out. UPML does not provide a way for specifying the operational description of a method and just uses textual descriptions (main drawback of UPML) to define the structures or blocks needed to coordinate the execution of the tasks that compose a method. Therefore, it is necessary to extend the UPML ontology to describe the operational description of composite methods. For it, a workflow (specified through the instances of the stack of ontologies) is introduced as a new piece of knowledge (Fig. 2(b)), which is associated with PSM by means of its operational description.

## 2.2 Software Architecture

Framework architecture is depicted in Fig. 3 and is based on Workflow Reference Model architecture [6] proposed by the Workflow Management Coalition. The core of the framework is the workflow enactment service. This service runs the workflow engine which provides all the functions for the management and execution of workflows. This software interprets the process description and controls the instantiation of processes, the sequencing of activities, the scheduling of user work and the invocation of external applications as necessary. However, besides the usual functionalities of this kind of software, our implementation also must interpret the knowledge features specified during the process definition. This extension adds semantic to workflow specification and, as we will see in this section, also improves the system functionalities. Workflow engine is made up of four blocks:

- *Workflow Manager* facilitates the tools to manage the definition of workflows. Correct definition of tasks, methods, resources and domain knowledge of workflows is guaranteed through the *Knowledge Manager* component. Workflow Manager is also in charge of creating new workflow instances execution.
- *Workflow Instance Manager* manages the execution of the workflow instances. Each time a task is performed, this component performs the changes in the work-

flow instance execution. For example, it removes the input tokens of the task and produces new ones in its outputs. Execution semantics is also guaranteed by the *Knowledge Manager* component. In fact, the workflow ontology created in this framework provides a set of axioms to ensure its correct execution.

- *Work List Manager* manages the resources work list. Basically, a scheduler associates each task to the most suitable resource. This association is based on the resource workload, capabilities, but also in the knowledge needed by tasks and provided by resources. This feature avoids the rejection of work by unskilled resources and improves the workload balancing.
- *Knowledge Manager* provides the tools to access to the ontologies that define the workflows. The knowledge base is defined around the extended version of UPML defined in this framework and incorporates the set of ontologies that deal with the task, resource, domain, and workflow models definition.

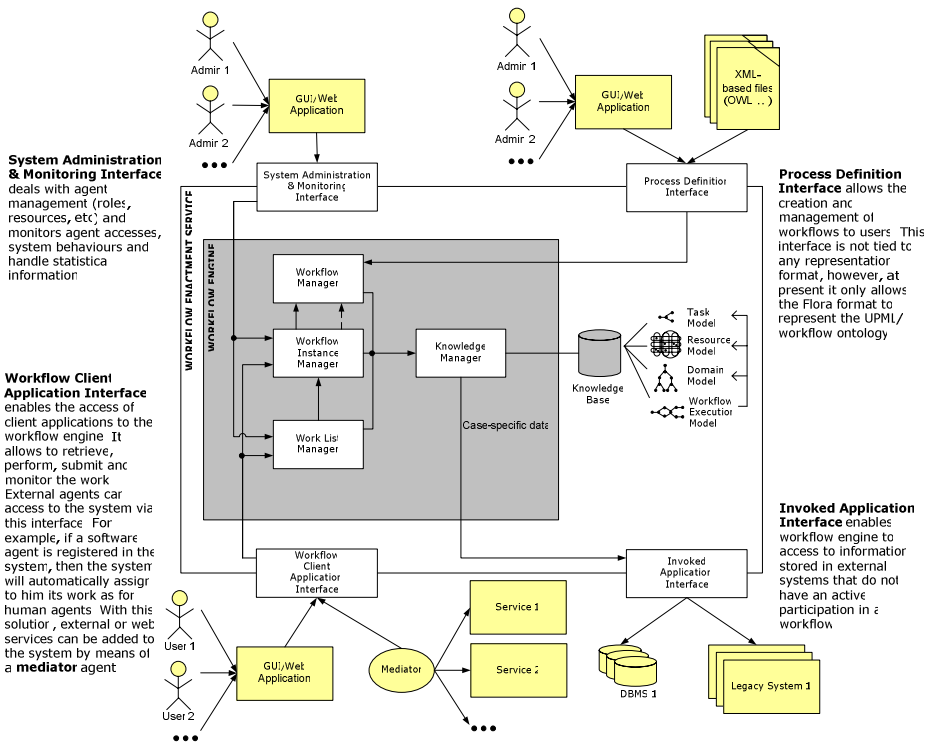


Fig. 3. Architecture of the workflow framework

### 3 Price Estimation in the Wood Furniture Industry

This section outlines a workflow developed by means of the framework described in this paper. This workflow deals with the price estimation task of wood-based furniture. The automation of this process is an important advance for improving productivity and

is a step forward to reducing product costs, getting new customers and defining a better manufacturing. Furthermore, the quality of its solutions will define the profits of the company. However, the automation of this manually-based process is not enough to get the desired improvements and needs: (i) to support a high level of collaboration, synchronize people, departments or resources involved in the process; and (ii) monitor its behavior and detect events in a defined way. In this sense, workflows and design for manufacturing and assembly (DFMA) guidelines [12] are the starting points to get these objectives.

Workflow structure is represented by means of the Petri net depicted in Fig. 4(a). It should be noted that the workflow herein described is a general solution for the price estimation task and could be applied to other domains. In fact, workflow reuse is one of the main motivations of this framework. In order to apply this workflow to our domain, the HLPN workflow specification is related to the desired domain model, in this case, the wood furniture manufacturing domain. For example, estimate evaluation depends on products design. If these products are clothes then sketches could be sufficient to define the conceptual design, however if they are wood furniture, CAD designs are necessary. For the sake of simplicity, we use a place-transition view to present the workflow structure, although the final workflow model is described with HLPN. The behavior of the workflow is modeled through a PSM known as Propose, Revise and Modify, which is a generic constraint-satisfaction method. This method obtains a viable and cost-optimized furniture design, through the evaluation and modification of the requirements and constraints of the conceptual design.

Figures 4(b) and 4(c) describes the workflows defined for create product description and product design tasks. These workflows solve the upper tasks defined in the workflow depicted in 4(a). Workflow composition is solved at design

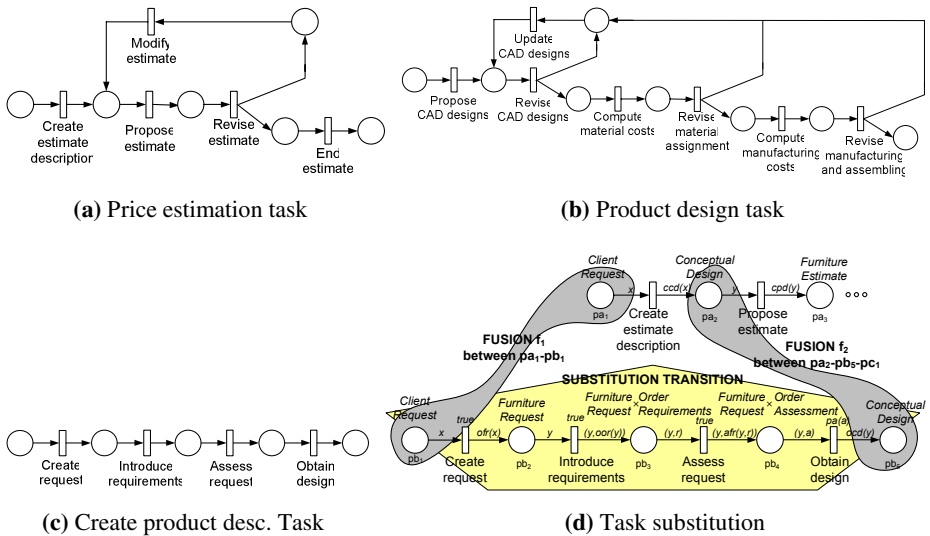


Fig. 4. Furniture price estimation workflow

time. The framework decides which PSM will solve composite tasks based on the assumptions of the task. Net composition is defined by means of hierarchical mechanisms defined by the HHLPO. For example, Fig. 4(d) depicts how the create estimate description transition is substituted by the create product description page. This substitution maps the input and output places of the substituted transition to the page places.

Within the architecture defined in section 2.2, PSM definition is performed through the *process definition interface*. When a PSM is defined in the framework, the adapters between this PSM (and its workflow operational description) and the tasks it can solve are also defined. Following with the example, the inputs of the create estimate description task are mapped to the places of the create product description page that defines the PSM operational description. We must remark that the places fused for the transition substitution must have a compatible type. This interface is also in charge of the definition of the resources involved in the estimate definition and the domain knowledge used to describe the wood-based furniture manufacturing. Agents (human and software agents) introduced in a workflow are classified within the resource model defined and related to the tasks they are able to perform and the domain knowledge they handle. PSM also define adapters with the

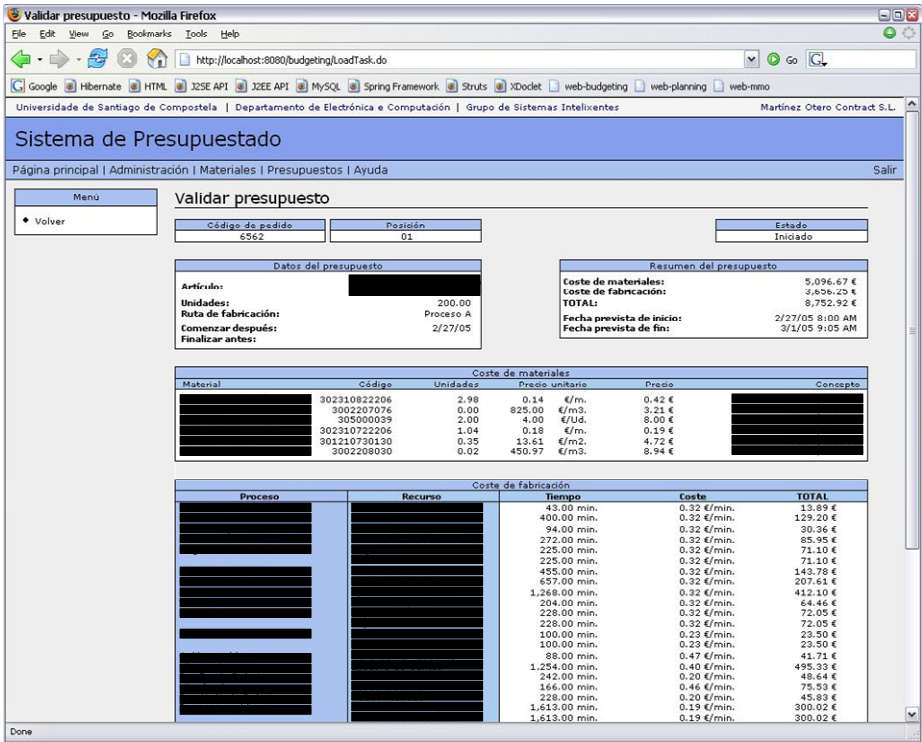


Fig. 5. Furniture estimate validation page framework

domain model needed to perform a task. We must remark that the consistency of the knowledge introduced in the workflow engine is always verified against the axioms of the ontology.

Once defined, agents (human or software agents) begin the execution of workflows through the *workflow client application interface*. The *workflow instance manager* creates an instance of the workflow and enables the execution of active tasks of the instance. The *work list manager* sorts the active tasks and schedules its execution between the users or group of users of the workflow. The same interface allows users to retrieve the tasks they must perform. For example, if the furniture price estimation workflow is started then the `create request` task become active (remember that the `create estimate description` task is substituted) and will be assigned to an agent that belongs to the technical direction. At this point, we must mention that the furniture price estimation system has been developed inside an application server that runs the workflow enactment service. The access of users is made through web pages. In this sense, a user can access its work list and select a task to perform. The system locks this task until the user finishes it or he exceeds a time out. Human tasks, like `create request`, `assess request` and `validate estimate` (Fig. 5), have been developed through web interfaces. Software tasks, like `compute material costs`, or `compute manufacturing costs`, have been developed through web services. The invocation of these services is made by means of software agents that mediate between the service and the workflow engine. In any case, both web interfaces and software agents follow the task definition taken from the workflow. They take their inputs from the instance of the task and produce the output specified in the task definition.

## 4 Conclusions and Future Work

This work describes how the modeling of business processes by means of workflows and knowledge can be integrated in a framework. This framework allows the representation of both the static and the dynamic knowledge of a process. The ontology for workflow description presented in this paper provides the basis for the reuse of dynamic knowledge and enriches the PSM operational description. Current PSM representation proposals describe PSM control and knowledge flow in a program like way. In our opinion, this kind of representation lack on real knowledge-level modeling support and approach the problem modeling from a perspective whose logical and implementation levels are closely related. The use of a formal model like Petri Nets for modeling the dynamic knowledge of PSM provide a knowledge-level way to construct a task model related to some (problem-solving) behavior. Furthermore, workflows provide a clear separation between logic and implementation and allow PSM to deal with concurrence in a graphical way, but with a formal semantics.

This work also describes how the framework has been implemented in the wood-based furniture manufacturing domain. An architecture has been proposed for that matter. Our solution integrates the common architecture of Workflow Management Systems with the knowledge components needed to support our framework. In fact, a knowledge engine supports the set of ontologies defined in the framework and is also in charge of verifying the consistency of the knowledge.

As future work, the workflow ontology stack will be extended in order to support the definition of services within the framework.

**Acknowledgments.** Authors wish to thank the Xunta de Galicia, Martínez Otero Contract, S.A., and the Ministerio de Ciencia y Tecnología for their financial support under the projects PGIDIT04SIN206003PR, PGIDIT04DPI096E, and TIC2003-09400-C04-03.

## References

1. W.M.P. van der Aalst: The application of Petri Nets to workflow management. *The Journal of Circuits, Systems and Computers* 8(1) (1998) 21–66
2. K. Jensen: Coloured Petri Nets. Basic concepts, analysis methods and practical use. EATCS monographs on Theoretical Computer Science. Springer-Verlag, Berlin (1992)
3. R. Milner: Communicating and mobile systems: The Pi-Calculus. Cambridge University Press, Cambridge, UK (1999)
4. W.M.P. van der Aalst: Three good reasons for using a Petri Net-based workflow management system. In: *Proceedings of the International Working Conference on Information and Process Integration in Enterprises (IPIC'96)*, Cambridge, Massachusetts (1996) 179–201
5. D. Fensel, E. Motta, V. R. Benjamins, et al.: The Unified Problem-solving Method Development Language UPML. *Knowledge and Information Systems* 5(1) (2002)
6. Workflow Management Coalition: The workflow reference model (1999)
7. P.V. Biron and A. Malhotra: XML Schema Part 2: Datatypes. (2001)
8. M. Kiefer, G. Lausen, and J. Wu: Logical foundations of object-oriented and frame-based languages. *Journal of ACM* (1995)
9. J. Billington, et al.: High-Level Petri Nets - Concepts, definitions and graphical notation. Final Draft International Standard ISO/IEC 15909 (2002)
10. L. Gomes and J.P. Barros: Structuring and composability issues in Petri Nets modeling. *IEEE Transactions on Industrial Informatics* 1(2) (2005) 112–123
11. W.M.P. van der Aalst and A.H.M. ter Hofstede: YAWL: Yet Another Workflow Language. QUT Technical report FIT-TR-2002-06, Queensland University of Technology (2002)
12. G. Boothroyd, P. Dewhurst, and K. A. Knight: *Product design for manufacture and assembly*. Marcel Dekker Inc., New York (1994)

# M-OntoMat-Annotizer: Image Annotation Linking Ontologies and Multimedia Low-Level Features<sup>\*</sup>

Kosmas Petridis<sup>1</sup>, Dionysios Anastasopoulos<sup>1</sup>, Carsten Saathoff<sup>2</sup>, Norman Timmermann<sup>2</sup>, Yiannis Kompatsiaris<sup>1</sup>, and Steffen Staab<sup>2</sup>

<sup>1</sup> Informatics and Telematics Institute, GR-57001 Thessaloniki, Greece

{kosmas, anastas, ikom}@iti.gr

<sup>2</sup> University of Koblenz, D-56016 Koblenz, Germany

{saathoff, normania, staab}@uni-koblenz.de

**Abstract.** Annotations of multimedia documents typically have been pursued in two different directions. Either previous approaches have focused on low level descriptors, such as *dominant color*, or they have focused on the content dimension and corresponding annotations, such as *person* or *vehicle*. In this paper, we present a software environment to bridge between the two directions. *M-OntoMat-Annotizer* allows for linking low level MPEG-7 visual descriptions to conventional Semantic Web ontologies and annotations. We use *M-OntoMat-Annotizer* in order to construct ontologies that include prototypical instances of high-level domain concepts together with a formal specification of corresponding visual descriptors. Thus, we formalize the interrelationship of high- and low-level multimedia concept descriptions allowing for new kinds of multimedia content analysis, reasoning and retrieval.

## 1 Introduction

Representation and semantic annotation of multimedia content have been identified as important steps towards more efficient manipulation and retrieval of visual media. Although new multimedia standards, such as MPEG-4 and MPEG-7 [1], provide important functionalities for the manipulation and transmission of objects and associated metadata, the extraction of semantic descriptions and annotation of the content with the corresponding metadata is out of the scope of these standards and is left to the content manager. This motivates heavy research efforts in the direction of automatic annotation of multimedia content.

Here, we recognize a broad chasm between current multimedia analysis methods and tools on the one hand and semantic annotation methods and tools on the other hand. State-of-the-art multimedia analysis systems are severely limiting themselves by resorting mostly to visual descriptions at a very low level, e.g.

---

<sup>\*</sup> This research was partially supported by the European Commission under contract FP6-001765 aceMedia. The expressed content is the view of the authors but not necessarily the view of the aceMedia project as a whole.

the *dominant color* of a picture. This may be observed even though the need for semantic descriptions that help to bridge the so called *semantic gap* has been acknowledged for a long time [2, 3]. At the same time, the semantic annotation community has only recently started to work into the direction of semantic annotation in the multimedia domain and still remains a long way to go. Work in semantic annotation currently addresses mainly textual resources [4] or simple annotation of photographs [5, 6].

Acknowledging both the relevance of low-level visual descriptions as well as a formal, uniform machine-processable representation [7], we here try to bridge the chasm by providing a semantic annotation framework and corresponding tool, *M-OntoMat Annotizer*, for eliciting and representing knowledge both about the *content domain* and the *visual characteristics* of multimedia data itself. In the framework we propose, this link between the MPEG-7 visual descriptors and domain concepts is made explicit by means of a conceptualization based on a *prototyping* approach. The core idea of our approach lies in a way to associate concepts with instances that are deemed to be prototypical by their annotators with regard to their visual characteristics. To establish this semantic link we have implemented our framework in a user-friendly annotation tool, *M-OntoMat-Annotizer*, extending our previous framework for semantic annotations of text [4]. The tool has been built in order to allow content providers to annotate visual descriptors without prior expertise in semantic web technologies or multimedia analysis. More specifically, it allows to extract MPEG-7 visual descriptors from both images and videos and to store these descriptors as so-called visual prototypes of ontology classes. The prototypes are stored as RDF instances using a RDF version of the MPEG-7 visual descriptors [8]. The prototype approach specifically provides an OWL-DL friendly way of linking classes to concrete visual characteristics.

The existence of such a knowledge base may be exploited in a variety of ways. In particular, we envision its exploitation in two modes:

(1) *Direct exploitation:* In this mode, an application uses the knowledge base directly. For instance, during the semantic annotation process one may gather information like *the blue cotton cloth 4711 in image 12 has a rippled texture described by values 12346546*. Such kind of semantic knowledge may be used later, e.g. for combined retrieval by semantics and similarity in an internet shop. Obviously, such kind of knowledge is expensive to be acquired manually, even when resorting to a user friendly tool. Thus, this kind of knowledge may only be provided for valuable data, such as images or videos of commercial products or of items from museum archives.

(2) *Indirect exploitation:* In this mode, the a-priori knowledge base ‘only’ serves as a data set provided to prepare an automatic multimedia analysis tool. For instance, consider the provider of a sports portal offering powerful access to his database on tennis, soccer, etc. He uses semantic annotation of multimedia images or videos in order to prepare an analysis system. For instance, he uses *M-OntoMat-Annotizer* in order to describe the shape and the texture of tennis balls,



rackets, nets, or courts and he feeds these descriptions into an analysis system. The system uses the descriptions in order to learn how to tag and relate segments of images and video keyframes with domain ontology concepts. A customer at the portal may then ask the system what it could derive about the images and the videos, e.g. he could ask for all the scenes in which a ball touches a line in a tennis court.

Our long term objectives are dedicated to the indirect exploitation of semantic multimedia annotation as presented in the second paragraph, which is an ongoing comprehensive and complex endeavor, providing a flexible infrastructure for further multimedia content analysis and reasoning, object recognition, metadata generation, indexing and retrieval. In the context of this paper, we only sketch the main steps of our approach.

The remainder of the paper is organized as follows: after briefly studying related work in section 2, we present in section 3 a description of the knowledge-assisted analysis process, which exploits the developed infrastructure and annotation framework. The annotation process needed for initializing the knowledge base with prototype instances of domain concepts in question, including a description of the actual implementation of the *M-OntoMat-Annotizer* tool are given in section 4. We conclude with a summary of our work and future directions in section 5.

## 2 Related Work

In the *multimedia analysis* area, knowledge about multimedia content domains, as for example reported in [9], is a promising approach by which higher level semantics can be incorporated into techniques that capture the semantics through automatic parsing of multimedia content.

Such techniques are turning to knowledge management approaches, including Semantic Web technologies to solve this problem [10]. In [11], semantic entities, in the context of the MPEG-7 standard, are used for knowledge-assisted video analysis and object detection, thus allowing for semantic level indexing. In [12] a framework for learning intermediate level visual descriptions of objects organized in an ontology is presented that aid the system to detect domain objects.

In [13], a-priori knowledge representation models are used as a knowledge base that assists semantic-based classification and clustering. MPEG-7 compliant low-level descriptors are automatically mapped to appropriate intermediate-level descriptors forming a simple vocabulary termed object ontology. Additionally, an object ontology is introduced to facilitate the mapping of low-level to high-level features and allow the definition of relationships between pieces of multimedia information. This ontology paradigm is coupled with a relevance feedback mechanism to allow for precision in retrieving the desired content.

Work in *semantic annotation* [14] has so far mainly focused on textual resources [4] or simple annotation of photographs [5, 6]. A presentation of an earlier version of *M-OntoMat-Annotizer* can be found in [15].

### 3 Knowledge Assisted Analysis

In order to handle the semantic gap in multimedia content interpretation, the aceMedia IST-FP6 Integrated project<sup>1</sup> proposed and implemented a comprehensive ontology infrastructure. An important part of this infrastructure is the Visual Descriptor Ontology (VDO), developed to link ontology concepts to low-level visual descriptors. It is based on MPEG-7 but modeled in RDFS, which allows for the direct integration with other RDF data used throughout the project. The descriptors are represented as so called *prototypes*, which are instances of the domain concepts linked to specific visual descriptors. The additional super-concept *Prototype* assures that prototypical instances can later be distinguished from the "real" metadata. By using the prototype approach to represent the visual features of concepts, we avoid direct linking of concepts to instances, and the ontologies are kept OWL DL compatible. Details about the aceMedia Knowledge Infrastructure and the VDO in particular can be found in [15].

We will shortly outline the analysis procedure for still images. Initially the image is segmented into a number of regions. For each region the MPEG-7 visual descriptors are extracted and then compared to the prototype instances stored in the active domain ontology. Using this approach, for each domain concept a distance to the descriptors of the region can be computed. This allows to decide which concept provides the best match for the specific region. Finally, the region is labeled with the concept providing the smallest distance. Apparently, the algorithm is domain independent, since it uses a generic distance computation which only relies on the visual descriptors. The concepts that can be detected, and especially the definition of the concepts, are completely defined in the ontologies and the extracted visual prototypes, so that switching the algorithm to another domain could be easily achieved by providing a different domain ontology and according prototypes.

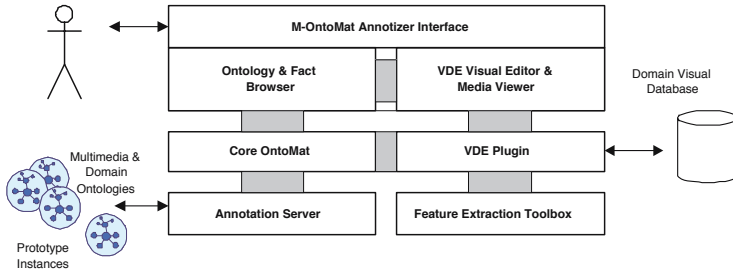
### 4 M-OntoMat-Annotizer

In order to exploit the ontology infrastructure mentioned above and enrich the domain ontologies with multimedia descriptors, we have developed the M-OntoMat-Annotizer (M stands for Multimedia), extending the CREAM (CREating Metadata for the Semantic Web) framework [4] and its reference implementation, OntoMat-Annotizer<sup>2</sup>, with the Visual Descriptor Extraction (VDE) tool, in order to allow low-level feature annotation. Figure 1 shows the integrated architecture the modules of which are explained in the following in more detail.

**Core OntoMat-Annotizer.** OntoMat-Annotizer supports two core applications: (i) it is used as an annotation tool for web pages and (ii) it acts as the basis of an ontology engineering environment. Also, by providing a flexible plugin interface it offers the possibility to implement new components and extend the core functionality of OntoMat-Annotizer.

<sup>1</sup> <http://www.acemedia.org>

<sup>2</sup> see <http://annotation.semanticweb.org/ontomat/>



**Fig. 1.** M-OntoMat-Annotizer and VDE plug-in design architecture

**Annotation Server.** The annotation server acts in the background and stores the entities of the knowledge base, maintains their mutual references and is responsible for maintaining the overall integrity of the stored entities.

**Domain Visual Database.** As easy content access is crucial for annotation and content analysis processes, a visual database containing content related to the domain examined and analyzed is always necessary.

**Feature Extraction Toolbox.** The actual extraction of the visual descriptors is performed using a feature extraction toolbox, namely the *aceToolbox*, a content pre-processing and feature extraction toolbox developed inside *aceMedia* project. The *aceToolbox* saves the extracted MPEG-7 Descriptors in XML format.

**VDE Visual Editor and Media Viewer.** The VDE Visual Editor and Media Viewer presents a graphical interface for loading and processing of visual content (images and videos), visual feature extraction and linking with domain ontology concepts. The interface, as shown in Figure 2, seamlessly integrates with the common OntoMat interfaces. Usually, the user needs to extract the features (multimedia descriptors) of a specific object inside the image/frame. For this reason, the VDE application lets the user draw a region of interest in the image/frame and apply the multimedia descriptors extraction procedure only to the specific selected region. Alternatively, M-OntoMat-Annotizer also supports automatic segmentation of the image/frame; whenever a new image/frame is loaded it is automatically segmented into regions. The user can then select a desired region or even merge two or more regions and proceed with the extraction. By selecting a specific concept in the OntoMat ontology browser and selecting a region of interest the user can extract and link concepts with appropriate prototype instances by means of the underlying functionalities of the VDE plugin.

**VDE Plug-in.** The *Visual Descriptor Extraction* (VDE) tool is implemented as a plug-in to OntoMat-Annotizer and is the core component for extending its capabilities and supporting the initialization of ontologies with low-level multimedia features. The VDE plugin manages the overall low-level feature extraction and linking process by communicating with the other components.



As seen above, these prototype instances are not only instances of the domain concept in question but are also stated to be instances of a separate Prototype concept. The created statements are added to the knowledge base and can be retrieved in a flexible way during analysis. The necessary conceptualizations can be seen as extensions to the VDO (*VDO-EXT ontology*) that link to the core ontology and are implemented in RDF:

```
<!-- Definition of concept "Prototype" as a subclass of Dolce's
"Physical-Object" -->
<rdfs:Class
  rdf:ID="http://www.acemedia.org/ontologies/VDO-EXT#Prototype">
  <rdfs:subClassOf rdf:resource=
    "http://ontology.ip.rm.cnr.it/ontologies/DOLCE-Lite#Physical-Object"/>
</rdfs:Class>

<!-- Definition of relation "hasDescriptor" having Dolce's
"Physical-Object" as domain and VDO's "VisualDescriptor" as range-->
<rdf:Property
  rdf:ID="http://www.acemedia.org/ontologies/VDO-EXT#hasDescriptor ">
  <rdfs:domain rdf:resource=
    "http://ontology.ip.rm.cnr.it/ontologies/DOLCE-Lite#Physical-Object"/>
  <rdfs:range
    rdf:resource="http://www.acemedia.org/ontologies/VDO#VisualDescriptor"/>
</rdf:Property>
```

All the prototype instances can be saved in a RDFS file. The VDE tool saves the domain concept prototype instances together with the corresponding transformed descriptors, *separately* from the ontology file, thus leaving the original domain ontology unmodified.

M-OntoMat-Annotizer is publicly available as free software through the aceMedia web site since May 2005<sup>3</sup>. An updated version of the tool is expected to be published during summer 2006.

## 5 Conclusions

In this paper we presented M-OntoMat-Annotizer, a tool for enriching domain ontologies with MPEG-7 visual descriptors expressed in RDF. We currently plan further extensions of the tool. One direction is the implementation of a high-level multimedia annotation tool based on M-OntoMat-Annotizer. Using the current plug-in, annotations could be made on a region level. Especially using the automatic segmentation capability of M-OntoMat-Annotizer, the detailed annotation would become less tedious. Furthermore, another direction is the extraction of spatial, topological and contextual knowledge from annotated content that can be used for multimedia reasoning and improve the automatic annotation significantly. Therefore, the tool can both be used by users to annotate their images for later retrieval or organization, but also as a means to generate a-priori knowledge useful for the knowledge-assisted analysis of multimedia content and multimedia reasoning.

---

<sup>3</sup> <http://www.acemedia.org/aceMedia/results/software/m-ontomat-annotizer.html>

## References

- [1] S.-F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6):688–695, June 2001.
- [2] O. Mich R. Brunelli and C.M. Modena. A survey on video indexing. *Journal of Visual Communications and Image Representation*, 10:78–112, 1999.
- [3] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12).
- [4] Siegfried Handschuh and Steffen Staab. Cream - creating metadata for the semantic web. *Computer Networks*, 42:579–598, AUG 2003. Elsevier.
- [5] J. Wielemaker A.Th. Schreiber, B. Dubbeldam and B.J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, May/June 2001.
- [6] L. Hollink, A.Th. Schreiber, J. Wielemaker, and B. Wielinga. Semantic annotation of image collections. In *Proceedings of the K-CAP 2003 Workshop on Knowledge Markup and Semantic Annotation, Florida*, 2003.
- [7] P. Wittenburg D. Thierry and H. Cunningham. The Automatic Generation of Formal Annotations in a Multimedia Indexing and Searching Environment. In *Proc. ACL/EACL Workshop on Human Language Technology and Knowledge Management*, Toulouse, France, 2001.
- [8] T. Athanasiadis, V. Tzouvaras, K. Petridis, F. Precioso, Y. Avrithis and I. Kompatsiaris. Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content. In *Proceedings of the 5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot 2005) at the 4th International Semantic Web Conference (ISWC 2005)*, Nov. 2005.
- [9] J. Hunter, J. Drennan, and S. Little. Realizing the hydrogen economy through semantic web technologies. *IEEE Intelligent Systems Journal - Special Issue on eScience*, 19:40–47, 2004.
- [10] A. Yoshitaka, S. Kishida, M. Hirakawa, and T. Ichikawa. Knowledge-assisted content-based retrieval for multimedia databases. *IEEE Multimedia*, 1(4):12–21, Winter 1994.
- [11] R. Tansley, C. Bird, W. Hall, P. Lewis, and M. Weal. Automating the linking of content and concept. In *Proc. ACM Int. Multimedia Conf. and Exhibition (ACM MM-2000)*, Oct./Nov. 2000.
- [12] Nicolas Maillot, Monique Thonnat, and Céline Hudelot. Ontology based object learning and recognition: Application to image retrieval. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004), 15-17 November 2004, Boca Raton, FL, USA*, pages 620–625, 2004.
- [13] I. Kompatsiaris, V. Mezaris, and M. G. Strintzis. *Multimedia content indexing and retrieval using an object ontology*. Multimedia Content and Semantic Web - Methods, Standards and Tools, Editor G.Stamou, Wiley, New York, NY, 2004.
- [14] Siegfried Handschuh and Steffen Staab, editors. *Annotation for the Semantic Web*. IOS Press, 2003.
- [15] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, I. Kompatsiaris, S. Staab, and M.G. Strintzis. Semantic Annotation of Images and Videos for Multimedia Analysis. In *Proceedings of the 2nd European Semantic Web Conference (ESWC 2005)*, May 2005.

# On the Profitability of Scalping Strategies Based on Neural Networks

Marina Resta

University of Genova, DIEM sez. di Matematica Finanziaria,  
via Vivaldi 5,16126 Genova, Italy  
`resta@economia.unige.it`

**Abstract.** We analyze the potential of unsupervised neural networks when they are employed to support intraday trading activity on financial markets. Several time frequencies have been considered: from five minutes to daily trades. At the current stage our major findings may be summarized as follows: a) unsupervised neural networks are helpful to localize profitable intraday patterns, and they make possible to achieve higher performances than common trading rules; b) trading strategies based on neural networks make exploitable with profits almost continuous trades (i.e. scalping), until transaction costs maintain below proper thresholds.

## 1 Outline of the Work

Starting from early 90's a plenty of academic work has been spent to analyze the potential of neural networks (NNs since now on) to trade financial markets. During those years almost every aspect of financial markets has been investigated by means of artificial neural networks: we have hence learned that some neural architectures seem to work better than others [7], and that NNs make possible to pick up trading opportunities by monitoring market volatility [4], or that some hybridisation with genetic algorithms can help to achieve better results [1]; additionally, NNs have been combined with almost every kind of more traditional tools, like moving averages [8].

It is then clear that at the present state, NNs applications in financial markets appear to have reached a fully developed stage, with smaller space for new contributions to the ongoing debate. Having this in mind, we have focused on a relatively lesser known topic, that is the profitability of scalping activity when this is driven by neural networks.

According to the definition given in [5], scalping consists in a set of more than one trade within a daily stock exchange session, so to take advantage of almost smallest markets movements. It is generally agreed that scalping activity is potentially profitable, however it should be carefully managed almost on very liquid markets, since it can increase traders exposure. Starting from this point, we are going to build a scalping trading system whose signals are generated through a set of unsupervised neural networks, and we will explore its profitability at various intraday time frequencies, and at different levels of transaction costs. Such

results will be then compared to those obtained at the same conditions in terms of both time frequencies and transaction costs, but without any neural network aid. What remains of the paper will be then organized as follows: Section 2 will give details about the neural trading system that we have implemented; after a brief description of the data employed, Section 3 will provide and discuss the results obtained; finally Section 4 will end the paper, giving some conclusions and outlooks on future activity.

## 2 Methodology

We have considered a set of 10 mono-dimensional wraparound Self Organizing Maps [6], each made up by 50 neurons.

After the random initialization, the nets have been trained over the input space. Such input space basically consists of the set of Open, High, Low, Close (OHLC) returns bars at various time frequencies: depending on the dataset in use, we have then managed OHLC returns bars at 5 minutes, 30 minutes and 1 day. Due to the high frequency nature of data, we have generally considered huge samples: to our purpose we have always used the earlier 70% to train the nets, while the remaining 30% has been shared out between validation (20%) and test set (latest 10% of the data sample).

The training stage was carried out for 125 epochs over the input set: the number of epochs has been chosen as the number at which the convergence indexes described in [3] and in [2] exceed the threshold of 60%.

Our trading system is based on a simple sequence of tasks:

- (a) Run the classical Kohonen maps learning stage on the training set, as previously described.
- (b) Classify OHLC patterns in the validation set by means of nets trained according to (a). In this way, we associate each OHLC pattern of the validation set ( $OHLC_V$ ) to the corresponding pattern  $OHLC_N$  in the neural space.
- (c) Cluster  $OHLC_N$  patterns associated through (b) to samples in the validation space into three groups of signals: Strong Sell (-1), Strong Buy (1), and Standby (0), and simulate to trade the market accordingly.
- (d) Evaluate the nets performances over the validation sets by means of the following financial criterion:

$$perf_{val} = \left\{ \prod_{k=2}^{Val_{dim}} [1 + \text{sign}(r_k \times \text{signal}) \times r_k] \right\} - 1 \quad (1)$$

where:  $Val_{dim}$  is the length of the validation set, and  $r_k$ , ( $k = 2, \dots, Val_{dim}$ ) are the close returns in each  $OHLC_V$  patterns.

The value returned by  $\text{sign}(r_k \times \text{signal})$  depends on the way  $r_k$  and  $\text{signal}$  reciprocally behave: whether they will be both positive and negative, the final exitus will be graded favourably; on the other hand, when their signs disagree, they will bring the final result of  $\text{sign}(r_k \times \text{signal})$  to unfavourable.



From a financial point of view, (1) considers the overall performance of each unit of available wealth (for instance: 1 \$), when such wealth is fully re-invested after each trade.

- (e) Choose the best performing nets according to (1).
- (f) Run steps (b)–(c) on the test set, using only the nets selected in (e).

The performances on the test data will be then evaluated according to a financial criterion similar to that suggested in (1), but comprehensive of transaction costs  $tc$ :

$$perf_{test} = \left\{ \prod_{k=2}^{Test_{dim}} [1 + sign(r_k \times signal) \times r_k - tc] \right\} - 1 \quad (2)$$

where  $Test_{dim}$  is the length of the test dataset. The variable  $tc$  is of particular relevance for scalpers, since the profitability is sensitively conditioned by transaction costs, as the number of trades increases. To such aim, we have assumed  $tc \in [0, 0.01\%]$ , and we have monitored how the overall performances change with  $tc$  varying in the given range.

### 3 Discussion of the Results

We report the results obtained with various data sets (whose main features are given in Table 1), at various time scales, and with different levels of transaction costs.

**Table 1.** The datasets employed in our work, with the related starting and ending days for training, validation and test sets

ID	Data Type	Starting bar (Training)	Ending bar (Training)	Starting bar (Validation)	Ending bar (Validation)	Starting bar (Test)	Ending bar (Test)
SPMib	5 min	09/17/2004	10/14/2005	10/17/2005	02/02/2006	02/03/2006	03/31/2006
Nasdaq1	5 min	04/30/2002	07/02/2003	07/03/2003	11/07/2003	11/08/2003	01/07/2004
Nasdaq2	30 min	04/10/1996	03/16/2001	03/19/2001	08/15/2002	08/16/2002	04/30/2003
Nasdaq3	Daily	02/05/1971	09/15/1995	09/18/1995	10/04/2002	10/07/2002	04/13/2006
Dax1	5 min	04/10/1996	03/16/2001	03/19/2001	08/15/2002	08/16/2002	04/30/2003
Dax2	Daily	11/26/1990	09/18/2001	09/19/2001	10/08/2004	10/09/2004	04/13/2006

The datasets in use are the financial indexes of various countries (namely: Italy, USA and Germany). During the period of observation, such indexes registered different performances: SPMib, Dax1 (5 minutes bars), Nasdaq3 and Dax1 (daily bars) for instance, exhibited positive trends, while Nasdaq1 (5 minutes bars) was characterized by strong corrections to the running positive trend, and Nasdaq2 (30 minutes bars) has shown substantially negative trend.

The choice of such different datasets, enabled us to test the capability of the system to work at various conditions over the markets. Due to the definition of scalping given in Section 1, daily records are assumed as border frequencies.

Table 2 reports the results of the neural trading system on the validation sets: for each data sample we have evidenced the nets best, worst and mean performance, and the related standard deviation. Finally, the last column shows the number of nets selected (according to their overall performance) to forecast the returns close bar of the patterns in each test set, following the procedure already described in the previous section. The results for the test sets are reported in Tables 3–6.

**Table 2.** The performances on the validation set

Name	Type	Min	Max	Mean	SD	Nr. Nets
SPMib	5 min	4.011%	75.991%	37.441%	31.850%	5
Nasdaq1	5 min	6.927%	17.436%	9.724%	3.079 %	10
Nasdaq2	30 min	0.213%	0.712%	0.453%	0.184 %	5
Nasdaq3	Daily	1.410%	10.033%	6.051%	4.35%	5
Dax1	5 min	0.009%	0.037 %	0.032%	2.748 %	6
Dax2	Daily	0.522%	3.412 %	1.905%	1.17 3%	5

**Table 3.** The performances on the test set: the case of SPMib, Nasdaq and Dax 5 minutes futures index

	SPMib		Nasdaq1		Dax1	
	tc=0 %	tc = 0.01%	tc = 0%	tc = 0.01%	tc = 0%	tc = 0.01%
min	16.29%	-28.79%	1.02%	-0.35%	0.62%	-7.65%
max	111.08%	29.25%	4.51%	2.06%	6.83%	2.67%
mean	59.74%	0.19%	3.29%	0.03%	3.22%	-0.40%
ctr	-37.10%	-61.49%	-35.34%	-40.75%	-17.68%	-34.01%

**Table 4.** The performances on the test set: the case of Nasdaq 30 minutes futures index

	tc = 0%	tc = 0.01%	tc = 0.02%	tc = 0.03%
min	6.48%	-14.63%	-31.56%	-45.13%
max	138.27%	91.04%	53.16%	22.80%
mean	63.16%	34.45%	11.44%	-7.01%
ctr	10.30%	10.29%	10.28%	10.27%

**Table 5.** The performances on the test set: the case of Nasdaq daily futures index

	tc = 0%	tc = 0.01%	tc = 0.02%	tc = 0.03%
min	0.70%	-3.12%	-5.80%	-10.34%
max	14.42%	10.08%	5.90%	1.88%
mean	6.20%	2.18%	1.71%	-5.44%
ctr	14.42%	10.08%	5.90%	1.88%

**Table 6.** The performances on the test set: the case of Dax daily futures index

	$tc = 0\%$	$tc = 0.01\%$	$tc = 0.02\%$	$tc = 0.03\%$
min	0.70%	2.64%	0.18%	-3.62%
max	11.35%	7.13%	3.06%	2.85%
mean	6.20%	4.16%	1.62%	0.23%
ctr	18.50%	14.00%	9.67%	5.51%

Tables 3–6 have to be interpreted as follows: like in the case of the validation set, we have evidenced the nets best, worst and mean performance. The last row is labelled *ctr* (the acronym for common trading rule) and it shows the results obtained by running the same trading rules without the aid of neural networks. For what concerning transaction costs, although  $tc$  was varying in the range  $[0, 0.01\%]$ , each table shows the results up to the maximum sustainable cost that guarantees favourable performances of the neural trading system.

With a look to the results, a number of remarks can be drawn out:

- in all the observed cases the neural system has got favourable results with zero transaction costs; however, in the cases of daily datasets, common trading rules have worked better. This result for daily data is also confirmed with non-zero transaction costs, and it can be supported by almost two different explanations: firstly, the rules thought to work at intraday frequencies, probably need to be better tuned to run on daily quotes; as second remark, it is interesting to observe that the nets trained on daily data exhibited convergence values closer to the lower bound (60%) assumed to stop the training procedure. This latter observation let us to think that the system could improve the overall daily performances, once trained for a greater number of epochs.
- assuming non-zero transaction costs leads us to a variety of results. SPMib, Nasdaq1 and Dax1 (5 minutes bars) remain profitable below  $tc = 0.02\%$ . It is interesting to observe that the corresponding operations without any neural aid get very unfavourable results (last row of Table 3). Trades on Nasdaq2 (30 minutes bars) are profitable below  $tc = 0.03\%$ ; finally, as previously stated, daily trades are satisfactory both on Nasdaq and Dax, but with lower performances than in the case of common trading rules.

## 4 Conclusions

We have built a trading system based on unsupervised neural networks to operate on financial markets at intraday trading frequencies. To such aim, we have considered futures indexes from different markets, and at various time scales (from 5 minutes to one day): 70% of the data sample has been employed to train the system, while the remaining 30% has been divided into validation set (to choose the best performing nets), and test set, where we have operated one bar in advance. The performances of the system are comprehensive of transaction

costs, varying in the range  $[0, 0.01\%]$ . The results obtained are quite promising, since they give evidence that unsupervised neural networks work well when they are called to discover intraday patterns. However, some problems still remain open. The first odd (of “technical” nature) concerns the way to decide when to stop the training procedure: we have assumed to monitor the level reached by some convergence indexes, but clearly other stopping solutions can be equally promising. The second problem is linked to the evaluation of how much the training and the validation sets are representative of the whole data sample (and of the test set in particular): it is evident that strong divergences among such blocks of data can seriously compromise the overall performance of the system. A possible solution (presently not explored in deepest detail) could be that to consider different neural systems tuned on various aspects of the market, with the possibility of switches from one to another according to the fluctuations in the market itself.

## References

1. Arifovic, J., and R. Genay (2001), *Using genetic algorithms to select architecture of a feedforward artificial neural network*, Physica A: Statistical Mechanics and its Applications, 289(3-4), 574-594.
2. Cattaneo Adorno, M. and M. Resta (2004), *Reliability and convergence on Kohonen maps: an empirical study*, in M. Gh. Negoita, R. J. Howlett and L. C. Jain, Proceedings of Knowledge-Based Intelligent Information and Engineering Systems, 8th International Conference, KES 2004, Wellington, New Zealand, September 20-25, Part I, Lecture Notes in Computer Science, Springer, 426-433.
3. De Bodt, E., Cottrell, M., and M. Verleysen (2002), *Statistical tools to assess the reliability of self-organizing maps*, Neural Networks, 15, 967-978.
4. Deboeck, G. J. (1999), *Modeling non-linear market dynamics for intra-day trading*, available on-line at: <http://www.dokus.com/PapersontheWeb/intradaymodel.htm>.
5. Graifer, V. (2005), *How to Scalp Any Market*, Reality Trading.
6. Kohonen T. (1997), *Self-Organizing Maps*, Springer Verlag, 2nd edition.
7. Refenes, A. P. N., Burgess, A. N., and Y. Bentz (1997), *Neural networks in financial engineering: A study in methodology*, IEEE Transactions on Neural Networks, 8(6), 1222-1267.
8. Refenes, A. P. N., Azema-Barac, M., Chen, J., and S. A. Karoussos (1993), *Currency exchange rate prediction and neural network design strategies*, Neural Computing & Applications, 1(1), 46-58.

# Effect of Moving Averages in the Tickwise Tradings in the Stock Market

Felix Streichert, Mieko Tanaka-Yamawaki, and Masayuki Iwata

Dept. of Information and Knowledge Engineering,  
Faculty of Engineering, Tottori University,  
101-4, Koyamacho-Minami, Tottori, 680-8552, Japan,  
`streichert@ike.tottori-u.ac.jp`

**Abstract.** In the recent years the automatic generation of trading rules for stock and currency markets by means of Evolutionary Algorithms has become a popular game. Although, it is disputed whether or not such evolved trading rules are able to generate reliable profit on out-of-sample sets, especially if trading costs are considered. In this paper we focus on tickwise data and introduce a simple trading scheme based on Learning Classifier like action rules. These rules have only access to the most recent time series history and are thus only able to exploit the short term memory effects of tickwise data. Rather than searching for profitable trading rules on tickwise data, we first concentrate on evaluating the predictive properties of alternative indices, namely moving averages.

**Keywords:** Trading Rules, Moving Averages, Evolutionary Algorithms.

## 1 Introduction

Ad hoc specification of trading rules and also the introduction of novel indices may suffer from human bias and selective thinking. While trading rules generated by means of Evolutionary Algorithms (EAs) do not suffer from such bias, they are unlikely to be generalizing and it is difficult to predict their performance on out-of-sample data. However, trading rules generated by EAs can be used to asses the predictive properties of indices for stock trading.

This paper introduces a simple scheme for trading rules which can be optimized by means of EAs. Since the performance of optimized rules depend on the indices allowed to be used as input, this scheme allows us to judge the predictive power of alternative index sets. This scheme is further extended by using an Meta-EA to select the the most predictive indices into the input set. The Meta-EA reliably selects the most predictive indices from a set of alternative indcies without introducing bias. This allows us to judge the predictive power of alternative definitions of moving averages on tickwise stock market data.

The next section will give a brief overview over the related work on evolving trading rules. Then Sec. 3 outlines the framework for the trading scheme used in this study. The experimental results regarding the properties of the proposed trading scheme and the predictive properties of alternative indices are given in Sec. 4. Finally, Sec. 5 gives the conclusions and an outlook on future work.

## 2 Related Work

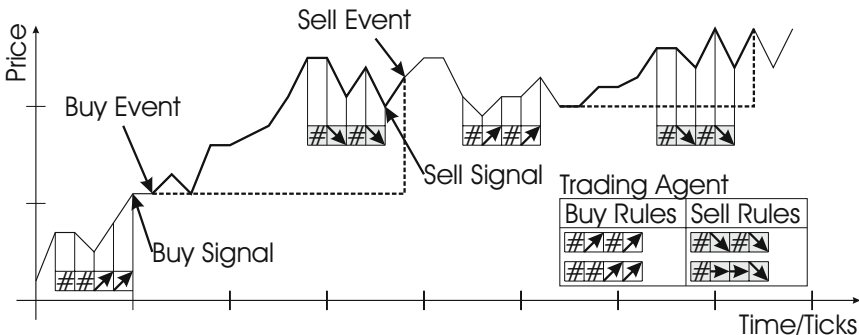
There are two alternative approaches to evolve trading rules for stock markets by means of EAs. One is based on Genetic Programming (GP) [4] and the other one is based on Learning Classifier Systems (LCSs) [3,12]. GP allows a most general approach to define trading rules, because GP is able to optimize most versatile program trees. Still, the initial results using GP hardly equaled the buy&hold strategy on the out-of-sample sets [1]. However, more recent publications are much more promising in this respect [7,8]. LCSs, on the other hand, use a fixed rule structure using multiple AND connected conditions per rule and allowing multiple OR connected rules. These are not as flexible as GP-based trading rules, but they are much easier to understand. Schulenberg *et al.* [9] and Lin *et al.* [6] applied the standard LCS to day trading in stock markets. An extended version of the LCSs was applied by Liao *et al.* [5] and Chen *et al.* [2] on stock markets. All authors using LCS typically stress the continuous adaption process of LCSs and don't use extra out-of-sample set. Thus, it is difficult to compare the LCS approach to the GP approach.

In this paper we don't claim to generate profitable trading rules, but we just want to judge alternative indices. Therefore, we can omit the out-of-sample test without reducing the significance of the results.

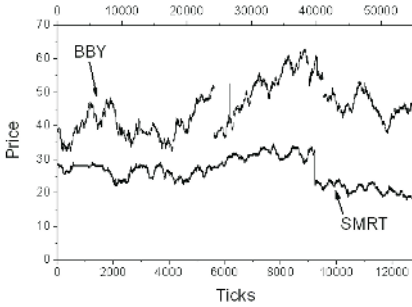
## 3 Experimental Settings

To evaluate the predictive properties of alternative indices we use a simplified version of the LCSs approach. While sticking with the static rule scheme based on multiple AND connected rules each consisting of multiple OR connected conditions, we remove the additional learning procedure of LCSs. Here, we treat the whole rule set as an virtual trading agent. The agent is evaluated on a given data set and it's profit is to be maximized using conventional EAs.

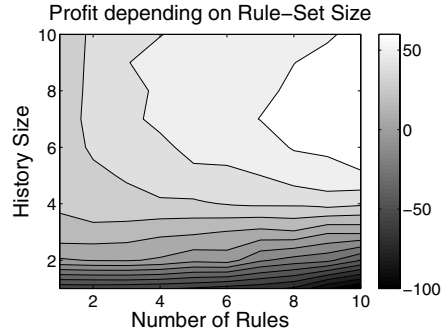
An exemplary trading agent is given in Fig. 1. It consists of two rule blocks: One rule block triggers the buy event, while the other triggers the sell event.



**Fig. 1.** Exemplary trading agent using two sets of two rules for buying and selling, a history size of four and distinguishing UP/DOWN/NO/DON'T CARE trends



**Fig. 2.** The data sets used in these experiments: SMRT and BBY



**Fig. 3.** Effect of history and rule size on the proposed trading scheme

Only one rule block is active at a time, thus reducing the number of states necessary to describe the state of the virtual trading agent.

A condition within a rule evaluates the relative value of an input whether it is larger, equal or smaller than zero and returns true if fulfilled. A DON'T CARE (#) element causes a condition always to return true regardless of the input. Since a condition requires relative inputs or trends, we use differences between the current index value and the previous index values as inputs. To give the trading agent access to the index history, the agent can evaluate a number of most recent index values depending on the allowed history size.

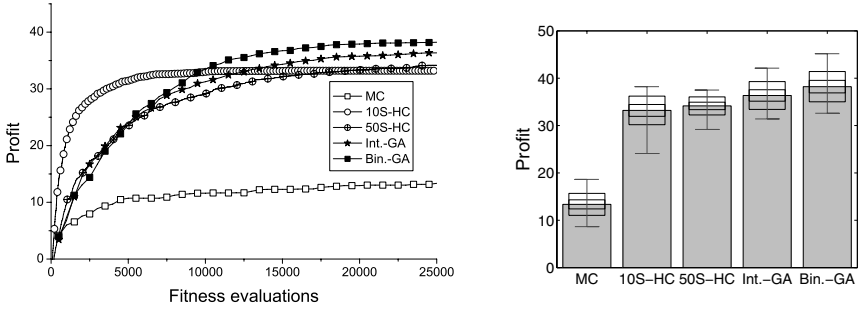
In the example given in Fig. 1 the agent is allowed two rules per block and a history size of four. After five ticks the second buy rule triggers a buy event. Allowing a trading lag of one the agent buys at the current price, subtracting trading cost of 0.1%, and is then waiting for a sell rule to trigger, while the buy rules are deactivated. The accumulated price differences between buy and sell prices gives the resulting profit in arbitrary units, which is then to be maximized.

## 4 Results

The following experiments have been performed on tickwise data of the year 1993 namely the Stein Mart (SMRT) and Best Buy Co. (BBY) data sets, see Fig. 2. To give statistically sound results for the optimized trading rules we have performed 25 independent runs over 25,000 fitness evaluations for each algorithm. Additional to the fitness plots, giving the averaged maximum profit over the required effort, we also give the final maximum profit averaged over the 25 multi-runs, the standard deviation, the maximum and minimum values and the 95%-confidence interval of the mean in separate plots.

### 4.1 History Size and the Number of Rules

To evaluate the general performance of the proposed trading scheme we optimize the trading rules for different values for the history size  $h$  and the number of



**Fig. 4.** Alternative optimization algorithms on the SMRT data set

rules per rule set  $r$ . For this proof of concept we use only price differences as inputs. We plot the best profit averaged over the 10 multi-runs on the SMRT data set over  $h$  and  $r$  in Fig. 3. As reference: an optimized golden/dead cross rule returns 22.5 units as profit on the SMRT data set.

Fig. 3 shows that the trading rules require at least a history size of four to generate profitable trading rules. Also the performance increases significantly with the number of rules per block. But since the number of variables increases with  $2 \cdot h \cdot r$  and the size of the search space increases proportional to  $4^{2hr}$ , we decided to use  $h = 5$  and  $r = 5$  for the following experiments.

### 4.2 Optimization Algorithms

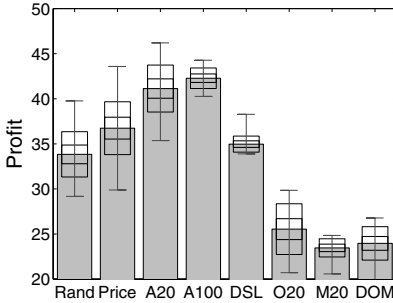
To evaluate alternative optimization algorithms we compare random search (MC), multi-start hill-climbing (MS-HC) and Genetic Algorithms (GAs) as implemented in JavaEVA [10]. The GAs use a population of 500 individuals and perform tournament selection with a tournament group size of four. We apply a GA using a discrete integer representation applying random integer mutation which randomly reinitializes a single integer variable with a probability of  $p_m = 0.1$  and three-point crossover with a probability of  $p_c = 0.7$  (Int.-GA). We compare it to a GA using a binary representation using one-bit mutation ( $p_m = 0.1$ ) and three-point crossover ( $p_c = 0.7$ ) (Bin.-GA).

Fig. 4 shows that the random search performs worst indicating that the optimization problem is non-trivial but likely to be causal. The MS-HC, on the other hand, converges very fast but prematurely, if only ten multi-starts are allowed. With increasing multi-starts the MS-HC converges much slower but also more reliable. In the end, the GA approaches with crossover outperform the local search strategies and the binary GA is slightly better than the integer GA. Therefore, we decided to use the binary GA for the following experiments.

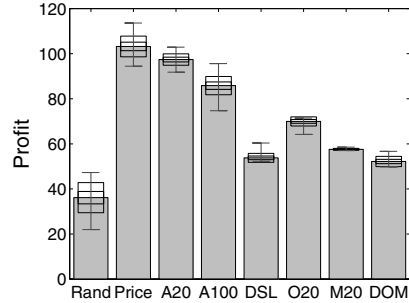
### 4.3 Using the Proposed Scheme to Evaluate Alternative Indices

Now we compare the performance of alternative indices: price difference, short and a long moving averages of the price with length 20 and 100 ( $A_{20}$  and  $A_{100}$ )





**Fig. 5.** The effect of the choice of indices on the SMRT data set



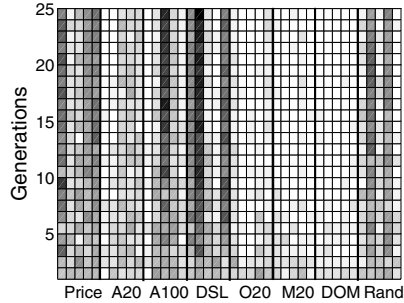
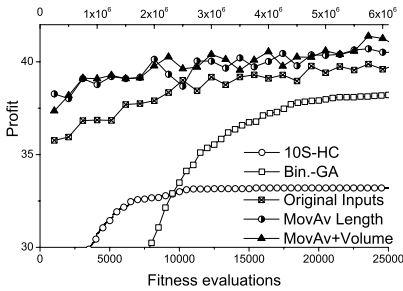
**Fig. 6.** The effect of the choice of indices on the BBY data set

and difference between the short and the long moving average ( $\Delta SL(t) = A_{20}(t) - A_{100}(t)$ ) (this allows the EA to evolve the golden/dead cross rule, if profitable). We also compute the optimal moving average  $O_{20}$  and the meta moving average  $M_{20}$  each with a length of 20. Additionally, we also evaluate the difference between the optimal and meta moving average ( $\Delta OM(t) = M_{20}(t) - O_{20}(t)$ ) [11]. As reference we also evaluate the performance of a random input ( $RAND$ ).

Since on the smaller SMRT data set ( $\approx 12,500$  data points) the random inputs give unique labels to the data, a trading rule memorizing these label can achieve rather good results, see Fig. 5. Because rules based on random inputs are obviously not generalizing, the achieved profit gives an estimate on the performance of non-predictive indices. Any indices outperforming the random inputs must be predictive.

Since the price index outperform the random inputs and the confidence intervals, inner thick black boxes in Fig. 5, do not overlap, the price must have some predictive properties. The same holds true for both moving averages, while the difference between the short and long moving average is not able to outperform the random inputs. Unfortunately, neither optimal nor meta moving averages and also their difference is able to outperform the random inputs. On closer inspection of the resulting trading rules these indices failed, because the trading rules were unable to predict the break down of the SMRT at tick 9100, see Fig. 2, and thus resulting in a loss of about 10 units in the resulting profit.

On the BBY data set ( $\approx 55,000$  data points) the random inputs perform not as well, because they don't give a unique labeling anymore. Thus the gap between the random inputs and the price differences and the moving averages on the price increases. Again the price differences,  $A_{20}$  and  $A_{100}$  perform best and seem to be most predictive, while again the optimal and meta moving average and the differences of moving averages do not perform well. However, on BBY these indices outperform the random input and are thus likely to have predictive power. In contrast to the SMRT data the BBY data seems to prefer signals with a shorter time scale. This can be accounted to the size of the BBY data set, because it limits the chance for overfitting.



**Fig. 7.** Comparing three Meta-EAs index sets to the GA and 10S-HC on SMRT

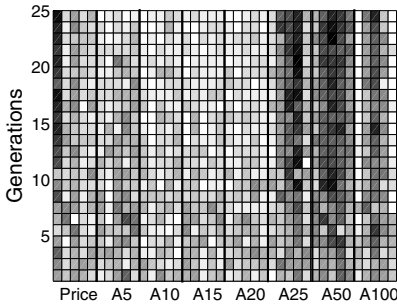
**Fig. 8.** Frequency of indices each with a history size of 5 (black = high freq.)

### 4.4 Meta Optimization on the Choice of Indices

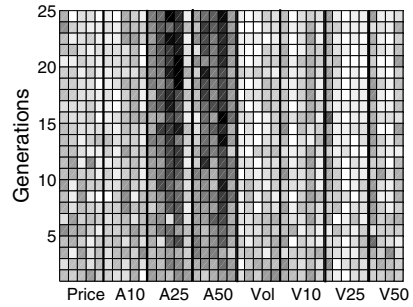
In the previous section we have investigated a static scheme for selecting and evaluating the predictive properties of alternative indices. In this section we use a Meta-EA decide which indices to use to generate most successful trading rules. Having eight alternative indices, each with a history size of five, results in forty alternative indices. Whether or not to use one of these indices for a trading rule can be coded in a 40-bit string. The quality of the selected index set can be tested by running the previous optimization procedure for this index set and using the profit of the best trading rule as fitness for the 40-bit string coding the index set. This two-level optimization procedure results in a Meta-EA. The top-level optimization procedures optimize the choice of indices, while a low-level optimization procedure is necessary to evaluate an index set by optimizing trading rules based on this set. Since the low-level optimization procedure is called several thousand times, we decided to use a 5-start hill-climber with 500 low-level fitness evaluations, because of it's fast rate of convergence. For the top-level optimizer we used the previously introduced binary GA with a population size of 100 and 2.500 top-level fitness evaluations.

Before comparing the results of the index selection to the previously obtained results it is important to note that the required fitness evaluations of the meta EA are significantly higher and are given on top of the graph in Fig. 7. Additionally, the curve of the meta EA is not monotonically increasing, since the fitness of a solution depends on the random low-level optimization process.

Still, the Meta-EA outperforms both the 10S-HC and the GA with static index sets, although based on a much simpler 5S-HC. Fig. 8 gives the frequency of the selected inputs averaged over 25 independent runs of the meta EA. In Fig. 8 dark color indicates a high probability of being selected in the best index sets, while bright color indicates a low probability and thus low predictive properties of the associated index. Unexpectedly, the Meta-EA combines several indices to generate most profitable trading rules. Successful combination typically include one price difference, a long moving average value and also the difference between the short and long moving average. This is astonishing, because the delta of



**Fig. 9.** Frequency of selected price moving averages (black = high freq.)



**Fig. 10.** Frequency of selected price and volume indices (black = high freq.)

moving averages did not perform well in Sec. 4.3. This indicates that the mix of different sources of information yields the best results. It is also interesting to note that although the meta EA could decide to use all input variables, it typically settles for using only four to five inputs. Unfortunately, due to the small size of the SMRT data set the random input is also selected rather frequently.

When comparing the choice of moving averages more closely, Fig. 9 shows that the proposed trading scheme prefers both the most recent price differences and larger moving averages  $A_{25}$ ,  $A_{50}$  and also with decreased frequency  $A_{100}$ .

In an additional experiment we added the trading volume (Vol) and also several moving averages of the trading volume ( $V_{10}$ ,  $V_{25}$ ,  $V_{50}$ ) to the index set and again used the Meta-EA to choose the most suitable indices. Fig. 7 shows that this index set performed even better than the previous two index sets. But in contrast to the previous experiments, the recent price fluctuations are less important there, compare Fig.9 to Fig.10. Instead, the moving averages on the price are combined with the volume or the moving average of the volume to generate successful trading rules. Unfortunately, the selection of the volume indices is not as specific such that no clear dark row occur in Fig.10. However, the best index set include 1.8 volume indices on average, even if the choice of the type of moving average on the volume is not as critical. This indicates that the volume has at least some predictive properties on tickwise trading data.

## 5 Conclusions and Outlook

We have presented a simple approach to evaluate alternative moving average indices for tickwise stock trading based on EAs. This approach yields the advantage that it allows a virtually unbiased approach to evaluate alternative or novel indices, since this approach is solely based on a rational selection criterion, the achieved profit. This approach has been checked against random rules and also against random inputs and the results show that the evolutionary approach is truly able to exploit predictive properties of the indices used. When comparing the predictive power of the individual indices we were able to show that moving averages have predictive power on tickwise data. Also the volume information

seems to yield predictive power, but here the choice of the time window seems to be not as critical as for the price indices. On the other hand, the optimal and meta moving averages seemed not to be as useful.

The proposed approach proved to be successful in identifying predictive indices. Using the Meta-EA approach we were even able to identify index ensembles which perform better than the isolated index sets. This or similar approach should be used on a regular basis to evaluate the predictive power of novel indices to remove human bias due to ad hoc index selection.

**Acknowledgments:** This research has been funded by the Canon Foundation in Europe.

## References

1. F. Allen and R. Karjalainen. Using genetic algorithms to find technical trading rules. *Journal of Finance and Economics*, 51:245–271, 1999.
2. A.-P. Chen, Y.-C. Chen, and W.-C. Tseng. Applying extending classifier system to develop an option-operation suggestion model of intraday trading - an example of Taiwan index option. In R. Khosla, R. J. Howlett, and L. C. Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 3681 of *LNCS*, pages 27–33, 2005.
3. J. Holland. *Adaption in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Systems*. The University Press of Michigan, Ann Arbor, 1975.
4. J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. 1992.
5. P. Y. Liao and J. Chen. Dynamic trading strategy learning model using learning classifier systems. pages 783–789, 2001.
6. J.-Y. Lin, C.-P. Cheng, W.-C. Tsai, and A.-P. Chen. Using learning classifier system for making investment strategies based on institutional analysis. In M. Hamza, editor, *Artificial Intelligence and Applications*, pages 765–769. ACTA Press, 2004.
7. C. J. Neely, P. A. Weller, and R. Dittmar. Is technical analysis in the foreign exchange market profitable? a genetic programming approach. *Journal of Financial and Quantitative Analysis*, 32(4):405–426, 1997.
8. J.-Y. Potvin, P. Soriano, and M. Vallee. Generating trading rules on the stock markets with genetic programming. *Computers and Operations Research*, 31:1033–1047, 2004.
9. S. Schulenburg and P. Ross. Explorations in LCS models of stock trading. In P. L. Lanzi, W. Stolzmann, and S. W. Wilson, editors, *Advances in Learning Classifier Systems*, volume 2321 of *LNAI*, pages 150–179, 2002.
10. F. Streichert and H. Ulmer. JavaEvA - a java framework for evolutionary algorithms. Technical Report WSI-2005-06, Centre for Bioinformatics Tübingen, University of Tübingen, 2005.
11. M. Takayasu, T. Mizuno, and H. Takayasu. Potentials of unbalanced complex kinetics observed in market time series, 2005.
12. S. Wilson. Classifier fitness based on accuracy. *Evolutionary Computation*, 3:149–175, 1995.

# A New Scheme for Interactive Multi-criteria Decision Making

Felix Streichert and Mieko Tanaka-Yamawaki

Dept. of Information and Knowledge Engineering,  
Faculty of Engineering, Tottori University,  
101-4, Koyamacho-Minami, Tottori, 680-8552, Japan  
streichert@ike.tottori-u.ac.jp

**Abstract.** Multi-objective optimization, also known as multi-criteria decision making in the field of operations research, is a common task in many financial engineering problems. Several alternative approaches to multi-objective optimization have been proposed in operations research. Depending on when the so-called decision maker introduces his preferences, three approaches to multi-criteria decision making can be distinguished: *a priori* decision making, interactive decision making, and finally *a posteriori* decision making. This paper suggests a new interactive multi-criteria decision making scheme which combines these three approaches in a single multi-objective optimization framework. In contrast to most operations research approaches, the new scheme is based on evolutionary algorithms due to their flexibility regarding the type of objectives and constraints. This way the new scheme allows *de novo* programming, which enables the decision maker to refine the problem definition and to reduce the size of the objective space iteratively.

**Keywords:** Multi-Objective Optimization, Multi-Criteria Decision Making.

## 1 Introduction

Many real-world optimization problems in financial engineering and operations research have not only one objective to optimize ( $\min(f(x))$ ), but multiple, often conflicting, objectives ( $\min(f_1(x)), \min(f_2(x)), \dots, \min(f_m(x))$ ). In contrast to conventional single-objective optimization problems (SOOPs), the multi-objective optimization problems (MOOPs) do not yield a single optimal solution  $x_{opt}$ , but yield a whole set of so-called Pareto optimal solutions  $x_{Pareto} \in \mathcal{PF}$ . A solution  $x_{Pareto}$  is Pareto optimal, if there exists no other solution  $x$  such that  $f_i(x) \leq f_i(x_{Pareto})$  for all  $i = 1, 2, \dots, m$  and  $f_j(x) < f_j(x_{Pareto})$  for at least one index  $j$  [8]. A solution  $x_{Pareto}$  is then also known as non-dominated.

The resulting MOOP is then a two-fold problem: On the one hand, how to identify non-dominated solutions. This gives an optimization problem which can be solved by means of optimization algorithms. And on the other hand, how to select the final solution alternative from a set of non-dominated solutions. The

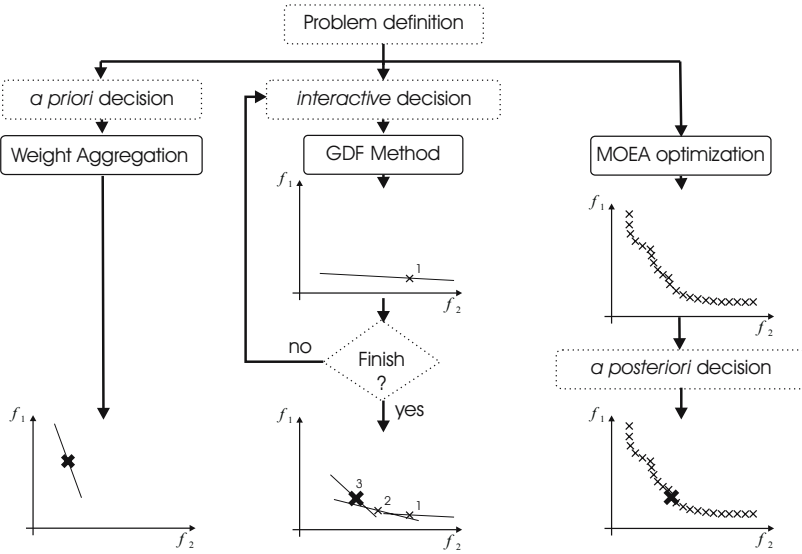


Fig. 1. The traditional *a priori*, interactive and *a posteriori* approaches to MOOPs

last problem can not be addressed without additional preference information from a so-called decision maker. Depending on when this preference information is introduced to the optimization process, three alternative approaches to multi-objective decision making can be distinguished: *a priori*, interactive and *a posteriori* decision making, see Fig. 1.

Based on evolutionary algorithms (EAs) and multi-objective evolutionary algorithms (MOEAs), this paper introduces a new iterative, interactive multi-objective optimization scheme, which combines the three alternative decision making approaches on the same hierarchical level. This scheme enables the decision maker to choose the most suitable approach for decision making depending on his problem specific knowledge, his personal intuition, and the optimization problem at hand. The flexibility of EAs and MOEAs enables this scheme to incorporate the *de novo* programming approach, which allows the decision maker to reconsider the problem definition itself based on the available solution alternatives. This allows the decision maker to incrementally reduce the number of objectives, through aggregation or transforming objectives into constraints.

The following section outlines the most common *a priori*, interactive and *a posteriori* approaches to MOOPs. Sec. 3 introduces the new iterative, interactive multi-objective optimization scheme and Sec. 4 gives a final discussion.

## 2 Multi-criteria Decision Making

In this section we outline the most common *a priori*, interactive and *a posteriori* multi-criteria decision making approaches. A more comprehensive overview on multi-criteria decision making is given in Ref. [7].

## 2.1 *A Priori* Decision Making

If the decision maker is able to formulate his preferences *a priori* based on a typically uninformed decision, the MOOP can immediately be converted into an SOOP, see left hand side of Fig. 1.

One example for this approach is weight aggregation which requires commensurable objectives and seeks to minimize a weighted sum of all objectives [14]. The decision maker needs to make an *a priori* decision on the weights for each objective. Another approach is the  $\epsilon$ -constraint method [5]. Here the decision maker gives constraints for  $m - 1$  objectives such that only one objective remains to be optimized. Goal programming requires the decision maker to give a goal vector  $v$  of desired objective values and the distance to  $v$  is to be minimized [10]. Without *a priori* knowledge  $v$  may be unattainable or trivial. Global criterion methods based on  $L_p$ -metrics or the Tchebycheff metric assume unattainable goal vectors and also minimize the distance to  $v$  [15].

## 2.2 Interactive Decision Making

In interactive decision making the decision maker gives an initial guess on his preferences and is then allowed to refine his preferences based on a set of solution alternatives. The optimization process is reiterated until a satisfactory solution is found, see center of Fig. 1. This way the decision maker is able to make an at least partially informed decision. Interactive approaches can be distinguished into search- and learning-oriented methods.

For example, the Geoffrion-Dyer-Feinberg (GDF) method gives a search-oriented approach [4]. This method assumes that the preferences of the decision maker are based on an implicitly known utility function. At each iteration the decision maker gives trade-off values for the objectives and thus approximates the utility function linearly. After optimizing the resulting SOOP, the decision maker is presented the currently best solution alternatives and he may refine the trade-off values or terminate the algorithm. The Tchebycheff method is also search-orientated and tries to minimize several randomly weighted Tchebycheff metrics to a goal vector selected by the decision maker [11]. The resulting solution set is then presented to the decision maker and the decision maker may refine the search by giving a new goal vector and a reduced range for the weights.

A learning-oriented approach is the STEP method [1]. After selecting a reference solution, the decision maker classifies the achieved objectives into satisfied and unsatisfied. Eventually allowing a certain amount of degradation for the satisfied objectives the SOOP is then given by an  $\epsilon$ -constrained problem minimizing a weighted sum for the remaining unsatisfied objectives. In the next step the decision maker may select a new reference solution and again classify the achieved objectives. Another learning-oriented approach is the Reference Point method [13]. Here the decision maker chooses a goal vector  $v$ , in this context also known as reference point of aspiration levels, and then multiple SOOPs are solved, minimizing an achievement scalarizing function for a number of perturbed goal vectors  $v'$ . The resulting set of solution alternatives is presented to the decision maker to select a final solution or a new goal vector.

### 2.3 A Posteriori Decision Making

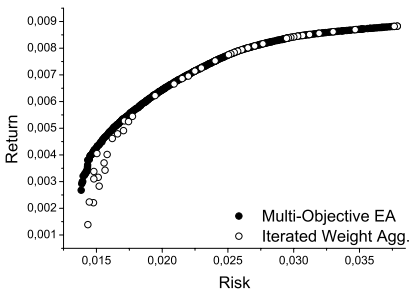
For a *a posteriori* decision making the whole Pareto front  $\mathcal{PF}$  needs to be approximated. Thus, the decision maker is able to make a fully informed decision based on all non-dominated solutions, see right hand side of Fig. 1.

One way to approximate the Pareto front is to iteratively solve multiple SOOPs based on aggregation methods and parameterizing the aggregation method differently for each SOOP instance. For example, different weight vectors for the previously mentioned weight aggregation approach or multiple alternative constraints for the  $\epsilon$ -constraint approach can be used [9].

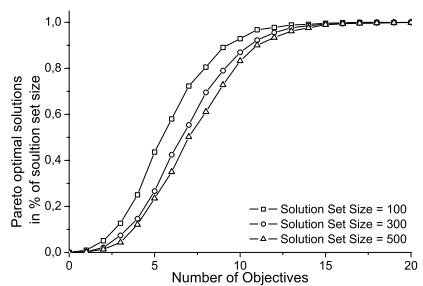
MOEAs, on the other hand, are able to approximate the Pareto front in a single optimization run. For this purpose the selection method must either be enhanced with aggregation methods for multiple objectives or it must be replaced with a truly multi-objective selection method based on the concept of Pareto dominance. For a more detailed description on MOEAs, refer to [2].

In case the search space is causal and the Pareto optimal solutions are similar to each other, the MOEA approaches can outperform the iterated aggregation. See Fig. 2 for exemplary solutions for a single MOEA optimization run and an iterated weight aggregation approach with 50 iterations on a portfolio selection problem with two objectives.

Although the *a posteriori* approaches offer the most information about the solution alternatives to the decision maker, they do not address the problem of how to select the final solution alternative. Additionally, approximating the whole Pareto set is much more time consuming than solving a single SOOP or even a limited number of SOOPs in case of interactive decision making. Also *a posteriori* approaches do not scale well with increased number of objectives, because the number of Pareto optimal solution typically increases dramatically, see Fig. 3. This can impair the selection mechanism of MOEAs. It also makes final decision making more complicated, regardless of the *a posteriori* approach used, unless special methods for aiding decision making are introduced.

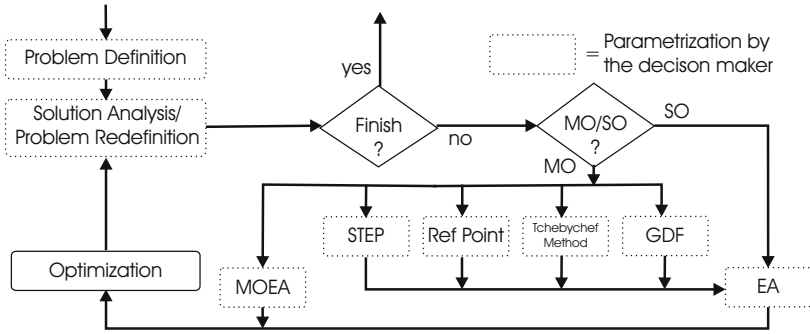


**Fig. 2.** MOEA and iterated aggregation for two-objective portfolio selection



**Fig. 3.** % of non-dominated solutions in a random solution set





**Fig. 4.** The iterative, interactive multi-objective optimization scheme (MOCCO). Grey boxes indicate interaction by the decision maker and black boxes optimization runs.

### 3 Interactive Multi-criteria Decision Making

All the previously mentioned approaches have their specific pros and cons. *A priori* methods suffer from the lack of *a priori* knowledge, while interactive approaches rely on specific properties of the search space like non-convex Pareto fronts or differentiable objective functions. Finally, *a posteriori* approaches suffer from increased objective space dimension. And even though intermediate solutions like the interactive evolutionary multi-objective optimization tool (I-EMO) have been suggested recently [3], such solutions often favor one approach and use alternative approaches only as post-processing kludge. Unfortunately, such simple approaches do not solve any of the above mentioned problems.

In this paper we propose a novel interactive optimization scheme which merges the alternative approaches on the same hierarchical level. In our initial implementation we utilize EAs and MOEAs as optimization algorithms. This allows us to deal with a wide range of constraints and virtually any type of search space and objectives. This way we are able to implement the *de novo* programming approach [16]. This approach calls for iterative reformulation of the original MOOP until only one solution alternative remains.

The iterative, interactive multi-objective optimization scheme as proposed in this paper starts with the initial problem definition by the decision maker, see Fig. 4. After generating an initial random solution set, the decision maker may reformulate the optimization problem by adding constraints, turning objectives into constraints, aggregating objectives and even introducing new objectives based on his analysis of the available solution alternatives. If the decision maker is not satisfied, the proposed scheme checks whether the optimization problem at hand is truly multi-objective or not. In case the problem is single-objective, because objectives have been removed, aggregated or turned into constraints, the scheme proceeds optimizing the problem using an EA. If the optimization problem is truly multi-objective the decision maker can choose from a number of interactive approaches including the Geoffrion-Dyer-Feinberg method, the

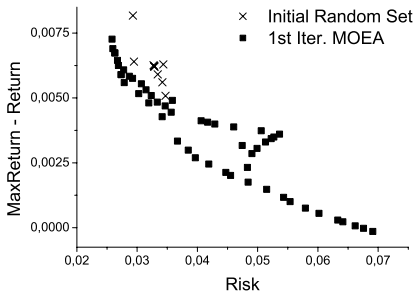


Fig. 5. First iteration on a portfolio selection problem with three objectives

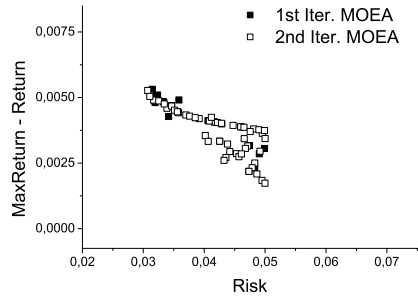


Fig. 6. Second iteration with applied constraints on all three objectives

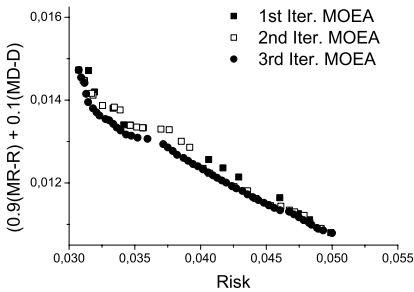


Fig. 7. Third iteration with aggregated return and dividends

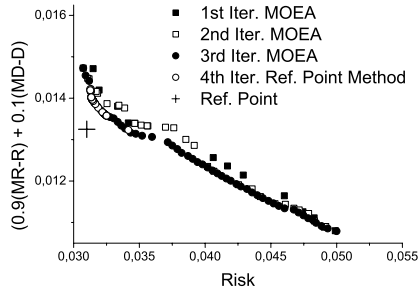


Fig. 8. Final iteration based on the reference point method with 4 perturbations

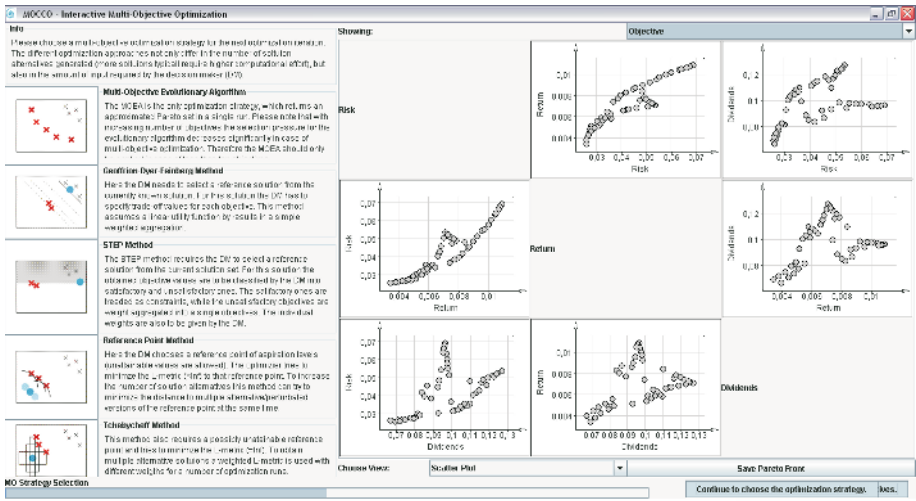


Fig. 9. GUI of the proposed MOCCO scheme showing the Pareto front in a scatter plot in objective space (in contrast to the plots in fitness space in Fig. 5 - 8)

Tchebycheff method, the STEP method or the Reference Point method or he may also use an MOEA to approximate the whole Pareto front for *a posteriori* decision making. After optimizing the resulting single-objective or multi-objective optimization problem the decision maker is again allowed to reformulate the problem and continue the optimization process or to terminate if satisfied.

### 3.1 Example on a Three Objective Portfolio Selection Problem

To give an illustrative example we use the new scheme on a three-objective portfolio selection problem minimizing the risk of the portfolio, while maximizing the return and the dividends for the portfolio based on the Markowitz portfolio selection model [6]. Note that Fig. 9 gives a scatter plot in objective space, while Fig. 5 - 8 give only two-dimensional plots in 'fitness' space. This 'fitness' is always to be minimized in the JavaEvA framework used [12].

For the first iteration the decision maker introduced all three objective and performed an initial optimization run using an MOEA to get a first impression on the objective space, see Fig. 5. For the second iteration the decision maker added constraints to each objective ( $Risk < 0.05$ ,  $Return > 0.005$  and  $Dividends > 0.1$  in objective space) and performed an additional MOEA optimization run, see Fig. 6. The resulting Pareto front is much more focused on the region most interesting for the decision maker. Since return and dividends are given in the same unit and are thus commensurable, the decision maker continues to aggregate these two objective by weight aggregation using  $w_{Return} = 0.9$  and  $w_{Dividends} = 0.1$ . This results in an only two-dimensional optimization problem and again the decision maker decides to use an MOEA to get an impression on the new objective space, see Fig. 7. Based on the obtained approximation of the local Pareto front the decision maker then utilizes the Reference Point method to generate most suited solutions based on his favored aspiration level, see reference point in Fig. 7. The resulting solution set generated for four perturbed reference points gives the final set of solution alternatives from which the final solution is selected by the decision maker.

## 4 Discussion and Outlook

The proposed framework allows an intuitive approach to high dimensional multi-objective optimization problems based on *de novo* programming. Using the *de novo* programming approach the decision maker is able to reduce the objective space based on his preferences by adding constraints, turning objectives into constraints and by aggregating commensurable objectives. Utilizing EAs and MOEAs as optimization algorithms this new approach is able to deal with arbitrary constraints and is able to address a wide range of optimization problems including continuous and combinatorial optimization problems. Depending on the problem type and the available problem specific knowledge at hand the decision maker may choose the most suitable multi-objective decision making approach for each iteration step. This way the new scheme enables both search- and also learning-oriented decision making.

**Acknowledgments.** This research has been funded by the Canon Foundation in Europe and the ALTANA Pharma AG, Konstanz, Germany.

## References

1. R. Benayoun, J. de Montgolfier, J. Tergny, and O. Laritchev. Linear Programming with Multiple Objective Functions: Step Method (STEM). *Mathematical Programming*, 1(3):366–375, 1971.
2. K. Deb. *Multi-Objective Optimization using Evolutionary Algorithms*. 2001.
3. K. Deb and S. Chaudhuri. I-EMO: An interactive evolutionary multi-objective optimization tool. KanGAL Report 2005003, Kanpur Genetic Algorithms Laboratory (KanGAL), Indian Institute of Technology, Kanpur, PIN 208016, India, 2005.
4. A. M. Geoffrion, J. S. Dyer, and A. Feinberg. An interactive approach for multi-criterion optimization, with an application to the operation of an academic department. *Management Science*, 12(4):357–368, 1972.
5. Y. Y. Haimes, L. S. Lasdon, and D. A. Wismer. On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE Transactions on Systems, Man and Cybernetics*, 1(3):296–297, 1971.
6. H. M. Markowitz. *Portfolio Selection: efficient diversification of investments*. 1959.
7. K. M. Miettinen. *Nonlinear Multiobjective Optimization*. 1998.
8. V. Pareto. *Manual of Political Economy*. The MacMillan Press Ltd, 1971. (the original edition is French in 1927).
9. S. R. Ranjithan, S. K. Chetan, and H. K. Dakshina. Constraint method-based evolutionary algorithm (CMEA) for multiobjective optimization. In E. Zitzler, K. Deb, and L. Thiele et al., editors, *Evolutionary Multi-Criterion Optimization*, volume 1993 of *LNCS*, pages 299–313, 2001.
10. C. Romero. *Handbook of Critical Issues in Goal Programming*. Pergamon Press, Oxford, UK, 1991.
11. R. E. Steuer. *Multiple Criteria Decision Making and Risk Analysis Using Microcomputers*, chapter The Tchebycheff Procedure of Interactive Multiple Objective Programming, pages 235–249. 1989.
12. F. Streichert and H. Ulmer. JavaEvA - a java framework for evolutionary algorithms. Technical Report WSI-2005-06, Centre for Bioinformatics Tübingen, University of Tübingen, 2005.
13. A. P. Wierzbicki. *Multiple Criteria Decision Making Theory and Applications*, chapter The Use of Reference Objectives in Multiobjective Optimization, pages 468–486. Number 177 in *LNEMS*. 1980.
14. L. Zadeh. Optimality and non-scalar-valued performance criteria. *IEEE Transactions on Automatic Control*, 8:59–60, 1963.
15. M. Zeleny. *Multiple Criteria Decision Making*, chapter Compromise Programming, pages 262–301. University of South Carolina, Columbia, South Carolina, USA, 1973.
16. M. Zeleny. The evolution of optimality: De novo programming. In *Proceedings of the Conference on Evolutionary Multi-Criterion Optimization*, volume 3410 of *LNCS*, pages 1–13, 2005.

# Utilization of NNs for Improving the Traditional Technical Analysis in the Financial Markets

Norio Baba

Information Science, Osaka Kyoiku University,  
Asahiga-Oka, 4-698-1, Kashiwara City, Osaka Prefecture, 582-8582, Japan

**Abstract.** In this paper, we propose a new decision support system (DSS) for dealing stocks which improves the traditional technical analysis by using neural networks. In the proposed system, neural networks are utilized in order to predict the “Golden Cross” and the “Dead Cross” several weeks before they occur. Computer simulation results concerning the dealings in the “TOPIX” and the “Nikkei-225” confirm the effectiveness of the proposed system.

**Keywords:** neural networks, traditional technical analysis, golden cross, dead cross, TOPIX, Nikkei-225.

## 1 Introduction

The widespread popularity of neural networks (NNs) in many different fields is mainly due to their ability to build complex nonlinear relationships between input variables and output variables directly from the training data [1]. NNs can provide models for a large class of real systems which are difficult to handle using traditional approaches. Recently, because of NNs’ characteristic property, a large number of people have paid particular attention to apply NNs to the financial market [2]-[5].

In this paper, we shall utilize NNs in order to predict the “Golden Cross (GC)” and the “Dead Cross (DC)” several weeks before they occur. The structure of this paper is as follows. In Section 2, we briefly touch upon the traditional technical analysis which utilizes the measure “long term moving average” versus “short term moving average” for finding the current tendency of the stock market. In Section 3, we suggest that predicting GC & DC several weeks before they occur by NNs could contribute a lot in dealing stocks. Further, we show how the prediction of GC & DC can be done by NNs. In Section 4, we give computer simulation results concerning dealings in the TOPIX and the Nikkei-225 which confirm that prediction of GC & DC by NNs can contribute a lot in dealing stocks. Finally, the paper concludes with a brief summary.

## 2 Traditional Technical Analysis Utilizing the Measure “Long Term Moving Average (LTMA)” Versus “Short Term Moving Average (STMA)”

In order to detect the current tendency of the stock market, many stock traders have often used the traditional technical analysis which utilizes the measure “Long Term

Moving Average (LTMA) versus Short Term Moving Average (STMA)”. Particularly, they have often fixed their eyes upon the “Golden Cross (GC)” and the “Dead Cross (DC)” which are the crossings of LTMA & STMA. (Fig.1 illustrates a typical example of GC and DC which appeared in the two moving averages of the stock price of “Nissan Motors” listed in the Tokyo Stock Market. In Fig.1, the point B (A) denotes GC (DC) where STMA cuts LTMA upwards (downwards).) They have believed that GC (DC) is a clear indicator that predicts the upward (downward ) moving of the stock price.

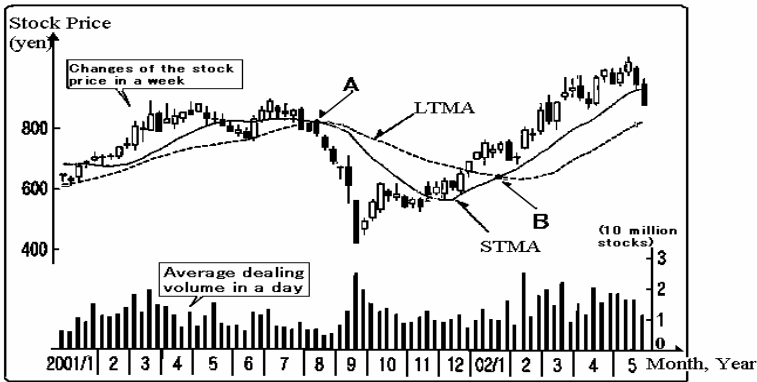


Fig. 1. Golden Cross and Dead Cross of the LTMA & STMA (Nissan Motors)

### 3 Utilization of NNs for Improving the Traditional Technical Analysis

#### 3.1 Prediction of GC & DC by NNs Could Contribute a Lot in Dealing Stocks

Many stock traders have often relied upon the traditional technical analysis which utilizes the measure LTMA versus STMA in dealing stocks. However, we consider that the traditional technical analysis could further be improved by the use of NNs. The prototype DSS utilizing the traditional technical analysis executes dealings after golden cross or dead cross has been observed. However, such a way of dealings often loses an important chance to execute dealings in a reasonable price in the stock market. If one could deal stocks several weeks before real crossing occurs, one could save considerable amount of money. We shall explain this by using the real data which illustrates the changes of the price of “Nissan Motors” during the period January 2001-May 2002. In Fig.1, the solid line and the broken line indicate the STMA(13MA) and the LTMA(26MA), respectively. If we utilize the traditional technical analysis, we have to sell the stock at the price around 750 yen and buy it at the price around 730 yen. On the other hand, we can sell the stock at the price higher than 800 yen and buy it at the price around 700 yen if we could predict the dead cross and the golden cross several weeks before those crossings occur.

Remark 3.1: *As the measure “LTMA versus STMA”, “26 weeks moving average (26 MA) versus 13 weeks moving average (13 MA)” has often been used in order to detect the current tendency of each individual stock price. On the other hand, the measure “13 MA versus 6 MA” has been used for checking the current tendency of the several indexes such as TOPIX and Nikkei-225. In this paper, we shall use the measure “13 MA versus 6 MA”.*

### 3.2 Input Variables of the Neural Network Model

Due to various causes, price of a stock changes. Therefore, in order to construct NNs models for making appropriate prediction of the crossings GC & DC between the STMA and the LTMA, one has to choose input variables carefully which may significant effect upon the changes of a stock price. Table 1 shows the candidates of the input variables of the neural network model for making the prediction of GC & DC of the TOPIX. (Due to space, we don't go into details concerning input variables of the Nikkei-225. Interested readers are kindly asked to attend our presentation.)

**Table 1.** Input Variables into the neural network model of TOPIX

-----  
 1) Rate of change of the TOPIX 2) Trend of changes of the TOPIX 3) Rate of change of the total amount of money used for dealing 4) Trend of changes of the total amount of money used for dealing 5) Rate of change of the PBR in the Tokyo Stock Market 6) Rate of change of the New York Dow 7) Trend of changes of the New York Dow 8) Rate of change of the amount of excess credit buying 9) Trend of changes of the amount of excess credit buying 10) Rate of change of the excess buying volume by foreign traders 11) Trend of changes of the excess buying volume by foreign traders 12) Rate of change of the Nasdaq 13) Rate of change of the bond 10 14) Trend of changes of the bond 10 15) Rate of change of dollars 16) Trend of changes of dollars 17) Rate of change of (13MA-6MA) of the TOPIX 18) Trend of changes of (13MA-6MA) of the TOPIX 19) Trend of changes of Dubai Oil 20) Sign of (13MA-6MA) of the TOPIX 21) Rate of change of the TOPIX during 4 weeks 22) Trend of changes of the Nasdaq 23) Rate of the absolute value of (13MA-6MA) for its maximum  
 -----

### 3.3 Teacher Signals and Neural Network Training

In order to predict GC & DC, we need to carry out training of the weight vectors in the neural network model. Fig.2 illustrates the teacher signals relating GC. (Due to space, we don't go into details of the teacher signals concerning DC.) Training of the neural network model has been carried out by using MATLAB Toolbox in order to make a prediction of the GC and DC several weeks before they occur.

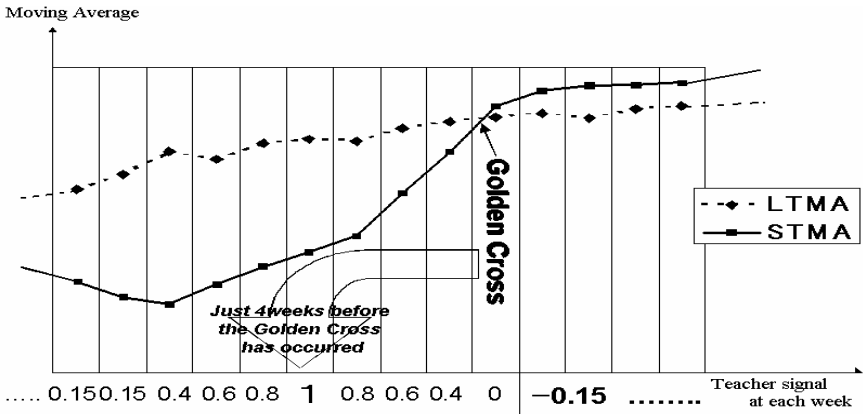


Fig. 2. Teacher signals concerning Golden Cross

### 3.4 Learning Periods and Prediction Periods

In order to predict GC & DC for a year, we have carried out neural network training by using the past data for three years. Table 2 shows the learning periods and the prediction periods being used for the prediction of GC & DC concerning the TOPIX.

Table 2. Learning Periods and Prediction Periods

	Learning Period	Prediction Period
A1	January 1991 - December 1993	January 1994 - December 1994
A2	January 1992 - December 1994	January 1995 - December 1995
.	.	.
.	.	.
.	.	.
A12	January 2002 - December 2004	January 2005 - May 2005

### 3.5 Neural Network Models and Output Function

As the neural network model, we have used  $N \times N \times 1$  model where  $N$  denotes the number of all the available input variables for the prediction of GC & DC. We have also used  $m \times m \times 1$  ( $m < N$ ) model where  $m$  denotes the number of input variables having been chosen by the sensitivity analysis [6], [7]. As the output function of NNs, we have used the bipolar-sigmoid function.

### 3.6 A Trial for Finding GC & DC by Using the Several Neural Network Models

In order to predict GC & DC accurately, we have tried to utilize several neural network models. By using the sensitivity analysis, we have arranged all of the available input variables in the order concerning the rate of the sensitivity. As the



input variables of the neural network model  $m \times m \times 1$ , we have chosen  $m$  variables which have been ranked from the top to the  $m$ th.

In order to predict GC & DC of the TOPIX several weeks before they occur, we have used the following 7 neural network models:  $23 \times 23 \times 1$ ,  $20 \times 20 \times 1$ ,  $18 \times 18 \times 1$ ,  $15 \times 15 \times 1$ ,  $12 \times 12 \times 1$ ,  $10 \times 10 \times 1$ ,  $8 \times 8 \times 1$ .

We have set the following decision rules:

- (1) If all of the outputs from 7 NNs are positive (negative) and the average of their absolute values is above 0.5, then we judge that GC (DC) will occur in several weeks.
- (2) If all of the following three conditions (2-a), (2-b), and (2-c) are satisfied, then we judge that GC (DC) will occur in several weeks.
  - (2-a) The greater part of the outputs from the 7 NNs is positive (negative).
  - (2-b) The rate of the outputs having values over 0.7 (below - 0.7) exceeds 50% of the number of the outputs having the same sign.
  - (2-c) The average of the outputs from 7 NNs is above 0.5 (below - 0.5).

Using the data from 1991 to 2005, we have carried out computer simulations in order to check how the above decision rules help well in making a prediction of GC & DC. Table 3 shows the success rate of the prediction concerning GC & DC among the cases where increase (decrease) rate of the TOPIX and the Nikkei-225 after GC (DC) is above 10 %. The computer simulation results shown in Table 3 confirm the effectiveness of the proposed decision rules.

**Table 3.** Success rate of prediction regarding GC & DC

T O P I X			
GC	5 / 6 (83.3 %)	DC	6 / 6 (100 %)
Nikkei-225			
GC	7 / 8 (87.5 %)	DC	5 / 5 (100 %)

## 4 Computer Simulations

We have carried out computer simulations concerning dealings in the TOPIX and Nikkei-225 from 1994 to 2005. Table 4 and Table 5 show the changes of the initial money (10 billion yen) during each year by the dealings in the TOPIX and Nikkei-225, respectively, utilizing the new DSS (NDSS) which used the new decision rule shown in Section 3.6, the traditional technical analysis (TTA), and the Buy-and-Hold method (BHM). The simulation results confirm the effectiveness of the NDSS.

*Remark 4.1: In the above simulations, we have taken the charge for dealing into account by subtracting  $(0.001 * (\text{total money used for dealing}) + 250,000)$  yen from the total fund for dealing.*

*Remark 4.2: In the above simulations, we have used the rule which allows “dealing on credit”. Due to space, we don’t go into details. Interested readers are kindly asked to attend our presentation.*

Remark 4.3: *In the above simulations, we have carried out dealings as soon as GC (DC) has been predicted by the new decision rule shown in Section 3.6.*

Remark 4.4: *Prediction periods in B1 and B2( in Table 5) are April 2003 – March 2004 and April 2004 – March 2005, respectively.*

Remark 4.5: *In the Table 5, simulations from January 2001 to March 2003 are lacked. This comes from the fact that a large number of stocks (more than 30) which constituted Nikkei-225 were replaced at the end of March 2000.*

**Table 4.** Dealing Result by the Three Methods (TOPIX) (Billion yen; Numbers below 10 million yen are rounded.)

	BHM	TTA	NDSS		BHM	TTA	NDSS
A1	10.80	9.85	11.05	A7	7.43	10.77	6.73
A2	10.25	10.44	9.45	A8	7.94	9.93	12.26
A3	8.96	9.81	9.27	A9	7.98	9.32	10.89
A4	7.75	9.95	8.82	A10	12.60	11.77	8.45
A5	9.19	7.96	13.13	A11	10.76	10.23	9.66
A6	15.33	9.51	13.84	A12	9.86	9.95	10.69
Total Return : BHM: - 1.07 ; TTA: - 0.45 ; NDSS: + 4.31							

**Table 5.** Dealing Result by the Three Methods (Nikkei-225) (Billion yen; Numbers below 10 million yen are rounded.)

	BHM	TTA	NDSS		BHM	TTA	NDSS
A1	11.29	10.26	8.62	A6	13.39	9.33	16.68
A2	10.15	10.03	12.04	A7	7.30	11.43	8.16
A3	9.35	10.19	9.46	B1	14.52	11.97	10.77
A4	7.65	9.54	11.22	B2	9.93	8.93	10.00
A5	9.03	8.29	13.90				
Total Return: BHM: + 2.64 ; TTA: + 0.01 ; NDSS: + 10.88							

## 5 Concluding Remarks

A decision support system for dealing stocks which improves the traditional technical analysis by utilizing NNs has been proposed. Computer simulation results suggest the effectiveness of the proposed DSS. However, these simulations have been done only for the TOPIX and the Nikkei-225. In order to execute full confirmation concerning the developed NDSS, we need to check whether it can be successfully applied for dealing in the other indexes such as S&P 500, DAX, and etc. For this purpose, we also need to carry out computer simulations concerning various individual stocks.

## Acknowledgements

The author would like to thank Prof. Y. Kai (Kwanseigakuin Univ.) and QUICK Corporation for their kind support in giving them various financial data. He would also like to express his thank to his students A. Matsuda, M. Nishida, M. Uematsu for their effort concerning computer simulations. He also gratefully acknowledges the partial financial support of the Foundation for Fusion of Science & Technology (FOST) and the Grant – In –Aid for Scientific Research (C) by Ministry of Education, Science, Sports and Culture, Japan.

## References

1. Rumelhart, D.E. et al.: Parallel Distributed Processing. MIT Press(1986).
2. Azoff, E.M.: Neural Network Time Series Forecasting of Financial Markets Wiley(1994)
3. Baba, N. and Kozaki, M.: An intelligent forecasting system of stock price using neural network. in Proceedings of IJCNN 1992, (1992)371-377.
4. Baba, N. and Suto, H.: Utilization of artificial neural networks and TD-Learning method for constructing intelligent decision support systems. European Journal of Operational Research 122(2000)501-508.
5. Baba, N. and Nomura, T.: An Intelligent Utilization of Neural Networks for Improving the Traditional Technical Analysis in the Stock Markets. Proceedings of KES2005, (2005)8-14.
6. Zurada, J.M. et al.: Sensitivity analysis for minimization of joint data dimension for feedforward neural network. in Proceedings of the IEEE International Symposium on Circuits and Systems.(1994)447-450
7. Viktor, H.L. et al.: Reduction of symbolic rules from artificial neural network using sensitivity analysis. in Proceedings of the IEEE International Symposium on Circuits and Systems. (1995)1788-1793.

# Use of Support Vector Machines: Synergism to Intelligent Humanoid Robot Walking Down on a Slope

Dongwon Kim and Gwi-Tae Park

Department of Electrical Engineering, Korea University, 1, 5-ka, Anam-dong, Seongbuk-ku,  
Seoul 136-701, Korea  
{upground, gtpark}@korea.ac.kr

**Abstract.** In this paper intelligent humanoid robot walking down on a slope with support vector machines is presented. Humanoid robots can be used as proxies or assistants to humans in performing tasks in real world environments, including rough terrain, steep stairs, and obstacles. But the dynamics involved are highly nonlinear and unstable. So the humanoid robot can not get the stable and reliable biped walking easily. As a significant dynamic equilibrium criterion, zero moment point (ZMP) is usually employed and we are establishing empirical relationships based on the ZMP trajectory as dynamic stability of motion. Support vector machines (SVM) are applied to model a ZMP trajectory of a practical humanoid robot. The SVMs' performance can vary considerably depending on the type of kernels adopted by the networks. The experimental results show that the SVM based on the kernel substitution provides a promising alternative to model robot movements but also to control actual humanoid robots.

## 1 Introduction

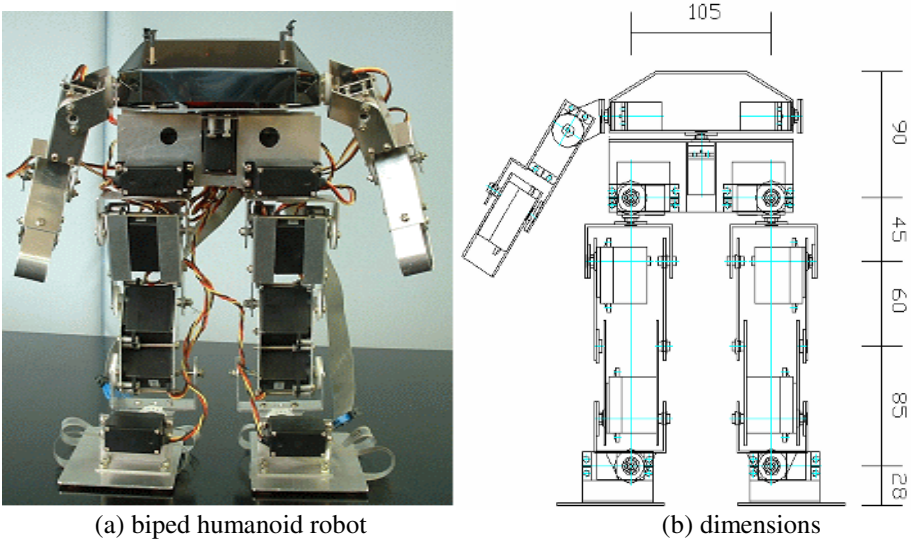
Machines of physical form similar to humans, humanoid robots can be used as proxies or assistants to humans in performing tasks in real world environments, including rough terrain, steep stairs, and obstacles. Humanoid robots have recently evolved into an active research area with the creation of several humanoid robot systems. Although there are many technical issues in the humanoid robot system, stable and reliable biped walking is the most fundamental issue that has yet to be solved with a high degree of reliability.

A zero moment point (ZMP) was originally defined for ground contacts of legs by Vukobratovic [1] as the point on the ground plane about which the total moment due to ground contacts become zero in the plane and is usually employed as a basic component for dynamically stable motion. These days, studies on the ZMP modeling for feasible walking motion of humanoid robot are increased. Kim *et al.* [2-3] proposed a method to generate a smooth walking pattern by using fuzzy and adaptive neuro-fuzzy systems. However, other intelligent systems like the support vector machines and walking down a slope have not yet been evaluated. Their applicability to humanoid robots should be investigated since they may provide better predictions than typical fuzzy systems, and thereby, provide better insight into human-like walking mechanisms.

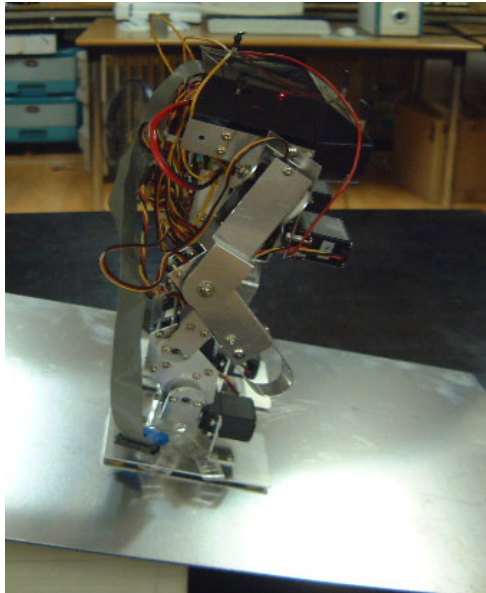
Support vector machines (SVM) are applied to model a ZMP trajectory of a practical humanoid robot in this paper. The SVMs' performance can vary considerably depending on the type of kernels adopted by the networks. The experimental results show that the SVM based on the kernel substitution provides a promising alternative to model robot movements but also to control actual humanoid robots.

## 2 Humanoid Robot Walking Down on a Slope and Its ZMP

In practice, we have design and implement a biped humanoid robot as shown in Fig. 1(a). The robot has 19 joints and the key dimensions of the robot are also shown in Fig. 1(b).



**Fig. 1.** Actual humanoid robot and its key dimensions

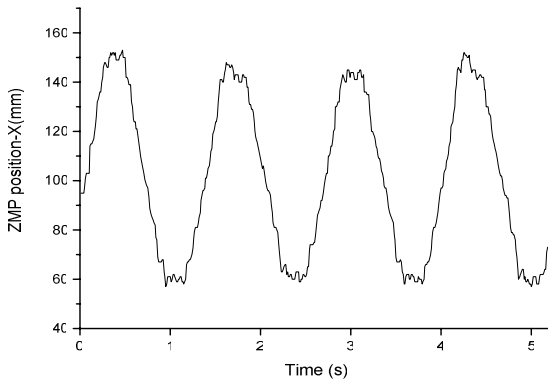


**Fig. 2.** Side view of the humanoid robot walking down a  $10^\circ$  slope

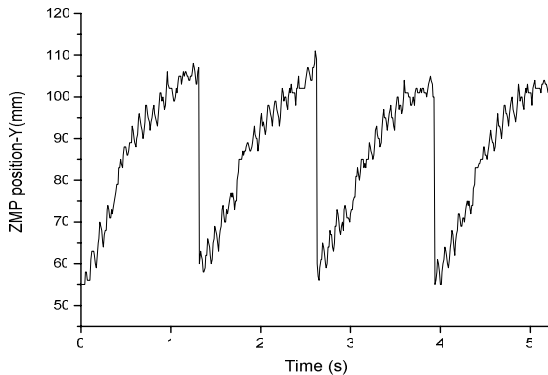
The height and the total weight are about 308mm and 1700g including batteries. Each joint is driven by the RC servomotor that consists of a DC motor, gear, and simple controller. Each of the RC servomotors is mounted in the link structure. Our biped walking robot is able to walk under the condition which one step is 48 mm per 1.4 s on the flat floor. More detailed information about specification and block diagram of the robotic system are in [2-3]. The walking motion of the robot is shown in Fig. 2 which shows snapshot of the robot walking down a slope.

Based on the data from the force sensors equipped on each foot, significant stability criterion, the real ZMP positions,  $x$ -coordinate and  $y$ -coordinate, of the four-step motion of the biped walking robot are calculated and depicted in Fig. 3. The corresponding ZMP trajectories of Fig. 3 are also shown in Fig. 4.

The complex dynamics involved in the biped walking robot make robot control a challenging task. But, if the highly nonlinear and complex dynamics are modeled well, it can be explained by empirical laws by incorporating these laws into the biped walking robot. So, human-like walking robots can be realized more easily. Support vector machine (SVM) to be studied in next section is applied to model the ZMP trajectory data, to present the nonlinearities, and to control complex real biped walking robot.



(a)  $x$ -coordinate



(b)  $y$ -coordinate

**Fig. 3.** ZMP positions of our humanoid robot

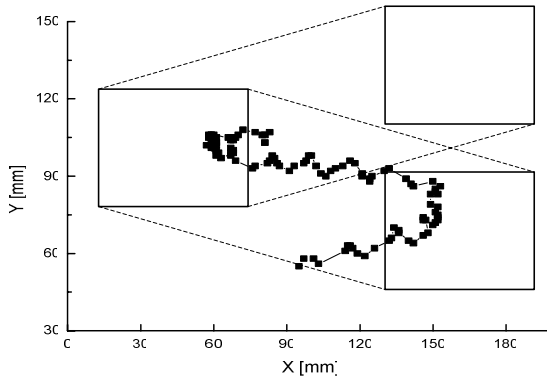


Fig. 4. ZMP trajectories of the humanoid robot corresponding to Fig. 3

### 3 Support Vector Machines and Its Usage

The following presents some basic concepts of SVM as described by prior research. A detailed explanation may be found in [4-5]. The SVMs can be applied to regression problems by the introduction of  $\epsilon$ -insensitive loss function: support vector regression(SVR) [6]. The loss function defines a band around the true outputs sometimes referred to as a tube. The idea is that errors smaller than a certain threshold  $\epsilon > 0$  are ignored. That is, errors inside the band are considered to be zero. On the other hand, errors caused by points outside the band are measured by slack variables,  $\zeta, \zeta^*$ .

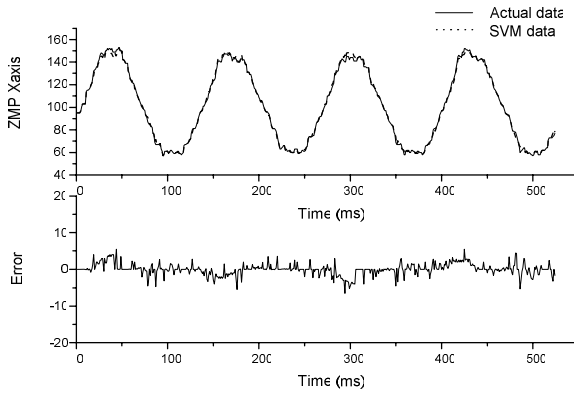
In the case of SVR for linear regression,  $f(x)$  is given by  $f(x) = \langle w, x \rangle + b$ , with  $w \in \mathbb{R}^n, b \in \mathbb{R}$ .  $\langle \cdot, \cdot \rangle$  denotes the dot product. For the case of nonlinear regression,  $f(x) = \langle w, \phi(x) \rangle + b$ , where  $\phi$  is some nonlinear function which maps the input space to a higher dimensional feature space. In  $\epsilon$ -insensitive loss function, the weight vector  $w$  and the threshold  $b$  are chosen to optimize the following problem

$$\begin{aligned} \text{Minimize} \quad & R(w, b, \zeta, \zeta^*) = \frac{1}{2} \langle w, w \rangle + C \sum_{i=1}^l (\zeta_i + \zeta_i^*) \\ & (\langle w, \phi(x_i) \rangle + b) - y_i \leq \epsilon + \zeta_i, \\ \text{Subject to} \quad & y_i - (\langle w, \phi(x_i) \rangle + b) \leq \epsilon + \zeta_i^*, \\ & \zeta_i, \zeta_i^* \geq 0. \end{aligned}$$

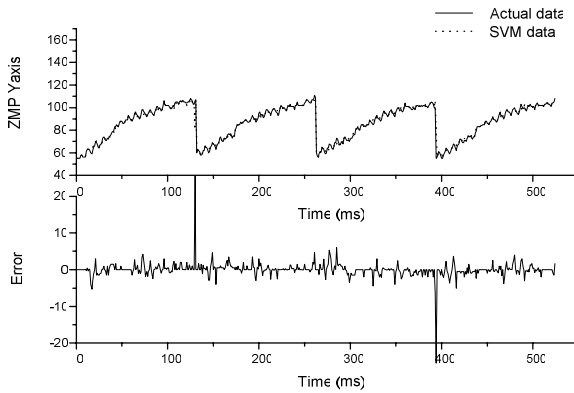
The constant  $C > 0$  determines the trade-off between the flatness of  $f$  and the amount up to which deviations larger than  $\epsilon$  are tolerated.  $\zeta, \zeta^*$  are called slack variables and measure the cost of the errors on the training points.  $\zeta$  measures deviations exceeding the target value by more than  $\epsilon$  and  $\zeta^*$  measures deviations which are more than  $\epsilon$  below the target value.

**Table 1.** Kernel functions and corresponding accuracy of humanoid robot

kernel type	x-coordinate	y-coordinate
linear	51.05	52.69
polynomial	10.29	19.95
RBF	<b>2.70</b>	<b>4.25</b>



(a) x-coordinate and its error



(b) y-coordinate and its error

**Fig. 5.** Generated ZMP xy-coordinate and error

The idea of SVR is to minimize an objective function which considers both the norm of the weight vector  $w$  and the losses measured by the slack variables. The minimization of the norm of  $w$  is one of the ways to ensure the flatness of  $f$ . The SVR algorithm involves the use of Lagrangian multipliers, which rely solely on dot



product of  $\phi(x)$ . This can be accomplished via kernel functions, defined as  $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ . Thus, the method avoids computing the transformation  $\phi(x)$  explicitly.

In this paper, we consider SVR with three types of kernels as follows.

*linear*:  $K(x_i, x_j) = x_i^T x_j$

*polynomial*:  $K(x_i, x_j) = (x_i x_j + 1)^d$

*RBF* :  $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\delta^2)$

Using the three types of kernel functions such as linear, polynomial, and radial basis function for SVR, approximated models are constructed and their results are obtained. The accuracy was quantified in terms of mean squared error (MSE) values.

MSE values corresponding three types of kernel functions are listed in Table 1 and we can also compare them. Regarding to the constant C, we set the value as 1000 and the degree of polynomial and width of RBF are set to 2.

From the Table1, the RBF kernel provides the best results so we can analysis the results more detail based on the generated ZMP trajectory from SVR. The generated ZMP positions from the RBF kernel, and its errors between actual data and generated data are shown in Fig 5. In Fig. 6, we can see the corresponding ZMP trajectories that are generated from the RBF kernel. From the figure, the generated ZMP is very similar to actual ZMP trajectory of the biped humanoid robot.

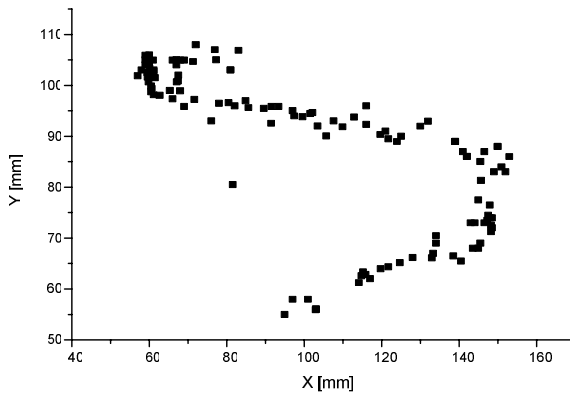


Fig. 6. Generated ZMP trajectories of the humanoid robot corresponding to Fig. 5

## 4 Conclusions

In this paper intelligent humanoid robot walking down on a slope with support vector machines is presented. As a significant dynamic equilibrium criterion, zero moment point (ZMP) of the robot is employed and we are establishing empirical relationships based on the ZMP trajectory as dynamic stability of motion. Support vector machines

(SVM) are applied to model a ZMP trajectory of a practical humanoid robot. We employed three types of kernel function and their results are obtained. Among these results, SVR with RBF function has the best accuracy. The experimental results show that the SVM based on the kernel substitution provides a promising alternative to model robot movements but also to control actual humanoid robots.

## Acknowledgment

The authors would like to thank the financial support of the Korea Science & Engineering Foundation. This work was supported by grant No. R01-2005-000-11044-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

## References

- [1] Vukobratovic, M., Brovac, B.: Zero-Moment Point-Thirty Five Years of Its Life. *Int. J. Humanoid Robotics*. 1 (2004) 157-173.
- [2] Kim, D., Kim, N.H., Seo, S.J., Park, G.T.: Fuzzy Modeling of Zero Moment Point Trajectory for a Biped Walking Robot. *Lect. Notes Artif. Int.* 3214 (2005) 716-722
- [3] Kim, D., Seo, S.J., Park, G.T.: Zero-moment point trajectory modeling of a biped walking robot using an adaptive neuro-fuzzy systems. *IEE Proc.-Control Theory Appl.* 152 (2005) 411-426
- [4] Vapnik, V.: *The Nature of Statistical Learning Theory*. John Wiley, New York (1995)
- [5] Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery*. 2 (1998) 121-167.
- [6] Oliveira, A.L.I.: Estimation of software project effort with support vector regression. *Neurocomputing*. (in press)
- [7] Gunn, S.: *Support vector machines for classification and regression*. ISIS technical report, Image Speech & Intelligent Systems Group University of Southampton (1998).

# Evolutionary Elementary Cooperative Strategy for Global Optimization

Crina Grosan<sup>1</sup>, Ajith Abraham<sup>2</sup>, Monica Chis<sup>3</sup>, and Tae-Gyu Chang<sup>2</sup>

<sup>1</sup> Department of Computer Science

Babeş-Bolyai University, Cluj-Napoca, 3400, Romania

<sup>2</sup> School of Computer Science and Engineering Chung-Ang University,  
Seoul 156-756, Korea

<sup>3</sup> Avram Iancu University, Cluj-Napoca 3400, Romania

cgrosan@cs.ubbcluj.ro, ajith.abraham@ieee.org, mchis@artelecom.net,  
tgchang@cau.ac.kr

**Abstract.** Nonlinear functions optimization is still a challenging problem of great importance. This paper proposes a novel optimization technique called Evolutionary Elementary Cooperative Strategy (EECS) that integrates ideas from interval division in an evolutionary scheme. We compare the performances of the proposed algorithm with the performances of three well established global optimization techniques namely Interval Branch and Bound with Local Sampling (IVL), Advanced Scatter Search (ASS) and Simplex Coding Genetic Algorithm (SCGA). We also present the results obtained by EECS for higher dimension functions. Empirical results for the functions considered reveal that the proposed method is promising.

## 1 Introduction

Global optimization of nonlinear functions is still a well studied problem if we take into account the amount of work published on this field during the last few decades [1], [2], [6], [8]. Therefore, the global optimization of functions with continuous variables is a challenging problem because of its immense practical applications. Recently, several approaches for global optimization have been developed. Stochastic techniques such as genetic algorithms are very popular and frequently used for global optimization due to their simplicity in implementation and quickness in finding an approximate solution. Most of the recent approaches hybridize local search methods with meta-heuristics in order to obtain more efficient methods for faster convergence. Simplex Coding Genetic Algorithm (SCGA) is a hybrid method combining genetic algorithms with a local search method called Nelder-Mead [5]. Scatter Search ([6], [7]) is an evolutionary method that operates on a small set of solutions and consists of five steps: diversification generation method, improvement method, reference set update method, subset generation method and the solution combination method. The scatter search methodology is very flexible since each of its elements can be implemented in a variety of ways and degree of sophistication.

This paper is an extension of the work proposed by Grosan and Abraham [4] dealing with higher nonlinear optimization functions. Evolutionary Elementary Cooperative Strategy (EECS) uses a small set of solutions and makes only a limited use of randomization within an evolutionary framework. EECS performance was tested by considering a set of 25 test functions which are well known benchmarks for global optimization problems. The paper is organized as follows: Section 2 describes EECS technique. In Section 3, some numerical experiments and comparisons with some existing approaches for global optimization are performed. Section 4 contains discussions and conclusions.

## 2 Evolutionary Elementary Cooperative Strategy (EECS)

EECS [4] uses a small set of solutions and a particular way to initialize the population. EECS is based on two important strategies: (1) the way the solutions are initialized and (2) the control of the evolutionary process. At the initialization stage, the population contains  $N^D$  individuals, where  $N$  is the number of space divisions for each variable and  $D$  represents the number of variables. We considered the domain of each variable being divided into the same number of divisions (sub-intervals) even if these domains of definition have different sizes.

Population initialization is achieved by following the steps given below:

*Equidistant domain division:* Divide the domain of definition of each variable into  $N$  subdivisions ( $N$  is the given number of divisions). For this, the *step* or distance between divisions is to be computed. The step is given by:  $step = \frac{\text{definition domain length}}{N}$ . A set of points are obtained in the following way:

first point is equal to the lower limit of the defined domain.

the point  $i$  ( $i = 2 \dots N$ ), will be  $i-1+step$ .

*Population construction:* An EECS individual of the population is represented as a string of size equal to the number of decision variables. The population consists of all the possible combinations of the points of each set obtained by applying the procedure described to the domain of definition of each variable.

Each individual from the population is able to interact only with its neighbors. By neighbors, we refer to the precedent and successor individual in the population (except for first and last individual which will only have successor or precedent). This interaction is in the form of a convex crossover. All the offspring are kept in a separate population. From the unified population of parents and offspring at the end of each generation a number of individuals are selected, which is equal to the initial population size. Selection is based on the fitness value.

## 3 Experiment Setup and Results

EECS is tested on several benchmark functions having between 1 and 8 variables. For the functions having 2 variables, results are compared with three new proposed mathematical models for global optimization. Advanced Scatter Search (ASS) [7] is an evolutionary method designed for nonlinear optimization.

Interval Branch and Bound with Local Sampling (IVL) [9] integrates the idea of branch and bound and several local sampling strategies. In [9], IVL is proved to be very efficient when compared with several other techniques including genetic algorithms. A Simplex Coding Genetic Algorithm (SCGA) is proposed in [5]. SCGA is hybridized version of genetic algorithm and simplex coding based search method (called Nelder-Mead).

### 3.1 Test Functions

We used 25 test problems based on a set of nonlinear objective functions. Test functions are well known benchmark problems and most of them are found in [3], [12] and [13]. Although the objective functions are build in a way that the optimal solutions are known, the optimization problems cannot be trivially solved by search procedures that do not exploit the special structure associated with each function [7].

The following are the test functions considered:

**F1:** defined as:

$$f(x) = 5.0 + \sin(x) + \sin(10x/3) + \ln(x) - 0.84x$$

Number of variables  $n = 1$ ; Range of initial points:  $2.7 \leq x \leq 7.5$ ; Global minimum:  $x^* = 5.199778$ ,  $f(x^*) = 0.398693$

**F2:** Six hump camel back; **F3:** Matyas; **F4:** Bohachevsky  $B_1$ ; **F5:** Bohachevsky  $B_2$ ; **F6:** Bohachevsky  $B_3$ ; **F7:** Easom; **F8:** Goldstein and Price; **F9:** Shubert; **F10:** Beale; **F11:** Booth; **F12:** Branin; **F13:** Hump; **F14:** Michalewicz; **F15:** defined as:

$$f(x) = x_1^2 + x_2^2 - \cos(18x_1) - \cos(18x_2)$$

Number of variables  $n = 2$ ; Range of initial points:  $-1 \leq x_1, x_2 \leq 1$ ; Global minimum:  $x^*=(0, 0)$ ,  $f(x^*)=-2$

**F16:** Schwefel; **F17:** Rosenbrock; **F18:** Zakharov; **F19:** De Jong; **F20:** Rastigrin; **F21:** Griewank; **F22:** Sum Squares; **F23:** Dixon and Price; **F24:** Levy and **F25:** Ackley.

### 3.2 Numerical Results and Discussions

For the F1 and F13 EECS is compared with Interval Branch and Bound with Local Sampling (IVL) technique. IVL was not applied for the other considered functions. In [9], IVL results are compared with a classical Genetic Algorithm (GA) and Simulated Annealing (SA). Results obtained in 40 runs are presented in Table 1. Results obtained by GA, SA and IVL are adapted from [9].

As evident from Table 1, for single and two objective test functions EECS clearly outperformed all the other considered techniques. The classical simulated annealing approach is not able to converge for test function F13. EECS is compared with ASS and SCGA for functions having 2 and more than 2 objectives. We adapted the results obtained by ASS and SCGA from [7] and [5].

**Table 1.** Comparison of results obtained for test functions F1 and F13

Algorithm	f-best(Mean, standard deviation)	No. of function evaluations	Time
<b>F1</b>			
GA	0.39869, 4.8e-9	38235	0.84s
SA	0.4189, 0.00817	55	7.5e-3s
IVL	0.398693	281	0.02s
EECS	0.398693	206	0.05s
<b>F13</b>			
GA	1.93e-6, 1.19e-7	61106	1.5s
SA	357.34, 77.5478	104	2.5e-3s
IVL	2.22992e-6	59075	15.1s
EECS	3.387e-4	742	13s

The performance of EECS is presented for all the 25 test functions. The authors of ASS [7] consider the following measure for defining the algorithm convergence:

Optimality GAP is defined as:

$$GAP = | f(x) - f(x^*)|,$$

Where  $x$  is an heuristic solution and  $x^*$  is the optimal solution. Then the heuristic solution  $x$  is optimal if:

$$GAP \leq \begin{cases} \varepsilon, & f(x^*) = 0 \\ \varepsilon |f(x^*)|, & f(x^*) \neq 0 \end{cases} .$$

In the considered experiments the value of  $\varepsilon$  is set as 0.001. The scatter search procedure is executed for 360 runs of up to 10000 objective functions evaluations. From the authors experiments for 20 objective functions [7], using different combinations of some proposed procedures, they still obtained a  $GAP$  (which means for some functions the algorithm did not converged). The minimum average  $GAP$  obtained (summarized for all the 20 test functions) was 9.90 (with a standard deviation  $GAP$  of 30.93) and the maximum average  $GAP$  obtained was 77.90 (with 206.21 for standard deviation for  $GAP$ ).

For testing SCGA performances, the authors used the conditions given in ([5]):

$$|f^*-f| < \varepsilon_1|f^*| + \varepsilon_2,$$

Where  $f$  refers to the best function value obtained by SCGA,  $f^*$  refers to the known exact global minimum and  $\varepsilon_1, \varepsilon_2$  are small positive numbers set to  $10^{-4}$  and  $10^{-6}$  respectively.

In Table 2, empirical results obtained by EECS for all the considered test functions are presented. Results (number of function evaluations and success rate) obtained in 100 independent runs are averaged and summarized in Table 2. For some of the test functions, SCGA is also applied. For EECS, the number of space divisions considered are displayed for each dimension and the number of generations in which the function converged to the optimal value.

**Table 2.** Results obtained by EECS for two dimension test functions. Comparisons with SCGA.

Fct. name	Fct.	name	No. of var	Div/Dim	Gen	Average no. of fct. eval			
						EECS	SCGA	EECS	SCGA
F2	Six	Hump	2	15	10	6282	-	100	-
		Camel Back							
F3	Matyas		2	10	10	2782	-	100	-
F4	Bohachevsky		2	6	2	106	460	100	99
		B1							
F5	Bohachevsky		2	6	2	106	471	100	99
		B2							
F6	Bohachevsky		2	6	2	106	468	100	100
		B3							
F7	Easom		2	10	10	2782	715	100	100
F8	Goldstein and Price		2	20	5	5192	191	100	100
F9	Shubert		2	20	10	11182	742	100	98
F10	Beale		2	10	10	2782	-	100	100
F11	Booth		2	10	10	2782	-	100	-
F12	Branin		2	20	10	11182	173	100	100
F13	Hump		2	20	20	23162	176	100	100
F14	Michalewicz		2	10	10	11182	179	100	100
F15	Function F		2	6	2	106	-	100	-
F16	Schwefel		2	20	10	11182	222	100	98
F17	Rosenbrock		2	10	15	4272	-	100	-
F18	Zakharov		2	10	10	2782	170	100	-
F19	De Jong		2	6	2	106	-	100	-
F20	Rastrigin		2	6	2	106	-	100	-
F21	Griewank		2	6	2	106	-	100	-
F22	Sum Squares		2	6	2	106	-	100	-
F23	Dixon and Price		2	10	10	11182	-	100	-
F24	Levi		2	15	5	9647	-	100	-
F25	Ackley		2	10	5	1292	-	100	-

In Table 3 results obtained by applying EECS for functions having a higher number of variables are presented.

As evident from Tables 2 and 3, EECS is always able to converge to the optimal solution (solutions). SCGA required some more information about the function. Function to optimize has to be a derivative function etc. For SCGA the success rate was not 100% for all the considered problems. Even if the required number of functions evaluations is smaller for some of the test functions than the number of functions evaluations required by EECS, the running time is almost same and sometime greater than the running time required by EECS. Another advantage of EECS is that the number of generations and space division does not depend on the domain of definition (minimum and maximum values for

**Table 3.** Results obtained by EECS for functions having three and more variables

		Function			
		DeJoung (F19)	Rastrigin (F20)	Griewank (F21)	Sum Squares(F22)
<b>3 Dimensions</b>					
Number of divisions/dimension		6	6	6	6
Number of generations		2	2	2	2
<b>4 Dimensions</b>					
Number of divisions/dimension		6	6	6	6
Number of generations		2	2	2	2
<b>6 Dimensions</b>					
Number of divisions/dimension		6	6	6	6
Number of generations		2	4	4	4
<b>8 Dimensions</b>					
Number of divisions/dimension		6	8	8	8
Number of generations		2	4	4	4

all the variables). For instance, for almost all the test functions considered in the experiments (F4, F5, F6, F10, F11, F17-F25), EECS will require the same number of space divisions and the same number of generations even for bigger range of values (for instance, [-1000, 1000] instead of [-10, 10]). This means that the search space is covered very well even if we consider only a small number of solutions for the initial populations.

## 4 Conclusions

The major advantage of EECS is its fast convergence. Both initial population generation and the way in which solutions are recombined contributes in covering the search space very effectively and ensuring a faster convergence. There are several test functions (considered in the above experiments) for which Evolutionary Elementary Cooperative Strategy converged very fast even by considering only few divisions of the search space (which implies a small population size). All these features - space division and equidistant population initialization, crossover only between neighbors and selection of the best individuals from the unified population of parents and offspring - generates a well exploration of the search space and guarantees a fast convergence. This technique can be also extended and applied to multiobjective optimization problems. Since EECS complexity increases



for larger number of variables, we are also exploring some ways to adapt the algorithm to make it more efficient in dealing with complex functions.

## Acknowledgements

This research was supported by the International Joint Research Grant of the Institute of Information Technology Assessment foreign professor invitation program of the Ministry of Information and Communication, Korea.

## References

1. Clausen, J., Zilinskas, A. Subdivision, sampling and initialization strategies for simplicial branch and bound in global optimization, *Computers and Mathematics with Applications*, 44, pp. 943-955, 2002
2. Csallner, A.E. Lipschitz continuity and the termination of interval methods for global optimization, *Computers and Mathematics with Applications*, 42, pp. 1035-1042, 2001
3. Floudas, C.A., Pardalos, P.M. A collection of test problems for constraint global optimization algorithms, Springer-Verlag, Berlin Heidelberg, 1990.
4. Grosan, C., Abraham, A. A simple strategy for nonlinear optimization, In *Proceedings of the Third International Conference on Neural, Parallel and Scientific Computation*, Atlanta, USA, 2006 (in press).
5. Hedar, A.R., Fukushima, M. Simplex coding genetic algorithm for the global optimization of nonlinear functions, In *Multi-Objective Programming and Goal Programming*, T. Tanino, T. Tanaka and M. Inuiguchi (Eds.), Springer-Verlag, Berlin-Heidelberg, pp. 135-140, 2003.
6. Laguna, M., Marti, R. Scatter search: methodology and implementations, Kluwer Academic Publishers, Dordrecht, 2003.
7. Laguna, M., Marti, R. Experimental testing of advanced scatter search designs for global optimization of multimodal functions, *Journal of Global Optimization*, 33, pp. 235-255, 2005.
8. Pardalos, P.M., Romejin, H.E. Handbook of global optimization, Kluwer Academic Publishers, Boston, MA, 2002.
9. Sun, M., Johnson, A.W. Interval branch and bound with local sampling for constrained global optimization, *Journal of Global Optimization*, 33, pp61-82, 2005.
10. Tsoulos, I.G., Lagaris, I.E. MinFinder: Locating all the local minima of a function, *Computer Physisc Communications*, 174, pp. 166-179, 2006.
11. Van Voorhis, T. A global optimization algorithm using Lagrangian underestimates and the interval Newton method, *Journal of Global Optimization*, 24, pp349-370, 2002.
12. [www.cyberiad.net/realbench.htm](http://www.cyberiad.net/realbench.htm) (accessed on May 20, 2006)
13. [www.solon.cma.univ.ie.ac.at/~neum/glopt/my\\_problems.html](http://www.solon.cma.univ.ie.ac.at/~neum/glopt/my_problems.html) (accessed on May 20, 2006)

# Human Hand Detection Using Evolutionary Computation for Gestures Recognition of a Partner Robot

Setsuo Hashimoto<sup>1</sup>, Naoyuki Kubota<sup>2</sup>, and Fumio Kojima<sup>3</sup>

<sup>1</sup> Faculty of Economics, Kyoto Gakuen University,  
1-1 Ootani, Nanjyo, Sogabe-cho, Kameoka, Kyoto, 621-8551, Japan  
setsuo-h@kyotogakuen.ac.jp

<sup>2</sup> Dept. of System Design, Tokyo Metropolitan University,  
1-1 Minami-Osawa, Hachioji, Tokyo 192-0397, Japan  
& PRESTO, Japan Science and Technology Agency  
kubota@comp.metro-u.ac.jp

<sup>3</sup> Dept. of Mechanical and Systems Engineering,  
Graduate School of Science and Technology, Kobe University,  
1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan  
kojima@cs.kobe-u.ac.jp

**Abstract.** This paper proposes a human hand detection method for gesture recognition used for communication of a robot with a human. A human hand motion can be recognized as a different meaning according to a situation in the communication between the human and robot. We propose a steady-state genetic algorithm for extracting a time series of human hand position. Finally, we discuss the effectiveness of the proposed method through several experimental results.

## 1 Introduction

Recently, various types of image processing methods have been developed as computational capabilities increase [1-4]. Thanks to real-time image processing, we can recognize both human expression according to the motion of body, face and hand. Furthermore, not only image processing, but also speech recognition technology has been improved. The integration of human motion recognition and speech recognition enables a robot to do a lot of things in human-friendly communication. Actually, various kinds of human-friendly robots has been cheaply produced and sold. However, a human hand motion, or a gesture can be recognized as a different meaning according to a situation in the communication between the human and the robot. In general, the gesture plays an important role in interaction and communication with a human, because the meanings of utterance can be emphasized by using gestures [8,9]. Furthermore, it is very natural and useful for a human to use a gesture in order to give the robot a specific task to share their attention. Therefore, a gesture is regarded as a symbolic action for communicating intention to the other. Accordingly, the robot needs to recognize gestures.

Various types of gesture recognition methods have been proposed so far [5-7]. The research stream can be classified into hand shape recognition and hand motion recognition. Since a hand is a complex object, it is very difficult to detect a hand from an image. Therefore, many researchers have used a simple background, colored gloves, and others. In general, skin color information is used for detecting a hand, we must distinguish the hand from its similar color of objects. Therefore, a hierarchical detection method of skin color region extraction, hand extraction, and hand shape recognition has been used in order to reduce computational cost. Hand shape recognition is performed by using hand contour or 3D hand model. On the other hand, hidden Markov Model has been used for extracting gesture sequence as one of hand motion recognition. If knowledge database of gesture sequence patterns is available in hand motion recognition, template candidates used in gesture recognition can be reduced. We should use both of hand shape recognition and hand motion recognition simultaneously. We also proposed alternative method of hand extraction and gesture recognition based on computational intelligence [10,11]. In our method, we used a blue glove to reduce the computational cost. Therefore, this paper proposes a method of human hand detection based on skin color and edge information. Next, we propose two methods of gesture recognition. The first one is used to extract the posture of the human arm. The position and posture of the human arm includes the meaning to be communicated. We apply a steady-state genetic algorithm [12,13] to identify the posture or configuration of the human arm. The other is used to extract a human hand motion. Finally, we show experimental results of gesture recognition for a partner robot, and discuss the effectiveness of the proposed method.

## 2 A Partner Robot

We developed a partner robot; MOBiMac as shown in Figure 1. Two CPUs are used for PC and robotic behaviors. The robot has two servo motors, eight ultrasonic sensors, and CCD camera. Therefore, the robot can take various actions such as collision avoiding, human approaching, and line tracing. The behavior modes used for this robot are human detection, human communication, behavior learning, behavioral interaction. The communication with a human is performed by the utterance as the result of voice recognition and human motion recognition. The behavior learning includes the reinforcement learning through interaction with the environment, and imitative learning through interaction with the human. The behavioral interaction includes the soccer and games with a human. In the following, we focus on image processing for gesture recognition.

The robot takes an image from the CCD camera, and extracts a human. If the robot detects the human, the robot extracts the motion of the human hand. According to the human hand motion, the robot decides the action outputs. Furthermore, the robot expresses the internal or perceptual state by utterance. A behavior of the robot can be represented using fuzzy rules based on simplified fuzzy inference [14,15]. The logical

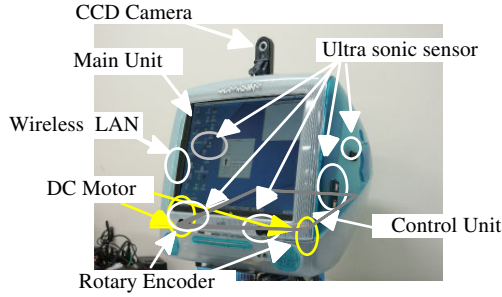


Fig. 1. A Partner Robot; MOBiMac

structure written by fuzzy rules is easy for humans to understand and to design. In general, a fuzzy if-then rule is described as follows,

**If**  $x_1$  is  $A_{i,1}$  and ... and  $x_m$  is  $A_{i,m}$  **Then**  $y_1$  is  $w_{i,1}$  and ... and  $y_n$  is  $w_{i,n}$

where  $A_{i,j}$  and  $w_{i,k}$  are a symmetric triangular membership function for the  $j$ th input and a singleton for the  $k$ th output of the  $i$ th rule;  $m$  and  $n$  are the numbers of inputs and outputs, respectively. Fuzzy inference is generally described by,

$$\mu_{A_{i,j}}(x_j) = \begin{cases} 1 - \frac{|x_j - a_{i,j}|}{b_{i,j}} & |x_j - a_{i,j}| \leq b_{i,j} \\ 0 & otherwise \end{cases} \tag{1}$$

$$\mu_i = \prod_{j=1}^m \mu_{A_{i,j}}(x_j) \tag{2}$$

$$y_k = \frac{\sum_{i=1}^R \mu_i w_{i,k}}{\sum_{i=1}^R \mu_i} \tag{3}$$

where  $a_{i,j}$  and  $b_{i,j}$  are the central value and the width of the membership function  $A_{i,j}$ ;  $R$  is the number of rules. Outputs of the robot are motor output levels. Fuzzy controller is used for collision avoidance and target tracing behaviors. The inputs to the fuzzy controller for collision avoidance and target tracing are the measured distance to the obstacle by ultrasonic sensors, and the relative direction and to a target point, respectively. Basically, a target point is generated by using the humans and objects on the image. The gesture recognition is composed of three stages; (1) human hand detection from a single image, (2) human arm posture extraction from a single image, and (3) human hand motion extraction from temporally sequential images (Fig 2). We explain these method in the following sections.

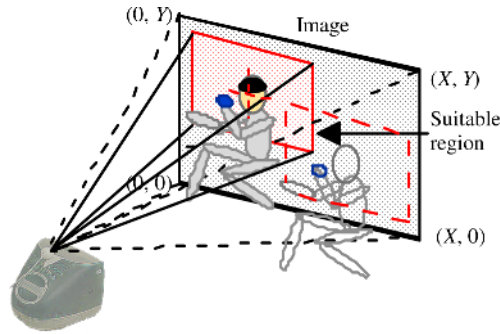


Fig. 2. Human detection by SSGA

### 3 Hand Detection for Gesture Recognition

The image of RGB color space is taken by CCD camera. Because the image processing takes much time and computational cost, the full size of image processing to every image is not reasonable. Therefore, we use the reduced size of image to detect a moving object, that is recognizable as a human. First, the robot calculates the center of gravity of the pixels different from the previous image as the differential extraction. The size of image used in the differential extraction is  $20 \times 15$ . The attention range is formed according the center of gravity. Next, the colors corresponding to human hair and skin are extracted by using thresholds. The region including the human face and hair colors are detected by using a steady-state genetic algorithm (SSGA) based on template matching. Figure 3 (a) shows a candidate solution of the template used for detecting a human as the target object. A template is composed of numerical parameters of  $g_{i,1}$ ,  $g_{i,2}$ ,  $g_{i,3}$  and  $g_{i,4}$ . The number of individuals is  $G$ . In SSGA, only few existing solutions are replaced with the candidate generated by genetic operators in each generation. In this paper, the worst candidate solution is eliminated ("Delete least fitness" selection), and it is replaced with the candidate solution generated by the crossover and the mutation. The fitness value is calculated by the following equation,

$$f_i^F = C^S + C^H + \eta_1 \cdot C^S \cdot C^H - \eta_2 \cdot C^{Other} \quad (4)$$

where  $C^S$ ,  $C^H$ , and  $C^{Other}$  indicate the numbers of pixels of the colors corresponding to human hair, human face, and other colors, respectively;  $\eta_1$  and  $\eta_2$  are coefficients. Therefore, this problem result in the maximization problem. We use the elitist crossover and adaptive mutation. The elitist crossover randomly selects one individual and generates an individual by combining genetic information from the randomly selected individual and the best individual. Next, the following adaptive mutation is performed to the generated individual,

$$g_{i,j} \leftarrow g_{i,j} + \left( \alpha_j \cdot \frac{f_{\max} - f_i}{f_{\max} - f_{\min}} + \beta_j \right) \cdot N(0,1) \quad (5)$$

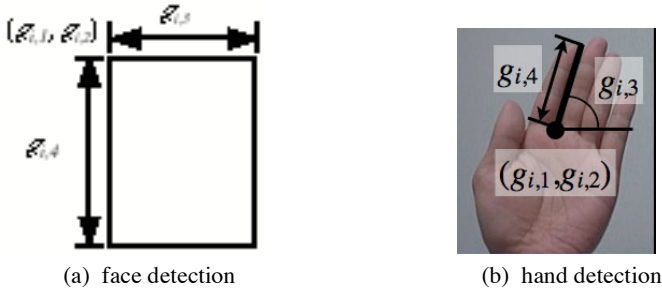


Fig. 3. A candidate solution of SSGA

where  $f_i$  is the fitness value of the  $i$ th individual,  $f_{max}$  and  $f_{min}$  are the maximum and minimum of fitness values in the population;  $N(0,1)$  indicates a normal random value;  $\alpha_i$  and  $\beta_j$  are the coefficient and offset, respectively. In the adaptive mutation, the variance of the normal random number is relatively changed according to the fitness values of the population. By using SSGA, the robot roughly detects a human face candidate. The SSGA for detecting a human face is called SSGA-1.

After detecting human face, the human hand detection is performed. A human hand is detected under the lighting condition of a finger edge. Figure 3 (b) shows a candidate solution for detecting a human hand where  $g_{i,1}$ ,  $g_{i,2}$ ,  $g_{i,3}$ , and  $g_{i,4}$  indicate the starting point of the finger edge, the angle and the length of the edge, respectively. First, the pixel corresponding to an edge is extracted by calculating difference among neighboring pixels as preprocessing. Here the length of edge is gradually extended according to the change of fitness values. The fitness value is calculated by the following equation,

$$f_i^H = C^S + C^E + \eta_1 \cdot C^S \cdot C^E - \eta_2 \cdot C^{Other} \tag{6}$$

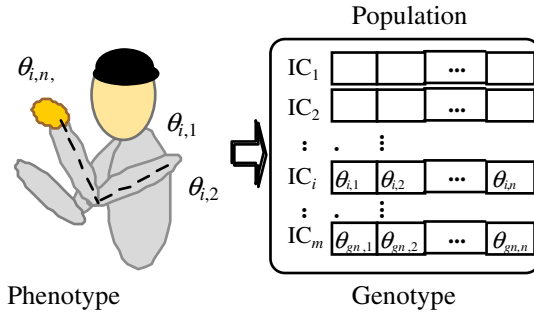
where  $C^E$  indicate the number of pixels of the colors corresponding to the edge. The pixel corresponding to an edge is extracted by calculating difference among neighboring pixels. The SSGA for detecting a human hand is called SSGA-2.

A 3D posture of a human arm is mapped into a 2D image. Therefore, we must solve an inverse problem from the 2D image to 3D posture. Here a configuration  $\theta$  of a human arm is expressed by a set of joint angles, because all joints are revolute,

$$\theta = (\theta_1, \theta_2, \dots, \theta_n)^T \in R^n \tag{7}$$

where  $n$  denotes the DOF of a robot arm. The number of DOF of a human arm is assumed to be 5 ( $n = 5$ ) in this paper. In addition, the position of the human hand is expressed as  $P = (p_x, p_y, p_z)^T$  on the base frame. SSGA is applied to detect the joint angles corresponding to the human arm position on the image. Here the SSGA for detecting the joint angles of the human arm is called SSGA-3.

An individual is composed of all joint variables (Fig 4). Initialization updates the population based on the previous best configuration to the next image. The  $j$ th joint angle of the  $i$ th configuration  $\theta_{i,j}$ , which is represented as a real number, is generated as follows ( $i=1, 2, \dots, g_n$ ),



**Fig. 4.** A candidate solution for extracting arm posture

$$\theta_{i,j} \leftarrow \theta_j^* + \beta_j^t \cdot N(0,1) \quad (8)$$

where  $\theta_j^*$  is the previous best joint angles;  $\beta_j^t$  is a coefficient for the  $j$ th joint angle. We use a following fitness function,

$$f_i = f_p + \eta^T f_d \quad (9)$$

where  $\eta^T$  is a weight coefficient. The first term,  $f_p$ , denotes the distance between the hand position and the target point. The second term,  $f_d$ , denotes the sum of squares of the difference between each joint angle between two configurations of  $t$  and  $t-1$ . A selection removes the worst individual from the current population. Next, an elitist crossover is performed. Consequently, the worst individual is replaced with the individual generated by the elitist crossover. Furthermore, we use the adaptive mutation.

## 4 Experimental Results

This section shows experimental results of the detection of human hand using steady-state genetic algorithm (SSGA). The size  $(X,Y)$  of an image is  $(160, 120)$ . The number of individuals and iterations in SSGA are 50 and 3000, respectively. Figure 5 show an image processing result of extraction of a human hand. The human hand moves in the original image as shown in Fig.5 (a). By using SSGA, the robot extracts human hand continuously.

## 5 Summary

This paper proposes a human hand detection method for gestures recognition of a partner robot based on human arm posture and human hand motion. We applied steady-state genetic algorithm for image processing. Human arm posture is static information, while the human hand motion is dynamic position. These can be used for expressing the human feeling and pointing direction, respectively. Experimental results show that the robot can extract and classify the human hand motion.

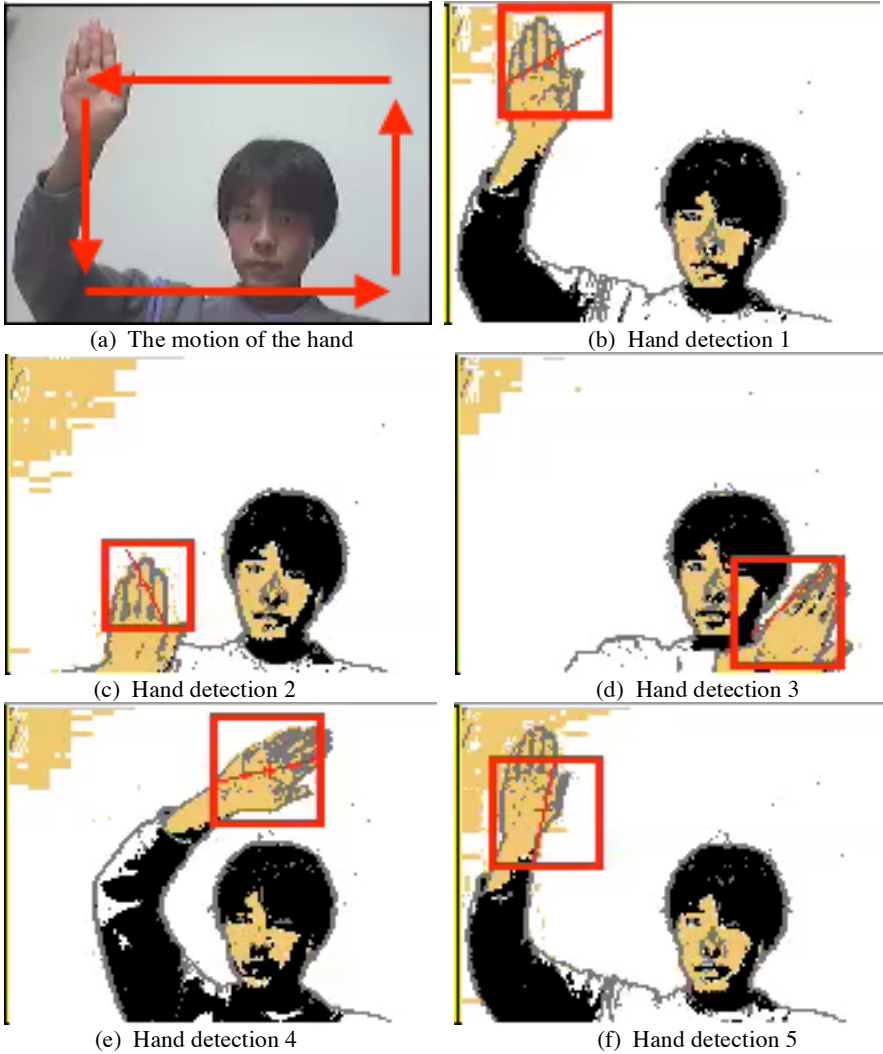


Fig. 5. Image processing results

We will develop the method for extracting the meanings of gestures through interaction with a human. Furthermore, as a future work, we intend to incorporate the obtained action patterns into reactive motions of the robot interacting with human.

## References

1. D. Marr, Vision, W. H. Freeman, San Francisco (1982)
2. K. Fukushima: Neural network model for selective attention in visual pattern recognition and associative recall, Applied Optics, 26 (1987) pp. 4985-4992.



3. P. Dayan, S. Kakade, P.R. Montague: Learning and selective attention, *Nature Neuroscience* 3 (2000) pp. 1218-1223.
4. H. H. Bulthoff, S.W. Lee, T.A. Poggio, & C. Wallraven: Biologically Motivated Computer Vision, Springer-Verlag (2002)
5. Y. Hamada, N. Shamada, Y. Shirai: Hand Shape Estimation Using Image Transition Network, *Proc. of Workshop on Human Motion* (2000) pp.161--166
6. E. Ong, R. Bowden: A Boosted Classifier Tree for Hand Shape Detection, *Sixth IEEE International Conference on Automatic Face and Gesture Recognition* (2004) pp. 889-894
7. V. Athitsos, S. Sclaroff: An Appearance-Based Framework for 3D Hand Shape Classification and Camera Viewpoint Estimation, *Proc. of Face and Gesture Recognition* (2002)
8. R.P.N.Rao, A.N.Meltzoff: Imitation Learning in Infants and Robots: Towards Probabilistic Computational Models, "Proceedings of Artificial Intelligence and Simulation of Behaviors" (2003)
9. S. Calinon and A. Billard: Stochastic Gesture Production and Recognition Model for a Humanoid Robot, *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems* (2004) pp. 2769-2774
10. N. Kubota: Computational Intelligence for structured Learning of A Partner Robot Based on Imitation, *Information Sciences*, No.171 (2005) pp. 403-429
11. N. Kubota and M. Abe: Interactive Learning for A Partner Robot Based on Cyclic Gestures, *Proc. (CD-ROM) of The 5th International Conference on Simulated Evolution And Learning* (2004)
12. D. B. Fogel: *Evolutionary computation*, IEEE Press (1995)
13. G. Syswerda: A study of reproduction in generational and steady-state genetic algorithms, *In foundations of genetic algorithms*, Morgan Kaufmann Publishers, Inc. (1991)
14. J-SR. Jang, C-T. Sun, E. Mizutani: *Neuro-fuzzy and soft computing*, Prentice-Hall, Inc. (1997)
15. T. Fukuda, N. Kubota: An intelligent robotic system based on a fuzzy approach, *Proc. of IEEE* 87(9) (1999) pp.1448-1470

# A Simple 3D Edge Template for Pose Invariant Face Detection

Stephen Karungaru<sup>1</sup>, Minoru Fukumi<sup>1</sup>, Norio Akamatsu<sup>1</sup>, and Takuya Akashi<sup>2</sup>

<sup>1</sup> University of Tokushima, 2-1, Minami-Josanjima, Tokushima 770-8506, Japan  
karunga@is.tokushima-u.ac.jp

<sup>2</sup> Yamaguchi University 2-16-1 Tokiwadai, Ube, Yamaguchi, 755-8611, Japan  
akashi@eee.yamaguchi-u.ac.jp

**Abstract.** In this paper, a simple 3D edge based template for pose invariant face detection creation using side and front profiles of a general face is proposed. The 3D template is created using the edges that are always most likely to be extracted from a face given a certain pose. Bezier curves are used to create the template. When testing the template, genetic algorithms are used to guide the matching process thereby greatly reducing the total computation time required. The genetic algorithm automatically calculates the angle and the size of the template during the matching. An average pose invariant face detection accuracy of 84.6% was achieved.

**Keywords:** 3D edge template, genetic algorithms, bezier curves and template matching.

## 1 Introduction

Face detection is an interesting topic in computer vision. It is the preprocessing step required in most face image oriented applications like face recognition, gesture recognition, camera mouse, etc. Therefore, if face detection does not achieve high detection accuracy, then the reliability of the systems using face detection as a preprocessing step cannot be guaranteed. Research in frontal face detection has produced almost 100% results especially in off-line images. However, reproducing the results in on-line images has not been as successful. In real time face detection, many issues, e.g. scene illumination, camera focus and shaking, pose, must be resolved. Of these, the major handle that must be cleared is pose invariance. We think that the perfect face detection preprocessor should be able to detect face of all poses and also rotate and transform them to the most likely frontal pose. If such a face detector were realized, then the face-based applications mentioned earlier would also be easier to design and implement since they only need to handle frontal faces. As a first step towards such a face detector, this paper proposes a high-speed face pose detection using a 3D edge template. The template is created using the edges extracted from the front and side views of a general face. In this design, not all the edges in the face are used as the core points in the design of the 3D template. We base our design on the edges that can be detected most of the times given any face orientation. To

determine what edges to use, 100 faces of different orientation were used. The edge detection is carried out using the Laplacian of Gaussian (LOG) filter. The bezier function is used to construct the 3D template. During testing, a genetic algorithm is used to guide the template matching process.

Some related works in image segmentation and edge extraction include the development of a framework for image segmentation based on the intuition that there should be evidence for a boundary between each pair of neighboring regions [1], unsupervised texture segmentation that relies on statistical tests as a measure of homogeneity [2], a spectral graph method that combines segmentation and detection [3], etc. in [4] a linear algorithm that uniquely recovers the 3D non-rigid shapes and poses of a human face from a 2D monocular video is described.

## 2 Core Edges Determination

To create a 3D edge template for use in pose invariant face detection, it is important to decide the core edges that should be used. These edges should always be detected in a face regardless of its pose. The edges will also be used during the genetic algorithm guided template matching process.

### 2.1 Face Database

The database used to estimate the average face contains 100 faces of various orientation and sizes. The faces were collected from a variety of place including the FERET [5], University of Oulu [6] face color databases and some taken using a digital camera in our lab. Some of the images used are shown below.



**Fig. 1.** Example faces from the face database

## 2.2 Face Edges Extraction

The edges inside the faces are extracted using the Laplacian of Gaussian filter, followed by a median filter to clean out the noise generated. The filters are only applied inside the face regions.

The first filter applied is the Laplacian of Gaussian (LoG), to smooth and detect the edges of the facial features. A large kernel (e.g. 15x15) when used with this filter produces good edge detection, Fig.3 (a), but the computation time rises exponential with the size of the kernel. After several experiments, a trade-off kernel of 7x7 was employed. However, a lot of noise is produced, Fig.3 (b). Therefore, a 3x3 medium filter was then applied to reduce this noise. Fortunately, the cost of calculation when using a 7x7 Laplacian of Gaussian filter followed by a 3x3 medium filter is less than using a 15x15 Laplacian of Gaussian that produced similar results, Fig.3 (c).

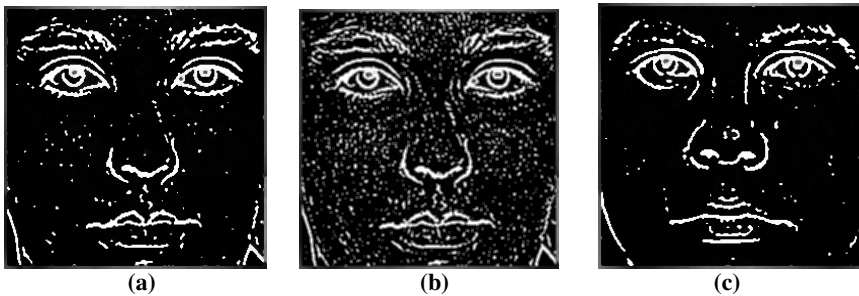


Fig. 2. Edge detection (a) 15x15 LOG, (b) 7x7 LOG, (c) 7x7 LOG and median filter

However, not all edges can be extracted from the face in all orientations. Using the 100 face images of different orientations, edge extraction was carried out to determine which of the edges could be extracted most often (over 90%). These edges, called the core edges, are the full or parts of eclipses around the eyes and mouth, the area around the nose and the line extending from the tip of the nose to the center of the eyes (nose line). The core edges are used during matching.

## 3 3D Face Model

A 3D facemask is constructed from the average of 100 subjects frontal and side profile's edges. The objective is not to create an accurate 3D face model but a mask that can help estimate the pose of the face. Using the edge information, it is possible to estimate the orientation of a face by matching the edges to the 3D facemask using a genetic algorithm.

The 3D facemask is then constructed using the two averaged face profiles (side profile is used for depth), the core points, the general relationships between them and Bezier curves [7], Fig. 3. Based on the core edges, there are 35 control points and 84 interconnection points. The size of the 3D template created is 30x30pixels. This size is selected because it is small enough to ensure high-speed operation and contains enough data for accurate template matching.

For any given set of orientation angles along the x, y and z-axis, the 3D edge template can be reconstructed in only 15 milliseconds. This high reconstruction speed contributes heavily to the total system time.

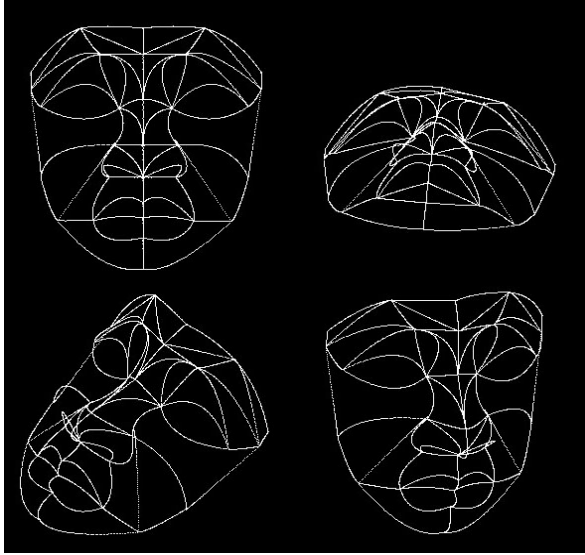


Fig. 3. Some views of the 3D edge template constructed

Note that not all the edges shown in Fig. 3 are used during matching. The orientation angle and the core edges determine the “visible” edges and thus can be used during matching.

### 3.1 GA Guided Template Matching

Template matching using the 3D edge template constructed can be performed using a genetic algorithm. The genetic algorithm is necessary because the size and orientation of the faces can vary greatly while the template constructed is of fixed size. The idea is, instead of matching the template detector pixel by pixel throughout the image and re-sampling (size and orientation), the GA is used to optimize the search by automatically selecting the position, orientation and the size of the 3D edge template. The test sample is then matched to the 3D edge template.

The frame size used is 320x240 pixels. This size is also the largest size of a face to be encountered in this system. This information is used to determine the length of the GA chromosome. This GA is binary coded. The reason of using binary representation is to be able to manipulate the parameters of the chromosomes. The chromosome bits represent:

1. Position: The x-axis position needs 9 (bits needed to code 320 (width) in binary) chromosomes and the y-axis 8 (to code 240 (height)).
2. Orientation: 7 bits to code 100 degrees in each direction. Fractional changes are not considered. For orientations above 100 degrees, accurate extraction of the core edges cannot be guaranteed because most of the core edges are not visible.

Note that the orientations are along the x, y and z-axis. Therefore in total 21 bits are required.

3. Scale: Needs eight bits. The scale is made up of a whole and decimal part. The shorter image dimension divided by the default training image size gives the maximum scale. That is 240/30 giving 8.0 as the maximum scale. This requires four bits to code in binary. The fraction is in one decimal place with a maximum of nine and thus needs four bits to code as well.

Therefore, the GA chromosome length is 46. The size, position and orientation of the samples can thus be reduced or increased, normalized and input into the basic face detector. The GA parameters are shown in table 1.

**Table 1.** GA parameters

Genetic Algorithm Parameters	
Unit	Bits
X-axis position	9
Y-axis position	8
Orientation	21
Scale	8

The mutation rate is 0.001, the rank selection method is used and there are multiple crossover points. During crossover, the child does not inherit all the parents' information. There is an element of random filling of chromosome bits. This method was introduced because under some circumstances, the search is trapped in local minima.

The initial population is set at 20 individuals assuming one person per frame. There are multiple crossover points randomly determined depending on the total fitness of the parents. The higher the fitness the fewer the number of crossover points are.

The roulette wheel selection method is used. The fitness function of the genetic algorithm is;

$$Fitness = \left( 1 - \frac{(3C_e - IE_e)}{(3C_e^2 + IE_e^2)} \right) \tag{1}$$

Where:  $IE_e$  is the accuracy of the target image edges template and  $3C_e$  is the 3D edge template.

During reproduction, the top 50% of the fittest individuals were used to reproduce 75% of the next population. The remaining 25% of the next population is reproduced by selection of their parents at random from all of the initial population. This method improves the search by ensuring that not only the best individuals are selected for

reproduction but also that the rest of the population is explored for other possible candidates.

## 4 Results

The test set contains 300 images. The size of the images we used is 320x240pixels. The images used in this work are from the physics-based face database from the University of Oulu, the FERET color database and others we took ourselves.

Each image is first passed through a skin color detector [8] to detect the skin color regions. We assume that faces can only be found in these regions. The edges of the face regions are then extracted using the LOG filter. After this, genetic algorithm guided template matching is applied. The results are shown in table 2.

**Table 2.** GA parameters

	Accuracy (%)
Core edges Extraction	98
Face Detection Result	84.6

Of the 300 images in our test database, in 294 of them, the core edges were accurately detected. Note that this accuracy is base on our visual observation. The result for face detection was 84.6% (253/300).

## 5 Conclusion

In this paper, a simple 3D edge template for pose invariant face detection was constructed. We achieved an average face detection accuracy of 84.6% using the images in the test database. Although the accuracy is not high, based on the simplicity of the 3D model created, we think that this result is acceptable. Problems with this method could have arisen from our verification experiment using the genetic algorithm. In future, we will further increase the images in our database and use other edge detection filters to improve the extraction of the core edges to about 100%.

## References

1. Hofmann, T., Puzicha, J., Buhmann, J. M., Unsupervised Texture Segmentation in a Deterministic Annealing Framework, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 20(8).
2. Felzenszwalb, P., Huttenlocher, D., Image Segmentation Using Local Variation, Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pages 98-104,1998.

3. Yu, S., Gross, R. and Shi, J., Concurrent Object Segmentation and Recognition with Graph Partitioning, Neural Information Processing Systems, NIPS,
4. Xiao, J, Baker S., Matthews, I., and Kanade, T., Real-Time Combined 2D+3D Active Appearance Models, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June, 2004.
5. Phillips, P. J., Moon, H., Rauss, P. J., and Rizvi, S., The FERET evaluation methodology for face recognition algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 10, October 2000.
6. Soriano, M., Marszalec, E., Pietikainen, M., Physics-based face database for color research, Journal of Electronic Imaging, 9(1) 32-38, 2000.
7. Demidov, E. An Interactive Introduction to Splines, Retrieved April 10 2005 from <http://www.ibiblio.org/e-notes/Splines/Intro.htm>, 2003.
8. Karungaru, S., Fukumi, M. and Akamatsu, N., Detection of human face in visual scenes” Proc of ANZIIS, pp.165-170, 2001.



# Emergence of Flocking Behavior Based on Reinforcement Learning

Koichiro Morihiro<sup>1,2</sup>, Tejiro Isokawa<sup>2,3</sup>, Haruhiko Nishimura<sup>4</sup>,  
and Nobuyuki Matsui<sup>2,3</sup>

<sup>1</sup> Hyogo University of Teacher Education, Hyogo 673-1494, Japan  
`mori@info.hyogo-u.ac.jp`

<sup>2</sup> Himeji Institute of Technology, Hyogo 671-2201, Japan

<sup>3</sup> Graduate School of Engineering, University of Hyogo, Hyogo 671-2201, Japan  
`isokawa@eng.u-hyogo.ac.jp`, `matsui@eng.u-hyogo.ac.jp`

<sup>4</sup> Graduate School of Applied Informatics, University of Hyogo,  
Hyogo 650-0044, Japan  
`haru@ai.u-hyogo.ac.jp`

**Abstract.** Grouping motion, such as bird flocking, land animal herding, and fish schooling, is well-known in nature. Many observations have shown that there are no leading agents to control the behavior of the group. Several models have been proposed for describing the flocking behavior, which we regard as a distinctive example of the aggregate motions. In these models, some fixed rule is given to each of the individuals a priori for their interactions in reductive and rigid manner. Instead of this, we have proposed a new framework for self-organized flocking of agents by reinforcement learning. It will become important to introduce a learning scheme for making collective behavior in artificial autonomous distributed systems. In this paper, anti-predator behaviors of agents are examined by our scheme through computer simulations. We demonstrate the feature of behavior under two learning modes against agents of the same kind and predators.

## 1 Introduction

Grouping motion of creatures is observed in various scenes in nature. As its typical cases, bird flocking, land animal herding, and fish schooling are well-known. Many previous observations suggest that there are no leaders to control the behavior of the group; rather it emerges from the local interactions among individuals in the group[1,2,3]. Several models have been proposed for describing the flocking behavior. In these models, a fixed rule is given to each of individuals a priori for their interactions[4,5,6]. This reductive and rigid approach is plausible for modeling flocks of biological organisms, for they seem to inherit the ability of making a flock. However what is more, it will become important to introduce a learning scheme for making collective behavior. In a design of artificial autonomous distributed system, fixed interactive relationships among agents (individuals) lose the robustness against nonstationary environments. It

is necessary for agents to be able to adjust their parameters of the ways of interactions. Some learning framework to form individual interaction will be of importance. In addition to securing the robustness of systems, this framework will give a possibility to design systems easier, because it determines the local interactions of agents adaptively as a certain function of the system.

Reinforcement learning[7,8] characterizes its feature of the unsupervised learning introducing a process of trial and error called exploration to maximize the reward obtained from environment. Introducing appropriate relations between the agent behavior (action) and its reward, we could make a new scheme for flocking behavior emergence by reinforcement learning. We have proposed an adaptive scheme for self-organized making flock of agents[9]. Each of agents is trained in its perceptual internal space by Q-learning, which is a typical reinforcement learning algorithm[10]. In this paper, anti-predator behaviors of agents are examined by our scheme through computer simulations. We demonstrate the feature of behavior under two learning modes against agents of the same kind and predators.

## 2 Reinforcement Learning

### 2.1 Q-Learning

Almost all reinforcement learning algorithms are based on estimating value functions. The system gets only an evaluative scalar feedback of a value function from its environment, not an instructive one as in supervised learning. Q-learning is known as the best-understood reinforcement learning algorithm. The value function in Q-learning consists of values decided from a state and an action, which is called Q-value. In Q-learning, proceedings on learning consist of acquiring a state ( $s_t$ ), deciding an action ( $a_t$ ), receiving a reward ( $r$ ) from an environment, and updating Q-value ( $Q(s_t, a_t)$ ). Q-value is updated by the equation written as follows:

$$Q(s_{t+1}, a_{t+1}) = Q(s_t, a_t) + \alpha[r + \gamma \max_{a' \in A(s')} Q(s', a') - Q(s_t, a_t)] \quad (1)$$

where  $A$  denotes a set of actions,  $\alpha$  is the learning rate ( $0 < \alpha \leq 1$ ),  $\gamma$  is the discount rate ( $0 \leq \gamma \leq 1$ ). Q-learning is one of the reinforcement learning algorithms to maximize the sum of the reward received. It attempts to learn the optimal policy by building a table of Q-value  $Q(s, a)$  according to the above update equation.  $Q(s, a)$  shows the estimated value of expected return taking action  $a$  in state  $s$ . Once these Q-values have been learned, the optimal action from any state is the one with the highest Q-value. In the original Q-learning algorithm, the greedy policy with pure exploitation is used. By employing this policy, however, it is generally difficult to obtain satisfactory results. So at present, some exploration which allows to adopt a non-optimal action is introduced.

### 2.2 Action Choice Generator

In the reinforcement learning, many kinds of exploration policies have been proposed as a process of trial and error such as  $\epsilon$ -greedy, softmax, and weighted

roulette action selection. Here, we adopt softmax action selection, and the rule is given as follows:

$$p(a|s) = \frac{\exp\{Q(s, a)/T\}}{\sum_{a_i \in A} \exp\{Q(s, a_i)/T\}} \tag{2}$$

where  $T$  is a positive parameter called the temperature. High temperatures cause the actions to be all (nearly) equi-probable, and low temperatures cause a greater difference in selection probability for actions that differ in their value estimates.

### 3 Model and Method

In this section, we introduce a scheme of perceptual internal space as the Q-value coordinates in the situation that an agent perceives (finds) another one among the others.

#### 3.1 Perceptual Internal Space for Each Agent

We employ a configuration where  $N$  agents that can move to any direction are placed in a two-dimensional field. The agents act in discrete time, and at each time-step an agent (agent  $i$ ) finds other agent (agent  $j$ ) among  $N - 1$  agents. In the perceptual internal space, the state  $s_t$  of  $Q(s_t, a_t)$  for the agent  $i$  is defined as  $[R]$  by Gauss' notation, the maximum integer not surpassing the Euclidean distance from agent  $i$  to agent  $j$ ,  $R$ . For the action  $a_t$  of  $Q(s_t, a_t)$ , four kinds of action patterns ( $a_1, a_2, a_3, a_4$ ) are taken as follows, illustrated in Fig.1.

- $a_1$  : Attraction to agent  $j$
- $a_2$  : Parallel positive orientation to agent  $j$  ( $\mathbf{m}_a \cdot (\mathbf{m}_i + \mathbf{m}_j) \geq 0$ )
- $a_3$  : Parallel negative orientation to agent  $j$  ( $\mathbf{m}_a \cdot (\mathbf{m}_i + \mathbf{m}_j) < 0$ )
- $a_4$  : Repulsion to agent  $j$

where  $\mathbf{m}_a$  is the directional vector of  $a_t$ ,  $\mathbf{m}_i$  and  $\mathbf{m}_j$  are the velocity vectors of agent  $i$  and agent  $j$ , respectively. When the velocities of agents are set to be one body-length (1 BL), then  $|\mathbf{m}_a|=|\mathbf{m}_i|=|\mathbf{m}_j|=1(\text{BL})$ . Agent  $i$  moves according to  $\mathbf{m}_i$  at each time-step, and  $\mathbf{m}_i$  is updated by

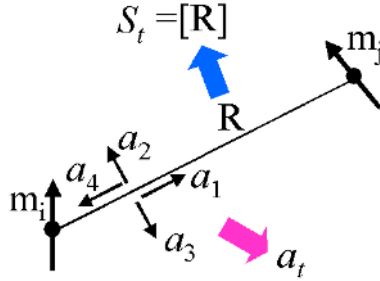
$$\mathbf{m}_i \leftarrow \frac{(1 - \kappa)\mathbf{m}_i + \kappa\mathbf{m}_a}{|(1 - \kappa)\mathbf{m}_i + \kappa\mathbf{m}_a|} \tag{3}$$

where  $\kappa$  is a positive parameter( $0 \leq \kappa \leq 1$ ), called the inertia parameter.

In this work, as we consider the same kind of agent and a predator as the perceptual objects, two sorts of the corresponding Q-value should be introduced.

#### 3.2 Learning Mode Against Agents of the Same Kind

In our proposed model, we prepare the reward  $r$  for  $(s_t, a_t)$  of each agent according to the distance  $R$  from the perceived agent of same kind. The learning



**Fig. 1.** Constitution of perceptual internal space for each agent

of the agents proceeds according to a positive or negative reward, as shown in Table 1, where  $R_1$ ,  $R_2$ , and  $R_3$  have the relationship of  $R_1 < R_2 < R_3$ . In case  $0 < [R] \leq R_3$ , agent  $i$  can perceive another agent of same kind with the probability in proportion to  $R^{-\beta}$ , where  $\beta$  is a positive parameter. This means that the smaller  $R$  value is, the easier the agent at that position is selected. When  $0 < [R] \leq R_1$ , the agent gets the positive reward (+1) if it takes the repulsive action against the perceived agent ( $a_4$ ); otherwise it gets the penalty (-1). In the cases of  $R_1 < [R] \leq R_2$  and  $R_2 < [R] \leq R_3$ , the agent also gets the reward or penalty defined in Table 1 with respect to the actions. In case  $[R] > R_3$ , agent  $i$  cannot perceive agent  $j$ , and then receives no reward and chooses an action from the four action patterns ( $a_1, a_2, a_3, a_4$ ) randomly.

**Table 1.** Reward  $r$  for the selected action  $a_t$  in the state  $s_t = [R]$  against the same kind of agent

$s_t$	$0 < [R] \leq R_1$	$R_1 < [R] \leq R_2$	$R_2 < [R] \leq R_3$	$R_3 < [R]$			
$a_t$	$a_4$	$a_1, a_2, a_3$	$a_2$	$a_1, a_3, a_4$	$a_1$	$a_2, a_3, a_4$	$a_1, a_2, a_3, a_4$
$r$	1	-1	1	-1	1	-1	0

### 3.3 Learning Mode Against Predators

When there is a predator within  $R_3$ , agent  $i$  perceives the predator with the probability 1 and the above learning mode is switched to this mode. In this case, the agent  $i$  gets the positive reward (+1) if it takes the repulsive action to evade the predator ( $a_4$ ); otherwise it gets the penalty (-1) as defined in Table 2.

## 4 Simulations and Results

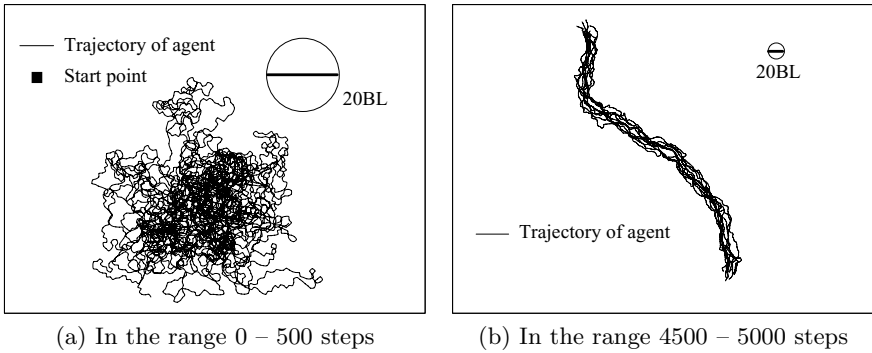
To demonstrate our proposed scheme in computer simulations, we take the following experimental conditions:  $\alpha = 0.1$ ,  $\gamma = 0.7$  in Eq.(1),  $T = 0.5$  (under learning) in Eq.(2),  $\kappa = 0.5$  in Eq.(3) and  $\beta = 0.5$  for the distance dependence of  $R^{-\beta}$ . All the velocities of the same kind of agents are set to one body-length (1BL) and that of the predator is set to two body-length (2BL).

**Table 2.** Reward  $r$  for the selected action  $a_t$  in the state  $s_t = [R]$  against predator

$s_t$	$0 < [R] \leq R_3$	$R_3 < [R]$
$a_t$	$a_4 \vdots a_1, a_2, a_3$	$a_1, a_2, a_3, a_4$
$r$	$1 \vdots -1$	0

**4.1 No Predator Case**

We simulated our model in the case of the number of agents  $N = 10$ , and  $R_1 = 4$  (BL),  $R_2 = 20$  (BL) and  $R_3 = 50$  (BL). Figures 2(a) and 2(b) show the trajectories in the range 0–500 steps and 4500–5000 steps under learning by Table 1, respectively. In Fig. 2(a) each of the agents changes its direction very often without regard to the other agents’ behavior, but it becomes to keep the direction for long time-step with the others as shown in Fig. 2(b). This indicates that the learning mode against the same kind of agent works well in flocking.



**Fig. 2.** The trajectories of agents under learning by Table 1 in the case of  $N = 10$ ,  $(R_1, R_2, R_3) = (4, 20, 50)$  with no predator

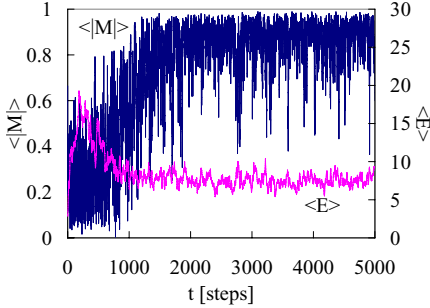
In order to evaluate how the agents make flocking behavior quantitatively, we introduce a measure  $|\mathbf{M}|$  of the uniformity in direction and a measure  $E$  of the spread of agents:

$$|\mathbf{M}| = \frac{1}{N} \left| \sum_{i=1}^N \mathbf{m}_i \right|, \tag{4}$$

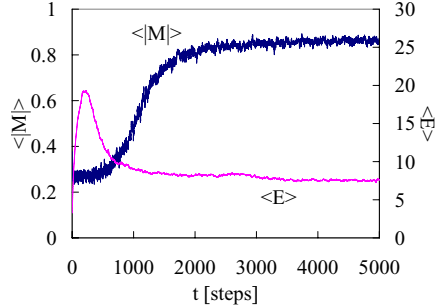
$$E = \frac{1}{N} \sum_{i=1}^N \sqrt{(x_A^i - x_G)^2 + (y_A^i - y_G)^2}, \tag{5}$$

where  $(x_A^i, y_A^i)$  and  $(x_G, y_G)$  are the two-dimensional coordinate of the agent  $i$  and the barycentric coordinate among the agents, respectively. The value of  $|\mathbf{M}|$  becomes closer to 1 when the directions of agents increase their correspondence. The agents come close together when the value of  $E$  gets small.

Figure 3 shows the time-step dependences of  $|\mathbf{M}|$  and  $E$  for this case. The transition of  $|\mathbf{M}|$  evolves good except for the fluctuation owing to the exploration effect in every time-step. The value of  $E$  takes a large value at the early stage of learning, and then it decreases with fluctuation to the value around 8 with the proceeding of learning. We further take the average of 100 events by repeating the above simulation with various random series in exploration. As a result, Fig. 4 is obtained in which the value of  $|\mathbf{M}|$  increases up to near 0.9 and the value of  $E$  decreases to 8.



**Fig. 3.** The time-step dependence of  $|\mathbf{M}|$  (Eq.(4)) and  $E$  (Eq.(5)) for the case in Fig. 2



**Fig. 4.** The step-time dependence of the averaged  $|\mathbf{M}|$  and  $E$  in 100 events of Fig. 3 case

### 4.2 The Case Predator Appears Later

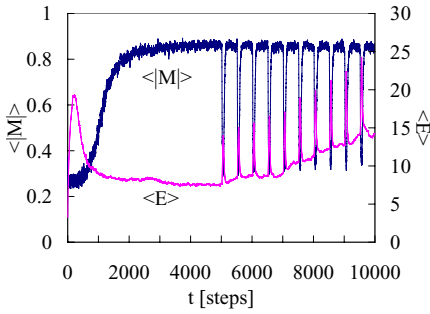
Each agent learns its flocking behavior without predator up to first 5000 time-steps and then a predator appears first. The predator approaches the center of agents from behind of them and passes straight. The predator appears in every 500 time-steps up to 10000 time-steps. The velocity of the predator is twice of the agent. Figure 5 shows the average of 100 events as like as in Fig. 4.

When the predator appears, the learning mode is changed. So,  $E$  takes a large value and  $|\mathbf{M}|$  decreases down to near 0.3. This means that the agents do not make flocking behavior. When the predator disappears, the learning mode is changed back.  $E$  takes a small value and  $|\mathbf{M}|$  increases up again to near 0.9 because of the re-flocking behavior of agents.

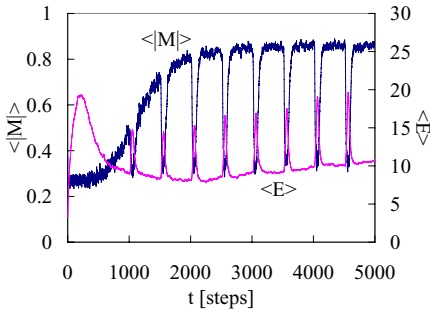
Figure 6 shows the trajectories of agents including the range of the 10th learning stage against predator (9500–9599 steps) of an event in Fig. 5. The agents in the first learning stage (5000–5099 steps) were panicked, but they learned to evade the predator in the 10th stage.

### 4.3 The Case Predator Appears from the Beginning

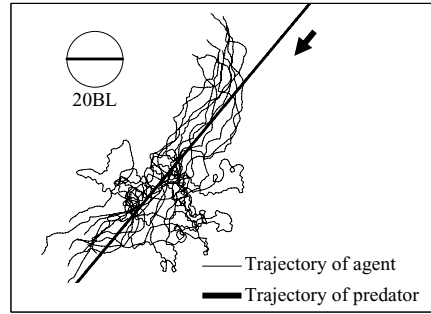
Figure 7 shows the case the predator appears from the beginning at the 1000th time-step in Fig. 4. Before the agents learn the flocking behavior enough, they begin to learn the escape from the predator. In this case, similar tendency as in Fig. 5 can be seen.



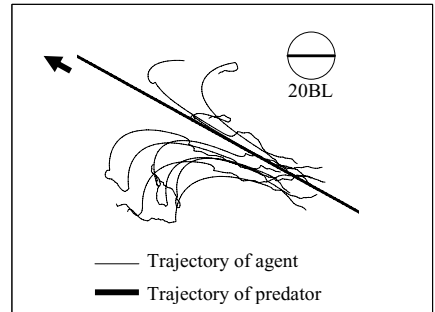
**Fig. 5.** The step-time dependence of the averaged  $|M|$  and  $E$  in 100 events of the case predator appears later



**Fig. 7.** The case the predator appears from the beginning at the 1000th time-step of Fig. 4



**Fig. 6.** The trajectories of agents in the range of 9450-9649 steps of an event in Fig. 5



**Fig. 8.** The magnified trajectories of agents after the learning in the range of 1500 – 1599 steps of an event in Fig. 7

Figure 8 is the magnification in the range of 1500 – 1599 steps of the trajectories of agents after the learning of an event in Fig. 7. After the learning, there is no fluctuation necessary for the exploration under the learning ( $T=0$ ). Finding the predator, the agents draw the shape like a fountain. This suggests that adaptive behaviors of agents including the escape from the predator emerge as a result of two mode learning.

## 5 Conclusion

We demonstrated a scheme for autonomously making flock of agents by reinforcement Q-learning. In addition to the flocking behavior of agents, the anti-predator behavior to escape from predator can emerge as a result of learning. This indicates the adaptive flexibility of our proposed scheme. In order to confirm whether our scheme is effective for various situations as to the patterns of escaping behavior or the heterogeneity of the group, we proceed further investigations.

## References

1. E. Shaw: "Schooling Fishes", *American Scientist*, 66:166–175, 1978.
2. B. L. Partridge: "The structure and function of fish schools", *Scientific American*, 246:90–99, 1982.
3. T. J. Pitcher and C. J. Wyche: "Predator avoidance behaviour of sand-eel schools: why schools seldom split." In *Predators and Prey in Fishes*, D.L.G. Noakes, B. G. Lindquist, G. S. Helfman and J. A. Ward (eds). The Hague: Junk, pp.193–204, 1983.
4. I. Aoki: "A Simulation Study on the Schooling Mechanism in Fish", *Bulletin of the Japanese Society of Scientific Fisheries*, 48(8):1081–1088, 1982.
5. C. W. Reynolds: "Flocks, Herds, and Schools: A Distributed Behavioral Model", *Computer Graphics*, 21(4):25–34, 1987.
6. A. Huth and C. Wissel: "The Simulation of the Movement of Fish Schools", *Journal of Theoretical Biology*, 156:365–385, 1992.
7. L. P. Kaelbling, M. L. Littman, and A. W. Moore: "Reinforcement Learning: A Survey", *Journal of Artificial Intelligence Research*, 4:237–285, 1996.
8. R. S. Sutton and A. G. Barto: *Reinforcement Learning*, MIT Press, Cambridge, MA, 1982.
9. M. Tomimasa, K. Morihiro, H. Nishimura, T. Isokawa, and N. Matsui: "A Reinforcement Learning Scheme of Adaptive Flocking Behavior", Proceedings of the 10th International Conference on Artificial Life and Robotics (AROB), GS1-4, Oita, Japan, Feb. 4-6, 2005.
10. C. J. C. H. Watkins and P. Dayan: "Q-learning", *Machine Learning*, 8:279–292, 1992.



# Real Time Head Nod and Shake Detection Using HMMs

Yeon Gu Kang, Hyun Jea Joo, and Phill Kyu Rhee

Dept. of Computer Science & Engineering Inha University,  
253, Yong-Hyun Dong , Nam-Gu, Incheon, South Korea  
cky@inhatc.ac.kr, hj0235@im.inha.ac.kr, pkrhee@inha.ac.kr

**Abstract.** This paper discusses a technique of detecting a head nod and shake. The proposed system is composed of face detection, eye detection and head nod and head shake detection. We use motion segmentation algorithm that makes use of differencing to detect moving people's faces. The novelty of this paper comes from the differencing in real time input images, preprocessing to remove noises (morphological operator and so on), detecting edge lines and restoration, finding the face area and cutting the head candidate. Moreover, we adopt K-means algorithm for finding head. Eye detection extracts the location of eyes from the detected face region. It is performed at the region close to a pair of eyes for real-time eye detecting. Head nod and shake can be detected by HMMs those are adapted by a directional vector. The HMMs vector can also be used to determine neutral as well as head nod and head shake. These techniques are implemented on a lot of images and a notable success is notified.

## 1 Introduction

For communication between person and computer, many researchers have studied on face processing such as face tracking, face detection, recognizing face and facial expression, lip reading, eye blinking, etc. Therefore, many algorithms and techniques are invented, but it remains a difficult problem yet.

Among head gestures, nodding and shaking can be used as a gesture to fulfill a semantic function (e.g., nod head instead of saying yes), to communicate emotions (e.g., nodding enthusiastically with approval) and as conversational feedback (e.g., to keep the conversation moving). A system that could detect head nods and head shakes would be an important component in an interface that is expected to interact naturally with people[2][3].

A lot of work has been done on facial pose estimation as well as face tracking. The related work presented in Davis and Vakes[4] is relevant to recognizing head gestures for interfaces. Recognition of gestures is achieved using Timed Finite State Machine, and a simple dialog-box agent is implemented to acquire yes/no acknowledgement from the user. But the detection of human face needs a hardware equipment, IBM PupilCam, so its application is limited. Another relevant work is done by Toyama [5] that uses a robust 3D face tracking system employing color, intensity templates, and point features for positioning a cursor on the computer monitor. 3D facial pose estimation with the regions of skin and hair have been reported in Chen, Wu, Fukumoto, and Yachida[6] and Heinzman and Zelinsky[7]. In their papers, area information are used instead of feature points, so the algorithms seem to be more robust. However, as

the head movements during nodding or head are small, the detection resolution they provide is insufficient. Kawato and Ohya[8] have described a system to detect head nods and head shakes in real time by directly detecting and tracking the ‘between-eyes’ region. The ‘between-eyes’ region is detected and tracked using a ‘circle frequency filter’ together with skin color information and templates. Pre-defined rules in consecutive frames are applied to detect head nodding and shaking. Because of its simple rule-based approach, some non-regular head nods and shakes cannot be detected. Our purposed system architecture is shown in Fig. 1.

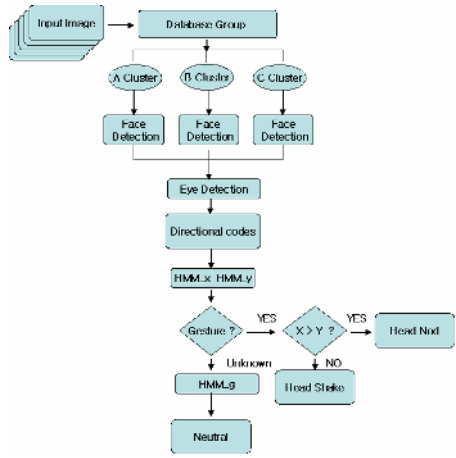


Fig. 1. The architecture of the proposed system

The organization of this paper as follows: section 2 describe face detection and eye detection, section 3 discusses head nod and shake detection, and section 4 illustrates the experimental results. Finally conclusion is resided.

## 2 Face Detection and Eye Detection

### 2.1 Real Time Face Detection

Input images are of  $320 \times 240$  sizes. Then the difference between the current image (N) frame and previous frame (N-1) is calculated and evaluated. The difference is then compared with certain threshold. The system proceeds towards edge detection only when the difference is significant enough to find the body segment. But when the result is not changed i.e. difference is less than the threshold value (person is not moving), current frame (N) is compared with the second previous (N-2) and difference is evaluated. Therefore the maximum difference is measured between two frames by comparing three image frames. If again the difference is not significant then previously segmented head part is used for face detection .Otherwise other steps are proceeded out which results in new head segmentation. The steps include edge detection or restoration. Here accurate edge line across the body is found by performing gradient operators

and interpolation. Next, the projection of horizontal distance between the edges is calculated for each row. Local minima and maxima point is calculated on the projection vector to find the head and body location respectively. To overcome the error due to noise, the projection vector is clustered around the two regions (one for head and one for body) using K-means algorithm. After the head segment is extracted from the image, Bayesian Discriminating feature method is used to search the target's face in the region. Fig. 2. is shown the result of face detection.

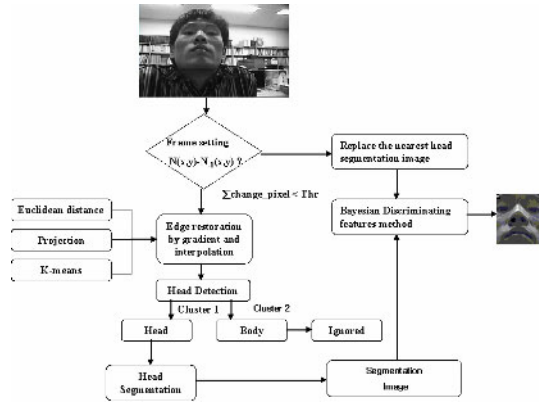


Fig. 2. The detection of the face

## 2.2 Eye Detection

When an image contains an object having homogeneous intensity and a background with a different intensity level, the image can be segmented in two regions by using image segmentation by thresholding[9].

The labeling algorithm examines a state of connection between a pixel and its 8-neighborhood pixels, and labels the objects in the image. The labeling algorithm is used to obtain connected and isolated regions[10]. The labeled facial components are recorded as eye candidates.

Candidates of eye and a pair of eyes are detected by heuristic rules based on eye's geometrical information in face. Eye candidate's width has much correlation with face candidate's height. However, because the height of eye candidate has an extreme changeability on condition of blinking one's eyes, it is difficult to set a correlation between eye candidate's height and face candidate region. Eye detection has a tendency to acquire the boundary of eye. However, image segmentation have a more concern in pupil regions than the boundary of eyes, so the connected regions by labeling have a tendency to become more smaller and more nearer a square than ones obtained by edge detection. Therefore, rules used in eye location differ as to binarization methods[11]. First, because two eyes locate in a similar position by y-axis, eye candidate regions are sorted by y-axis. Candidates of a pair of eyes satisfy the following rules- the gradient of two eyes, comparison between two eye's size and distance between two eyes. These rules also differ as to binarization method similar to the rule of

eye candidate regions. After eye location process, Ranker calculates an energy of each candidate of a pair of eyes detecting in each face region  $Fk$ . Each energy obtained by Ranker inputs to  $Max$  and it selects a max value among those. A candidate of a pair of eyes whose energy is equal to this selected value become a pair of eyes.

There can exist more than one candidate for a pair of eyes obtained in each candidate face region. Now, determined eyes have a maximum value of energy by Ranker. Therefore if there exist  $m$  candidates for a pair of eyes obtained in a face candidate region  $Fk$ , each candidate for a pair of eyes is expressed as,  $Eye_k^i, 1 \leq i \leq m$  and its energy calculated by Ranker is define as  $P(Eye_k^i)$ . Therefore, the energy of a determined pair of eyes in face candidate region  $Fk$  are defined as below equation,

$$E(k) = \underset{1 \leq i \leq m}{MAX} \{P(Eye_k^i)\} \tag{1}$$

and finally, the determined eyes of this facial image should satisfy equation.  $n$  is the number of the face candidate regions.

$$\underset{1 \leq k \leq n}{MAX} (E(k)) \tag{2}$$

The results of executing the binarization, connected region segmentation by labeling and conditioning the rule that will be satisfied by a candidate for a pair of eyes. we determine as eye regions because of it having maximum value among them[11].

### 3 Head Nod and Shake Detection

The coordinates of eye are transformed into a vector suitable for head nod and shake detection. Head nod and shake is detected by direct observation of the variation of the vector. HMMs are used to absorb a few errors which can be generated in eye detecting process, for more accurate detection. HMMs are simply introduced, and the technique of head nod and shake detection is described in this section.

#### 3.1 Hidden Markov Models

HMMs are a set of statistical models used to characterize the statistical properties of a signal. They have an immediate and obvious application in speech processing, particularly recognition, where the signal of interest is naturally represented as a time-varying sequence of spectral estimate[12]. In recent years, HMMs have been applied in many new fields, such as face recognition [13], handwriting recognition and so on.

The elements of HMMs are defined as follows.  $N$  is the number of states in the model.  $M$  is the number of different observation symbols.  $A$  is the state transition probability matrix.  $B$  is the observation symbol probability matrix.  $\Pi = \{\pi_i : 1 \leq i \leq N\}$  is the initial state distribution. Using shorthand notation, an HMM is defined as:  $\lambda = (A, B, \Pi)$ .

We adopt left-right models of HMMs. In left-right models, the state transition coefficients have the property  $a_{ij} = 0$ , for  $j < i$ . These models have more than two states and are trained by the Baum-Welch procedure[14].

### 3.2 Head Node and Head Shake Detection

The system acquires consecutive images from video camera. In our developed system, three head gestures, i.e. head nod, head shake, and neutral gesture, are detected. After performing eye location, the system calculates a vector representing the user's head nod and shake. Owing to the difference of people's head motions, the vector must be normalized. The normalized vector can be denoted in the charts with the x-axis, the y-axis, and the time-axis as follows. When a user responds head nod, the normalized vector of the x-coordinate is distributed in the 0~63 range value.

Head nod has a little variation of the x-coordinate, but much variation of the y-coordinate. Head shake has much variation of the x-coordinate, but a little variation of the y-coordinate. If there exists much variation of the x-coordinate and the y-coordinate at the same time, it can be considered as a neutral gesture.

The average of the normalized vector of the x-coordinate or the y-coordinate is used to roughly determine head nod or head shake. As comparing the average of the x-coordinate with that of the y-coordinate, we can know that the average of which coordinate is higher. A vector with higher average has sinusoid pattern. Simple detection of head gesture is possible by determining which one has sinusoid pattern.

HMMs with 64 symbols and 4 states are used to produce more accurate recognition result. The input to HMMs is a normalized vector obtained in section 3.1. These are trained by the Baum-Weiss procedure[14]. One detects sinusoid pattern of the vector; the other detects irregular linear-like pattern of the vector. Two HMMs are trained respectively with appropriate pattern, i.e sinusoid or irregular linear-like pattern. If the vector at the x-axis has sinusoid pattern and the vector at the y-axis has irregular linear-like pattern, the system detects a user acting head shake.

One of HMMs will output a result of detection of sinusoid pattern or irregular linear-like pattern of vector. If HMMs detects the sinusoid pattern, we can believe the fact that the selected axis by the first method has activity of gesture. Suppose that the x-axis is selected, we now have strong confidence that the head shake. Max selector selects HMM-x or HMM-y. If the output value of HMM-x or HMM-y is less than the predefined threshold, the system outputs head shake. Otherwise, according to the result of the comparison of HMM-x value with HMM-y value, the system reports head nod or head shake.

It is difficult for the previous methods to detect neutral gesture, because neutral gesture is detected based on the threshold only. This paper proposes another method using HMMs to clearly detect neutral gesture. These are designed using the directional vector of the center of two eyes as input. The HMMs generate the appropriate result of head nod and shake.

The movement of the face is converted into 8-directional vector codes. The number stands for the code representing the direction of face movement. The directional codes from 1 to 8 indicate that a face moves to each direction, the directional code 0 indicates that a face does not move at all. 8 directional codes obtained from the result of video image sequences are converted into a vector code. The vector code is inputted into the HMMs to decide the head shake, head nod, or neutral gesture of the user. Fig. 3. shows the block diagram of the head nod and shake detection.

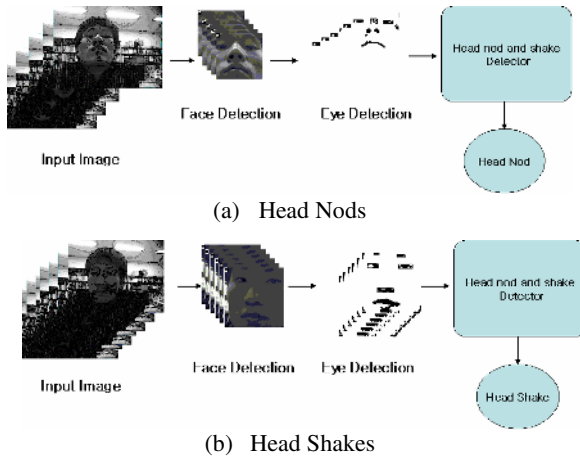


Fig. 3. The detection of the head nod and head shake

### 4 Experimental Results

Experiments have been performed with sequential images which are captured from a cheap LG USB camera is used as a video capturing device. The sequence frames of 320 x 240 pixels were acquired at 13 fps(frame per second) in 2.4GHzx Pentium 4 PC computer and stored in the temporary folder in hard disk. Database is separated into 3 groups A, B and C. Group A contains normal people without any spectacles and disguise form. Group B group includes people wearing spectacle. The last group C contains the people having beard and wearing spectacle in changing illumination condition. One group consists of 5 people dataset, and each person has twenty second movie clips at 13 fps. So each group has 1300 images sequence Table 1 shows the experimental results. After the head segmentation, the system finds Bayesian classifier to discriminate face and non face part of image. FERET database is used to learn the face Bayesian face model where as no-face Bayesian model is generated from the CMU database set.

Table 1 shows the face detection result for different scenario. Here it can see that face detection without segmentation gives high accuracy rate but it found to be time consuming. The third row of table 1 shows the result of our proposed method, which uses head segmentation and k-means for noise reduction. The result of our proposed method is found to be superior in terms of about 95% and very fast in terms of speed. K-means for noise reduction is not considered during the time of experiment and the recognition rate in this case is found to be very poor. Hence the credit of higher recognition rate goes to the error reduction technique employed by k-means algorithm. And Table 1 shows the result of experiment that's shows the accuracy of the head location. Same database and experiments were conducted. The table suggest that our method is almost perfect to find the head location with the accuracy higher than 97%. But, when we do not contain the nested K-means algorithm, the result is very low because our experiment in real-time visual surveillance system contains many noise effects like as changing illumination.

**Table 1.** Experiments about face detection rate about head region segmentation methods

	Division	Database A	Database B	Database C	Average
Without Nested k-means head detection +BDF (Bayesian Discriminating feature)	Face Detection rate	73.45%	69.25%	65.86%	69.52%
Head seg +BDF (proposed method)		97.39%	96.34%	91.31%	95.01%
Without Nested k-means head detection +BDF	Head Location rate	73.13%	72.17%	69.11%	71.47%
Head seg +BDF (proposed method)		97.52%	97.15%	96.51%	97.06%

**Table 2.** The result of head nod and head shake detection

	Data	Head nod	Head shake	Neutral	Total
Database A	The number of data	100	100	100	300
	Success	97	96	90	283
	Recognition rate (%)	97.0	96.0	90.0	94.3
Database B	The number of data	100	100	100	300
	Success	96	95	89	280
	Recognition rate (%)	96.0	95.0	89.0	89.0
Database C	The number of data	100	100	100	300
	Success	94	93	81	268
	Recognition rate (%)	94.0	93.	81.	89.3

We use 5 training data for HMM learning for head node, head shake and neutral gestures. Each data contains 180 frames. We use 900 data for testing our system. Table 2 show the result of head nod and head shake detection.

## 5 Concluding Remarks

This paper discusses head nod and head shake detection system consisting of face detection, eye detection, and HMMs. We have proposed real time head region detection based on difference of image frames from camera. The main component of the system is the use of vertical projection of edge contours to find head location and nested K-means algorithm for error minimization during differencing and threshold. Overall the system shows low computational cost and high detection rate which suits for the real time system. Head detection rate of about 97% is achieved while minimizing the computational cost and time.

The future enhancement to the system can be to overcome the occlusion and to find automatic threshold for frames evaluation using multiple cameras. Since HMMs absorb some errors of eye location, very encouraging experimental result is obtained.

Thus, this detection system will be able to effectively applied to visual user interface. One weak point of our system is that when a user moves his face right and left, detection result is head shake. Likewise, when moving his face up and down, detection result is head nod. Other features of face and face rotation are observed to overcome this fault. Another weak point is that a pair of eyes of an extremely tilted face cannot be located in many cases. It is necessary to focus head nod area rather head shake.

## References

1. P.Ekman, T.Huang, T.Sejnowski, and J. Hager.: Final Report to NSF of the planning Workshop on Facial Expression Understanding, Technical report, National Science Foundation, Human Interaction Lab., UCSF, CA 94143, (1993)
2. Ashish Kapoor, Rosalind W. Picard.: A Real-Time Head Nod and Shake Detector, Proceedings from the Workshop on Perceptive User Interfaces, November (2001)
3. Wenzhao Tan, Gang Rong: A real-time head nod and shake detector using HMMs, Expert Systems with Applications 25, pp.461-466 (2003)
4. Davis, J. W., & Vakes, S.: A perceptual user interface for recognizing head gesture acknowledgements, ACM workshop on perceptual user interfaces (2001)
5. K. Toyama.: Look, ma—no hands! Hands free cursor control with real-time 3D face tracking, Proceedings of workshop on perceptual user interfaces, pp.49-54 (1998)
6. Q. Chen, H. Wu, T. Fukumoto, and M. Yachida.: 3D head pose estimation without feature tracking, IEEE International Conference on Automatic Face and Gesture Recognition(1998)
7. J. Heinzman, and A. Zelinsky.: 3-D facial pose and gaze point estimation using a robust real-time tracking paradigm, IEEE International Conference on Automatic Face and Gesture Recognition (1998)
8. S. Kawato, J. Ohya.: Real-time detection of nodding and head-shaking by directly detecting and tracking the 'between-eyes'. Fourth IEEE international conference on automatic face and gesture recognition (2000)
9. I. Pitas, Digital image processing algorithms, Prentice Hall (1993)
10. R. L. Lumina, G. Shapiro, and O. Zuniga.: A New Connected Components Algorithm for Virtual Memory Computers, Computer Vision, Graphics, and Image Processing, Vol. 22, pp. 287-300 (1983)
11. Jo Nam Jung, Mi Yung Nam and Phill Kyu Rhee.: Adaptive Eye Location using Fuzzy ART, ICNC2005 (2005)
12. L. Rabiner.: A tutorial on hidden markov models and Selected applications in speech recognition, Proceedings of the IEEE, 77(2):257-285, February (1989)
13. Nefian, A. V., and Hayes, M. H., III.: Face detection and recognition using hidden Markov models, IEEE International Conference on Image Processing. (1998)
14. L. Baum, T. Petrie, S. G. and N. Weiss.: A maximization techniques occurring in the statistical analysis of probabilistic functions of Markov chains, Ann. Math. Stat., 41(1):164-171, (1970)



# Construction and Evaluation of Text-Dialog Corpus with Emotion Tags Focusing on Facial Expression in Comics

Masato Tokuhisa, Jin'ichi Murakami, and Satoru Ikehara

Dept. of Information and Knowledge Engineering, Tottori University,  
4-101, Koyama-Minami, Tottori 680-8550, Japan  
{tokuhisa, murakami, ikehara}@ike.tottori-u.ac.jp

**Abstract.** Large-scale text-dialog corpora with emotion tags are required to generate a knowledge base for emotional reasoning from text. Annotating emotion tags is known to suffer from problems with instability. These are caused by the lack of non-linguistic expressions (e.g. speech and facial expressions) in the text dialog. We aimed to construct a stable, usable text-dialog corpus with emotion tags. We first focused on facial expression in comics. Some comics contain many text dialogs that are similar to everyday conversation, and it is worth analyzing their text. We therefore extracted 29,538 sentences from 10 comic books and annotated face tags and emotion tags. Two annotators independently placed “temporary face/emotion tags” on stories and then decided what the “correct face/emotion tags” were by discussing them with each other. They acquired 16,635 correct emotion tags as a result. We evaluated the stability and usability of the corpus. We evaluated the correspondence between temporary and correct tags to assess stability, and found precision was 83.8% and recall was 78.8%. These were higher than for annotation without facial expressions (precision = 56.2%, recall = 51.5%). We extracted emotional suffix expressions from the corpus using a probabilistic method to evaluate usability. We could thus construct a text-dialog corpus with emotion tags and confirm its stability and usability.

## 1 Introduction

Large-scale text-dialog corpora with emotion tags are required to generate a knowledge base for emotional reasoning from text. Annotating emotion tags is known to suffer from problems with instability, for instance, differences between the timing of emotional reasoning and its depth. This is caused by the lack of non-linguistic expressions (e.g. speech and facial expressions) in text dialogs. It is also difficult to collect many dialog texts with real speech or facial expressions because of cost. However, if pseudo-dialogs are acceptable for research, dialogs in comics are useful because some contain a great variety of facial expressions and dialogs similar to everyday conversations.

We aim at constructing a stable, usable text-dialog corpus with emotion tags created by focusing on facial expression in comics. We first annotate facial tags

and emotion tags into dialog text extracted from comics. For instance, two annotators place “temporary face/emotion tags” on stories before deciding which were “correct face/emotion tags” by discussing these with each other.

Next, we evaluate the stability and usability of the corpus. For instance, we evaluate the correspondence between temporary tags and correct tags to confirm what effect facial expressions have on emotion tagging stability, and compare this correspondence with annotation done without referring to facial expressions. Then, we extract emotional suffix expressions from the corpus to evaluate its usability by using a probabilistic method in a trial experiment.

## 2 Related Work

Rosis and Grasso proposed an affective natural language generator that could plan affective discourses[7]. They investigated a corpus of explanations on drug prescriptions while focusing on affect-conveying techniques. It was thus necessary to annotate emotions to dialogs.

Tokuhsa et al. proposed a very rich tag set, which they attempted to annotate to dialogs[8]. They annotated not only emotion tags but also non-emotional mental state tags into dialogs. This rich tag set provided knowledge about deep emotional reasoning, but it was difficult for inexperienced annotators to annotate tags stably.

Simple emotion tag sets, on the other hand, for instance, *Positive*, *Neutral* and *Negative*, have been used by other researchers. Litman and Forbes tried to predict emotions in their spoken dialog tutoring system[4]. Two annotators manually annotated three types of tags in their corpus to achieve 81.75% agreement.

Speech expressions might be of great help in annotation. Non-linguistic information is indispensable for acquiring stable emotion tags. However, speech expressions are not readily obtained and we cannot start collecting or annotating them easily. Therefore, in this paper, emotional annotation would be done by two annotators referring to facial expressions who avoid the cost of collecting dialogs by using comics.

## 3 Construction of Emotion Annotated Corpus

### 3.1 Steps in Annotation

The target comic for annotation was “Chibi Maruko-chan”, which is one of the most famous in Japan. It describes an elementary school student’s life. The language expressions are ordinary so that reasonable emotional language expressions can be extracted. We extracted all narrations, monologues, utterances and onomatopoeia done by the characters in the story from 10 comic books.

“Face tags” and “emotion tags” can be used in our corpus. The face-tag category consists of *happiness*, *disgust*, *sadness*, *surprise*, *fear*, *anger*, and *back*. All face tags except *back* are based on Ekman’s classification[2]. The *back* represents

the character's appearance from the rear, which sometimes expresses emotions with *cold sweat* and *trembling out of fear* in comics.

The emotion-tag category consists of *gladness*, *sadness*, *liking*, *disliking*, *surprise*, *expectancy*, *fear*, *anger*, and *no-emotion*. They are based on Plutchik's classification[6], because his system covers complex emotions by combining these eight pure emotions and such compositional ideal is suitable for computation. If an annotator needs a complex emotion tag, the combination of several emotion tags can be used instead. Furthermore, three classification like *positive*, *negative* and *neutral* is so poor that it cannot distinguish emotions appearing in ordinary conversation. For example, anger shouldn't be aroused to keep friendly relationships, even if sadness might be aroused. However, although OCC model especially provides rich classification on human relationships[5], it is so various that annotators would not be able to perform stably.

Two annotators annotate tags to a story using the following steps.

**Step 1:** Each annotator independently annotates face tags to story's characters per block of comics.

**Step 2:** Each annotator independently annotates emotion tags to characters per block of comics annotated with face tags. These face and emotion tags are annotated temporarily and there are no confirmation that these tags are correct, so they are called "temporary face/emotion tags."

**Step 3:** Annotators review their temporary tags and decide the "correct face/emotion tags" through discussions with each other.

### 3.2 Results of Annotation

Table 1 lists part of the corpus with the correct tags. "Maruko" expresses sadness in her face in the first line and her emotion is sadness. "Saki," who is Maruko's elder sister, smiles in the second line, but annotators decided sadness was her emotion. "Maruko" makes no utterances in the fourth line but expresses emotion in her face. The face and emotion tags are annotated to the two utterances in the eighth and ninth lines.

The target comics had 104 stories, consisting of 10,213 blocks. The utterance text was divided into 29,538 sentences (388,809 letters).

The annotation was divided among six students at our laboratory. It took approximately 2 hours per story to perform the three steps except for typing the descriptions of the comics. As a result, 14,040 correct face tags and 16,635 correct emotion tags were acquired. Table 2 lists the number of correct emotion tags in detail. The *liking* tags appeared so few times. The annotators leaned toward *gladness* when a comic character expressed a *happiness* face, or it might be really few to change character's mind toward liking on the stories.

## 4 Evaluation of Stability

The corpus contained temporary and correct tags. The temporary tags were decided from both the face tags and context in comics. The correct tags were

**Table 1.** Sample from corpus

#	Page	Block	Speaker	Utterance*	Face tag	Emo. tag
1	22	3	Maruko	Uchi-no Mominoki-wa Chiisai-ne (Our Xmas tree is small.)	<i>sadness</i>	<i>sadness</i>
2	22	3	Saki	Shikata Nai-jan (We have no choice.)	<i>happiness</i>	<i>sadness</i>
3	22	4	Saki	Gyaaa!	<i>surprise</i> <i>fear</i>	<i>surprise</i> <i>fear</i>
4	22	4	Maruko		<i>surprise</i>	<i>surprise</i>
5	23	1	Saki	Maruko, Anta Mominoki-no Hachi-ni Kingyo-no Shigai Ume-ta-desho (Maruko, you buried a gold fish in the pot of the tree, didn't you?)		
6	23	2	Maruko	Sou-da-yo (Yes, I did.)		
7	23	2	Maruko	Datte Hiryou-ni Naru-to Omotte (Because I thought it would become manure.)		
8	23	3	Saki	Yame-te-yo (Oh, no.)	<i>disgust(sweat)</i>	<i>disliking</i>
9	23	3	Saki	Kimochi Warui (That's terrible.)		

\* English utterances in this table have been translated by authors of this paper for reference.

decided from temporary tags, but the set of correct tags was not a simple combination of temporary tags.

In this section, we first evaluate the correspondence between the set of temporary and set of correct tags. We annotate new temporary tags without referring to face tags and then evaluate the correspondence to confirm what effect referring to face tags had.

#### 4.1 Correspondence Between Temporary and Correct Tags Where Face Referred to

Correspondence was measured by Precision( $P$ ) and Recall( $R$ ).  $N_{corresponding}$  was the number of corresponding between temporary tags and correct tags,  $N_{temporary}$  was the number of temporary tags, and  $N_{correct}$  was the number of correct tags. Precision and recall can be defined as:

$$P = N_{corresponding} / N_{temporary} * 100(\%)$$

$$R = N_{corresponding} / N_{correct} * 100(\%)$$

**Table 2.** Number of correct emotion tags in corpus

Emotion	Percentage	Number
<i>gladness</i>	26.9 %	( 4,469 )
<i>disliking</i>	18.0 %	( 2,990 )
<i>expectancy</i>	13.4 %	( 2,237 )
<i>surprise</i>	12.1 %	( 2,010 )
<i>fear</i>	10.6 %	( 1,757 )
<i>sadness</i>	8.6 %	( 1,428 )
<i>anger</i>	8.1 %	( 1,347 )
<i>no-emotion</i>	1.2 %	( 207 )
<i>liking</i>	1.1 %	( 190 )

**Table 3.** Correspondence between tags where face referred to

Annotator	Precision	Recall	Corresp. tags
Person-A	93.1%	87.6%	4,579
Person-B	88.9%	84.8%	8,435
Person-C	83.2%	76.6%	7,069
Person-D	90.1%	82.2%	8,051
Person-E	69.0%	67.0%	6,865
Person-F	84.3%	79.9%	3,961
Total	83.8%	78.8%	38,960

Table 3 summarizes the precision and the recall results for the six annotators. Because the precision and the recall for the total were 83.8% and 78.8%, annotations by Person-A, B, D, and F were stable but annotations by Person-C and E were not.

#### 4.2 Correspondence Between Temporary and Correct Tags Where Face Not Referred to

Although non-linguistic information effectively stabilized annotation, its effect needed to be quantitatively clarified. Therefore, Person-G annotated new temporary emotion tags to the corpus without referring to face tags but reading utterance text in the corpus.

Person-G annotated four stories that were not annotated by Person-E, for instance, two stories in Volume 2 and two stories in Volume 6. They contained 1,217 sentences. Persons-A and B had annotated the stories in Volume 2, and Persons-C and D had annotated these in Volume 6 while referring to face tags. Table 4 compares them. The precision and recall of Person-G are 56.2% and 51.5%, which are 30.7/27.7 lower than for Persons-A, B, C, and D.

The 200 temporary tags by Person-G that did not correspond to the correct tags were randomly extracted and checked. The reasons for lack of correspondence were classified as:

**Table 4.** Correspondence between tags where face not referred to

Annotator	Precision	Recall	Corresp. tags
Person-G	56.0%	55.8%	216
Person-A	92.9%	84.0%	325
Person-B	84.0%	81.4%	315
Person-G	56.4%	48.5%	276
Person-C	82.9%	73.3%	417
Person-D	88.9%	80.3%	457
Total-G (no face)	56.2%	51.5%	492
Total-else (face)	86.9%	79.2%	1,514

- (1) Emotions in the correct tag sets were only decided by facial expressions (37%),
- (2) Because no face tags were annotated, no emotion tags were annotated in the correct tag sets (26%),
- (3) Emotions were ambiguous (26%), and
- (4) Others (11%).

This means that facial expressions might help define the timing when tags are annotated ((2)) and discriminate between ambiguous emotions ((1)).

This corpus will be used for test data on emotional reasoning in the future. If the facial expressions are not used in reasoning, the target score for  $P$  and  $R$  in tests will be approximately 60% because of the total correspondence by Person-G.

## 5 Evaluation of Usability

This corpus was constructed from comics. How useful is the corpus for analyzing language? This section explains a trial experiment to extract Japanese emotional expressions. If emotional expressions are acquired, the usability of this corpus will be clarified. So the accuracy of the emotion reasoning by the expressions would not be pursued in this paper<sup>1</sup>.

### 5.1 Extracting Emotional Suffix Expressions

The speaker’s introspection in a Japanese sentence is expressed in the suffix, which consists of post-positional words, auxiliary verbs, and some independent words<sup>2</sup>. The suffix for introspection, however, has not been clarified in terms of either the combination of such words or the emotional classification.

<sup>1</sup> In our preliminary experiment, the accuracy of automatic emotion reasoning explained in this section was 53%(236/441 sentences), where the test set was the same parts in Table 4 and the training set was extracted from the corpus except for the test set. Person-G’s accuracy was 57%(256/441 sentences). In order to improve the automatic reasoning, the main expression, i.e. subject-verb-object relations [9], should be taken account of in the reasoning.

<sup>2</sup> e.g. the Japanese sentence “Zettaini Ika-nai-yo” means “SURELY GO-NOT-will(= I won’t go).” “-yo” is a suffix expression discussed in this paper.

The following describes the procedure for extracting co-occurrences between suffixes for emotional introspection and their emotional categories.

**Procedure 1:** Divide the corpus into a training set and a test set.

**Procedure 2:** Enter the pair of the utterance sentence and the emotion tag in the training set into “a suffix trie structure” where each word in the sentence is entered in reverse order and whose node stores emotion tags.

**Procedure 3:** Parse the utterance sentence in the test set by using the suffix trie structure, and acquired the suffix expression and co-occurring emotions. Emotions can be probabilistically selected from them as a result of emotional reasoning from the sentence. If the selected emotion corresponds to the correct emotion tag of the utterance, it is possible for the acquired suffix expression to express emotional introspection.

For example, Figure 1 shows a suffix trie structure created from the example training set in Table 5. As node  $n10$  is only created with sentence  $s2$ ,  $n10$  stores “ $angry = 1$ ” as the frequency of co-occurrences between the sentence and the emotion. As node  $n7$  is created by sentences  $s2$  and  $s3$ ,  $n7$  stores “ $angry = 2$ .” In the same way,  $n6$  stores “ $angry = 2$  and  $no-emotion = 1$ ,” and  $n1$  stores “ $disliking = 1$ ,  $angry = 2$  and  $no-emotion = 1$ .” Therefore, if the sentence “**Sonna Koto Dekiru-mon-ka .**” is input as test data, the suffix expression “**mon-ka .**” would be the parsing result to acquire the degree of emotion “ $angry = 2$ ,” which is denoted as “ $C(\text{mon-ka.}, angry) = 2$ .”

As we can see from Table 2, the frequency of emotions is different. Therefore, emotions co-occurring with a suffix expression should be probabilistically selected from those stored in the node. Where “emotion expressing probability  $P_{EX}$ ” is the probability when a person feeling emotion  $e$  utters sentences  $U$ , emotion  $E$  inferred from utterance  $U$  can be defined as:

$$E = \arg \max_e P_{EX}(U|e)$$

As utterance  $U$  contains one or more sentences, if the  $i$ -th. sentence is  $s_i$ ,  $E$  can be changed to:

$$E = \arg \max_e \sum_i P_{EX}(s_i|e)$$

Focusing on the suffix expression, sentence  $s_i$  is approximately changed by  $suffix(s_i)$ , which means the suffix expression acquired from  $s_i$ . The following is an approximated equation for  $P_{EX}(s_i|e)$ .

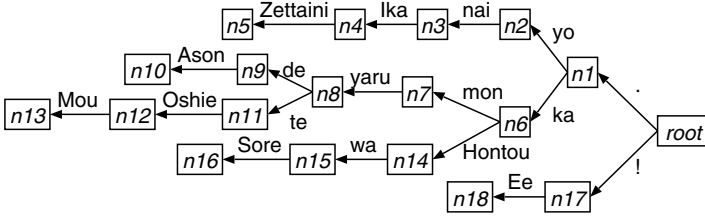
$$P_{EX}(s_i|e) \approx P_{EX}(suffix(s_i)|e) = C(suffix(s_i), e)/C(e)$$

$C(suffix(s_i), e)$  represents the frequency mentioned in the previous paragraph<sup>3</sup>.

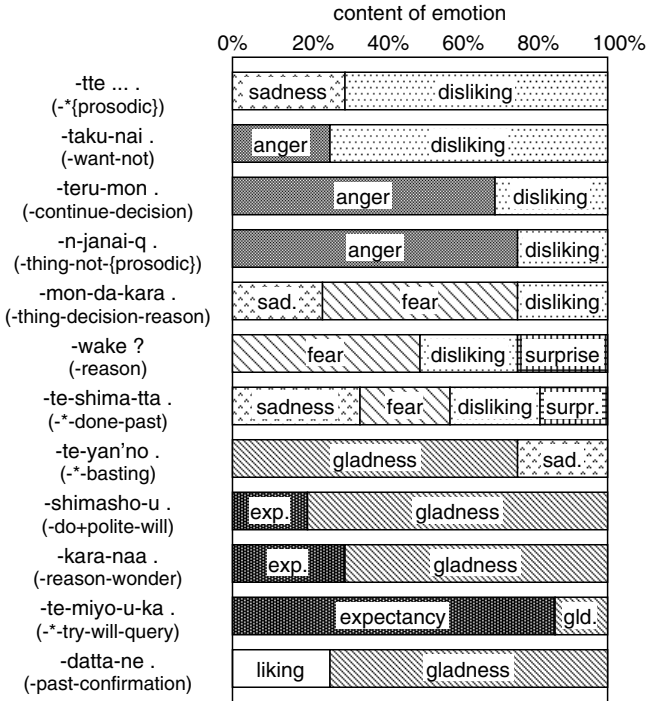
<sup>3</sup> [1] proposed a similar model for automatically tagging affect, which used an  $N$ -gram model as a language model. In our preliminary experiment, uni-gram, bi-gram, and tri-gram models were less effective than the suffix expression model because of the characteristics of Japanese.

**Table 5.** Example of training set

Sentences	Emotion tags
s1: Zettaini Ika-nai-yo .	<i>disliking</i>
s2: Ason-de yaru-mon-ka .	<i>angry</i>
s3: Mou Oshie-te-yaru-mon-ka .	<i>angry</i>
s4: Sore-wa Hontou-ka .	<i>no-emotion</i>
s5: Ee !	<i>surprise</i>



**Fig. 1.** Suffix trie structure



**Fig. 2.** Samples of emotional suffix expressions



### 5.2 Results of Extraction

The procedure described in Section 5.1 extracted 3,652 kinds of suffix expressions from the corpus. The suffix expressions were mapped into an average of 2.85 kinds of emotions. Figure 2 lists samples of these, where *content of emotion* represents the ratio of the correct emotion tags in test data. The emotions in the figure are suitable for suffix expressions. These suffix expressions are a kind of “emotive pattern” as discussed by [3], but our suffix expressions would adapt to the target sentence of emotion reasoning by using the longest matching suffix expression from the trie structure to refer more specific emotional co-occurrence. Figure 3 shows variety of the suffix expressions including “kedo,” which means “the speaker explicitly says so, but implicitly thinks other things.”

The descriptions in parentheses in Figure 2 represents combinations of tense/aspect/modality decomposed from suffix expressions. Some suffix expressions successfully seems to be explained by combinations that parse the syntactic side. However, such decomposition may be meaningless to infer emotional side, because emotional nuances can be understood from the combinations, i.e., the words themselves or their prosodic sense. Therefore, it is important to analyze suffix expression based on examples for emotion reasoning. We could thus confirm this proposed corpus was usable.

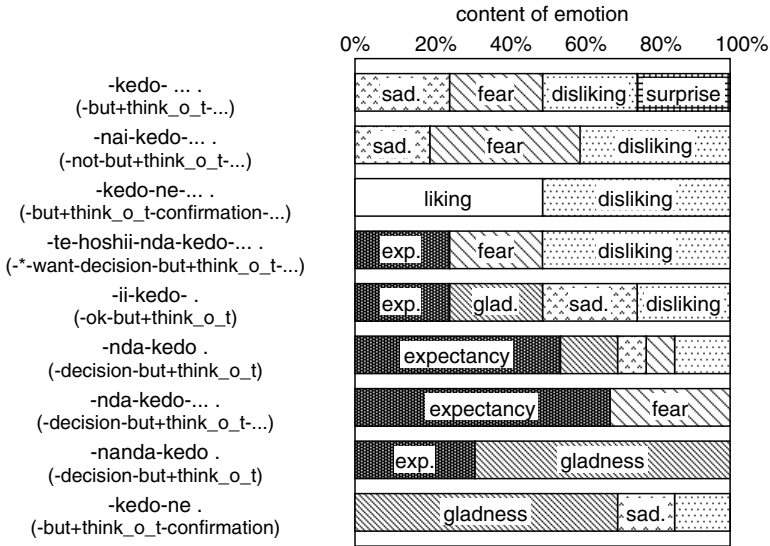


Fig. 3. Samples of emotional suffix expressions including “kedo (but+think\_other\_things)”

## 6 Conclusion

We constructed a text-dialog corpus with emotion tags and evaluated the stability of annotation and the usability for generating a knowledge base for emotional

reasoning. The corpus contained 29,538 sentences and 16,635 correct emotion tags. The annotation was done from text dialog in comics while referring to facial expressions to improve stability, which was evaluated by criteria on precision ( $P = 83.8\%$ ) and recall ( $R = 78.8\%$ ). These were both higher than for annotation when facial expressions were not referred to. Usability was evaluated in a trial experiment to extract reasonable emotional suffix expressions from the corpus.

## Acknowledgments

This research was partially supported by “the Ministry of Education, Science, Sports and Culture, a Grant-in-Aid for Young Scientists (B), 17700151, 2005-2007” and “Japan Science and Technology Agency (JST), the Core Research for Evolution Science and Technology (CREST) project.”

## References

1. N. Chambers, J. Tetreault and J. Allen, “Approaches for automatically tagging affect,” in *Exploring attitude and affect in text: Theories and applications*, pp. 36–43, AAAI Press, 2004.
2. P. Ekman and W. V. Friesen, *Unmasking the face*. Prentice-Hall, Inc., Englewood Cliffs, 1975. (Japanese translation version: T. Kudo, D. Matsumoto, Y. Shimomura, and E. Ichimura, Seishin Shobou, 1990).
3. G. Grefenstette, Y. Qu, D. A. Evans and J. G. Shanahan, “Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes,” in *Exploring attitude and affect in text: Theories and applications*, pp. 63–70, AAAI Press, 2004.
4. D. Litman and K. Forbes, “Recognizing emotions from student speech in tutoring dialogues,” in *Automatic Speech Recognition and Understanding Workshop*, 2003.
5. A. Ortony, G. L. Clore and A. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, 1988.
6. R. Plutchik, “The multifactor-analytic theory of emotion,” in *The Journal of Psychology*, vol. 50, pp. 153–171, 1960.
7. F. de Rosis and F. Grasso, “Affective natural language generation,” in *Affective Interactions* (A. Paiva, ed.), vol. 1814 of *Lecture Notes in Artificial Intelligence*, pp. 204–218, Springer, 2000.
8. M. Tokuhisa, R. Tokuhisa, K. Inui and N. Okada, “Emotion recognition in dialogue,” in *Affective Minds* (G. Hatano et al. eds.), pp. 221–229, Elsevier Science, 2000.
9. M. Tokuhisa, T. Tanaka, S. Ikehara and J. Murakami, “Emotion reasoning based on valency patterns - prototype annotation of causal relationships,” in *Human and Artificial Intelligence Systems*, pp.534–539, 2004.

# Human Support Network System Using Friendly Robot

Masako Miyaji<sup>1</sup>, Toru Yamaguchi<sup>1</sup>, Eri Sato<sup>2</sup>, and Koji Kanagawa<sup>1</sup>

<sup>1</sup> Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo 191-0065 Japan

<sup>2</sup> Tokyo Metropolitan Institute of Technology, 6-6 Asahigaoka, Hino, Tokyo 191-0065 Japan

**Abstract.** Because the traditional driver support system warns unilaterally by a display and a sound, it often addles a driver. If you feel burden about a present system using keyboard, mouse and so on, you might feel familiar with a system using a pet-robot. A Pet-robot, which pleases, helps and stays close to human, is called Friendly robot. So we construct a warning system using Friendly robot. In the evaluative experiment 1, we compare the understandability, the familiarity, and the ability of the notification between Friendly robot and the traditional system. Furthermore, from the result of the evaluation experiment 1, we improve the driver support system about “easy to notify” using several time warnings by Friendly robot. We show the effectiveness of the driver support system using Friendly robot.

## 1 Introduction

A various types of information tools are used with advancement of information society. Moreover, robot technology is progressing and pet-robots at home are becoming more popular. We think that pet-robots are capable of becoming nonverbal interface tool that doesn't require keyboard or mouse pointer and so on [1]. As the pet-robots are often used for animal therapy, pet-robots have a high affinity for human. Meanwhile, in the automotive industry, the car navigation system is becoming popular. But current driver support system warns unilaterally by a display and a sound, it could be addle a driver.

If you feel burden about a present system such as using keyboard, mouse and so on, you might feel familiarity with a system using a pet-robot [2]. It is same for the driver support system, so the pet-robots could be used as the interface of the support system A Pet-robot, which pleases, helps and stays close to human, is called Friendly robot. We use Friendly robot for warn system to a driver. Moreover, a high affinity system should be like an interpersonal communication.

We construct a warning system, which is easy to use and to avoid confusion for drivers using Friendly robot.

## 2 Human Support Network Robot System

In this study, we use i-spot for detecting cars or pedestrians as shown in Fig. 1. We set i-spot for each street and in a car [3]. i-spot is composed camera, wireless

LAN device and PC. Each i-spot connect wireless LAN, and share information. At first, when camera of i-spot on the street catches a car or pedestrians in certain distance, warning is transmitted to a server on the car in several patterns, according to its circumstance. i-spot is able to detect a driver's face direction using a camera inside the car.

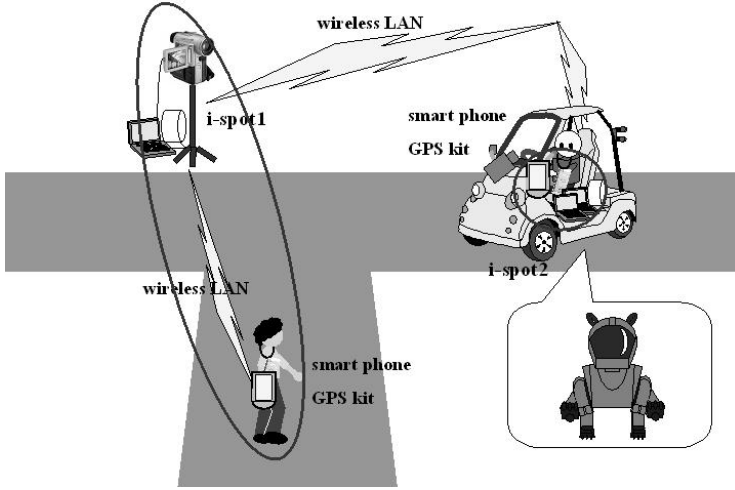


Fig. 1. Human support network robot system

Next, I explain inside a car system. PC of i-spot is main PC of car. Additionally, we set three PCs inside a car as show in Fig. 2. The middle PC assists warning. The Left and the right PC display a condition of backside, and transmit a presence of obstacle to the main PC. Display of middle PC and Friendly robot's motion caution a warning for a driver. This warning is decided by main PC.

In this experiment, we use these PCs. In the future, it is expected to use the car with backward monitoring. In addition, GPS kit is not well developed yet, it will be incorporated mobile telephone.

This system doesn't warn a danger in same direction of a driver's face to prevent overabundance of information. By using driver's face direction, main PC decides whether the system warns the danger or not. Using location information of an obstacle on the street from i-spot and location information of the car from GPS data, the main PC choose Friendly robot's motion.

Friendly robot's motions are made in advance. We made nine different motions that 'an obstacle to the left', 'an obstacle to the right', 'an obstacle in front', 'stop', 'slow', 'keep out', 'confirming right display', 'confirming left display' and 'Accidents when turning left ( at pedestrian crossing )'. In addition, when we make Friendly robot's motions, we use the method of individualized expression [4].

We use i-space in this system [3]. i-space is a software that is able to track movement and position of an object by color information as shown in Fig. 3.

### 3 Experiment

#### 3.1 Primary Experiment

We experimented on some situations using above nine motions. We show one of the experiments, ‘an obstacle to the left’. In this experiment, we use two motions that ‘an obstacle to the right’ and ‘stop’. First, Friendly robot points out the left direction to the driver. This motion means that there is an obstacle on the left. Next, Friendly robot sticks out both hands slowly, this motion means stop. It is shown in Fig. 4 and Fig. 5.

#### 3.2 Evaluation Experiment 1

We use warning of display as a metaphor for current driver support systems. We made a comparative experiment between display and Friendly robot with ten participants. In this experiment “there is an obstacle to the left”, when assuming display as a standard, Friendly robot’s ‘Easy to understand’, ‘Feel familiar’ and ‘Easy to notify’ are evaluated in five rankings as shown in Fig. 6. 1 is ‘very bad’ and 5 is ‘very good’.

“Easy to understand”: 1:0, 2:2, 3:5, 4:1, 5:2

“Easy to understand” average is 3.3.

“Feel familiar”: 1:0, 2:0, 3:1, 4:2, 5:7

“Feel familiar” average is 4.6.

“Easy to notify”: 1:0, 2:0, 3:4, 4:2, 5:4

“Easy to notify” average is 4.0.

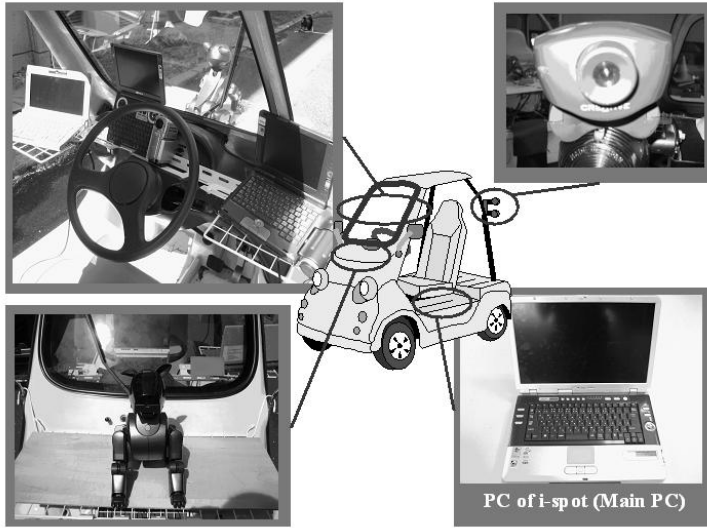
It means Friendly robot is more familiar than using display. The average of rating “Easy to understand” was not good. The reason is that the caution system using display was easy to understand in the first place.

#### 3.3 Evaluation Experiment 2

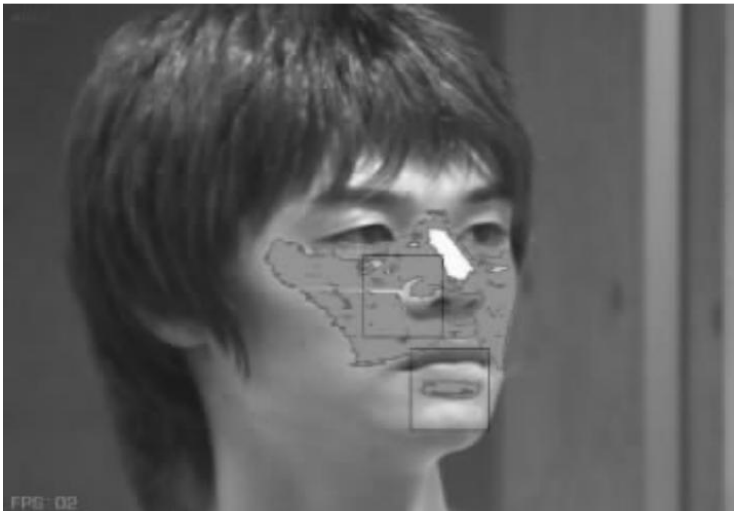
In the second experiment “there is an obstacle to the left”, we developed information presentation system which Friendly robot warns several times according to the circumstance. And we did a comparative experiment with this. First, Friendly robot warned a driver when the system detects a danger, and then Friendly robot warns concretely that an obstacle is coming. When assuming using display as a standard, Friendly robot’s ‘Easy to understand’, ‘Feel familiar’ and ‘Easy to notify’ are evaluated in five rankings as shown in Fig. 7.

“Easy to understand”: 1:0, 2:0, 3:6, 4:2, 5:2

“Easy to understand” average is 3.6.



**Fig. 2.** The middle PC assists warning. The Left and the right PC display a condition of backside, and transmit a presence of obstacle to the main PC.



**Fig. 3.** This is the appearance of tracking using I-space

“Feel familiar”: 1:0, 2:0, 3:1, 4:2, and 5:7

“Feel familiar” average is 4.6.

“Easy to notify”: 1:0, 2:0, 3:1, 4:4, 5:5

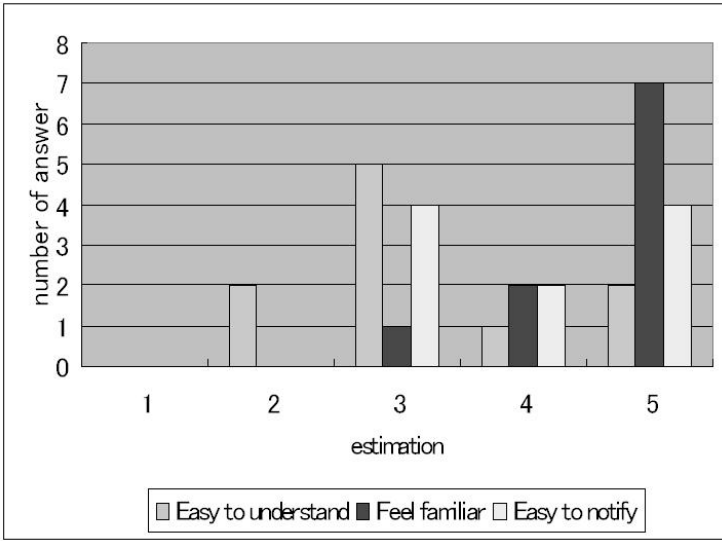
“Easy to notify” average is 4.4.



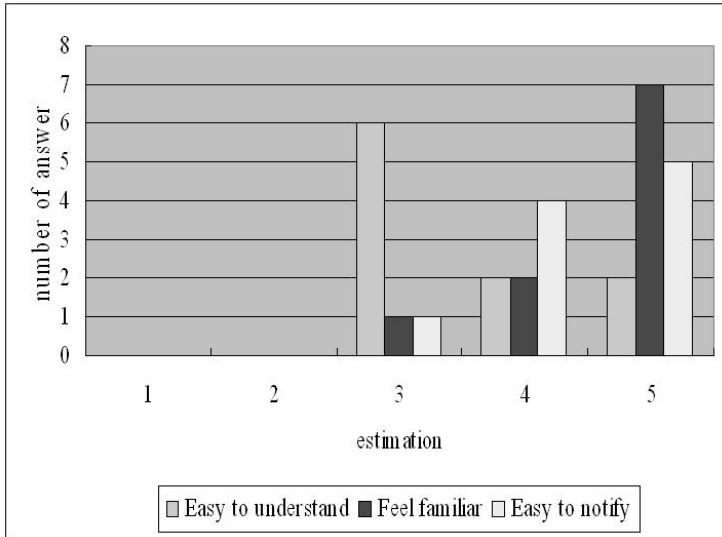
**Fig. 4.** This image shows Friendly robot's instruction that "there is an obstacle to the left". Friendly robot points to the left of the driver, this means there is an obstacle on the left.



**Fig. 5.** This image shows Friendly Robot's instruction that "stop". Friendly robot sticks out both hands slowly, this means stop.



**Fig. 6.** This is a result of the evaluation questionnaire1. Gray is Friendly robot’s ‘Easy to understand’, Black is ‘Feel familiar’ and white is ‘Easy to notify’.



**Fig. 7.** This is a result of questionnaire2 Gray is Friendly robot’s ‘Easy to understand’, Black is ‘Feel familiar’ and white is ‘Easy to notify’

Comparing with experiment 1, average of “Easy to notify” increased. From this result, drivers can pay more attention to Friendly robot’s warning easily, that is, the drivers’ judging abilities have been risen. In these circumstances, the





**Fig. 8.** Friendly robot points out to PC's display

result of evaluation experiment 2 shows that Friendly robot warns several times better than caution only once.

## 4 Conclusion

In this paper, we showed three experiments. In the primary experiment, we constructed driver support system using Friendly robot. The result of evaluation experiment 1 and 2 showed the effectiveness of warning system.

In evaluation experiment 1, the driver support system using Friendly robot was felt more familiar than current driver support. But average of "Easy to notify" was not good. So we developed our system to warn several times. In evaluation experiment 2, average of "Easy to notify" rose. The result shows that Friendly robot warns several times better than caution only once.

In the future, we will experiment about driver support system under more realistic circumstances.

In addition, it will construct a system based on the interpersonal communication in order for more familiar system. We will construct driver support system based on joint attention. Joint Attention is a process of sharing observing an object or event, by following others gaze or pointing gestures.

For example there is an obstacle in front of the car. If the driver doesn't see, Friendly robot points to the obstacle. And Friendly robot changes angle of shoulder by the obstacle's location information from GPS data. There is an

obstacle in back of the car. An obstacle is captured by cameras connected to two PCs. At the same time, this system finds an obstacle. If an obstacle is caught by PC's display, Friendly robot points out to PC's display as shown in Fig. 8. We will construct a system, which responds according to the circumstances.

## References

- [1] Toru Yamaguchi, Eri Sato, Yasufumi Takama: Intelligent Space and Human Centered Robotics. *IEEE Transactions on Industrial Electronics*. **50** No.5 (2003) 881–889
- [2] Yamasaki, N. and Anzai, Y.: Active Interface for Human-Robot Interaction. *Proceedings of IEEE International Conference on Robotics and Automation*. **3** (1995) 3103–3109
- [3] Kunihiro Ohashi, Toru Yamguchi, Ikuo Tamai: Humane Automotive System Using Driver Intention Recognition. *SICE Annual Conference 2004*. CD-ROM (2004) 1164–1167
- [4] Kouji Kanagawa, Eri Sato, Toru Yamaguchi, Masako Miyaji: Acquisition of Kansei expression by study of mimicry and application to human support network robot system. *The 6th International Symposium on Advanced Intelligent Systems (ISIS2005)* (2005) 171–174

# Developing a Decision Support System for a Dove's Voice Competition

Chotirat Ann Ratanamahatana

Dept. of Computer Engineering, Chulalongkorn University, Bangkok 10330 Thailand  
ann@cp.eng.chula.ac.th

**Abstract.** Zebra dove's voice competition has become more popular in Thailand and many other South East Asian countries. Even though there are specific judging rules to follow, the results could still be somewhat subjective since each judge has various levels of expertise, knowledge, feeling, and experiences. Some criteria are quite straightforward and not very difficult to judge, but some could be very challenging and subjective due to several reasons including the limitations of human ears and personal preferences—aesthetic appreciation. This paper proposes a preliminary framework that could be used in developing a decision support system that exploits time series representation and signal processing that could potentially help alleviate this problem in judging the voice's quality of each zebra dove.

## 1 Introduction

Zebra doves' voice competition has become more popular among interested breeders in Thailand and many other South East Asian countries. The activity has first begun to enhance relationship among the participating countries by organizing a meeting and a zebra dove's voice competition. It has expanded from only a small group to an association, both nationally and internationally. A picture of zebra doves are shown in **Fig. 1**. In the competition each year, there usually are more than 1,500 contestants from Thailand, Malaysia, Singapore, Indonesia, etc. participating in the voice competition, as shown in **Fig. 2**.



**Fig. 1.** Zebra doves

Even though there are specific rules to judge each of the contestants, the results could still be somewhat subjective since each judge has various levels of expertise, knowledge, feeling, and experiences. However, each of the judges is responsible to select a dove with ‘best-quality’ voices. Some criteria such as volume, continuity, and the length of each syllable in the song are not very difficult to judge; some such as ‘sweetness’ and ‘mellowness’ are subjective to some extent due to several reasons including limitations of human ears and personal preferences. This paper proposes a preliminary framework for a decision support system using time series representation and signal processing techniques that could potentially help overcome this problem in judging the quality of the voice made by each dove.



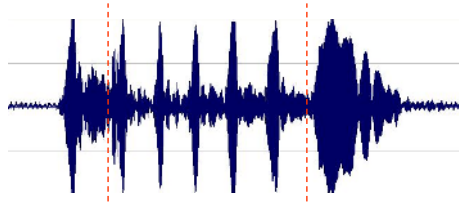
**Fig. 2.** An annual zebra dove competition at Yala province, Thailand

The rest of the paper is organized as follows. Section 2 gives a background about the competition, including the judging criteria. Section 3 shows the proposed method that could potentially help the evaluation. Section 4 provides some experiment, and Section 5 gives conclusions and offers a direction for future work.

## 2 Competition Rules and Judging Criteria

First and foremost, before we can work on analyzing the data, we need to understand the problem and the actual rules of the voice competition of the zebra dove.

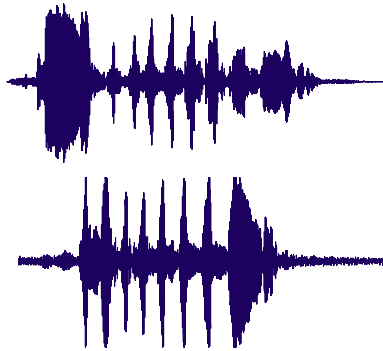
When searching for potential winners, judges will look for doves that have distinguished vocal quality. The competition is usually divided into three categories: Class A (Big/Bass voices), Class B (Medium/Tenor voices), and Class C (Small/Soprano voices.) The criteria in judging the quality of a dove’s coo is based on pitch, melody, volume, and continuity. When the competition starts, the doves have to sing/coo for about 3 hours at a stretch. The judges will look at the details in three stages of the song – the opening, the mid-section, and the closing. Fig. 3 shows a visualization of a song’s wave form plotted from the raw ‘.wav’ audio file.



**Fig. 3.** The three stages of the song by a typical zebra dove in the voice competition--the opening, the mid-section, and the closing stages

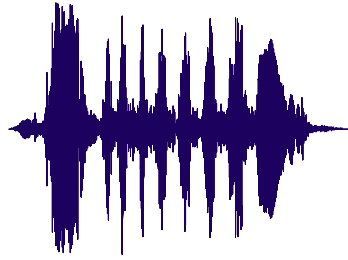
Specifically, the details in each stage are explained below:

1. The opening/ first note. This is the note that draws people's attention. A good first note has to be long, clear, and emphasized. A perfect overall song should have long first and last notes. However, the one with a short first note but a long last note is still better than the one with a long first note but extremely short last note or none at all. **Fig. 4** shows some examples of two different coos, one with a nice long first note and another with a very short first note.



**Fig. 4.** A comparison between the coo with a nice long first note (Top) and the coo with a very short first note (Bottom). The one with longer first note is generally considered a better coo, given that the last note is not too short as well.

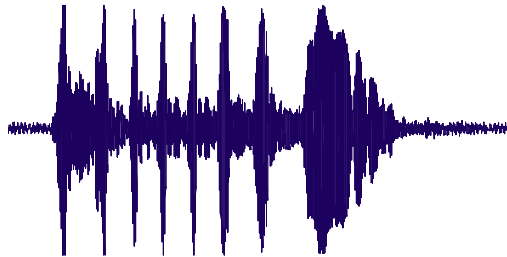
2. Middle note quality. Good middle notes must be slow, clear, and placed in a good rhythm. Even though we allow the middle notes to be either fast or slow, but the slow beats have become more and more popular. We count the middle notes together with the first and last note to be an overall beat. Some good zebra dove can call up to 6 beats, as in "Wow-ra-ka-ta-ta-kooong," and some only sing in 3 beats, as in "Wow-ra-kooong." **Fig. 5** shows an example of a coo that would get a lower score since the middle notes contain too many beats, and are relatively faster than coos by others. Though this criteria may seem reasonably straightforward to judge, in practice, the coo goes very fast and with human ears, it is extremely hard to count how many beats/syllable the bird actually made, where it is very evident from the audio plot in **Fig. 5**.



**Fig. 5.** An example of a 'too-frequent' coo. The middle notes are a bit too fast and have as high as 7 syllables/beats, which is considered too many by a judge.

3. The last note quality. The one with a long mellow note will get better score than a short one. A good quality last note should be long and powerful, similar to a cymbal sound.
4. Voice quality. A good quality voice can leave a good impression to a person hearing it; it should be "sweet," "mellow," "powerful," "not dry," and in a wrong tone. These voice's qualities are somewhat subjective and are much more difficult to judge, but we could use some signal processing procedure to analyze the sound, as will be explained in the next section.
5. Volume. A good volume should be loud enough to be audible from a long distance, or even when the bird is placed on a very high pole (7-9 meters). This characteristic can simply be analyzed by looking at the amplitude of the audio's wave form.
6. Continuity. This is to show the bird's strength and energy, and it is a good indicator of having good health, good lung, and high endurance. A god-quality dove should be able to coo continuously for more than 20 times in a row. But if the bird coos continuously for more than 50 times in a row, it may not be suitable for a competition because it tends to be exhausted and probably cannot last the coos for the whole 3 hours duration of the competition.

**Fig. 6** shows an example of so called the 'best quality' overall coo (the evaluation was made by human judges).



**Fig. 6.** An example of a very good overall coo, exhibited by high volume (high amplitude) and clearness with high consistency of each syllable with long and emphasized first and last notes

We can see that many of the criteria mentioned here are very difficult to be judged by 'untrained' human ears; some are restricted by our ears' limitation (such as counting the number of beats in a very fast song.) For completeness, the actual sound samples are placed at <http://www.cp.eng.chula.ac.th/~ann/dove/doves.zip> for interested readers. The next section will explain how we could exploit signal processing technique to help develop a decision support system to assist the judges' scoring process.

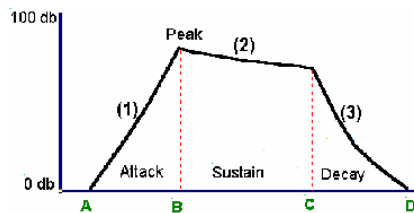
### 3 Audio Processing for a Decision Support System

From the previous section, we can see that several criteria can be readily judged by looking at the raw audio wave form alone. This includes volume, beat consistency, number of beats/syllables, and the stretch/length of each note. However, many other criteria can not be observed by looking at the raw audio alone; some are very subjective to each specific judge's preferences and experiences. In this section, we will suggest an approach for a decision support system that could be used in scoring process.

#### 3.1 Sound Components

Before further analysis, we need to first understand the components of sound we perceived. It can be categorized into Pitch, Timbre, Harmonics, Loudness, Rhythm, Attack, Sustain, Decay, and speed [3]. A desirable pitch for a competition should be relatively high. For highly skilled judges (or ones with 'perfect pitch' ears), identifying the pitch may not be difficult. For those who are not, pitch can easily be observed by looking at its frequencies obtained from a signal and Fourier analysis in our proposed framework. However, a good sound must consist of all ranges of frequencies, where low frequencies will make the sound warm and powerful, mid-range frequencies will give the sound its energy, and high frequencies will give a sound its "presences" and life like quality [3]. As mentioned in the previous section, loudness and rhythm of a sound can be observed from the raw audio waveform. The ones that are more difficult to judge by our ears are "timbre" or "quality of the sound" and harmonics. Harmonics or overtones give the pleasantness to the sound identified by the number and relative intensity of the upper harmonics (multiples of the fundamental frequency) present in the sound. The greater the number of harmonics, the more interesting and more pleasant the sound would be. This can also be seen in the Fourier analysis.

Timbre or quality of sound, on the other hand, is not clearly present in the Fourier analysis. The 'quality' of a sound generally describes sound characteristics which allow our ears to distinguish sounds with the same pitch and loudness. This is mainly determined by the attach-sustain-decay envelope of the sound, as shown in **Fig. 7**.



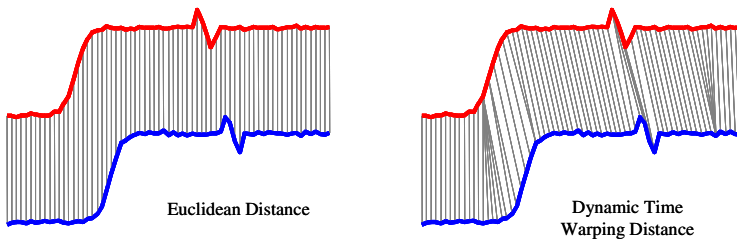
**Fig. 7.** An envelope of sound is composed of a sound's attach, sustain, and decay

A ‘mellow’ sound, generally referred to as a sound which is pleasantly smooth and free from any harsh qualities and does not draw the attention of a listener away from other sounds [1], therefore should not have a ‘sharp’ or sudden attack. Once the tone gradually reaches the desired pitch at B, it should drop slightly and remain steady until C, and then gradually decay to silence at D or until the new sound production. Generally speaking, we can look at the mellowness of the sound simply by looking the overall shape of the audio waveform.

### 3.2 Audio Processing and Similarity Measures

To give a score to a particular sound of a coo, we need to look at both raw audio waveform and the frequency content obtained from the Fourier analysis, each of which can be considered as a time series of numerical values that changes over time. This method is similar to query by humming [7], music retrieval [4][5], or bird calls recognition [6] methods that exploit the information from the audio’s spectrums. We then calculate the score by calculating the distance using similarity measures – We could either use a classic Euclidean distance metric, or a more computationally expensive but accurate Dynamic Time Warping distance measure.

The Euclidean distance is very simple to implement and is widely used; it is calculated by the sum of the squared difference between the two signals using one-to-one mapping. However, if there are some discrepancies in the time axis (x-axis), Euclidean distance may not give an intuitive mapping between the two. Dynamic Time Warping similarity measure can instead be used to alleviate this problem, as illustrated in **Fig. 8** below.

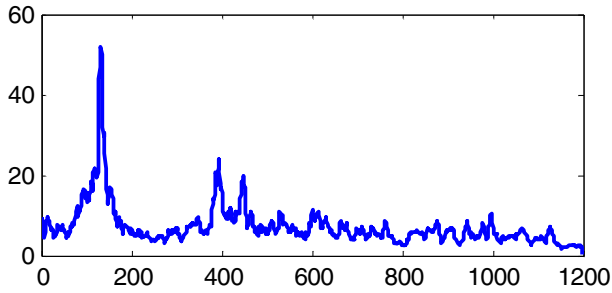


**Fig. 8.** Note that while the two time series have an overall similar shape, they are not aligned in the time axis. Euclidean distance, which assumes the  $i^{\text{th}}$  point in one sequence is aligned with the  $i^{\text{th}}$  point in the other, will produce a pessimistic dissimilarity measure. The non-linear Dynamic Time Warped alignment allows a more intuitive distance measure to be calculated.

Due to space limitations, we refer the interested readers to [2] for more details about Dynamic Time Warping distance measures.

Before we could use any distance measures above, a Fourier analysis must be first calculated; in this work, a standard Fast (Discrete) Fourier transform is used. **Fig. 9** illustrates a plot after Fourier transformation of the sound in **Fig. 6** with some smoothing effect to remove noise and outliers.





**Fig. 9.** A Fourier transformation (only subsection is shown) of the sound in **Fig. 6**

### 3.3 Algorithm

To develop a decision support system that gives scoring suggestion to the judges, we propose the following algorithm to calculate the score for each instance in the test set according to the judging rules, as explained below.

```

Input: z-normalized training_data, test_data
Output: score(n)-final score for each of n instance in the test data
for j = 1 to size(test_data)
    best_so_far = Inf;
    for i = 1 to size(training_data)
        score1 = dist1(env(training_data(i)), env(test_data(j)));
        score2 = dist2(FFT(training_data(i)), FFT(test_data(j)));
        total_score = ((score1*0.6) + (score2*0.4))*weight(i);
        if (total_score < best_so_far)
            best_so_far = total_score;
        end;
        score(j) = best_so_far;
    end;
end;
end;

```

All data must first be z-normalized so that they have zero mean and standard deviation of one. For each instance in the test data, it would try to find a best match in the training data (where the ranking of those sounds are already known). This is done by calculating the two scores: - the first one (dist1) is by looking at the envelope of the raw audio; the audio is then segmented into 3 parts (opening, mid-section, and closing, as mentioned in Section 2), then we compute the Euclidean distance (or DTW) of each part, before summing them up with weights 0.4, 0.3, 0.3, respectively (according to the judging rules) then add to the overall distance of the *whole* envelope. This score is to capture the overall quality of the sound, volume, rhythm, aggressiveness, and timbre. Score2 is calculated by computing the Euclidean distance (or DTW) between the Fourier transformations of the two transformed signals. This score is to determine pitches, melodies, and harmonic content. The total scores are then added up with 60% of Score1 and 40% of Score2, and then are rescaled by the weight (associated to the known ranking in the training data) of that particular training instance. At the end, we will end up with the best possible score for each test instance. Then the final ranking can be determined by sorting these final scores. However, we can also show all the raw scores breakdown if needed.

## 4 Experiment

We begin our experiment by first collecting the data as a training set. This involves evaluation from experts, ranking each example by the quality of the overall sound and the weights are given to each example accordingly. We have collected 37 sound samples from the 2005 competition with ranking scores ranging from the best quality sound to the unpleasant sound. We use another disjoint set of 8 examples for testing and give the ranking scores. Both good and “not-so-good” examples are included in the train/test sets. We are fully aware that this smaller training/test sets are quite small, but we would like to first see the quality of the evaluation of our preliminary framework before developing and extending it further to a functional decision support system. The results are reported in Table 1 below.

**Table 1.** The competition result predicted by our decision support system

Dove ID	Real Ranking	Predicted Ranking
A	3	3
B	8	8
C	1	2
D	4	5
E	7	7
F	2	1
G	5	4
H	6	6

## 5 Conclusion

From the result, we can see that our framework can give suggestive rankings that closely follow the real competition outcome, especially for the ones that do not have good quality sound. However, to better evaluate our system or to fully develop it to be used in practice, much larger training set is needed. Similar to an evaluation of any artwork, evaluation of voice or sound depends on human perception and his/her aesthetic appreciation. In this work, we therefore do not intend to replace a human judge with a machine. Instead, we are proposing a tool that could help facilitate the judging/scoring process, which can potentially be unintentionally overlooked by human.

## Acknowledgment

I would like to thank Chaiwat Eimthanakul and Pakit Kanchana for their assistances in the competition details. The facts about zebra doves and the competition rules are from <http://www.nokkhao.com> and <http://www.nokkaochava.com>.

## References

- [1] Erickson, R. (1975). Sound Structure in Music. Univ. of Calif. Press. ISBN 0520023765
- [2] Kruskal, J. B. & Liberman, M. (1983). The symmetric time warping algorithm: From continuous to discrete. In Time Warps, String Edits and Macromolecules.

- [3] Mott, Robert L (1990). *Sound Effects – Radio, TV, and Film*, Focal Press, pp. 53-70.
- [4] Sato, A., Hiraki, K., & Ichimura, T. (2001). A proposal of an estimation method for the affective value of music work using fuzzy regression analysis. In *Proc. of 5<sup>th</sup> Int'l Conf. on Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies*, pp.993-997.
- [5] Sato, A., J.Ogawa, J., & Kitakami, H. (2000). An impression-based retrieval system of music collection. In *Proc. of 4<sup>th</sup> Int'l Conf. on Knowledge-Based Intelligent System & Allied Technologies*, pp.856-859.
- [6] Wilde, M., Menon, V. (2003). Bird call recognition using Hidden Markov Models. Technical Report. EECS Dept., Tulane University.
- [7] Zhu, Y. & Shasha, D. (2003). Query by Humming: a Time Series Database Approach. *ACM SIGMOD '03*.

# Time Series Data Classification Using Recurrent Neural Network with Ensemble Learning

Shinichi Oeda<sup>1</sup>, Ikusaburo Kurimoto<sup>1</sup>, and Takumi Ichimura<sup>2</sup>

<sup>1</sup> Department of Information and Computer Engineering, Kisarazu National College of Technology, 2-11-1 Kiyomidai-higashi Kisarazu, Chiba 292-0041, Japan  
{oeda, kurimoto}@j.kisarazu.ac.jp

<sup>2</sup> Faculty of Information Sciences, Hiroshima City University, 3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194, Japan  
ichimura@its.hiroshima-cu.ac.jp

**Abstract.** In statistics and signal processing, a time series is a sequence of data points, measured typically at successive times, spaced apart at uniform time intervals. Time series prediction is the use of a model to predict future events based on known past events; to predict future data points before they are measured. Solutions in such cases can be provided by non-parametric regression methods, of which each neural network based predictor is a class. As a learning method of time series data with neural network, Elman type Recurrent Neural Network has been known. In this paper, we propose the multi RNN. In order to verify the effectiveness of our proposed method, we experimented by the simple artificial data and the heart pulse wave data.

## 1 Introduction

Various natural phenomena change at time. For instance, there are changes in temperature, the fluctuation of the economy, and human's brain or heart pulse wave, etc. Such information is usually measured as time series data that changes at intervals of time. Thus, the time series prediction is a major task in many areas of research, biology, physics, business and engineering, etc. Since the information technology has made great advances in decade, it came to be able to gather a large amount of various time series data easily through the Internet. However, it is difficult to acquire effective information and knowledge from a large amount of data.

To resolve such a problem, several authors have given a method of different types of RNNs(Recurrent Neural Networks) for time series processing. RNN that proposed Jordan[1] involves the use of recurrent links in order to provide networks with a dynamic memory. In this approach, hidden unit patterns are fed back to themselves; the internal representations which develop thus reflect task demands in the context of prior internal states. There are many ways in which this can be accomplished, and a number of interesting method have appeared in the literature[2]. In this paper, we propose a classification method using some recurrent neural networks for time series data with the mechanism of ensemble learning.

In the section 2, we briefly present some theoretical background of recurrent neural network. The section 3 describes the approach used in this work. An experimentation setup and results are presented in the section 4, and some conclusions are drawn towards the end. regardless

## 2 Recurrent Neural Network

Neural networks are developed primarily for the purpose of pattern recognition or classification. However, these are not well suited for modeling time series data. Because the original applications of neural networks were concerned with detection of patterns in arrays of measurements which do not change in time[2]. The dynamic nature of spatiotemporal data as time series requires introduction of additional mechanisms. In particular, a neural network used for time series processing must possess memory. The way to provide memory to the neural network is to store past values of output (state layer) or hidden (context layer) nodes in additional layers. These additional layers are connected to the hidden layer in a similar way as the input layer as shown in Figure 1 and 2.

### 2.1 Jordan Network

Figure 1 shows the approach for processing time series using neural networks, the so-called “Jordan network”[1]. It is a multi-layer neural network with one hidden layer and a feedback loop from the output layer to an additional input layer called “state layer”. Each node in the state layer is connected to itself via self-recurrent loops with a weight smaller than 1. Thus, a Jordan network takes into account not only past time series elements, but also its own forecasts.

### 2.2 Elman Network

The Elman recurrent neural network is a subclass of recurrent neural networks[3]. They are multilayer neural networks augmented with one or more additional context layers storing output values of one of the layers delayed by one step, as shown in Figure 2. The Elman network has context units, which store delayed the value of hidden layer output and present these as additional inputs to the network. The Elman network can learn sequences that cannot be learned with other recurrent neural network such as Jordan recurrent neural network, since networks with only output memory cannot recall inputs that are not reflected in the output. Several training algorithms for calculation of error gradient in general recurrent networks exist.

Both the input units and context units activate the hidden units; the hidden units then feed forward to activate the output units. The hidden units also feed back to activate the context units. Depending upon the task, there may or may not be a learning phase in this time cycle. In case of, the output is compared with a teaching signal, and back propagation of error is used to adjust connection strengths incrementally. The values of recurrent connection weight are fixed at 1.0 and are not subject to adjustment. At the next time step,  $t+1$ , the above sequence is repeated. In this calculation, the context units contain values which are equal to exactly the value of hidden units at time  $t$ . These context units thus provide a learning method of the network with memory.

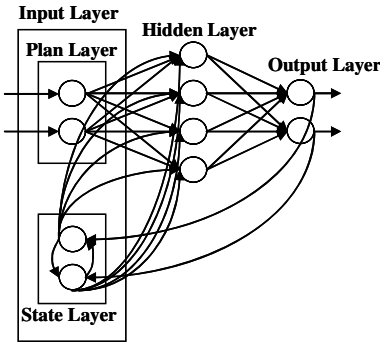


Fig. 1. Jordan network

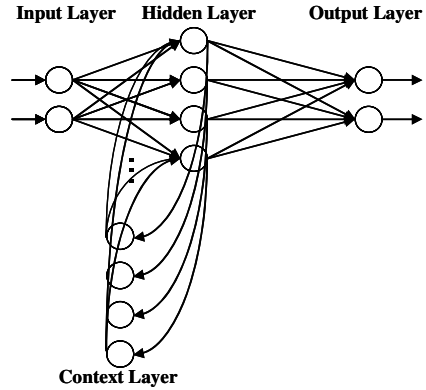


Fig. 2. Elman network

### 3 Recurrent Neural Network with Ensemble Learning

Many ensemble learning algorithms, such as bagging [4] and the Ada-boost [5] algorithm, try to improve upon the performance of a single learning machine by using many learning machines; such an approach has recently received considerable attention in the field of machine learning. Ensemble learning that combines the decisions of multiple hypotheses is some of the strongest existing machine learning methods.

A general strategy for designing ensemble learning systems is to treat the problem as one of combining multiple models, each of which is defined over a local region of the input space. Jacobs et al.[6] introduced such a strategy with their “mixture of experts (MoE)” architecture for supervised learning. The architecture involves a set of function approximates (“expert network”) that are combined by a classifier (“gating network”). These networks are trained simultaneously so as to split the input space into regions where particular experts can specialize. As shown in Figure 3, the mixture of experts (MoE) architecture is comprised of  $N$  expert networks. Each expert network solves a function approximation problem over a local region of the input space. The total output  $y$  which is the weighted sum of the expert network outputs, is given by

$$y = \sum_{n=1}^N w_n y_n \tag{1}$$

where  $w_n$  is the connection weights and  $y_n$  is the gating network output.

In this paper, we propose a classification method using some recurrent neural networks for time series data with the mechanism of ensemble learning. Some Elman networks corresponding to them is prepared for some kinds of time series data, then each of Elman network learns own time series data. In Elman network after it learns, when the time series data until time  $t$  now is input, the value at time  $t+1$  of the future can be obtained. Therefore, the target time series data to be classified until time  $t$  now is inputted each of Elman network, and each network calculate the value at time  $t+1$

of the future. Then, the time series data of  $t+1$  is actually observed, and an error of between the prediction outputs is obtained. Finally, the kind of time series data is classified to one network which has a minimum error as shown in Figure 4.

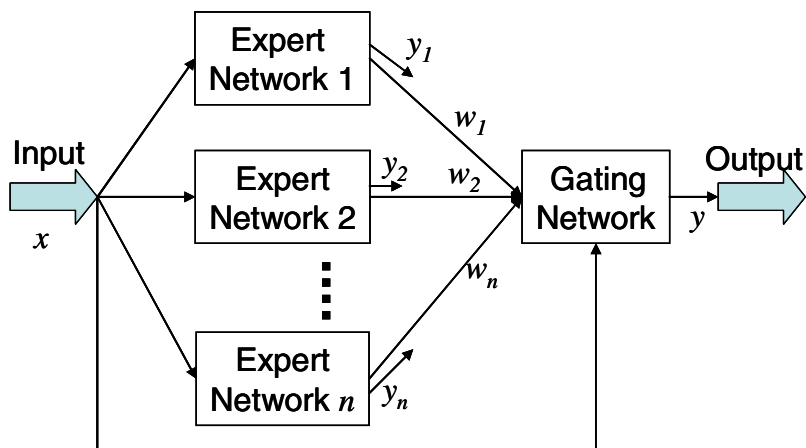


Fig. 3. The mixture of experts

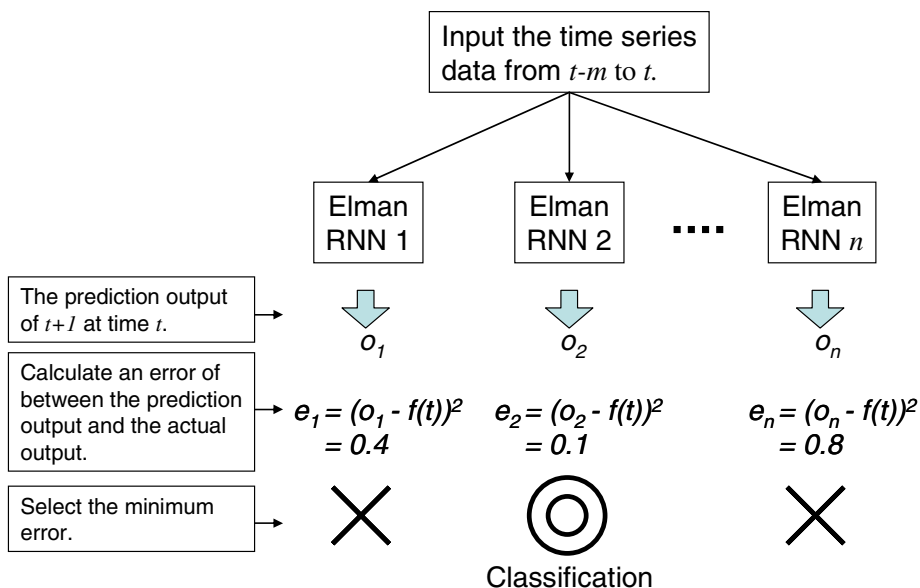


Fig. 4. The proposed classification method (RNN with ensemble learning)

## 4 Experimental Results

In order to verify the validity and effectiveness of proposal method, we experimented by the simple artificial data and the heart pulse wave data.

### 4.1 The Results for the Simple Artificial Data

We make the experimental time series data as the simple artificial data by  $f_A(x) = \sin(x)(0 \leq x \leq 2\pi)$  and  $f_B(x) = \log(x)(0 \leq x \leq 100)$ . Then, the range of values  $x$  are normalized into  $[0.0,1.0]$ . We prepared two Elman networks A and B to identify such two time series data. Network A and Network B are learned about the shape of  $f_A(x)$ ,  $f_B(x)$ , respectively.

In these experimental parameters, the input layer has 20 neurons, the hidden layer has 10 and the context layer has 10 neurons. The BP learning trains the network until the mean square error becomes 0.001. Figure 7 depicts the test data and the classification result. The classification result is represented as “1” is output when identified with  $f_A(x)$  and “0” is output when identified with  $f_B(x)$ . The identified result is 82.2%(684/784) correct answer ratio of test dataset.

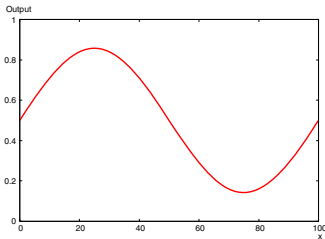


Fig. 5. The signal of  $f_A(x)$

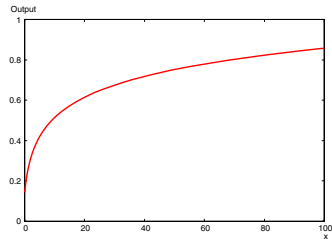


Fig. 6. The signal in  $f_B(x)$

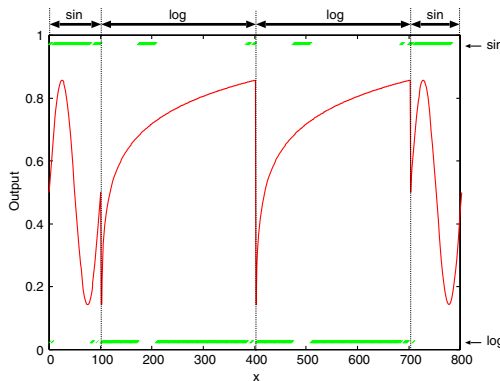


Fig. 7. The mixture of experts



## 4.2 The Results for Heart Pulse Wave Data

In order to investigate the classification capability for the practical use, we tried to apply the identification of person classification from heart pulse wave. In this paper, we prepared two people's heart pulse wave data. Figure 8 and 9 show the pulse wave of person A and the one of B, respectively.

In these experimental parameters, the input layer has 20 neurons, the hidden layer has 10 and the context layer has 10 neurons. The BP learning trains the network until the mean square error becomes 1.0. Figure 10 depicts the test data and the classification result. The classification result is represented as "1" is output when identified with person A and "0" is output when identified with person B. The identified result is 74.7%(2975/3980) correct answer ratio of test dataset.

We have the data of four total person of the data C and D besides A and B. Then, it experimented on the combinations other than A and B; e.g. (A,C), (A,D), (B,C), (B,D), (C,D). When a same experiment was done, the correct answer ratio from

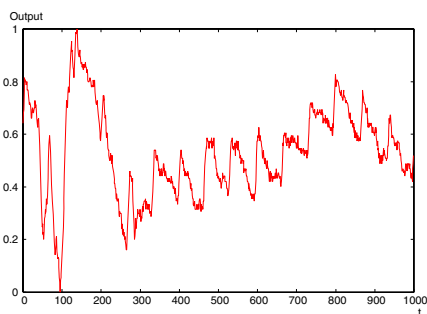


Fig. 8. The pulse wave of person A

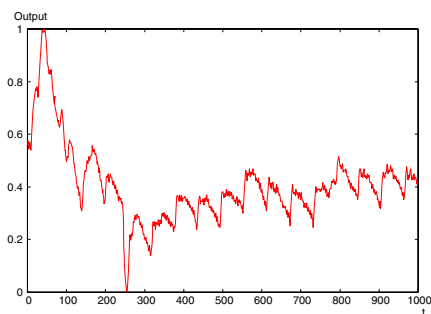


Fig. 9. The pulse wave of person B

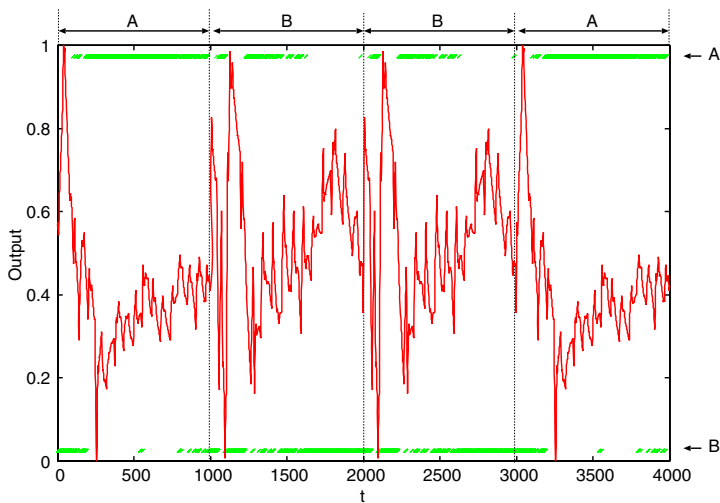


Fig. 10. The identification result for heart pulse wave

68 to 74% was able to be obtained. Physiologic data is considerably different depending on the individual. We will experiment by more data.

## 5 Conclusion and Future Work

In this paper, we proposed a classification method using some recurrent neural networks for time series data by imitating mechanisms of ensemble learning.

We applied the proposed method to classify the simple artificial data and the heart pulse wave data as experimental data. The experiments show that the proposed method obtained identification at a high correct accuracy ratio.

However, more experiments in complex situations are also required for applying it to actual problems. For instance, we consider the proposed method is applied to NIRS(Near-Infrared Spectroscopy) signal. NIRS optical topography was developed in recent years for cerebral function analysis[7]. This equipment is caught as an amount of oxygenated and deoxygenated hemoglobin of the blood flow within cerebral cortex, measuring transmitted near-infrared absorption signal. This procedure is called near-infrared spectroscopy, it abbreviates to Following NIRS. The measuring method and the analysis procedure are not yet established. Since we expect that the proposed method will become an effective method to classify NIRS signal, we apply to NIRS signal.

## Acknowledgements

Thanks to Yukari Suzuki and Nozomi Toyozaki implementations of several key components of the system described in this paper.

## References

1. Jordan M. I., "A Parallel Distributed Processing Approach", ICS-USCD, No. 8604, 1986.
2. G. Dorffner, "Neural Networks for time series processing", *Neural Network World*, 6(4), pp.447-468, 1996.
3. Elman, J. L., "Finding Structure in Time", *Cognitive Science*, Vol. 14, pp. 179-211, 1990
4. Breiman L., *Machine Learning*, Vol. 24, No. 123, 1996.
5. Freund Y., Shapire R. E., *Journal of Comp. and Sys. Sci.*, Vol. 55, No.119, 1997
6. Jacobs, R. A. Jordan, M. I. Nowlan, S. J. and Hinton, G. E., "Adaptive mixtures of local experts", *Neural Computation*, Vol.3, pp.79-87, 1991.
7. Hideaki Koizumi et-al, "Non-stress Higher Brain functional imaging, The Institute of Electronics, Information and Communication Engineering, Vol.87 No3, pp207-214, March 2004.

# Expressed Emotion Calculation Method According to the User's Personality

Kazuya Mera and Takumi Ichimura

Faculty of Information Sciences, Hiroshima City University,  
3-4-1, Ozuka-Higashi, Asa-Minami-Ku, Hiroshima, Japan  
{mera, ichimura}@its.hiroshima-cu.ac.jp  
<http://mera.ariake.jp/>

**Abstract.** In human communication, people do not always express the emotions they feel inside. In this paper, we propose a method to calculate “expressed emotions” from “aroused emotions” based on the personality of the agent, the situation of the conversation, and relationship with the partner. We consider five types of personality factors (extroversion, agreeableness, conscientiousness, neuroticism, and openness to experience) based on the “Big 5” model. The effect of each personality factor is calculated to amplify/suppress the aroused emotions. The amplification/suppression effects from five personality factors are collated and the degree of aroused emotion is calculated.

## 1 Introduction

There are a lot of computer systems in our daily life and we often encounter the situation of human-human communication via computer and human-computer communication. In order to realize comfortable communication, the interface tools have been developed such as GUI, avatar, facial expression, voice, eye contact, and so on. One of the methods is considering the emotional process of the user. If the computer can guess the user's emotion, it will avoid irritating that user, express sympathy to the user, and sometimes cheer up the user. On the other hand, if the computer can calculate its own emotions and expresses the emotions to the user, the user can enjoy communicating and they may be tolerant to the computer's behavior.

In order to calculate a human's emotion, we proposed “Emotion Generating Calculations (EGC)” method [2]. The EGC method calculates pleasure/displeasure and classifies it into 20 types of emotions using personal preference information. The EGC method is applied into e-mail software [3] and chat system [4]. These interactive systems express aroused emotions by synthesized facial expression.

However, when people communicate with each other, they do not always express the real emotions that they feel or that are aroused. For example, people who are outgoing will amplify their emotions while shy people will suppress their emotions. Furthermore, the situations of the conversation and the relationship with the partner are also amplification and suppression of the aroused emotions. Such adjustment by personality, situation, and relationship is very important among human communication and if the computer conversation system does not have such ability, it will give the impression of being detached, unsympathetic, and unemotional to the user.

In this paper, we propose a method to calculate “expressed emotions” from “aroused emotions” considering the personality of the agent, the situation of the conversation, and relationship with the partner. The aroused emotions calculated by the EGC method are amplified or suppressed based on the personality. Some personalities suppress all emotions and the other personalities amplify specified emotions. The degree of amplification and suppression is calculated using the function for amplification or suppression. Each emotion has different amplifying/suppressing rates. We consider five types of personality factors (extroversion, agreeableness, conscientiousness, neuroticism, and openness to experience) based on the “Big 5” model.

In Section 2, we briefly explain about the EGC method used to calculate aroused emotions. Section 3 introduces the expressed emotion calculation method for each personality factor. The conclusion is presented in Section 4.

## 2 Emotion Generating Calculation Method

### 2.1 Process of Emotion Generating Calculation Method

In our method, we calculate expressed emotion from aroused emotion. In order to realize this process, we have to obtain “aroused emotion” first. We propose the EGC method which can calculate 20 types of emotions from an input sentence based on the preference information for the object in the sentence. In this section, we briefly explain about the process of the EGC method.

Firstly, the EGC method calculates pleasure/displeasure. The EGC method assumes three-dimensional space to calculate pleasure/displeasure. We define three important elements for each type of event concept. The system constructs the synthetic vector based on the elements’ preference values named “Favorite Value (FV)” and distinguishes pleasure/displeasure by judging which area the vector is in. The FVs are given a real number on a ratio of  $[-1.0, 1.0]$ . The method also calculates the degree of extracted emotion from the length of the synthetic vector.

Next, the system classifies the pleasure/displeasure into 20 various emotions. It requires judging some conditions as follows; “feeling for another,” “prospect and confirmation,” “approval/disapproval.” We consider 20 emotion types which are classified into six emotional groups as follows; “joy” and “distress” grouped under “Well-Being”; “happy-for,” “gloating,” “resentment,” and “sorry-for” grouped under “Fortunes-of-Others”; “hope” and “fear” grouped under “Prospect-based”; “satisfaction,” “relief,” “fears-confirmed,” and “disappointment” grouped under “Confirmation”; “pride,” “admiration,” “shame,” and “disliking” grouped under “Attribution”; “gratitude,” “anger,” “gratification,” and “remorse” grouped under “Well-Being/ Attribution.”

### 2.2 Application Using Emotion Generating Calculation Method

The EGC method calculates the agent’s emotion from sentences. We applied the EGC method to a mail system and a chat system. Both of them analyze the content of the mail/chat and calculate the emotion from the point of view of the agent. Successively, calculated emotion is expressed by synthesized facial expression as shown in Fig. 1. In order to generate facial expression, firstly the 20 types of emotions are classified

into 6 types of groups. Each group corresponds to the typical facial expression; anger, disgust, fear, happiness, sadness, and surprise. Then, the degree of each group is calculated from the sum of the emotion degrees. Thirdly, the facial expression is synthesized by a sand glass type neural network trained using real face images [5].

This process synthesizes the facial expression based on the EGC output. However, expressing aroused emotion directly may cause some trouble as I described before. Therefore, we add a new process to transfer aroused emotions into the emotions regarded for personality and situation between the EGC process and the facial expression generating process.

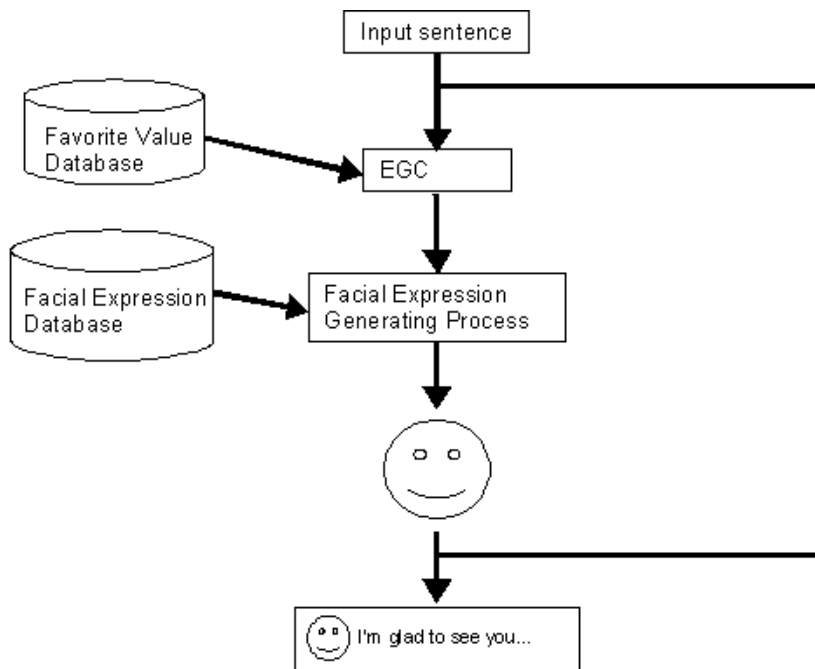


Fig. 1. Example of Face-to-Face Interaction System

### 3 Expressed Emotion Calculating Method from Aroused Emotion

#### 3.1 Aroused Emotion and Expressed Emotion

The EGC method calculates the emotions from the viewpoint of the agent. But, the calculated emotions only consider the agent's preference information. So, if the agent expresses the emotions, it may cause some troubles because it does not consider the relationship with the partner, the situation of the conversation, the feeling of the partner, and so on. In face-to-face communication, we rarely express our emotions directly unless we know the partner well or the situation of the conversation is relaxed.

In this paper, we define the emotions calculated by using the EGC method as “aroused emotion (AE)” and the emotions to be expressed to the partner as “expressed emotion (EE).”

The EGC method calculates the degree for each of the 20 emotions from an input sentence. Therefore, we define AE as a 20-dimensional vector and each axis indicates each emotion calculated from the EGC method. When the emotions are expressed, each emotion is amplified or suppressed by the agent’s personality, the situation of the conversation, the feelings of the partner. The amplified or suppressed emotions are synthesized to a 20-dimensional vector and we define it as EE. We show the relationship between AE and EE as follows;

$$EE_{anger} = f_{anger}(AE_{anger}) \tag{1}$$

$AE_{anger}$  and  $EE_{anger}$  show the value of the emotion type “anger” in AE and EE, respectively.  $f_{anger}$  shows the translated function for the emotional type “anger.” The function is different among individuals. We explain the detail of the function in Section 3.4. By applying these functions into each emotion in AE, EE is calculated as shown in Fig. 2. Then, the calculated EE is used to express the agent’s emotion as shown in Fig. 1.

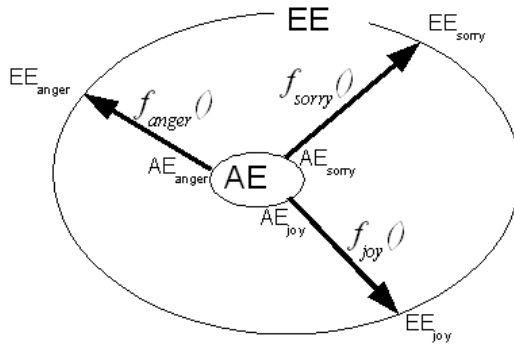


Fig. 2. Relationship between AE and EE

### 3.2 Personality Model

Many researchers over the years tried to classify human personality but it is difficult to define them clearly. After nearly 50 years of sophisticated factor analysis research, there is something of a consensus on what these basis factors are. The key factors in this so-called “Big 5” model are described below [6].

1. *Extroversion.* This dimension describes whether a person is active, sociable, outgoing, and optimistic at one end of the continuum or passive, quiet, unsociable, and careful at the other.
2. *Agreeableness.* This dimension describes whether a person is goodnatured, gentle, cooperative, trusting, and helpful at one end of the continuum or irritable, ruthless, suspicious, uncooperative, and headstrong at the other.

3. *Conscientiousness*. This dimension ranges from responsible, self-disciplined, and achieving at one end to irresponsible, careless, and undependable at the other.
4. *Neuroticism*. This dimension describes whether a person is touchy, nervous, and anxious at one end or calm, even-tempered, and carefree at the other.
5. *Openness to experience*. People who rate high in this dimension are intelligent, open to new ideas, and interested in cultural pursuits. Low scores tend to be conventional, down-to-earth, narrower in their interests and inartistic.

We propose a method to calculate emotion amplification/suppression for each personality factor in the “Big 5” model.

### 3.3 Effect from Personality Factors

#### 3.3.1 Extroversion/Introversion

In general, extrovert people tend to express their emotions in a more outward manner than actually felt/aroused. On the other hand, introvert people tend to suppress their aroused/felt emotions. The tendency gives the effect for all aroused emotions. Therefore, we define the degree of amplification/suppression as follows.

$$\begin{cases} \text{if } p \geq 0 & AMP_x^{extro} = f_1(p) \\ \text{if } p < 0 & SUP_x^{extro} = f_1(-p) \end{cases} \tag{2}$$

$p$  is the degree of extroversion and its value is in  $[-1.0, 1.0]$ .  $x$  is the type of the amplified/suppressed emotion. In this personality, all emotions are applied into  $x$ . Then,  $AMP_x^{extro}$  indicates the amplification degree of the aroused emotion  $x$  from the factor of extroversion, and  $SUP_x^{extro}$  indicates the degree of suppression. In this paper, we define the function  $f$  as a direct proportion as shown in Fig. 3.

Furthermore, the introverted people are often shy towards unfamiliar people. Their behavior can be expressed by the following functions.

$$SUP_x^{extro2} = f_2(fam) \tag{3}$$

$$f_2(x) = \frac{1}{1 + e^{(0.5-x)\alpha}} \tag{4}$$

$fam$  is the degree of familiarity to the partner and it is in  $[0.0, 1.0]$ .  $AMP_x^{extro2}$  also indicates the suppressive effect caused by introversion. The function  $f_2$  is defined as shown in Fig. 4.  $\alpha$  is defined as 4 to suit the range of  $AMP_x^{extro2}$   $[0.0, 1.0]$ . The function indicates that the attitudes for close people and for unfamiliar people are quite different. In human communication, this situation is called “glass wall.” This function simulates it using a sigmoid function, so that when the degree of familiarity increases and rises above the threshold required to break the “glass wall” of the mind, he/she suddenly becomes friendly and expresses their true feelings/emotions. Our method can simulate such a “glass wall” of the mind. By adjusting the sigmoid function, we can simulate the detailed introversion personality.

When we apply this method into the communication tool, the degree of familiarity can be calculated from preference to the partner, communication frequency, and the

log of aroused emotion from the partner. This data can be collected easily from the user’s profile, system log, and parsing result of the utterances. Table 1 shows an example to calculate the degree of familiarity from such data if our method is applied into a chat system.

**Table 1.** Example of the Degree of Familiarity Calculation

familiarity	relationship	calculated from
increase	user knows partner well	quantity of partner’s data
	user is familiar with partner	frequency of chat with partner
	partner is user’s superior	partner’s profile
	user likes partner	FV for partner
	user wants to be liked by partner	user’s input or user’s conversation tactics
decrease	user doesn’t knows partner well	quantity of partner’s data
	user isn’t familiar with partner	frequency of chat with partner
	partner is user’s inferior	partner’s profile
	user dislikes partner	FV for partner
	user wants to be liked by partner	user’s input or user’s conversation tactics

The calculated amplification and suppression degree from five factors are collated at the end of the calculation and applied to calculate the degree of expressed emotions.

**3.3.2 Agreeableness**

In general, a person who is agreeable tends to take care of others and express sympathy to others. The EGC method which is used to calculate aroused emotions can extract the emotion which relates to others; “happy-for,” “sorry-for,” “gloating,” and “resentment.” Gentle and helpful people express sympathetic emotions like “happy” and “sorry” more than other people in general. On the contrary, ruthless and headstrong people express “gloating” and “resentment” more than other. Furthermore, sympathetic people express emotions more outwardly than usual when their partner feels the same emotion, i.e. “sympathy.” However, people generally also have such a tendency more or less. Therefore, we define the degree of amplification/suppression for agreeableness as follows.

$$AMP_x^{agree} = f_3(p) \quad x \in \{happy, sorry, same\ as\ partners'\} \tag{5}$$

$$SUP_x^{agree} = f_3(p) \quad x \in \{gloating, resentment, different\ to\ partners'\}$$

$$f_3(x) = \alpha x + \beta \tag{6}$$

The emotion which is “the same as the partner feels” does not indicate a complete match. All the calculated emotions from the EGC method can be classified into pleasure/displeasure. So, if the aroused pleasure/displeasure of the partner and the agent is the same, they are recognized as the same emotions in this process. Fig 5 shows the



function (6). In this function,  $\beta$  is the degree of general people's amplification/suppression. The value is different for each target emotion.

### 3.3.3 Conscientiousness

Conscientious people can consider the situation of the conversation and can suppress unsuitable emotions. For example, we should not laugh loudly at funeral and yell in anger at a wedding ceremony. The effect is also calculated by using a function the same as in (5). In this case, the target emotion  $x$  is different for each conversational situation.

$$\begin{aligned}
 AMP_x^{conscient} &= f_3(p) \\
 SUP_x^{conscient} &= f_3(p)
 \end{aligned}
 \tag{7}$$

### 3.3.4 Neuroticism

The people who have this tendency are usually touchy and nervous, and such people express the displeasure emotion more outwardly than other people. Therefore, we amplify such displeasure emotions based on the degree of neuroticism as follows.

$$\text{if } p \geq 0 \quad AMP_x^{neuro} = f_1(p)
 \tag{8}$$

In this paper, we do not give any effect to non-neurotic (stability) personality because it will not cause any effect to emotion amplification/suppression.

### 3.3.5 Openness to Experience

Although the definition of this factor is different from extroversion, there are not significant differences between these two factors from the point of view of emotion amplification/suppression. Therefore, we apply the same calculation to this factor, too.

$$\begin{cases}
 \text{if } p \geq 0 & AMP_x^{open} = f_1(p) \\
 \text{if } p < 0 & SUP_x^{open} = f_1(-p)
 \end{cases}
 \tag{9}$$

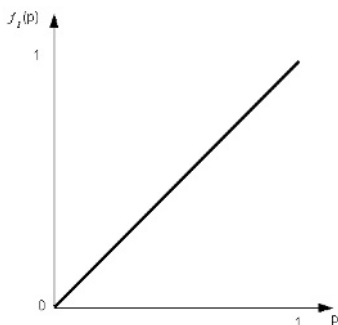


Fig. 3. Function Type 1

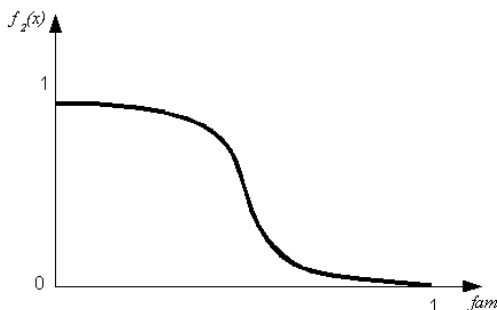


Fig. 4. Function Type 2

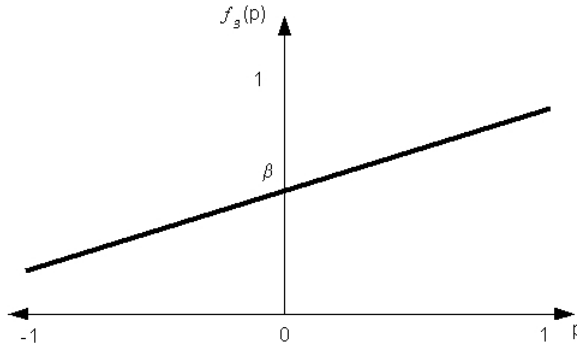


Fig. 5. Function Type 3

### 3.4 Expressed Emotions Calculation

We proposed eight functions to calculate the effect from five personality factors in the “Big 5” model. However, individuals can have multiple personalities and it often causes conflicts. In order to unify these effects, we define the following calculations to calculate the degree of expressed emotion from the aroused emotion as follows.

$$EE_x = AE_x \times (1 + (AMP_x - SUP_x))$$

$$AMP_x = \max(AMP_x^{extro}, AMP_x^{agree}, AMP_x^{conscient}, AMP_x^{neuro}, AMP_x^{open}) \quad (10)$$

$$SUP_x = \max(SUP_x^{extro}, SUP_x^{agree}, SUP_x^{conscient}, SUP_x^{neuro}, SUP_x^{open})$$

In general, when the amplification factor and suppression factor are aroused at the same time, we feel “conflict” and some special behavior like a defense mechanism may be caused. However in this paper, we just calculate the difference between the amplification degree and suppression degree.

## 4 Conclusion

We proposed a method to calculate expressed emotions from aroused emotions according to the agent’s personality, the situation of the conversation, and the relationship with the partner. At first, we defined AE (Aroused Emotion) and EE (Expressed Emotion). AE is a 20-dimensional vector where each axis consists of the output emotion of the EGC method. Each emotional value of AE is amplified or suppressed and the results consist of a 20-dimensional vector that is defined as EE. The amplification/suppression of the emotion is decided from personality and the situation of conversation. We defined the emotion amplification/suppression calculations for each personality factor in the “Big 5” model. Extroversion amplifies all emotions and Introversion suppresses all emotions. Introversion also changes attitudes to unfamiliar people. Next, Agreeableness and Conscientiousness change the degree of amplification/suppression when the partner feels/arouses some emotions or the conversation is in a special situation. Then, Neuroticism amplifies all the displeasure emotions and

Openness to experience has the same effect as Extroversion. The effects from the personality factors are collated and then used to calculate expressed emotions from felt/aroused emotions.

In this paper, we defined the calculation for the factors in the "Big 5" model. However, by extracting and paraphrasing some factors, our method can be applied into many other personality models such as Hippocrates, Eysenck [7], and Goldberg [8].

We can simulate various personalities by applying this method into interactive systems like mailer and chat systems when we prepare personality input interface and feature extracting methods. In future work, we will apply this method in communication systems such as e-mail systems or chat systems and research the effect of the new system on the user.

## References

1. Ichimura T., Mera K., "Dialogue system with emotion through computer network", Japan Patent No.2005-237668 (2005)
2. Mera, K.: Emotion Orientated Intelligent Interface, Doctoral Dissertation, Tokyo Metropolitan Institute of Technology, Graduate School of Engineering (2003).
3. Ichimura T., Mera K., et al., "An Emotional Interface with Facial Expression by Sand Glass Type Neural Network and Emotion Generating Calculations Method", Proc. of The International Symposium on Measurement, Analysis and Modeling of Human Functions, (2001) pp.275-280
4. Ichimura, T., Mera, K., et al., "Emotional Interface for Human Feelings by Mobile Phone", Proc. of KES2002, Vol. 1 (2002) pp.708-712
5. Ishida, H., Ichimura, T., et al., "Classification of Facial Expressions using Sandglass-type Neural Networks", Proc. of FAN2000, (2000) pp.201-204 (in Japanese)
6. Huffman, K., et al., "Psychology in Action 3rd edition," Wiley and Sons Inc. (1994).
7. Eysenck, H. J., "Biological dimensions of personality", in L. A. Pervin (Ed.), "Handbook of personality: theory and research", New York: Guilford Press (1990)
8. Goldberg, L. R., "The structure of phenotypic personality traits", American Psychologist, 48, (1993) pp.26-34

# Integrated Agent-Based Approach for Ontology-Driven Web Filtering

David Sánchez, David Isern, and Antonio Moreno

Universitat Rovira i Virgili (URV)  
Computer Science and Mathematics Department  
Artificial Intelligence Research Group, BANZAI  
43007 Tarragona, Catalonia, Spain  
{david.sanchez, david.isern, antonio.moreno}@urv.net

**Abstract.** For knowledge-intensive industries it is of paramount importance to keep an up-to-date knowledge map of their domain in order to take the most appropriate strategic decisions. The Web offers a huge amount of valuable information, but its interaction is very hard and time consuming for humans because it requires to filter, analyse all related web pages and integrate it in a knowledge repository. This paper describes an integrated agent-based ontology-driven approach to retrieve web pages that contain data relevant to each of the main concepts of the domain of interest in a completely automatic, unsupervised and domain independent way.

## 1 Introduction

The Web offers a huge amount of valuable information, but it is scattered, unstructured and impossible to analyse manually. It is usually searched by means of keyword-based search engines, allowing a user to retrieve information by stating a combination of keywords. The results of this type of search usually suffer from two problems derived from the nature of the query and the lack of structure in the documents: low precision and recall ratios. Furthermore, while search engines provide support for automatic information retrieval, the tasks of extracting relevant data and its further processing remain to be done by the user.

In the last years it has been argued that the performance of a search engine can be improved by using ontologies [1]. Ontologies allow organizing and centralizing knowledge in a formal, machine, and human understandable way, making themselves an essential component to many knowledge-intensive environments like the Semantic Web, knowledge management, and electronic commerce. In consequence, they provide a semantic ground that can help to sort out web pages with relevant information about a concept from web pages that contain data with just syntactic similarities to the concept. However, ontologies are traditionally built entirely by hand and, in consequence, their creation and management requires a significant amount of human effort that can compromise the performance and applicability of knowledge based tools.

To tackle those problems, this paper presents an integrated approach for web information retrieval and filtering, providing a tool for obtaining web pages that contain relevant information about the main concepts of a domain, expressed by an automatically obtained domain ontology. This ontology has a hierarchical tree structure that contains the basic classes (concepts) of the domain and the main characteristics (attributes) of each concept.

The system uses two previously developed tools for knowledge acquisition and information retrieval. Concretely, in [2] an automatic, domain independent, web based ontology learning method is presented. Its results (machine readable ontologies for any domain) can be used as input for the system described in [3], which implements methods and techniques that allow the use of the information contained in the domain ontology in order to move from a purely syntactic keyword-based web search to a semantically grounded search. The final result is a structured representation (in an ontological fashion) of the main concepts for a certain domain, which is used to retrieve, filter and classify the most relevant web resources available in the Web. As the processing required to treat with a huge repository like the Web is a very time consuming task, the full system is presented in a distributed approach. More concretely, in order to provide a scalable solution, the agent paradigm is a promising technology for information retrieval. *Multi-agent* systems provide advantages with respect to traditional systems such as scalability, flexibility and autonomy [4] and they are very suitable for implementing dynamic and distributed systems like the presented.

As a summary, the main features of our contribution are: *i)* unsupervised operation during the analysis, learning process and filtering of Web resources. This is important due to the amount of resources available, avoiding the need of a human expert on the searched domain. *ii)* automatic operation, allowing to perform easily executions at any time in order to maintain the results updated. This characteristic fits very well with the dynamically changing nature of the Web. *iii)* domain independent solution, because no domain related assumptions are formulated and no predefined knowledge is needed. This is interesting when dealing with technological domains where specific concepts may appear.

The rest of the paper is organised as follows. Section 2 introduces the learning methodology used to obtain basic ontologies. Section 3 describes the ontology driven web information retrieval and filtering tool. Section 4 presents the agent-based integration of those two complementary approaches and discusses the evaluation of the results. The last section discusses related approaches and introduces some lines of future research.

## 2 Ontology Learning

The base for obtaining the basic ontologies for a domain, is the intensive use of a methodology [3] for acquiring knowledge from the Web. The most important characteristic of the method is that the whole process is performed in an automatic, unsupervised and domain independent way, allowing to obtain results without user's intervention.

The algorithm is based on analysing a significant number of web sites in order to find important concepts for a domain by studying the neighbourhood of an initial keyword. Concretely, in the English language, the immediate anterior word for a keyword is frequently classifying it (expressing a semantic specialization of the meaning) [5]. So, the *previous word* for a specific *keyword* is used for obtaining the taxonomical hierarchy of terms (*e.g. breast cancer* will be a subclass of *cancer*). The process is repeated recursively in order to create deeper-level subclasses (*e.g. metastatic breast cancer* will be a subclass of *breast cancer*).

In order to extract and select the most relevant concepts for a domain from the Web, the method relies on a search engine for accessing the available web resources. It constructs dynamically the appropriate queries for the search engine, obtaining the most adequate corpus of web resources at each time. Moreover, the Web search engine is also used to select the most appropriate concepts, checking their relevance for the specific domain (the strength of the taxonomical relationship) through a statistical analysis based on the number of estimated results returned by the search engine. This approach allows obtaining robust statistical measures (as they are based in the whole Web) in a very efficient and scalable way, and has been proved to be an effective strategy for inferring the degree of relatedness between concepts [6].

As an additional step, the taxonomy is filtered in order to detect if some of those selected concepts can be considered as properties or attributes of a specific class. More concretely, we consider that those terms that appear in several branches of the obtained hierarchical structure, are a common property of the immediate superclass. For example, if we find that among the different types of discovered cancers (*e.g. breast cancer, lung cancer, etc.*), many of them can have a common attribute (*e.g. metastatic breast cancer, metastatic lung cancer*), we consider this attribute as a property of the common superclass (*cancer*).

The result of the process is a hierarchical organization of the main concepts available for a domain according to the information contained in the Web, enriched with some automatically discovered attributes (see an example in Fig. 1). This structure is presented in an ontological fashion using a standard machine readable language that allows an easy integration with the rest of the system.

### 3 Ontology-Based Web Filtering and Ranking

Here, the ontology-based search system that explores the Web to find relevant pages related to the different concepts in an ontology is introduced (see [7, 2]). The retrieved pages are textual instances of the concepts, but conditioned to the meaning of the concept in the whole ontology. It means that the same concept in a different ontology would produce different results because it is in a different context. The content of the pages related to a particular concept are analysed in order to rank them according to a relevance function which takes into account the properties describing the user desired profile of such concept.

The search system wraps traditional search engines by an intelligent system that cuts the domain ontology into pieces (sub-ontologies) according to the

degree of concurrence, and scatters these pieces between the available search processes running in the computers involved in the process. A sub-ontology is formed by concepts (and attributes) from a leaf to the root node. Then, each search process works to obtain as many relevant pages as possible.

As the pages are retrieved, their contents are analysed and the relevance of the page is calculated in terms of the sub-ontology concepts and properties appearing in the documents. The relevance value is used to rank the pages and also to discard those pages which are not of the expected quality. The most relevant ones will populate the domain ontology, and will be joined asynchronously later in order to be sorted and filtered appropriately (see and example in Fig. 1).

In some situations, the amount of ontological information available could be too restrictive for the keyword based search engine (*e.g. microinvasive endobronchial non-small cell lung cancer*). In these cases, a complementary component was designed in order to modify these problematic sub-ontologies by removing the least representative concepts and give alternative queries less constrained.

## 4 Agent-Based Ontology-Driven Web Filtering

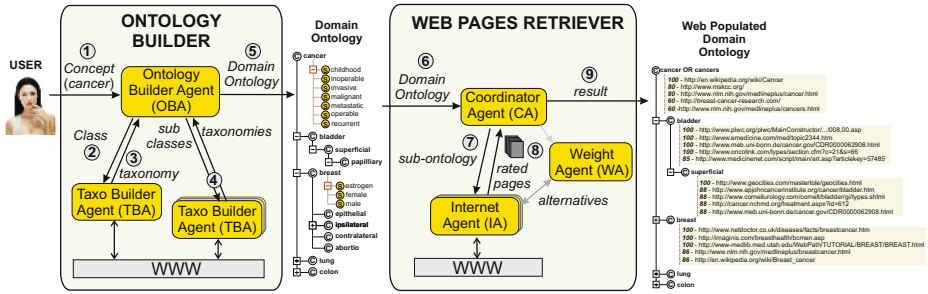
As has been introduced, due to the computational cost of the described Web based methodologies, we have opted for modelling them into a distributed agent-based platform that can be executed on the server side to which the user can access via a web interface. The full process has been divided into several tasks that are mapped into different types of agents. In this manner, those agents can be executed in parallel among nodes of a network, cooperating to achieve the common goal. Moreover, as shown in [4], agents provide the high degree of flexibility required for the dynamic management of the platform, and implement complex communication skills necessary for coordination.

As shown in Fig. 1, the multi-agent system is divided in two parts that correspond to each described methodology: the output of the ontology learning methodology is used as input of the web filtering process. As a consequence, different types of agents are created and managed dynamically (created, configured and finalized) according to the execution requirements at each moment, ensuring that the available computational resources (nodes of the computer network) are always efficiently maximizing the throughput of the system.

### 4.1 Ontology Learning

The *Ontology Builder* module is composed by two types of agents: an Ontology Builder Agent (OBA) and several Taxonomy Builder Agents (TBA).

The process starts when the OBA receives from the user the concept (*e.g. cancer*) that represents the domain to explore (step 1). Optionally, the user can specify some parameters to constrain the search process according to his desires as described in [3]. As OBA's goal is to construct a basic ontology that represents the available knowledge for this domain, it creates a first TBA that starts building a one level taxonomy using the methodology presented in §2



**Fig. 1.** Integrated agent based ontology driven web recommendation platform. Example results presented correspond to the *Cancer* domain.

(step 2). As a result, a set of immediate taxonomically related subclasses (e.g. *breast*, *lung*, *colon*) is returned to the OBA (step 3). The OBA incorporates this knowledge into the domain ontology and, for each new subclass, it creates a new TBA for exploring it.

Again, several sets of taxonomically related subclasses are returned to the OBA that incorporates them into the ontology (step 4). Repeating this process until no more subclasses are found, the OBA is able to compose recursively an ontology that taxonomically represents the available knowledge in the Web for the domain. As a final step, the OBA refines the ontology in order to detect attributes for each class (e.g. *metastatic cancer*) as described in §2 and outputs the result in a machine readable ontology representation language (step 5).

### 4.2 Web Retrieving and Filtering

The *Web Pages Retriever* module is composed by three types of agents: a *Coordinator Agent (CA)*, a *Weight Agent (WA)* and some *Internet Agents (IA)*.

As a result of the execution of the *Ontology Builder* module, the CA receives the automatically acquired *domain ontology* (step 6). Then, the CA divides the domain ontology and distributes the search work among the available IAs (step 7). It uses for this purpose a split operator that creates one ontology per class, keeping the whole path from the root node of the ontology and all the properties inherited from it [2] (*i.e.* all the superclasses of the concept, with all their properties). Each IA uses a standard keyword-based search engine to retrieve a set of web pages that are related to a concept of the domain. The agent uses the semantic information of the its subontology to filter and rank these pages, and sends them to the CA (step 8).

IAs may have problems for retrieving results if its subontology is excessively restrictive, as mentioned in §3. In this case, he can request the help of the WA. This agent is able to find less constrained sets of keywords that can be used by IAs to find more pages. This agent implements a Best First Search with all possible sequences of keywords to be considered by an IA. The agent maintains a sorted queue per IA with the alternatives to be sent when it is needed.



Parallely, the CA waits for all the IAs to supply the results. When those are provided, it incorporates the returned lists of web resources into the domain ontology that is presented to the user as the final result of her request (step 9).

At the end, for each automatically acquired concept, a set of 2-tuple formed by an URL and a rate is presented. This last value indicates the degree of relevance of the particular URL and its associated concept according to the ranking measure employed during the retrieval and filtering process. Note that due to a specificity policy implemented, no redundant results between classes and subclasses are presented.

It is important to note that the full system is not intended to provide an immediate response in a first moment as, depending on the queried domain, the computational resources available on the server side and the search engine response times, the full process can go from minutes to several hours. However, once a domain is explored, results can be consulted immediately and even updated in a very fast way (as previous results as the domain ontology can be reused).

### 4.3 Evaluation

As the present proposal is an integration of two previously developed tools, the quality of the final results depends on the performance of each methodology. Regarding to the evaluation of the taxonomies obtained by the first module, a discussion is offered in [3], offering a comparison against a gold standard and several taxonomical web search engines. With respect to the second module, in [2] are presented several evaluations against different technological domains starting from ontologies composed manually by experts.

The full platform has been tested in several domains as medicine, biotechnology and computer science. The evaluation has been performed by comparing the results against the web search engine used during the analysis (Google). More concretely, for the list of URLs retrieved for each automatically discovered concept,

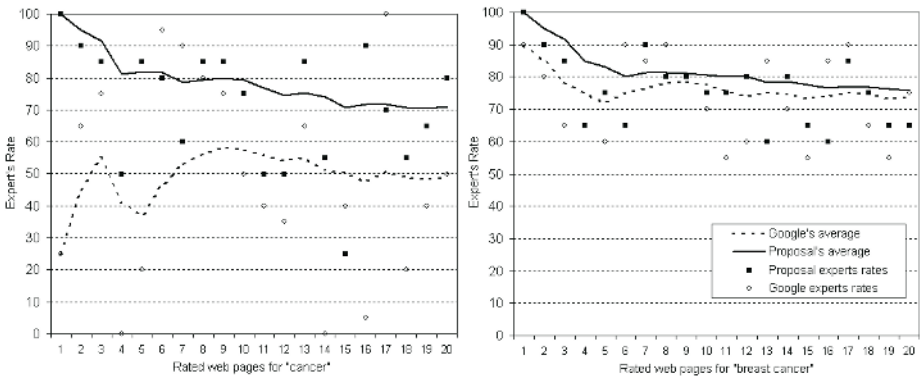


Fig. 2. Expert's rating for the first 20 web pages returned by our approach against the ones retrieved by Google for the *Cancer* and *Breast Cancer* concepts

two users are requested to rate each web site according to their degree of interest for the particular domain with a value between 0 and 100. The same process is repeated for the first web sites returned Google when manually querying the same acquired concept. These ratings will indicate which approach returns, in average, the most interesting set of web resources for the particular domain.

As an example, in Fig. 2, expert's rating for our results against Google for a pair of concepts of the *cancer* domain are presented. One can see that, for the most general concept (*cancer*), the quality of our results overpass significantly, in average, the ones presented by Google. This behaviour have been observed for several tested domains and it is caused by the higher contextualization that the presented approach can apply to the web sites analysis thanks to the automatically acquired knowledge for the domain. Observing the average rating for a more concrete concept (*breast cancer*), we can see that the quality of the returned web sites by each system is very similar. In this case, the search is, in both cases, contextualized enough to retrieve high quality resources.

## 5 Discussion and Future Work

On the one hand, some authors have been using the Web as a learning corpus for developing [8] or enriching knowledge structures [9], presenting techniques adapted to this environment. On the other hand, the use of knowledge structures (e.g. thesaurus like WordNet) is a common approach for improving the performance of Web information retrieval [10]. However, results for those approaches depend on the domain coverage of the knowledge base used.

On the contrary, the presented proposal does not start from any predefined knowledge and, in consequence, it can be applied over domains that are not typically considered in semantic repositories, but maintaining the semantic context provided by the automatically and unsupervisedly obtained domain ontology. At the end, we can bring the benefits of unsupervised and domain independent and, at the same time, semantically grounded Web information retrieval together into an integrated agent-based approach.

Moreover, the distributed and coordinated agent-based execution, improves the scalability and the throughput of the system, by taking profit of a parallel execution through several nodes of a computer network. In this sense, agent's flexibility (such as dynamic management and adaptation to the execution requirements of each moment) and communicative skills [4] have been crucial points modelling the presented platform. These facts result in a scalable and suitable method for acquiring knowledge and retrieving relevant web resources from a huge and dynamic repository as the Web.

Applications of the results obtained can be: *a)* the domain ontology builder can be a great tool for structuring automatically the Web's knowledge. The acquired ontologies are crucial in many knowledge intensive tasks such as electronic commerce, knowledge management or the Semantic Web; *b)* the automatic filtering, ranking and structuring of web resources can be considered as an improvement over the classical way of accessing web sites.

As future lines of research, some topics can be proposed:

- When dealing with natural language resources like web pages, problems about semantic ambiguity may arise. For that reason, we have developed complementary methods for dealing with polysemy and synonymy [11] specially adapted to our working environment. We plan to integrate those techniques into the learning methodology in order to improve the results.
- The multi-agent system versatility can be improved by incorporating communication and negotiation capabilities between agents that can allow them to share intermediate results in order to avoid redundant searches improving the learning performance.

## Acknowledgements

The work has been supported by *Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya i del Fons Social Europeu* of Catalonia.

## References

- [1] Fensel, D.: *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer Verlag (2001)
- [2] Moreno, A., Riaño, D., Isern, D., Bocio, J., Sánchez, D., Jiménez, L.: Knowledge Exploitation from the Web. In: 5th International Conference on Practical Aspects of Knowledge Management (PAKM'04). Volume 3336 of LNAI., Springer Verlag (2004) 175–185
- [3] Sánchez, D., Moreno, A.: Agent-Based Knowledge Acquisition Platform. In: 9th International Workshop on Cooperative Information Agents (MATES/CIA 2005). Volume 3550 of LNAI., Springer Verlag (2005) 118–129
- [4] Wooldridge, M.: *An Introduction to Multiagent Systems*. John Wiley and Sons, Ltd., West Sussex, England (2002)
- [5] Grefenstette, G.: SQLET: Short Query Linguistic Expansion Techniques: Palliating One-Word Queries by Providing Intermediate Structure to Text. In: *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, RIAO'97. Volume 1299 of LNAI., Springer Verlag (1997) 97–114
- [6] Turney, P.: Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: *Twelfth European Conference on Machine Learning*. (2001)
- [7] Bocio, J., Isern, D., Moreno, A., Riaño, D.: Semantically Grounded Information Search on the WWW. In: *Artificial Intelligence Research and Development*. Volume 100., IOS Press (2005) 349–356
- [8] Navigli, R., Velardi, P.: Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics* **30** (2004) 151–179
- [9] Agirre, E., Ansa, O., Hovy, E., Martínez, D.: Enriching very large ontologies using the WWW. In: *Workshop on Ontology Construction of the European Conference of AI (ECAI'00)*. (2000)
- [10] Abramowicz, W.: *Knowledge-Based Information Retrieval and Filtering from the Web*. Springer Verlag (2003)
- [11] Sánchez, D., Moreno, A.: Development of new techniques to improve web search. In: *9th International Joint Conference on Artificial Intelligence*. (2005) 1632–1633

# A Constrained Spreading Activation Approach to Collaborative Filtering

Josephine Griffith<sup>1</sup>, Colm O’Riordan<sup>1</sup>, and Humphrey Sorensen<sup>2</sup>

<sup>1</sup> Dept. of Information Technology,  
National University of Ireland, Galway, Ireland  
josephine.griffith@nuigalway.ie, colm.oriordan@nuigalway.ie  
<http://www.it.nuigalway.ie/cirg/>

<sup>2</sup> Dept. of Computer Science, University College Cork, Cork, Ireland  
h.sorensen@cs.ucc.ie

**Abstract.** In this paper, we describe a collaborative filtering approach that aims to use features of users and items to better represent the problem space and to provide better recommendations to users. The goal of the work is to show that a graph-based representation of the problem domain, and a constrained spreading activation approach to effect retrieval, has as good, or better, performance than a traditional collaborative filtering approach using Pearson Correlation. However, in addition, the representation and approach proposed can be easily extended to incorporate additional information.

## 1 Introduction

There is much evidence from various domains within Information Retrieval that the combination of sources of evidence leads to more effective retrieval [7]. With respect to recommender systems, Basu et al. claim that “there are many factors which may influence a person in making choices, and ideally, one would like to model as many of these factors as possible in a recommendation system” [3]. Although the types of information combined and the techniques used vary substantially across the domains, it has been shown that there are advantages to be gained from considering more than one source of information. Two of the primary issues to consider when combining multiple sources of information are that of the representation and the approach that can be used to model the problem.

The aims of this work are to show that not only does a graph representation and spreading activation approach provide good performance but furthermore that both the representation and approach can be easily extended to incorporate additional information. In this paper, the additional information considered is information that can be obtained from the collaborative filtering dataset itself. Future work will consider the incorporation of other information.

The collaborative filtering problem space is often viewed as a matrix consisting of the ratings given by each user for items in a collection. Using this matrix, the aim of collaborative filtering is to predict the ratings of a particular user,  $i$ , for

one or more items previously not rated by that user. The problem space can equivalently be viewed as a network where users and items are represented by nodes and the links between user nodes and item nodes represents a rating for the item by the user. Several researchers have adopted graph representations in order to develop recommendation algorithms to deal with particular aspects of the recommendation problem (e.g. sparsity [10]). A variety of graphs have been used and a number of graph algorithm approaches have been adopted [1], [10].

In this work, a graph model and spreading activation approach is proposed for the incorporation of additional sources of information. Users and items are represented as nodes in the graph and user nodes and item nodes are connected by weighted edges. A spreading activation approach is used to provide item recommendations. The additional information used is extracted from the analysis of certain features of the dataset. Based on these features, certain nodes (or users and items) of the graph are identified prior to retrieval. These nodes are used to constrain the spreading activation approach thus resulting in only some of the edges being considered and only portions of the graph being traversed. The idea is that this representation and approach can be easily extended to incorporate additional information via more constraints on nodes and via additional weighted edges between nodes.

The paper outline is as follows: Section 2 presents related work and the approach taken is outlined in Section 3. Section 4 details the experiments performed. Results are presented in Section 5 and Section 6 discusses conclusions and future work.

## 2 Related Work

Given a set of users, a set of items, and a set of ratings, collaborative filtering systems attempt to recommend items for some active user using the ratings of other users, where these users have similar preferences to the active user. The collaborative filtering system essentially automates the “word of mouth” process [17]. It differs from traditional retrieval and filtering systems which return items to some user based on a comparison between the content contained in items (documents) and the content of a user query (information need).

In various situations, collaborative filtering, on its own, is not adequate, e.g. when no ratings exist for an item; when new users join a system; or when a user is not similar to any other users in the dataset. Many different models and approaches have been proposed to overcome these problems. One proposed solution involves utilising any available additional information or evidence (collaborative, content, link, etc.). Work on the combination of different types of information is not restricted to the collaborative filtering domain and within the information retrieval and filtering domains, many approaches have been developed for combining different types of information and combining results from different information retrieval systems [7]. Several authors suggest methods for combining content with collaborative information [2], [14]. There has also been some work on extracting additional information (or features) from the collaborative filtering dataset and using this information to aid in recommendation [15].

## 2.1 Graph-Based Approaches for Recommendation

Several researchers in Information Retrieval and Collaborative Filtering have adopted graph representations of the problem domain [5], [16], [13]. A variety of graphs have been used (e.g. directed and two-layer) and a number of graph algorithm approaches have been adopted.

In the collaborative filtering domain, Aggarwal et al. present *horting*, a graph-based technique where nodes represent users and directed edges between nodes correspond to the notion of predictability [1]. Predictions are produced by traversing the graph to nearby nodes and combining the ratings of the nearby users. Huang et al. present a number of graph-based representations of the collaborative filtering space using binary ratings. They test a number of associated algorithms, including spreading activation and link prediction approaches [10], [11]. In [10] the goal of the work was to deal with the sparsity problem in collaborative filtering by investigating transitive relationships using a graph representation and spreading activation approaches. Users and items are represented using a bipartite graph representation where the transactions of users and user feedback are modelled as binary weighted edges connecting the nodes between the two sets of nodes. The authors noticed the effect of over-activation but did not constrain the spreading activation approach.

## 2.2 Spreading Activation

The idea of spreading activation originated from the field of Psychology and was first used in Computer Science in the area of Artificial Intelligence to process semantic networks. Spreading activation approaches have been used in many Information Retrieval applications [5] and more recently in the domain of collaborative filtering [10]. Spreading activation approaches have also been used to integrate sources of evidence and information [18], [6].

The spreading activation model involves a number of iterations where each iteration consists of one or more “hops” or “pulses” where a hop involves the spreading of activation from one node to all other nodes connected to it. A hop may also include a pre-adjustment stage and a post-adjustment stage which usually involves a decay factor. The activation of a node is calculated by summing the output from each of the nodes connected to it by the weight on the link connecting the two nodes. The output or activation of a node is usually a function of the activation value where many different functions can be used (e.g., a threshold function, a sigmoid function, etc.). A number of problems with this basic spreading activation approach have been identified, one of which is that activation can spread to all nodes in just a few hops. This problem can be overcome by the use of constraints in the pre- and post- adjustment stages. These constraints will specify which portions of the network should be traversed for some hop.

## 3 Methodology

In this paper, a graph representation is used consisting of a set of user nodes and a set of item nodes. User and item nodes are connected via weighted edges where

the weights on the edges represent ratings given to items by users (See Fig. 1). Each user node and item node has an associated activity, output, constrained flag and threshold. The activity of a user or item node  $a$ , for  $N$  nodes connected to the node  $a$  with non-zero weight, is calculated by:

$$activity_a = \sum_{i=1}^N x_i w_i$$

where  $x_i$  is the output of the node  $i$  that is connected to node  $a$  and  $w_i$  is the weight on the edge connecting node  $i$  to node  $a$ . The output,  $output_a$ , of a user or item node is calculated using a threshold function:

$$output_a = \begin{cases} activity_a & \text{if } (constrained_a \neq 0) \text{ and } (activity_a > \tau) \\ 0 & \text{otherwise.} \end{cases}$$

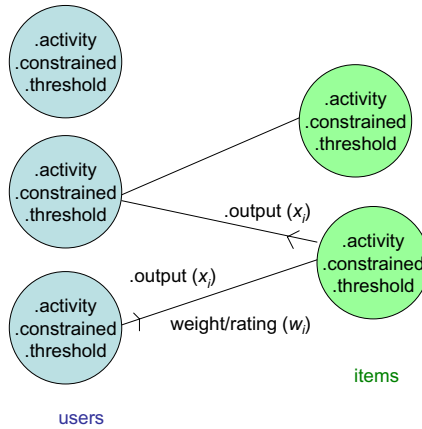
The threshold function uses the node’s activity, the constrained flag value (which can be on (1) or off (0)), and a threshold value,  $\tau$ , where each node may have its own threshold value depending on its importance.

Spreading activation involves moving activation from one set of nodes to a second set of nodes. The terminology of a hop is used in this paper to define the activation spreading from one set of nodes to a second set of nodes. A hop involves the calculation of all node outputs in either the user set or item set, updating the associated activities and outputs of the nodes. The steps involved are:

1. Hop 1: Calculate the activities of all item nodes connected, with non-zero weight, to the current active user node. For each activated item node, calculate the output of the node using the threshold function.
2. Hop 2: Calculate the activities of all user nodes connected, with non-zero weight, to item nodes where the item nodes have non-zero output. For each activated user node, calculate the output of the node using the threshold function.
3. Hop 3: Calculate the activities of all item nodes connected, with non-zero weight, to user nodes where the user nodes have non-zero output. For each activated item node, calculate the output of the node using the threshold function.
4. Following three hops, items with the top-20 highest positive activities are recommended to the active user.
5. Steps 2 and 3 can be repeated any number of times before recommendations are given (step 4).

Two hops result in activating a set of user nodes constituting a user neighbourhood of the original active user node. The third hop, from user nodes to item nodes, provides item recommendations for the active user.

Much implicit information about users and groups can be extracted from the collaborative filtering dataset and can be represented in this graph model via the threshold value, the constrained flag, and the values of the weighted edges. The initial experimental work in this paper only considers two of the simpler user and item features and incorporates information relating to these via the constrained flag. The user and item features considered are:



**Fig. 1.** Graph Representation of Users, Items and Ratings

- *rated*: the number of items rated by some user in comparison to the maximum and minimum number of items rated by users.
- *avg-item-popularity*: the number of ratings received by some item in comparison to the maximum and minimum ratings received by items.

Based on the values of these features certain user and item nodes are identified prior to filtering and the constrained flag value of these nodes is set to 1. This results in only some of the edges being considered and only portions of the graph being traversed.

## 4 Experiments

The experiments involve the comparison of two collaborative filtering approaches: a constrained spreading activation approach and a traditional memory-based collaborative filtering approach. The reason for choosing a traditional memory-based collaborative filtering approach is that it has been shown to perform well in comparison to many other collaborative filtering techniques [4], [8]. Also it is quite similar to the spreading activation approach outlined.

The aim is to show that the constrained spreading activation approach can perform as well, or better, than a traditional memory-based collaborative filtering approach. The important difference between the representations and approaches is the flexibility of the graph-based and spreading activation approach in allowing the incorporation of additional information. In this work some simple constraining of nodes is performed but future work will hope to show improved performance over a traditional memory-based collaborative filtering approach, and other collaborative filtering approaches, by the incorporation of additional information via further more complex constraints, varying values for  $\tau$  (the threshold), and further links between all sets of nodes.

A standard subset of the Movie Lens dataset is considered that contains 943 users and 1682 movies. Weights on the network edges indicate the strength of



like/dislike for an item where “dislike” can be viewed as an inhibitory or negative rating and “like” can be viewed as an excitatory or positive rating. Given that the original rating values in the Movie Lens dataset are all positive numbers, the approach adopted maps the ratings to positive and negative values to indicate positive and negative influences. The mapping chosen is to subtract 2.5 from all non-zero values which will give ratings around 0, giving:

$$\{0, 1, 2, 3, 4, 5\} \rightarrow \{0, -1.5, -0.5, 0.5, 1.5, 2.5\}$$

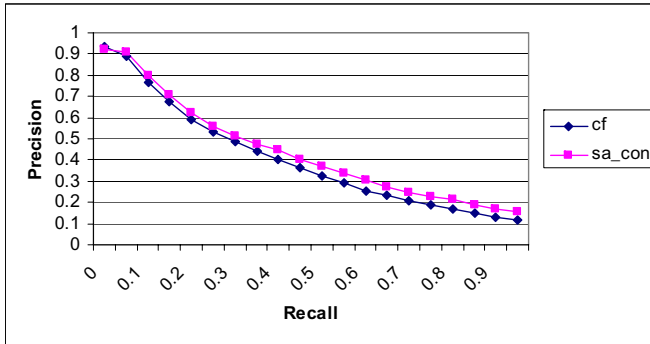
A proportion of the dataset is removed for testing and the metric of precision is used to compare the performance of the two approaches at different recall points. Precision is used because the spreading activation approach returns a ranking of recommended items, not prediction values that can be compared with actual values. In addition, work by Herlocker et al. [9] suggests that distinguishing between exact actual and predicted rating scores is not as important as measuring if the system correctly or incorrectly identifies “good” or “liked” items.

The traditional collaborative filtering approach uses Pearson Correlation to find correlated (similar) users. An adjustment is used in the Pearson Correlation calculation based on the number of items that users have rated in common (co-rated items). The adjustment ensures that users must have similar preferences over more than a few items to be considered highly correlated [12]. The “nearest” neighbours of a user are selected using a low neighbour selection threshold, with any correlation value greater than 0.01 being considered. Work by Herlocker [8] using the Movie Lens dataset found that experiments with no threshold (using all correlation values  $> 0$ ) always outperformed experiments with higher thresholds.

In the spreading activation approach to collaborative filtering, three stages corresponding to the three stages in the traditional memory-based collaborative filtering approach are used. Neighbours of some active user are found after two hops of the approach, at which stage user nodes that have non-zero activity are the neighbours of the active user. When activation is spread again, from user nodes to item nodes, items not rated by the active user will be highlighted. These items are recommended to the user if the activity is sufficiently high. At any stage, a node may be constrained in spreading its activation if it has been previously identified based on the analysis of user and item features. The threshold value used in these experiments is 0 for all nodes, i.e. all positive activities will result in a node outputting a value.

## 5 Results

Fig. 2 illustrates the precision recall graph for the two approaches. Results were averaged over 100 runs. It can be seen that even with the limited sources of additional information included in this representation, the constrained spreading activation approach outperforms the traditional memory-based approach at all recall points other than at the first returned recommendation. These results were shown to be statistically significant using a 2-tailed paired t-test at p-value  $< 0.05$ .



**Fig. 2.** Comparing Constrained Spreading Activation and Traditional Memory-based Approaches to Collaborative Filtering

## 6 Conclusions

We have presented a graph representation of the collaborative filtering space and a spreading activation approach to collaborative filtering using features of the dataset to constrain the activation approach. Such a flexible representation and approach allow for the incorporation of additional information. Results show that the performance of the approach is better than a traditional memory based approach. Future work will consider more complex user features, as well as group features, and will show that the graph representation and algorithm outlined and demonstrated in this paper can easily be extended to incorporate additional sources of evidence.

## References

1. C.C. Aggarwal, J.L. Wolf, K.-L. Wu, and P.S. Yu. Horting hatches an egg: A new graph-theoretic approach to collaborative filtering. In *Proceedings of the Fifth ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'99)*, San Diego, CA, pages 201–212, 1999.
2. M. Balabanovic and Y. Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.
3. C. Basu, H. Hirsh, and W. Cohen. Recommendation as classification: Using social and content-based information in recommendation. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, pages 714–721, 1998.
4. M.L. Calderon-Benavides, C.N. Gonzalez-Caro, J. Perez-Alcazar, J.C. Garcia-Diaz, and J. Delgado. A comparison of several predictive algorithms for collaborative filtering on multi-valued ratings. In *Proceedings of the 2004 ACM Symposium on Applied Computing*, pages 1033–1039, 2004.
5. P. Cohen and R. Kjeldsen. Information retrieval by constrained spreading activation on semantic networks. *Information Processing and Management*, 23(4):255–268, 1987.
6. F. Crestani and P.L. Lee. Searching the web by constrained spreading activation. *Information Processing and Management*, 36:585–605, 2000.

7. W.B. Croft. Combining approaches to information retrieval. In *Advances in Information Retrieval*, pages 1–36. Kluwer Academic Publishers, 2000.
8. J.L. Herlocker. *Understanding and Improving Automated Collaborative Filtering Systems*. Phd thesis, University of Minnesota, 2000.
9. J.L. Herlocker, J.A. Konstan, L.G. Terveen, and J.T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22:5–53, 2004.
10. Z. Huang, H. Chen, and D. Zeng. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems*, 22(1):116–142, 2004.
11. Z. Huang, W. Chung, and H. Chen. A graph model for e-commerce recommender systems. *Journal of the American Society for Information Science and Technology*, 55(3):259–274, 2004.
12. M.R. McLaughlin and J.L. Herlocker. A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In *Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 329–336, 2004.
13. B. Mirza, B. Keller, and N. Ramakrishnan. Studying recommendation algorithms by graph analysis. *Journal of Intelligent Information Systems*, 20:131–160, 2003.
14. C. O’Riordan and H. Sorensen. Multi-agent based collaborative filtering. In M. Klusch et al., editor, *Cooperative Information Agents 99, Lecture Notes in Artificial Intelligence*, 1999.
15. J. Palau, M. Montaner, and B. Lopez. Collaboration analysis in recommender systems using social networks. In *Cooperative Information Agents VIII: 8th International Workshop, CIA 2004*, pages 137–151, 2004.
16. M.F. Schwartz and C.M. Wood. Discovering shared interests using graph analysis. *Communications of the ACM*, 36:78–89, 1993.
17. U. Shardanand and P. Maes. Social information filtering: Algorithms for automating word of mouth. In *Proceedings of the Annual ACM SIGCHI on Human Factors in Computing Systems (CHI ’95)*, pages 210–217, 1995.
18. G.-R. Xue, S. Huang, Y. Y., H.-J. Zeng, Z. Chen, and W.-Y. Ma. Optimizing web search using spreading activation on the clickthrough data. In *Proceedings of the 5th International Conference on Web Information Systems*, 2004.

# Fuzzy Model for the Assessment of Operators' Work in a Cadastre Information System

Dariusz Król<sup>1</sup>, Grzegorz Stanisław Kukła<sup>1</sup>,  
Tadeusz Lasota<sup>2</sup>, and Bogdan Trawiński<sup>1</sup>

<sup>1</sup> Institute of Applied Informatics, Wrocław University of Technology,  
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

dariusz.krol@pwr.wroc.pl, grzegorz\_kukla@o2.pl, trawinski@pwr.wroc.pl

<sup>2</sup> Faculty of Environmental Engineering and Geodesy, Agricultural University of  
Wrocław, Norwida 25/27, 50-375 Wrocław, Poland

tadeusz.lasota@wp.pl

**Abstract.** One of critical tasks of cadastre system maintaining is the input of changes into its database. Managers of information centres often complain they have no adequate tools for the assessment of work of cadastre system operators. In the paper a fuzzy model is proposed which goal is to provide a useful tool for management of an information center. The architecture of the fuzzy system comprises five main modules of operators' work statistics, fuzzification, inference, defuzzification and visualization. For each input criterion i.e. productivity (P), complexity (C), time (T) and quality (Q) as well as for output assessment triangle and trapezoid membership functions have been defined. The statistics module provides initial parameters of the model and values of input criteria. The model based on change records saved in cadastre database produces the assessments of operators' work for defined periods of time automatically.

## 1 Introduction

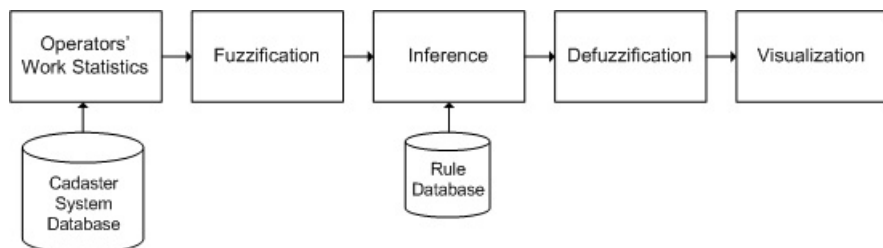
Cadastre systems are designed for the registration of parcels, buildings and apartments as well as their owners and users. Those systems have complex data structures and sophisticated procedures of data processing. The maintenance of real estate cadastre registers is dispersed in Poland. There are above 400 information centres located by district local self-governments as well as by the municipalities of bigger towns which exploit different cadastre systems. One of critical tasks of cadastre system maintaining is the input of changes into its database. The changes are allowed only on the basis of legal documents such as sale purchase agreements, extracts from perpetual books, results of surveyors' works and others. Managers of information centres often complain they have no adequate tools for the assessment of work of cadastre system operators. Reviews with the managers revealed that using only productivity expressed by the number of changes input within a given period is highly insufficient. Criteria that they mentioned first of all were complexity of changes calculated as a mean number of objects

which were changed in database falling per one change, average time of inputting one change and quality of work determined by the percentage of changes without any corrections. The assessment of operators' work is subjective and imprecise to some extent. If, for example, during one month an operator input 87 changes of mean complexity of 9.3 objects per change, with mean time of 13.4 minutes and without any corrections then what was his work: good, average or bad? Such evaluation has fuzzy nature. It is much easier for a manager to estimate each criterion separately using such linguistic values as high, medium and low. So it was the main reason for developing our fuzzy model of operators' work assessment.

Many books have been written on fuzzy logic and control [2], [8], numerous articles deal with fuzzy expert systems [1], [3], [6]. The fuzzy model presented in the paper is composed of typical elements. However having the statistics of cadastre system operators' work it is possible to apply them as initial parameters of fuzzification process. Based on managers' suggestions four following input variables have been designed: productivity, complexity, time and quality. Moreover using the statistics as the input of the model allows obtaining final assessments automatically. It has been assumed that experts will tune the initial parameters and the centre managers, who are the primary group of users of the system, will be able to modify such parameters as the average values of input criteria, which play the role of points of reference, or weights of rules.

## 2 The Model of the Assessment of Operators' Work

The fuzzy model proposed in the paper will constitute the basis of the fuzzy system which is intended to rationalize the management of information centres, to improve the organization of work and to determine wages of outsource workers. The architecture of the fuzzy system is shown in Fig. 1. It comprises five main modules of operators' work statistics, fuzzification, inference, defuzzification and visualization.



**Fig. 1.** Architecture of the fuzzy system for the assessment of operators' work

### 2.1 Input and Output of the Model

For each input criterion i.e. productivity (P), complexity (C), time (T) and quality (Q) as well as for output assessment triangle and trapezoid membership

functions have been defined (see Fig. 2). The statistics module provides initial parameters of the model and values of input criteria. The idea of obtaining the final assessment consists in calculating the average value of P, C and T criteria taking into account the change records saved in cadastre database for all operators and for long period of time, e.g. a year or a half of year. These average values are used as the reference values of 100% for calculating what percentage of corresponding average value a given operator achieved within the assessment period. Values of time variable are reversed in the scale form 0% to 200% in order to achieve lower input values for longer times of introducing changes into the system. Data for quality variable are applied directly, because this criterion is expressed in percents. Standard deviations, calculated separately for each criterion for a long period of time, determine the width of the basement of triangle and trapezoid fuzzy sets by adding or subtracting them from 100%. The domain for input P, C and T are 0-200 percentage values, with 0 being the lowest and 200 the highest mark. Data for quality variable are applied directly, because this criterion is expressed in percents. The domain for output, represented by five fuzzy sets, is an arbitrary assessment scale from 1 to 200, with 0 being the lowest and 200 the highest mark.

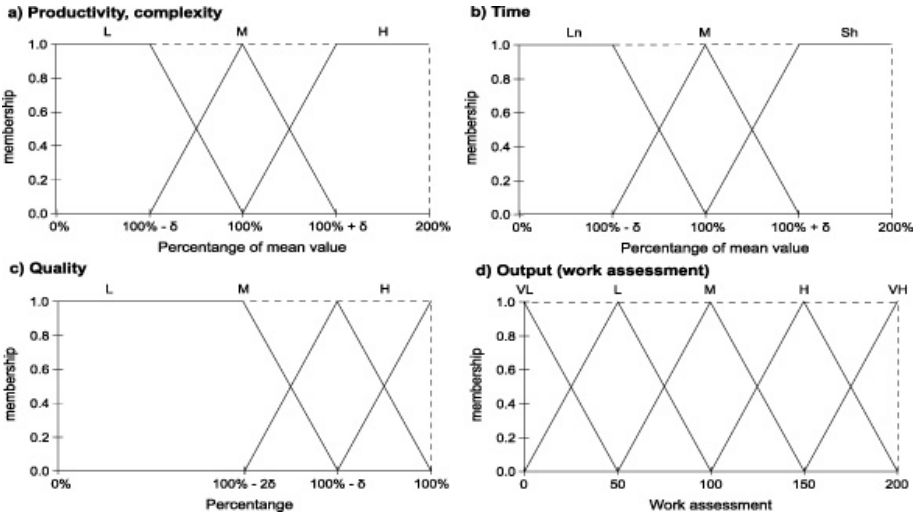


Fig. 2. Membership functions of input and output variables, where VL, L, M, H, VH denote very low, low, medium, high, very high respectively, Ln, Sh mean long and short

### 2.2 Formal Description of Linguistic Variables

Let  $x_P$  be a linguistic variable of productivity, and its term set be  $T_P = \{\text{low, medium, high}\}$ , where each term in  $T_P$  is characterized by a fuzzy set in a universe of discourse  $U_P = [0, 200]$ , then the membership functions of these fuzzy sets are as follows:

$$\begin{aligned}
 low(x_P) &= \begin{cases} 1 & \text{if } x_P \leq 100 - \delta_P \\ 1 - (x_P + \delta_P - 100)/\delta_P & \text{if } 100 - \delta_P < x_P < 100 \\ 0 & \text{if } x_P \geq 100 \end{cases} \\
 medium(x_P) &= \begin{cases} 0 & \text{if } x_P \leq 100 - \delta_P \\ 1 - |(x_P - 100)|/\delta_P & \text{if } 100 - \delta_P < x_P < 100 + \delta_P \\ 0 & \text{if } x_P \geq 100 + \delta_P \end{cases} \\
 high(x_P) &= \begin{cases} 0 & \text{if } x_P \leq 100 - \delta_P \\ 1 - (100 + \delta_P - x_P)/\delta_P & \text{if } 100 - \delta_P < x_P < 100 + \delta_P \\ 1 & \text{if } x_P \geq 100 + \delta_P \end{cases}
 \end{aligned}$$

where  $\delta_P$  is the standard deviation of productivity calculated for a long period of time.

Let  $x_C$  be a linguistic variable of complexity, and its term set be  $T_C = \{\text{low, medium, high}\}$ , where each term in  $T_C$  is characterized by a fuzzy set in a universe of discourse  $U_C = [0, 200]$ . Then the membership functions of these fuzzy sets are as follows:

$$\begin{aligned}
 low(x_C) &= \begin{cases} 1 & \text{if } x_C \leq 100 - \delta_C \\ 1 - (x_C + \delta_C - 100)/\delta_C & \text{if } 100 - \delta_C < x_C < 100 \\ 0 & \text{if } x_C \geq 100 \end{cases} \\
 medium(x_C) &= \begin{cases} 0 & \text{if } x_C \leq 100 - \delta_C \\ 1 - |(x_C - 100)|/\delta_C & \text{if } 100 - \delta_C < x_C < 100 + \delta_C \\ 0 & \text{if } x_C \geq 100 + \delta_C \end{cases} \\
 high(x_C) &= \begin{cases} 0 & \text{if } x_C \leq 100 - \delta_C \\ 1 - (100 + \delta_C - x_C)/\delta_C & \text{if } 100 - \delta_C < x_C < 100 + \delta_C \\ 1 & \text{if } x_C \geq 100 + \delta_C \end{cases}
 \end{aligned}$$

where  $\delta_C$  is the standard deviation of complexity calculated for a long period of time.

Let  $x_T$  be a linguistic variable of time, and its term set be  $T_T = \{\text{low, medium, high}\}$ , where each term in  $T_T$  is characterized by a fuzzy set in a universe of discourse  $U_T = [0, 200]$ . Then the membership functions of these fuzzy sets are as follows:

$$\begin{aligned}
 low(x_T) &= \begin{cases} 1 & \text{if } x_T \leq 100 - \delta_T \\ 1 - (x_T + \delta_T - 100)/\delta_T & \text{if } 100 - \delta_T < x_T < 100 \\ 0 & \text{if } x_T \geq 100 \end{cases} \\
 medium(x_T) &= \begin{cases} 0 & \text{if } x_T \leq 100 - \delta_T \\ 1 - |(x_T - 100)|/\delta_T & \text{if } 100 - \delta_T < x_T < 100 + \delta_T \\ 0 & \text{if } x_T \geq 100 + \delta_T \end{cases} \\
 high(x_T) &= \begin{cases} 0 & \text{if } x_T \leq 100 - \delta_T \\ 1 - (100 + \delta_T - x_T)/\delta_T & \text{if } 100 - \delta_T < x_T < 100 + \delta_T \\ 1 & \text{if } x_T \geq 100 + \delta_T \end{cases}
 \end{aligned}$$

where  $\delta_T$  is the standard deviation of time calculated for a long period of time.

Let  $x_Q$  be a linguistic variable of quality, and its term set be  $T_Q = \{\text{low, medium, high}\}$ , where each term in  $T_Q$  is characterized by a fuzzy set in a universe of discourse  $U_Q = [0, 100]$ . Then the membership functions of these fuzzy sets are as follows:

$$\text{low}(x_Q) = \begin{cases} 1 & \text{if } x_Q \leq 100 - 2\delta_Q \\ 1 - (x_Q + \delta_Q - 100)/\delta_Q & \text{if } 100 - 2\delta_Q < x_Q < 100 - \delta_Q \\ 0 & \text{if } x_Q \geq 100 - \delta_Q \end{cases}$$

$$\text{medium}(x_Q) = \begin{cases} 0 & \text{if } x_Q \leq 100 - 2\delta_Q \\ 1 - |(x_Q - 100)|/\delta_Q & \text{if } 100 - 2\delta_Q < x_Q \leq 100 \end{cases}$$

$$\text{high}(x_Q) = \begin{cases} 0 & \text{if } x_Q \leq 100 - \delta_Q \\ 1 - (100 + \delta_Q - x_Q)/\delta_Q & \text{if } 100 - \delta_Q < x_Q \leq 100 \end{cases}$$

where  $\delta_Q$  is the standard deviation of quality calculated for a long period of time.

Let  $y$  be a linguistic variable of assessment, and its term set be  $T_A = \{\text{very low, low, medium, high, very high}\}$ , where each term in  $T_A$  is characterized by a fuzzy set in a universe of discourse  $U_A = [0, 200]$ . Then the membership functions of these fuzzy sets are as follows:

$$\text{very low}(y) = \begin{cases} 1 - y/50 & \text{if } 0 \leq y < 50 \\ 0 & \text{if } y \geq 50 \end{cases}$$

$$\text{low}(y) = \begin{cases} 1 - |(y - 50)|/50 & \text{if } 0 \leq y < 100 \\ 0 & \text{if } y \geq 100 \end{cases}$$

$$\text{medium}(y) = \begin{cases} 0 & \text{if } y \leq 50 \\ 1 - |(y - 100)|/50 & \text{if } 50 < y < 150 \\ 0 & \text{if } y \geq 150 \end{cases}$$

$$\text{high}(y) = \begin{cases} 0 & \text{if } y \leq 100 \\ 1 - |(y - 150)|/50 & \text{if } 100 < y \leq 200 \end{cases}$$

$$\text{very high}(y) = \begin{cases} 0 & \text{if } y \leq 150 \\ 1 - |(y - 200)|/50 & \text{if } 150 < y \leq 200 \end{cases}$$

### 2.3 Rule Database and Inference Process

It has been assumed that the rule database should contain simple IF-THEN rules where the condition consists of only two input variables combined by AND operator and the conclusion is built by one variable. An example of a rule is as follows: **IF Productivity is medium AND Complexity is low THEN Assessment is low**. Thus the rules for one pair of input criteria can be given in the form of a matrix shown in Fig. 3. So 6 matrices for 9 rules has been designed for C-P, C-T, C-Q, P-T, P-Q, T-Q combinations. Having such form of rule database, the experts will not have difficulties to determine the output values for each pair



		Input variable C		
Input variable P		L	M	H
	L	VL	L	M
	M	L	M	H
	H	M	H	VH

– input values  
 – output values

VL – very low  
 L – low  
 M – medium  
 H – high  
 VH – very high

Fig. 3. Representation of rule database in matrix form for C and P input

of input values. In order to express the strength of rules belonging to particular combination, rule weights were designed with initial values of  $w_{C-P}=0.30$ ,  $w_{C-T}=0.25$ ,  $w_{C-Q}=0.20$ ,  $w_{P-T}=0.10$ ,  $w_{P-Q}=0.10$  and  $w_{T-Q}=0.05$  as the multipliers of rule conditions in aggregation step.

In order to assure that each rule will have influence on the final assessment following operators has been used: PROD for aggregation of rule conditions, PROD for activation of rule conclusions and ASUM for accumulation of output membership functions, where PROD means algebraic product and ASUM denotes algebraic sum [5]. In defuzzification step the center of gravity method is used.

### 3 Evaluation of the Fuzzy Model

Statistics module and fuzzy model have been programmed using C# and Java languages. Work statistics for the period of 6 months from January to June 2005 have been calculated for 13 operators in one of information centre (see Fig. 4).

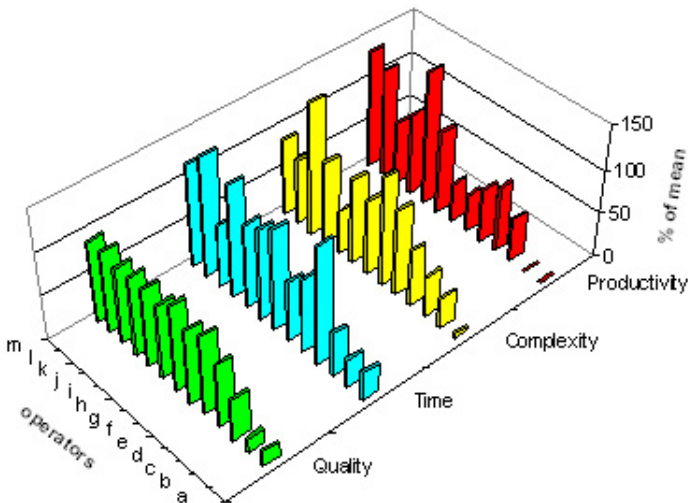


Fig. 4. Values of input criteria for 13 operators for the period of six months

The results of statistics module have been applied as model parameters and as input values. The data have been processed by fuzzy model using rule weights and final assessments for each operator and for each month have been obtained as the output.

The values of input criteria and the output produced by the model for one operator are presented in Fig. 5. The weights line denotes the result obtained using rule weights and the second one indicates assessments calculated without rule weights.

Data given in Fig. 5 provide the basis for a centre manager to take decisions.

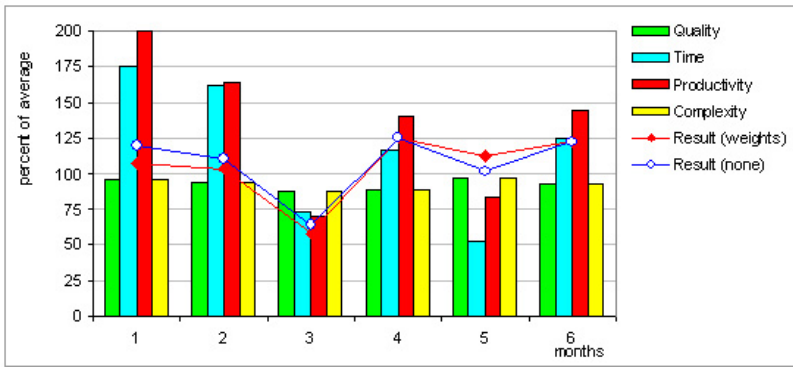


Fig. 5. Input criteria and the final assessment for one selected operator

## 4 Conclusions

The model for the multicriterial assessment of cadastre system has been designed and programmed with the aid of an expert and information centre managers. The evaluation of the model has proved that it is reasonable to use operators' work statistics as the initial parameters and the input values for the model. The model based on change records saved in cadastre database produces the assessments for defined periods of time automatically. The architecture of proposed model enables the centre managers to modify system parameters to customize them to local circumstances.

It is planned to carry out an evaluation experiment with the participation of the centre managers. The number of input criteria will be increased by incorporating additionally the statistics of system reports usage and the subjective managers' assessment of the usefulness of a given operator during selected period of time. Tuning tests will also be performed in order to determine how the output depends on the number of input linguistic values and rule weights and other model parameters. Finally, this idea will be used in the Web cadastre system.

## References

1. Bae S.M., Park S.C., Ha S.H.: Fuzzy Web Ad Selector Based on Web Usage Mining. IEEE Intelligent Systems November/December 2003
2. Driankov D, Hellendoorn H., Reinfrank M.: An Introduction to Fuzzy Control Springer-Verlag Berlin 1993
3. Drigas A., Kouremenos S., Vrettos S., Vrettaros J., Kouremenos D.: An expert system for job matching of the unemployed. Expert Systems with Applications **26** (2004) 217–224
4. Eirinaki M., Vazirgiannis M.: Web Mining for Web Personalization. ACM Transactions on Internet Technology, **3** (2003) 1-27
5. IEC 1131 - Programmable Controllers. Part 7 - Fuzzy Control Programming. Committee Draft CD 1.0 (Rel. 19 Jan 97)
6. Levary R.R., Lin C.Y.: Modelling the Software Development Process Using an Expert Simulation System Having Fuzzy Logic. Software-Practice And Experience **21** (1991)133–148
7. Saini H.S., Kamal R., Sharma A.N.: Web Based Fuzzy Expert System for Integrated Pest Management in Soybean. International Journal for Information Technology **8** (2002)
8. Yager R.R., Filev D.: Essentials of fuzzy modeling and control Wiley New York 1994

# Local Buffer as Source of Web Mining Data

Andrzej Siemiński

Institute for Applied Informatics,  
Technical University of Wrocław  
Wyb. Wyspińskiego 27  
50-370 Wrocław Poland  
andrzej.sieminski@pwr.wroc.pl

**Abstract.** The data for Web mining is usually extracted from the WWW server or proxy server log files. The paper examines the advantages and disadvantages of exploiting another source of input data – the browser buffer. The properties of data extracted from different types of sources are compared. The browser buffer contains data about user navigational habits as well as the formal properties and the content of all recently accessed WWW objects. The paper uses the data obtained from this source to examine the statistical properties of different types of texts extracted from HTML pages.

## 1 Introduction

Web data mining is a now the focus of study for an increasing number of researchers. It has been defined as the implementation of data mining techniques to automatically discover and extract information from the Web. The scope of applications of the research in this area ranges from improving the quality of service by reducing the user perceived latency to personalizing the Web content by generating user specific pages or supplying him/her potentially useful links. [9].

One of the most important issues in data mining is starting with the right data. The data is usually extracted from the web servers or proxy servers logs. This source of data is however limited in:

- availability: the logs are regarded to contain sensitive commercial data and are usually not available for research;
- scope of contained data: to conserve space only the basic data is kept in the logs, the potentially valuable content of WWW objects is not stored.

The paper proposes an alternative source of data – the local system cache for Internet objects. The cache contains all recently requested WWW objects along with the header control data supplied by the WWW server. Web data mining is subdivided into three sub areas: web formal features, web usage patterns and web content analysis [9]. Previous attempts to utilize local buffer data had concentrated upon the detection of web usage patterns [11] or Web objects formal features [8]. This paper analyses the statistical properties of different types of texts that are extracted from the buffer – it is located in the Web Content area. The aim of the analysis is to discover properties that could be useful for the design of prefetching algorithms.

The rest of the paper is organized as follows:

Section 2 compares the data that can be extracted from the different types of logs with the data obtained from the local cache. Section 3 describes the content and the properties of a local buffer. Section 4 contains the analysis of link text similarity. Section 5 concludes and outlines future research areas.

## 2 Sources of Data for Web Mining

The sheer size of Internet makes it impossible to analyze all of its objects. Therefore it is so important to start with the right data. The ideal set of data should be easy available, preferably user specific, complete and span over a reasonable long period of time. The usual source of data for web mining are logs collected at different levels of Internet infrastructure: at web servers, proxy caches or in local networks.

The Web server logs are collected at each Web server. They are widely considered to contain valuable commercial data and are usually not available for research purposes. Popular logs such as Music Machinelog [19] or World Cup log [22] are several years old. The log does not contain the body of the requested data, hence web content mining is not possible. The data is well suited for the analysis of one web site users preferences. The collection of reliable data on a single user preferences that span over several sessions is much difficult. The user has either to log in to a server or the cookies have to be used. This is not acceptable to many users as being bothersome or because it could infringe their privacy.

The data is limited to a single server so the identifying of general (encompassing many servers) user preferences is impossible. This makes the behavioral targeting not possible at all.

The availability of proxy logs is far greater than the web servers logs. Large proxy log files are free for research purposes, e.g. the popular logs from the IRCache project [17]. The log formats are normalized and even programs for the statistical analysis of popular log formats are available [14], [15]. That type of data has however serious disadvantages. The logs of proxy servers are not available for every region and they contain only basic header data, in some cases even the download time, so crucial in the caching/prefetching area, is not available. The requested objects are also not available so the direct study of web content is not possible. The data is anonymized to protect user privacy. As a result, the collecting of multi session single user data is impossible. The log data removes only one limitation of the Web Server logs: they cover a wide spectrum of servers.

The local network logs are even more scarce than the Web server logs. There are only a few publicly available log data of that type but they are generally outdated as e.g. the Boston University Log [2]. One session user identification does not pose any problem but due to the anonymization process the multi session information is still not available. The scope of the visited Web sites is complete but the data is not detailed enough.

The traditional sources are not capable of providing a complete picture of user browsing habits that span over a several days period. Precisely that type data is necessary for the behavior targeting. The behavior targeting is a relatively new and promising research topic in the area of Internet recommendation. The targeting requires data that is: specific for a single user, covers all visited Web servers and spans over a

reasonably long period of time. The usefulness of such data is appreciated not only in research but also in the industry. Software companies such as GAIN [16] or Google [18] can offer incentives to Web users prompting them to make reveal the way interact with the Web. The legislative and regulatory framework that governs behavioral marketing is described in [20].

### 3 Content and Properties of Browser Buffer Data

Some of the mentioned above problems could be solved by collecting data about user interaction with the Web directly at their source of origin – at the browser cache. The buffer has four advantages over the traditional data sources:

- completeness - it contains all object requested by a user;
- is user-centric – the data is not mixed with data of other users of the system;
- scope - the buffer includes not only description of an WWW object but also its body.
- privacy- the buffer is stored at a local system so as long as the data is processed locally there are no privacy concerns.

It should be stressed that the buffer contains a set of requested pages but not the sequence in which there were visited. Web mining procedures require intensive processing. Analyzing data locally disperses the processing. The algorithms applied locally could require intensive computations and thus are capable of data analysis that is deeper then in server based solutions.

In the experiment the most popular Windows operating system was used. For each user the system maintains a local buffer - the TIF (Temporary Internet Files) which is a simple data base. The database contains a table with all recently requested files and their attributes. A detailed analysis of the way the TIF operates performed for the study revealed that a buffer having a moderate size of 100MB contains virtually all objects requested during last several days.

During the experiment cache data were extracted from 29 computers in a student laboratory. The laboratory was not used for a regular classes. The students of the University worked there to prepare for their classes or used the computers to surf on Internet. In the preliminary stage of experiment the data was extracted from local buffers. In the process all of the headers send by WWW servers were copied out. During the next stage the text and links from HTML objects were extracted. It turned out, that the specification of the HTML format is not closely followed by many webmasters. As a result, the official SGML parser of the W3C consortium available on [21] could produce as many as 300 errors or warnings while analyzing a popular web page. The Lynx text browser was used to cope with the problem and parse the HTML files. In all, the test data contained some 50 000 web pages with over 950 000 links. The link texts consisted of just under 3 000 000 words.

### 4 Buffer Data Analysis

The diversity of words in natural language texts is described by the well known Zip law [3], [13]. The analysis presented in the paper is focused on the diversity of link

texts. It is generally believed that the link texts are very important for the Internet data processing. The texts are characterized by three features:

- They are regarded as a reliable summary of the text of a page they point to.
- They are much shorter than the page text itself.
- They enable to infer about the content a goal page even when the page is not loaded or even not available at all e.g. due to the failure of a WWW server.

### Link texts diversity

Two types of link texts were considered:

- internal: the texts of all links extracted from a given source page and
- external: texts that describe links leading to a given goal page from any page in the buffer.

As stated above link text are a good description of a page the link points to. The diversity of internal links is useful as it differentiates the individual links and it limits e.g. the number of pages accessed by prefetching algorithms. On the other hand, the diversity of external links texts hampers prefetching. Different texts appearing on links leading to the same goal page provide alternative descriptions of a page and thus make the selection of a link leading to a page more difficult.

### Text diversity coefficient

The link texts often have the same meaning, they look similar but they are not identical. The phenomenon occurs due to e.g. different lexical word forms used or even due to spelling errors. The usual way of processing such data involves the elimination of common words from the so called stop list and then reducing the remaining words to their stems. There are many freely available stop lists and stemmers for e.g. English and Polish [4], [5], [10], [13]. Such an approach was considered but finally rejected. It turned out that the link text were written in many natural languages, the texts were rather short what made the language identification prone to error and such an approach does not accommodate for spelling errors. Therefore a the text similarity measure that is based on the 3-grams was used.

Let  $n(s)$  be a set of  $n$  3-grams forming the string  $s$ , e.g.  $n(\text{test\_string}) = \{\text{tes, est, st\_}, \text{t\_s, \_st, str, tri, rig}\}$ . The similarity function  $\text{Sim}(s, t)$  of the two strings  $s$  and  $t$  is defined as follows:

$$\text{Sim}(s, t) = \frac{\overline{n(s) \cap n(t)}}{\overline{n(s) \cup n(t)}} \quad (1)$$

where  $\overline{n}$  is the number of elements in the set  $n$ .

Let  $S$  be a set of  $n$  text strings,  $S = \{s_1, \dots, s_m\}$ . The individual strings are grouped into packages. Each package contains only similar strings. Let  $P(S, x)$  denote a set of all packages that were built using the value  $x$  of the similarity function that is for all pairs  $r$  and  $t$  contained in each package the inequity holds:

$$\text{Sim}(r, t) \geq x \quad (2)$$

In an extreme case all strings in a sequence are treated as identical and they form a single package. In this case the diversity should be equal to 0. In the opposite case all of strings of a sequence are different. Each of text strings forms a separate package. The diversity should have maximal value of 1.

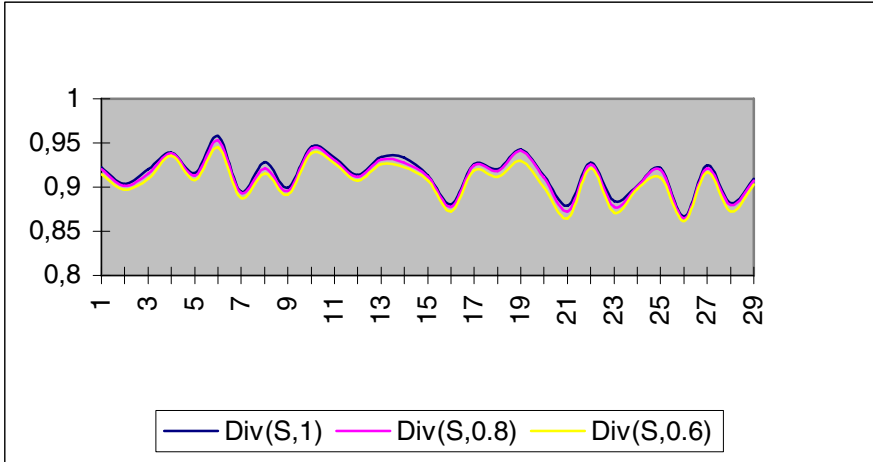


Fig. 1. Internal link text diversity for three selected similarity levels

The  $Div(S,x)$  defines the diversity coefficient of packages formed from the set  $S$  using the value  $x$  of the similarity function.

$$Div(S,x) = \sum_{i=1}^n \frac{\overline{pi}}{n} \ln\left(\frac{n}{\overline{pi}}\right) / \ln(n) \tag{3}$$

where:

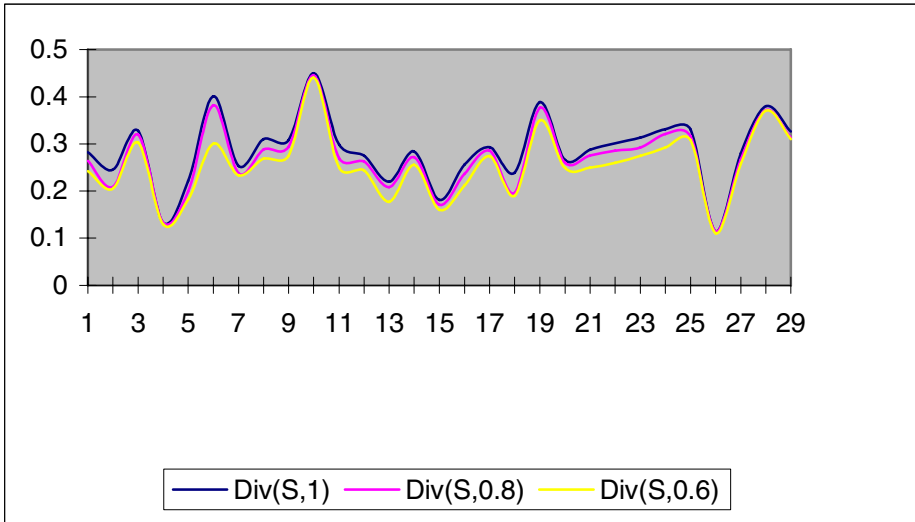
$n$  - the number of packages, it is equal to  $\overline{P(S,x)}$  ;

$\overline{pi}$  - the number of elements in the package  $pi$ .

It could be easily seen that  $Div(S,x)$  satisfies fulfills both stated above requirements. The average internal link text diversity for all individual buffers is shown on the fig. 2. Three diversity coefficients were used:  $Div(S,1)$ ,  $Div(S,0.8)$  and  $Div(S,0.6)$ . The relaxing of the criteria for package formation did not change the value of  $Div$  in any substantial manner. The average values for all buffers were:  $Div(S,1)=0.914$ ,  $Div(S,0.8)=0.912$  and  $Div(S,0.6)=0.906$ .

The average values of external text diversity for the same as above text similarity levels are shown on the Fig. 3. The values of the diversity coefficient are significantly lower then in the case of internal links. The average values for all buffers are:





**Fig. 2.** External link text diversity for three selected similarity levels

$Div(S,1)=0,28$ ;  $Div(S,0.8)=0.27$  and  $Div(S,0.6)= 0,25$ . The relaxing of the way in which packages are formed has more influence on the values of the coefficients.

The results indicate that link texts could be successfully used for prefetching of web pages. They internal link texts are highly diverse while internal link texts are exhibit a fair amount of similarity.

## 5 Conclusions and Future Work

The browser buffer is a useful source of web mining data. The data it contains is:

- complete - covers all Web servers;
- easily available – the buffer is used by most computers;
- detailed – contains both formal object features as well as the body of the object and
- user specific – system maintains a separate buffer foe each user.

The co-occurrence of such features is not to be found among other sources of data.

The first results clearly indicate that there are substantial differences in the properties of link texts that accompany visited and not visited links. A further research on that area, should result in the development of new prefetching algorithms. The main aim of the prefetching is to reduce the user perceived browser latency. The reduction is accomplished due to the downloading in advance of the web objects that a user might need in future.

## References

1. Ajiferuke I., Wolfram D.: Analysis of Web Page Image Tag Distribution, *Information Processing and Management*, 41 (2005), p 987-1002
2. Carlos A. Cunha, Azer Bestavros and Mark E. Crovella: "Characteristics of WWW Client Traces", Boston University Department of Computer Science, Technical Report TR-95-010, April 1995
3. Gelbukh A., Grigori Sidorov G: "Zipf and Heaps Laws' Coefficients Depend on Language" Proc. CILing-2001, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, p 332–335.
4. Lovins, J.B.: "Development of a Stemming Algorithm". *Mechanical Translation and computation Linguistics*. 11 (1) March 1968 p 23-31.
5. Porter M.F.: " An algorithm for suffix stripping", *Program*, 14 no. 3, pp 130-137, July 1980.
6. Rabinowich M., Spatschech O.: "Web Caching and Replication", Addison Wesley, USA, 2002
7. Siemiński A., *The Cacheability of WWW Pages*, in *Multimedia and Network Information Systems 2004*, Technical University of Wrocław, Poland, 2004
8. Sieminski A.: "Changeability of Web Objects", ISDA'05 -5th International Conference on Intelligent Systems Design and Implementation, Wrocław, 2005
9. Srivastava J., Desikan P., Kumar V.: "Web Mining: Accomplishments & Future Directions", National Science Foundation Workshop on Next Generation Data Mining (NGDM'02) , 2002
10. Szafran K.: .SAM 95 - Morphological Analyzer., TR 96-05 (226), Instytut Informatyki Uniwersytetu Warszawskiego, 1996.
11. Tran L, C. Moon, D. Le, Thoma G.: "Web Page Downloading and Classification", The Fourteenth IEEE Symposium on Computer-Based Medical Systems, July 2001.
12. Weiss D.: "A Survey of Freely Available Polish Stemmers and Evaluation of Their Applicability in Information Retrieval", 2nd Language and Technology Conference, Poznań, Poland, 2005, p 216-221
13. Zipf, G. K. *Human behavior and the principle of least effort*. Cambridge, MA, Addison-Wesley, 1949.

### Internet sources:

- [14] <http://www.web-caching.com/cacheability.html>
- [15] Common Log Format: <http://www.baculus.com/WsvlCLF.html>
- [16] Gain Network: <http://www.gainpublishing.com/>
- [17] log data: <http://www.ircache.net/Traces/>
- [18] [http://www.theregister.co.uk/2004/10/15/google\\_desktop\\_privacy/](http://www.theregister.co.uk/2004/10/15/google_desktop_privacy/)
- [19] Music Machines log data: <http://www.cs.washington.edu/ai/adaptive-data/>
- [20] Reed D.: Privacy and the Future of Behavioral Marketing,  
[http://www.claria.com/advertise/oas\\_archive/privacy.html?pub=imedia\\_module](http://www.claria.com/advertise/oas_archive/privacy.html?pub=imedia_module)
- [21] <http://validator.w3.org/>
- [22] WorldCup98 log data: <http://ita.ee.lbl.gov/html/contrib/WorldCup.html>

# Lessons from the Application of Domain-Independent Data Mining System for Discovering Web User Access Patterns

Leszek Borzemski and Adam Druszcz

Wroclaw University of Technology,  
Institute of Information Science and Engineering,  
Wybrzeze Wyspianskiego 27,  
50-370 Wroclaw, Poland  
leszek.borzemski@pwr.wroc.pl

**Abstract.** This paper presents the usage of a general domain-independent data mining system in discovering of the Web user access patterns. DB2 Intelligent Miner for Data was successfully used in data mining of a huge Web log which was collected during the World FIFA Cup 1998. The clustering, associations and sequential pattern mining functions were considered in the context of Web usage mining. The clustering method was found the most profitable and the discovered usage patterns can be used in Web personalization and recommendation systems.

## 1 Introduction

The Web site designers often are challenged how to satisfy the needs of large numbers of various users visiting their Web pages. The creation of a Web site which would fulfill expectations of all users is impossible in practice. This problem can be solved by the preparation of the Web site which could be adaptable to the users' pattern usage. Such adaptive Web site would learn users' behavior and improve its organization and presentation based upon certain pattern discovering techniques. Therefore current research is focused on the methods how to discover and model of user access patterns on the Web. The most promising research approach is web mining, i.e. application of data mining methods and algorithms to Web problems related to the Web content, Web usage, Web structure and Web user profile. The discovering process is based on the analysis of Web documents, Web services and Web logs. We are looking for regularity and dynamics of Web content, Web access patterns, Web (linkage) structures and user profile information, respectively [4, 6]. The data mining application showed in this paper corresponds to a Web usage mining. We want to discover general hidden patterns of user Web site usage and exploit this knowledge in the adaptation of the Web site design. Web usage mining has been studied from a several years; however it is still a vital research field. The surveys [8, 12] present comprehensive state-of-the-art of Web usage and personalization approaches and research perspectives in the field. Recent advances and various aspects of Web usage mining are discussed. The Web server logs are used in Web usage mining. The typical data collected in that log include IP user addresses, Web page URLs and request

completion time. Access pattern analysis was a goal of several projects however it is usually made by means of specific or domain-dependent tools, such as WEBMINER, WUM, Speed Tracer [1, 7, 9, 11]. WUM used sequential mining to analyze Web site navigational behavior of users. Speed Tracer system was applied to discover the most common traversal paths and groups of pages frequently visited together. Our work differs in that we use a general domain-independent data mining system instead of domain specific software. The goal is to show how such analysis can be effectively and friendly performed using a general and conventional stand-alone domain-independent data mining system. General data mining systems may have advantages including their productivity proved in many domain applications. We used here the IBM Intelligent Miner for Data (IM4D), which is one of the software leaders among high-end commercial data mining tools [13]. IM4D is a general framework for deploying the data mining applications for different domains with the minimum user skills required. Normal usage needs only a domain expert. It is assumed that the mining expert is needed only for the development of the mining scenarios. IM4D provides proven mining algorithms including clustering, association discovery, sequential patterns discovery and classification algorithms to address a wide range of problems. Each data mining algorithm is implemented as a function that can be used to perform a specific discovery task. For example, we used the clustering and decision tree functions in the Internet path performance model [3].

In this paper we analyze real-life Web server log from 1998 World FIFA Cup Web site (<http://www.france98.com>). [2] presents in-depth workload characterization based on traditional statistical data analysis. This data was not previously mined in the way as we are presenting now in this paper. The rest of this paper is organized as follows. Section 2 explains data preparation. Section 3 presents application of the Associations mining function. Section 4 shows application of the Sequential Patterns function. Section 5 talks about application of the Clustering mining function. Section 6 includes conclusions.

## 2 Data Preparation

The data mining (DM) application is beginning from the data preparation task. The goal of data preparation is to create a suitable target data set for mining algorithms. This involves some preparation of the original data for specific data mining operations. This process is time consuming and is crucial for further analysis.

The datasets used in our analysis were provided by Hewlett-Packard Labs [2]. They were originated from the Web server logs collected during the FIFA Cup in 1998. A Web site was serviced by 30 various HTTP servers distributed globally over four locations: Paris, France; Plano, Texas, U.S.; Herndon, Virginia, U.S., and Santa Clara, California, U.S. A number of load balancers (network dispatchers) were used to distribute the requests across these four locations and among the servers at each location. The servers recorded their activity using one second time stamps - the time was coordinated on every server with local time in France. During the World Cup the local time was 2 hours ahead of GMT. Data was collected from 30 April to 26 July of 1998. During this period 1,352,804,107 http requests were collected. The logs from all servers were merged in [2] into a single sequence of requests sorted by the time of

receiving of the request. The created set of data was not suitable for further analysis because of the size and it was divided into smaller data sets containing entries from one day and of the size below 50 MB (this corresponds to seven millions requests at most).

The Web server log is saved in the Common Log Format (CLF) [2]. Each record has the following structure: *remotehost rfc931 authuser [date] request status bytes*, where *remotehost* is the IP address or name of the client (remote host), *rfc931* is remote log name of user (almost always "-" meaning "unknown"), *authuser* is the authenticated username, *date* is the date and time of the completion of the request, *request* - first line of the request including the HTTP method and URL, *status* - HTTP response status code (200, 304, ...), and *bytes* is the number of bytes in the response.

The raw Web log is not suitable for data mining because it contains extra and redundant information. Therefore it must be prepared and cleaned before mining. The Perl scripts with the three-phase data preparation procedure were used for that.

First, we chose the entries referring to main documents. Web log keeps information about individual objects/files instead of direct information about the Web documents (pages) containing these objects. Therefore in the case of the Web page containing e.g. images, sounds or films, we obtain in the log entries referring to the main document (usually with *html* or *htm* extension), as well as entries of all objects embedded in that document. In the Web usage analysis we are interested in the main document entries. Therefore only such entries were selected. However it was shown that some documents were retrieved multiple times together with other. Because the page map of the site is not available (the FIFA site is not still under operation) we could not exactly define the role of these redundant documents. After further analysis of the names of files and their addresses we understood that probably they were frames used in the construction of the main documents. Since they did not enter any significant information for our study we deleted them. In conclusion, the filtering script selected entries referring to documents with the *htm* or *html* extensions, excluding *nav\_inet.html*, *nav\_top\_inet.html*, *splash\_inet.html* documents.

Next, we identified the users and user sessions. The users were recognized on the basis of the user unique identifiers. Each identifier refers to the original user IP address and was generated in [2] to make the log compact, and to protect user privacy. The definition of the user session is based on the idle time between consecutive requests. The session was assumed to be a new one if the idle time is greater than the assumed threshold. That threshold was determined to obtain the compromise between the number of possible sessions and their lengths within the request partitions, each including seven millions requests. The value of the idle time interval was set to 30 minutes and it is the same as in [6].

Lastly, we rejected Proxy server and NAT entries. In this phase we determined entries for the individual users. As we have already mentioned the user identification is based on the IP address used in the request. Such a solution does not fully guarantee the correct user identification because a single IP address can be used by many users. This concerns the proxy server and NAT technology. Entries concerning Proxy users should be rejected from our analysis. Such users were recognized as ones with too many open documents in the same session. Therefore the users with more than 15 open documents were classified as proxy users and appropriate entries were filter out

from the log. Additional Pearl scripts filled in the log some extra data such as the day number, the session number, and the request number in the session, as well as the session length. Each one-day log was formatted into two relational tables as IM4D uses the DB2 relational database: (i) type\_1 table with the records containing all data referring to a single request, and (ii) type\_2 table containing in a single record the whole sequence of requests from a user in a single session. Type\_1 tables were utilized when discovering association rules and revealing sequential patterns. Type\_2 tables were used in clustering.

### 3 Application of the Associations Mining Function

Uncovering of association rules is based on the search of pairs or groups of elements which are appearing together within subsets of the bigger set of data. The goal of the application of the Associations mining function to the analysis of the log of the Web server is to find the sets of pages that are accessed together in a single session. Association rule discovering can be done with a support value exceeding specified threshold. Generally, this technique can be used to discover unordered correlation between items found in a database of transactions. IM4D supports Apriori algorithm which finds groups of page views occurring frequently together in many transactions (i.e., satisfying a user specified minimum support threshold). Such groups of items are referred to as frequent itemsets. As the whole log was very large we prepared 88 subsets, each subset referred to a single day of observation. Such subsets were explored using the same set of parameter settings for the Associations mining function. In further description of case study of Web usage mining we will employ a general nomenclature of mining functions, transactions and items which is used by the IM4D [10].

The following input fields and parameter settings were specified in the Associations mining function. The input fields were: the session number as the Transaction ID field and the page name together with the access path as the Item ID field. The minimum support indicating the minimum relative occurrence of the detected association rules was set to 1 percent therefore we accepted rules which have been found not less than in 1% transactions. We think that such low support is acceptable due to the large number of Web documents (several thousands) and sessions (about 150,000). Similar support level (0.25% to 1.5%) was used in [5]. The minimum confidence was equal to 50 and the maximum rule length was unlimited. Items constraints were not defined.

Discussing the results we can say that the most often discovered rules informed about joint accessing the following pages */english/competition/matchstatXXXX.htm*, */english/competition/matchprogXXXX.htm* (a four-digit number was found in XXXX places). Based on the names of these pages we can guess that they are pages referring to specific football matches identified by four digit XXXX number. Next rules found show the associations between the pages of the type */english/teams/teambioXXX.htm* which probably contained the information about the teams identified by the three-digit XXX number. We also discovered the growing interest in pages containing the news. We found that the pages from the */english/news/* directory were often referenced together in a single server session. Some suggestions what pages we are dealing with can be presented on the basis on the name of the page, e.g. a page

*/english/teams/teambio124.htm* probably contains the biographic data of the football team #124 but we still do not know which team is about. Deeper analysis cannot be done without the knowledge of the page content.

Similar character of our findings for each day proves the correctness of the results returned by the Associations mining function. These findings can be used to tune the structure of hyperlinks on the pages in the way that the links to more popular pages should be found together in most visible places, making a site adaptive to user needs. This knowledge can be also useful in efficient storage organization of the pages that are accessed together.

## 4 Application of the Sequential Patterns Mining Function

Discovering sequential patterns assumes that in the user behavior we can meet repeating patterns of his/her transactions. The application of pattern sequence analysis is strictly involved with e-commerce. This can give information that, for example, 20 percent of users visited, at first the page */products/prod1.htm*, and then (in some next session) the page */products/prod2.htm*. The knowledge of this type of patterns is able to be helpful when planning marketing strategies. The mining function of uncovering of sequential patterns is very similar to the function to the associations revealing, the difference is only the fact that associations are being searched within the content of single sessions but sequential patterns in the groups of the sessions. This technique can be used to discover if a set of pages is followed by another page in time-order.

Similarly to the association rules revealing, we discovered sequential patterns on the data set from the specific day. The Transaction Group field was defined by the user ID, Transaction Field was defined by the session ID, whereas the Item Field was defined by the name of required Web document together with its URL path. The minimum support was assumed equal to 1%. The maximum pattern length was unlimited, and none items constraints on the values of the attributes were assumed.

Unfortunately, the results obtained by this mining function were very limited and it is hard to find new interesting relationships. The results from particular days were similar almost the same and followed the general pattern of visited pages: */english/competition/matchstatXXXX.htm*, */english/competition/matchprogXXXX.htm*. Another distinctive pattern often found was based on the page */english/competition/matchprogXXXX.htm/English/frntpage.htm* and page named */english/teams/teambioXXX.htm*, with the various numbers *X..X* in different days. The results showed as in case of the association rules that the users are interested in pages connected to the specific matches played in the day of the page downloading.

The weak sequential patterns mining result can be due to the vast number of users with a single session we observed in the log. Thus we filtered out all such users; however the character of the result did not change – only the share of each pattern increased. Another reason of such result could be connected with the problem of possible multiple (Proxy) users staying behind the same IP address. The huge number of accessed documents could imply such result, as well. Because, the results of both associations and pattern sequence mining speak with one voice, we think that the problem lays in too big diversity of input data for mining. Our recommendation is to collect the pages into thematic page groups and discover the patterns at the level of

accesses to the thematic page groups instead of individual pages. However, we could not prepare input data in such a way without the knowledge about the site page map and content of pages.

## 5 Application of the Clustering Mining Function

The goal of discovering clusters is to group records that have similar characteristics. Clustering in our case is performed to identify the group of users having similar interests in Web documents. Each such group defines a user usage profile which in turn can be used in the prediction of behavior of new users. In clustering we used the datasets of type 2 where each record includes successive requests within a single user session. The results of the clustering function show the number of detected clusters and the characteristics that make up each cluster. In addition, the results show how these characteristics are distributed within the clusters.

Clustering in IM4D can be accomplished by using either demographic (distribution-based) or neural (center-based) clustering algorithms [10]. The methods are distinguished by the data types of the input attributes that are allowed - demographic clustering methods function on records with categorical variables whereas neural clustering accepts only numeric input data, however categorical inputs can be transformed into quantitative variables. Here we used the demographic method. We mined the data subsets with particular day's observations using the following settings. We restricted the number of algorithm passes to 5. Specifying multiple passes through the input data improves the quality of the generated clusters but also extends the processing time required to perform clustering. We restricted the number of clusters to be generated up to 16, as well. In our previous work [3] we found that it is a good assumption in order to find accurate and homogenous groups of objects and simultaneously avoiding the production of too many small clusters. We also limited the number of passes and the processing time by setting an accuracy improvement to 2, i.e. the iteration process ended when the quality improvement between two phases was less than 2%. The percentage of improvement is measured at each pass over the data. If the actual improvement is less than the value specified, then no more passes occur. The smaller the value, the more accurate is the clustering. As the candidates of the active fields (attributes), i.e. record fields that participate in the creation of clusters, we chose the successive requests within the session, i.e. REQUEST1, ..., REQUEST15 (the maximum number of requests in the session was 15).

We mined for variable vs. fixed length sessions. First, we mined the datasets containing all sessions from one day of observation. There were sessions with the variable lengths, from 2 to 15 clicks. The result was not successful. The clusters contained only the repetitions of one or two specific documents as they were accessed in consecutive requests. This effect was due to the behavior of users who usually ended their sessions via accessing the most frequent documents. Therefore we decided to make the segmentation of input data to obtain datasets containing sessions of the same length. We got 14 tables for each observation day with, each having the sessions of the same length (2-15 clicks). Then the mining was successful and we obtained a vast amount of clustering results.



The analysis of clustering results confirms the majority of conclusions from the previous mining. We can fully claim that the greatest interest is in the Web pages referring to particular matches, after then in pages with teams' biographies and the news. Discovered user usage profiles like revealed association rules are keeping very similar characteristics between the results from various days, the differences most often come down to the changes in the match numbers, teams or news. These usage profiles not always wholly correspond to the sessions assigned to them. The most frequent request is being selected in most cases. There are rare situations when all requests in the sessions are the same as in the profile. The diversity is since the clustering was done based on individual documents. Then we have the big number of possible requests. Thus we recommend, as in the case of sequential pattern mining, to collect the pages into thematic page groups and discover the clusters at the level of accesses to these page groups instead of individual pages. Then the clusters and thus the user profiles would be more homogeneous.

In order to verify this proposition we cluster the documents based on their names in which we removed the names of the target pages. For example, the pages */english/competition/matchprog8880.htm*, */english/competition/matchprog8879.htm* matched the same cluster */english/competition/matchprog* after removing "the ending" *XXXX.htm*. Unfortunately, the cluster homogeneity does not increase. We think that the reason is the specific usage pattern of the site by the users. Namely, each day they accessed mainly one match page (the match of the day) and two or three team pages. However despite this result we think that the deeper thematic grouping may be successful.

## 6 Conclusions

This paper presents some lessons from the real-life application of a general domain-independent mining system in Web mining. The analysis was carried out on well known Internet archive containing the Web log collected during FIFA World Cup 1998. We showed how the Web log has to be transformed into a transactional form suitable for data mining as well as for processing based on a relational database. The data preparation process, which can be done only by the domain user, includes three-phase data transformation, Proxy server and NAT entries rejection, and transformation into relational tables.

We demonstrated that the clustering, associations and sequential patterns mining functions, as implemented in IM4D, can be successfully applied to Web usage mining. Using association mining we discovered pages that are accessed together. These findings can be used to tune the structure of hyperlinks on the pages in the way that the links to more popular pages should be found together in most visible places, making a site adaptive to user needs. Unfortunately, the sequential patterns mining generally showed weak results. We think that the problem lays in too big diversity of input data for mining due to an excessive number of much differentiated documents. We proposed the improvement based on the thematic page group concept. However, it could not be done without the knowledge of the site page map and content of the pages. We are satisfied best from the application of clustering where we discovered usage profiles within the user sessions. This knowledge can be used, for example, for

automated site personalization or recommendation. Then for each page request the Web server can return the set of URLs of the related pages that were most often referenced together with the requested page. Additional links can be provided from one or more previously discovered user profiles which match with an active user's access pattern.

We also learned from this case study that the designing of Web data archives with potential Web mining applications in mind is not a trivial task. This is the matter of detailed analysis of the context in which the mining applications are to be used. For example, the archive should store the Web site page map and content description to be used for the needs of insightful Web usage mining.

## References

1. Albanese, M., Picariello, A., Sansone, C., Sansone, L.: A Web Personalization System based on Web Usage Mining Techniques, WWW2004, ACM Press, New York, (2004)
2. Arlit M., Jin T., A Workload Characterization Study of the 1998 World Cup Web Site, IEEE Network, May/June (2000) 30-373
3. Borzowski, L.: Data Mining in Evaluation of Internet Path Performance. LNAI, Vol. 3029. Springer-Verlag, Berlin (2004) 643-652
4. Chakrabarti, S.: Mining the Web: Analysis of Hypertext and Semi Structured Data. Morgan Kaufmann, San Francisco (2003)
5. Chen, M.-S., J. S. Park, Yu, P. S.: Efficient Mining Date for Path Traversal Patterns in Distributed Systems. 16th IEEE Int. Conf. on Distributed Computing Systems (1996)
6. Fu Y., Sandhu K., Shi, M-Y.: Clustering of Web Users Based on the Access Patterns. LNAI, Vol. 1836, Springer-Verlag, Berlin (2000)
7. Fürnkranz, J.: Web mining. In: Maimon, O., Rokach, L., (eds.): Data Mining and Knowledge Discovery Handbook. Springer-Verlag, Berlin (2005) 899-920
8. Mobasher, B.: Web Usage Mining and Personalization. In: Practical Handbook of Internet Computing, Munindar P. Singh (ed.), CRC Press (2005)
9. Spiliopoulou, M., Lukas, C.: WUM: A Tool for Web Utilization Analysis. LNCS, Vol. 1590. Springer-Verlag, Berlin (1998) 184-203
10. Using Intelligent Miner for Data. V8 Rel. 1, IBM Redbooks, SH12-6750-00 (2002)
11. Wu K.L., Yu P.S., Ballman A.: Speed Tracer: A Web Usage Mining and Analysis Tool. IBM Systems Journal, Volume 37, Number 1 (1998)
12. Zhang, F., Chang, H.: Research and Development in Web Usage Mining System—Key Issues and Proposed Solutions: A Survey. Proc. First IEEE Int. Conf. on Machine Learning and Cybernetics (2002) 986–990
13. <http://www.kdnuggets.com>

# Application of Hybrid Recommendation in Web-Based Cooking Assistant

J. Sobecki, E. Babiak, and M. Ślanina

Institute of Applied Infoematics, Wrocław University of Technology  
Wyb. Wyspińskiego 27, 50-370 Wrocław, Poland  
{sobecki@pwr.wroc.pl, emilia-babiak@wp.pl,  
marta.slanina@neostrada.pl}

**Abstract.** The application of hybrid recommendation enables to overcome disadvantages of all three basic approaches: demographic, content-based, and collaborative ones. In this paper we present application of web-based cooking information system that recommends cooking recipes for different users. This work is continuation of previous works on hybrid recommendation that introduces application of fuzzy inference for demographic stereotype reasoning, which is the main new contribution of this paper.

## 1 Introduction

Recommender systems are one way of the success of the today's web-based information systems that is achieved by delivering of customized information for their users [6]. We can distinguish three basic types of recommendations: demographic, content-based and collaborative [6].

The central element of all recommender systems is the user model. The user model contains knowledge about the individual preferences which determine his or her behavior [8]. According to [5] the user model is usually built of the two elements: the user data and the usage data. The former contains different information such as: demographic data, users' knowledge, their skills and capabilities, their interests and preferences and also their plans and goals. The later element of the user model, the usage data, is observed and recorded during the user's interactions with web-based systems. The usage data may concern selective operations that express users' interests, unfamiliarity or preferences, temporal viewing behavior, as well as ratings concerning the relevance of these elements.

The methods for user model (user profile [5]) representation, initialization and acquisition are very differentiated [6]. The user profile in the recommender systems could be represented by: binary vectors, feature vectors, trees, decision trees, semantic networks, Bayesian networks, etc. The initial profile may be empty, may be created from the questionnaire that is filled in by the user or may be observed and recorded during the whole process of user's interactions with web-based systems. In this case it may concern selective operations that express users' interests, unfamiliarity or preferences, temporal viewing behavior, as well as ratings concerning the relevance of these elements that are expressed by different events, such as: opening a page, purchasing a product, sending feedback information to the system are stored.

The initial profile may be also modified according to the whole user population behavior, which is the case of collaborative recommendation. However this brings some problems with finding similar users. These problems may be solved by application of the clustering methods [5].

In this paper we will present a new hybrid approach for recommendation in web-based cooking assistant. In many previous works [7],[10],[11],[12] several hybrid approaches based primarily on consensus methods were presented. They were mainly combination of demographic and collaborative recommendations with some application of content-based recommendation. The main new contribution of the recommender system presented in this paper is application of fuzzy reasoning for demographic recommendation instead of standard stereotype reasoning [6]. In the paper the adaptation of these fuzzy rules is proposed as well as application of recipes ratings together with Pearson coefficient [1] for determining of similar users.

In the following chapter the general architecture for hybrid recommendation presented in several previous works [7],[10],[11],[12]. Then application of fuzzy rules in demographic recommendation in web-based cooking assistant is presented. Finally in the summary achieved results and future work is presented.

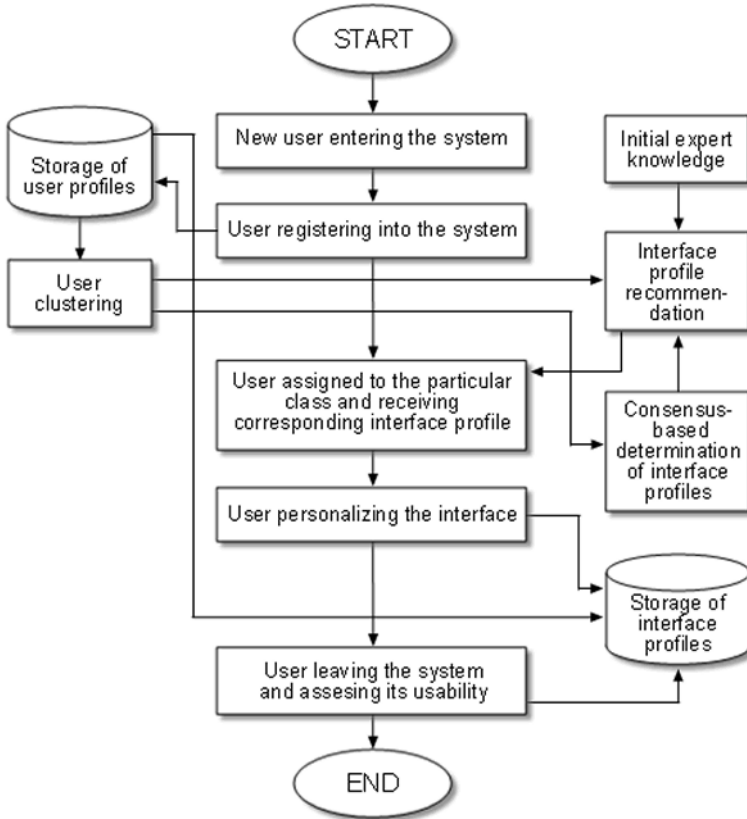
## 2 Architecture of the Hybrid Recommender System

In the early implementations, the system architecture was based on the general ideas of demographic and collaborative recommender systems. The demographic approach is mainly using stereotype reasoning [5] and is based on the information stored in the user profile that contains mainly different demographic features [6]. Stereotype reasoning is a classification problem that is aimed at generating initial predictions about the user. The demographic recommendations have however two basic disadvantages [6]: they may be too general and they do not provide any adaptation to user interests changing over time. The collaborative recommendation is able to overcome some disadvantages of the demographic approach because it may offer novel items, even such that user has never seen before. Collaborative recommended agents have also some disadvantages, such as: poor predictions when the number of other similar users is small, lack of transparency in the process of prediction and finally the user's personal dislike may be overcome by the number of other similar users' opinions.

The system adaptation (see Fig. 1) starts with registering each new user. The registration data are stored in the user profile that is represented by a tuple that is a function  $p:A \rightarrow V$ , where  $A$  is a set of attributes and  $V$  is a set of their elementary values and  $(\forall a \in A)(p(a) \in V_a)$ . Then demographic recommendation is applied that according to the user profile assigns the user to the appropriate user group. With each group of users there is associated a corresponding interface settings. The user registration is not obligatory but in this case the default interface profile is delivered.

According to the interface settings the actual user interface content, layout and structure is generated. The user may start to work with the system and if he or she wishes so they also may modify the interface settings. Finally, these settings together with subjective satisfaction evaluation given by user are stored in the user profile.

After registering the required number of users, the users are clustered using Dattola algorithm [3] and then according to these clusters using consensus methods [7] new interface settings are distinguished for recommendation. These procedures may be repeated from time to time in the following occasions: many new users registering to the system or interface recommendations become poor usability ratings.



**Fig. 1.** Architecture of the consensus-based user interface adaptation

The later applications [11] also applied the third approach, i.e. content-based recommendation that takes descriptions of the content of the previously evaluated items to learn the relationship between a single user and the description of the new items [6], or in other words a user is supposed to like a new item if the item is similar to other items that are liked by the user [2]. This approach however, has also some disadvantages, it tends to overspecialize recommendations and is based only on the particular user relevance. So the best way to overcome all disadvantages of each single approach is to use hybrid recommendation that is a combination of all three methods [11].

The rules for efficient content-based recommendations strongly depend on the goals of the web based system. For example for web-based information retrieval systems we can consider the previous relevant items as a basis for recommendation of further retrievals. In this case many different methods can be used: the cosine similarity function that is presented below, fuzzy retrieval, Bayesian networks or other intelligent information retrieval method. For quite many systems however, the logic used for the retrieval systems, does not hold. So, for each recommended item, no matter if an element of interface settings or a content item, we shall define precise relationship between the user profile (or also other users profiles) and this particular item, which may be implemented for example as ruled based system, Bayesian or neural networks. However the usage data may lead to many different recommendations so there is a place for consensus determination.

In the work [11] the following web-based recommender systems were presented: "The cooking assistant", "The Movies" and "Comp News". Besides the application of demographic and collaborative recommendation they also apply content-based recommendation. The content-based recommendation applied in these systems is constructed in the following way. First, the items (movies, recipes or news) that were highly ranked have an influence on the user profile elements representing user interests. Then each new item represented with the feature vector may be compared with the user profile interest vector with cosine similarity (1) and recommend those items that are the most similar to the interest vector:

$$\beta(a,b) = \frac{\sum_{i=1}^n a_i * b_i}{\sqrt{\sum_{i=1}^n a_i^2 * \sum_{i=1}^n b_i^2}} . \quad (1)$$

In the formula (1),  $a$ ,  $b$  are feature vectors,  $n$  is the number of features and  $a_i$ ,  $b_i$  are  $i$ -th feature of the vectors  $a$ ,  $b$  accordingly.

In "The cooking assistant" recommender system different types of recommendation were used: basic and hybrid ones. The demographic recommendation is based on the age and gender of the user offers different interface settings such as: music track, volume, font size and hints; and content settings: additional information and wine selection. These settings may be changed by the users so it is possible to find consensus among these settings and offer them as a new recommendation for similar users. This implementation also delivers situation recommendation that offers receipts for breakfast, lunch or dinner according to the daytime. Finally also content-based recommendation is applied according to the preferences explicitly stated by the users concerning preferred cuisine and specified food products. The system usability was tested with 12 users using the questionnaire method and the classical usability tests [11]. These tests proved that the system usability was satisfactory.

In "The Movies" recommender system also different types of recommendation were used. The system delivers information about movies stored in the system "Stopklatka" and current cinema repertoire from that information system. The user model is initialized during registration process, where each user is asked to give ratings of selected films. Then according these ratings each user is grouped and then according to the stereotype reasoning specific interface settings concerning layout, types of icons, color and background are recommended. The user profile settings

concerning different movie attributes such as: genre, director, writing, music, cast and cast are changed according to the explicit user ratings of films or specified attributes and implicit user searching and browsing of movies.

The system also delivers reach collaborative recommendation that concerns: interface settings according to the changes delivered by the similar users, sorting the current cinema repertoire for unregistered users according to the preferences of all the system users and optionally also for registered users according to the preferences of all similar users, three extra recommended movies according to the highest ranks given by the similar users. The content based recommendation for registered user that sorts the cinema repertoire according to the preferences reflected in the user profile. By default the hybrid recommendation that is mixture of collaborative and content-based repertoire sorting is applied. The conducted usability tests containing the questionnaire method and the classical usability tests showed some minor problems with interaction but most of the users appreciated implemented recommendation mechanisms.

Finally, the simplest system “Comp News” recommends computer news. At the beginning users delivers some demographic data, preferred interface style and their interests on four topics: purchase optimization, over-clocking, hardware modding and new products, in form of a feature vector. Then news are sorted according to the similarity with the interest vector, which is subject to constant changes according to reading consequent news. The user may also select an average interest vector determined for the group of similar users. Even that simple approach was assessed by the experimental users as being pretty useful.

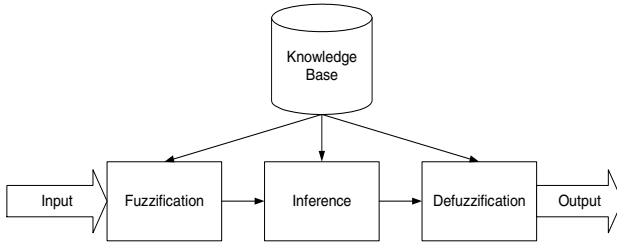
The main conclusion from these implementations was following: to deliver content-based recommendation for a particular user we must have sufficient usage data of that user and appropriate inductive rules that will transform this data into the user interface settings.

### 3 Demographic Recommendation Using Fuzzy Inference

Fuzzy logic was derived from the fuzzy set and possibility theory that was introduced by Lofti Zadeh in 1965. Despite its controversy it is widely accepted in many technology areas [4]. In this paper we present application of fuzzy inference in demographic recommendation in web-based system cooking assistant that was based on the system “The Cooking Assistant” presented in [11]. There are 340 cooking recipes available in the system. These recipes were also ranked cooking recipes according to their difficulty, fantasy and number of calories.

The main advantage of application of fuzzy reasoning over the standard stereotype reasoning is giving experts the more convenient tool for modeling different relationships from the real world. The general fuzzy inference system applied in our recommender system is shown in Fig. 2. In this model [9] the particular input characteristics is mapped to input membership functions, which is mapped to rules that are mapped into a set of output characteristics and consequently to output membership function, which finally is mapped to a single-valued output associated with the decision.

In our recommender system we selected the following demographic attributes describing each user for the input: age, gender, number of inhabitants in the place of living; user knowledge attributes: cooking experience and preferences: vegetarian or



**Fig. 2.** Fuzzy inference system [9]

not. The values of these attributes were used as the input to determine the membership function values of linguistic variables. All the membership functions applied in the system have the trapezoid shape, as shown in Fig. 3. The membership function values are defined as follows:

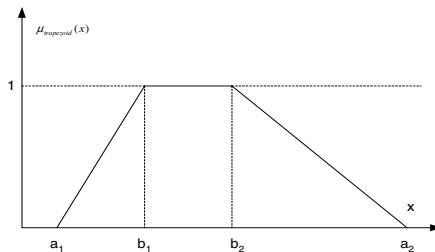
$$\mu_{trapezoid}(x) = \begin{cases} 1 & , x \in [b_1, b_2] \\ \frac{x - a_1}{b_1 - a_1} & , x \in [a_1, b_1] \\ \frac{a_2 - x}{a_2 - b_2} & , x \in (b_2, a_2) \\ 0 & , otherwise \end{cases} \quad (2)$$

For example for variable  $x$  representing age and membership function representing user's being of middle-ages, we have the following values of  $a_1=40, b_1=45, b_2=55, a_2=59$ .

These attributes were used to construct fuzzy inference rules that for some assumptions expressed in so called linguistic values (eg. medium age) assign conclusions concerning recipes description also expressed in the linguistic values. In the system we implemented over 170 fuzzy rules in the following form:

*If the age is medium and the cooking experience is high and user lives in a big city then the difficulty of meal may be high.*

Each rule has associated confidence factor ( $cf \in [0,1]$ ), which multiplies the membership function value of the outcome of the rule. This factor expresses the expert's confidence in the particular rule. By default this factor is set to 1.0, however it may be changed by some adaptation procedures that are described in the last section.



**Fig. 3.** Fuzzy inference system [9]



For the defuzzification we applied very simple method called Mean of Maximum (MOM), which takes the mean value of the set with maximum membership grade. According to the fuzzy inference system we determine values that represent cooking recipes: difficulty, fantasy and number of calories. Then using standard cosine similarity function we determine the most similar recipes with the value over the specified limit. These recipes are sorted with descending similarity value and the number is limited to 15 recipes.

“The Cooking Assistant” recommender system was implemented in the three tier technology using Macromedia Flash for user interface implementation, PHP for the system intelligence and MySQL for the database tier.

Besides demographic recommendation the system applied also collaborative, content-based and “season-based” recommendation in the way that was described in the section 2.

## 4 Conclusions and Future Work

The usability of the final version of the system has not been tested yet, however the earlier version of the system was tested with 12 users using the questionnaire method and the classical usability tests and the tests proved that the system usability was satisfactory. Usually it is quite difficult to test recommender systems, because many different users are necessary to show how collaborative method operates, as well as each user needs rather long time of working with the system to show how content-based method operates.

Application of fuzzy reasoning in demographic recommendation gave experts a great freedom in the knowledge modeling in the area of cooking recipes recommendation by means of fuzzy rules. We must also remember that user preferences in this area are very difficult to formalize.

Before we conduct more thorough experiments, at least as such presented in paper [10] that engaged 80 users, we should prepare more cooking recipes and introduce some modifications to the presented method. However the number of 340 recipes prepared so far seems to be pretty large, in case of very differentiated users with very differentiated cuisine preferences their number seems to be insufficient.

The first modification should concern the registration process because users are very reluctant to submit data about themselves, so the very well known mechanisms of ranking recipes that is known from many movie recommender systems should be applied. We should however consider the application of the following rankings: one for dishes they like, second for dishes they can prepare and third, of cooking recipes presented by the recommender system. Then having these ratings we can easily find similar users using for example Pearson coefficient is used [1]:

$$r(x, y) = \frac{\sum_{m \in \text{movie}} (rate_{x,m} - \overline{rate_x})(rate_{y,m} - \overline{rate_y})}{\sqrt{\sum_{m \in \text{movie}} (rate_{x,m} - \overline{rate_x})^2 \sum_{m \in \text{movie}} ((rate_{y,m} - \overline{rate_y})^2)}} . \quad (3)$$

In the formula (3)  $rate_{x,m}$  is the mean of ranks (for all six features) given by the user  $x$  for them movie  $m$ , and  $\overline{rate_x}$  is the mean value of  $rate_{x,m}$  for all movies evaluated by user  $x$ . The value of the correlation  $r(x,y)$  is a real number that lies within  $[-1, 1]$ . By defining the given threshold  $\tau$ , we can determine the most similar users. Then this group of similar users may be used for determination the user interface layout using consensus methods [7],[10] as well as to find some recommended cooking recipes using standard collaborative filtering methods as applied for example in the system “The Movies” presented in paragraph 2. The demographic recommendations even based on the fuzzy rules do not provide any adaptation to user interests changing over time so we should also consider adaptation of the confidence factors ( $cf$ ) of the fuzzy inference rules using for example neural networks or data mining methods.

## References

1. Christakou C, Stafylopatis A, “A Hybrid Movie Recommender System Based on Neural Networks”, in Proc. Fifth Int. Conf. on Intelligent Systems Design and Applications, (2005) 500-505.
2. Dastani M, Jacobs N., Jonker CM, Treur J, Modelling User Preferences and Mediating Agents in Electronic Commerce. LNCS 1991, (2001) 163-193.
3. Dattola RT, A fast algorithm for automatic classification. In: Report ISR-14 to the National Science Foundation, Section V, Cornell University, Dep. of Computer Science, (1968).
4. C. Elkan. The Paradoxical Success of Fuzzy Logic. IEEE Expert, pp. 3-8, August 1994. With fifteen responses on pp. 9-46. First version in AAAI'93 proceedings, (1994) 698-703.
5. Kobsa A, Koenemann J, Pohl W, Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. Knowledge Eng. Rev. 16(2), (2001) 111-155.
6. Montaner M, Lopez B, de la Rosa JP A, Taxonomy of Recommender Agents on the Internet. Artificial Intelligence Review 19, (2003) 285-330.
7. Nguyen NT, Sobecki J, Using Consensus Methods to Construct Adaptive Interfaces in Multimodal Web-based Systems. Universal Access in Inf. Society 2(4), (2003) 342-358.
8. Papatheodorou C., Machine Learning in User Modeling. Machine Learning and Its Applications (2001) 286-294
9. A.P.Papłoński, Neuro-Fuzzy Computing. Download date:10.03.2006 from <http://www.csse.monash.edu.au/courseware/cse5301/04/Lnts/L12fz.pdf>, (2006)
10. Sobecki J, Weihberg M, Consensus-based Adaptive User Interface Implementation in the Product Promotion. In Keates S., Clarkson J., Langdon P. and Robonson P.: *Design for a more inclusive world*, Springer-Verlag: London (2004) 111-122.
11. Sobecki J (2005) Implementation of consensus-based hybrid recommendation in different web-based, 5th International Conference on Intelligent Systems Design and Applications. Proceedings. Eds. H. Kwaśnicka, M. Paprzycki. Wrocław, September 8-10, 2005. pp. 494-499, Los Alamitos [i.in.]: IEEE Computer Society [Press]
12. Sobecki J, Consensus-based hybrid adaptation of web systems user interfaces, J. Univ. Comput. Sci. 11(2), (2005) 250-270.

# Using Representation Choice Methods for a Medical Diagnosis Problem

Kamila Aftarczuk, Adrianna Kozierekiewicz, and Ngoc Thanh Nguyen

Institute of Information Science and Engineering, Wrocław University of Technology, Poland  
kamila.aftarczuk@student.pwr.wroc.pl,  
adrianna.kozierekiewicz@student.pwr.wroc.pl,  
Ngoc-Thanh.Nguyen@pwr.wroc.pl

**Abstract.** This paper presents two solutions of a medical diagnosis problem. We present two algorithms of this problem: one of them is based on data mining methods and the second relies on representative choice methods. The analysis of these solutions and the comparison of both algorithms are presented.

## 1 Introduction

Every physician is directly responsible for health and life of his patients. Consequently his occupation is a difficult and responsible one. If only for that reason, giving a proper diagnosis or a referral to a suitable specialist is an extremely important, although difficult task.

More and more often, medical institutions are being equipped with sophisticated management systems. Over the period of the years medical institutions equipped with such systems may accumulate information of vast quantity and variety. Acquired data can then be processed in a specific way, in order to obtain nontrivial dependencies between patients and symptoms that were affirmed in the diagnosis given by a physician. Discovered rules, in considerable way facilitate doctors' work helping them during the diagnosing process. Efficiency of the diagnosis is based on their own knowledge and experience, as well as additional information content in the database.

Such medical decision making process considerably reduces the risk of mistakes, influences on time and precision of physician's work and patient's satisfaction from medical services. Decision making methods, which are the focus of this work, are based on two types of data: qualitative data (temperature, pulse, analytical laboratory tests, etc.) and what is more of a quantitative data, based on diagnosis from the previous years. The vast amount of quantitative data in medical systems is very difficult to review and extract important information from by a medical worker. It is then necessary to process and analyse them in a proper way. The methods described in this work permit to solve the introduced problem.

Our first task is to convert data to a decision table, which then will be used during the analysis. A row in this table refers to a patient. This table has the structure shown in Table 1, where attributes  $s_1, s_2, \dots, s_l$  represent symptoms, and  $d_1, d_2, \dots, d_K$  represents the results of diagnosis. Value 1 of an attribute  $s_i$  means that the symptom was observed by the physician when diagnosis was given, value 0, on the other hand, means the lack of the symptom.

**Table 1.** The decision table

<i>Patient</i>	$s_1$	$s_2$	...	$s_l$	$d_1$	$d_2$	...	$d_K$
$p_1$	1	0		0	1	1		1
$p_2$	1	1		0	1	1		0
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Referring to an attribute  $d_j$  values 0 and 1 mean the negative and positive diagnosis for disease  $d_i$ , respectively. For example: referring to patient  $p_2$  the second row means that symptoms  $s_1$  and  $s_2$  were among others observed and diseases  $d_1$  and  $d_2$  were among others diagnosed.

Some chosen methods of exploration from long-term medical databases are presented in this paper. The solution of the problem was obtained on the basis of the data mining method and a representation choice method. Data mining methods are already being used in medicine, for example in radiology to diagnose cancer cells. The representation choice method is an innovatory solution in this field. The topic of consideration is a question of knowledge discovery about patients, symptoms and diagnosis from long-term medical data sets. This information is then processed and gathered in the decision table. Our task is to discover nontrivial dependencies between specific symptoms and a diagnosis.

In reality, the structure of data can be very complex. The data can be gathered in many tables instead of one. For example, patients table (name, address, sex, age, etc.), symptoms table, diseases table, visits table, etc. The solution to our problem requires the concrete data structure. Before the suggested algorithms can be used, gathered data ought to be converted into one table. On this basis it is possible to find some dependences useful in the process of diagnosis.

The most popular method of knowledge representation in data mining are decision tables and decision trees. In this paper we concentrate on methodology uncovering dependencies accumulated in data contained within decision table in which rows represent objects, and the columns represent attributes. Every element of the table characterises an object using a suitable attribute. There exist two types of attributes: conditional and decision. The first of them, set  $S = \{s_1, s_2, \dots, s_l\}$ , represents concrete symptoms, and the second, set  $D = \{d_1, d_2, \dots, d_K\}$  represents the diseases. Let  $P = \{p_1, p_2, \dots, p_N\}$  be the set of patients. Then the decision table may be defined as the following quadruple:

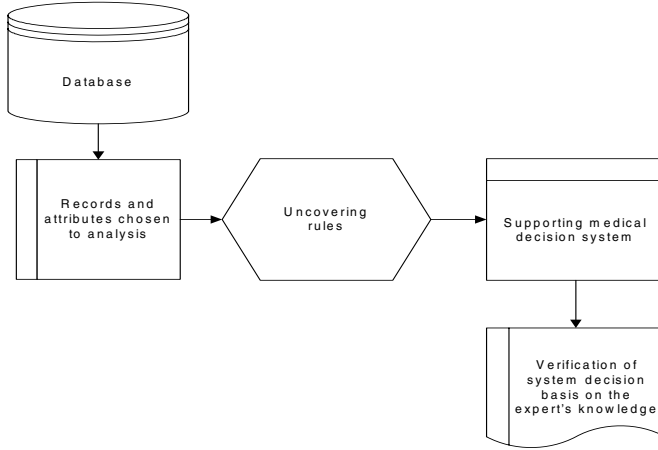
$$T = (P, S, D, \rho)$$

where  $\rho$  is a function

$$\rho: P \times \{S \cup D\} \rightarrow \{0, 1\}.$$

There exist several methodologies for discovering dependencies. Frequently, these methods are taken from artificial intelligence and are connected with learning machines.

The simplicity of description and legibility of represented knowledge is the basic requirement as far as the representation for the user goes. Data mining methods are helpful in a knowledge discovery process. In Fig. 1 a process for knowledge discovery process is presented.



**Fig. 1.** Knowledge discovery process

Very important element in data mining is the analysis of the association, the aim of which is to find hidden connections, that could be presented as rules, in large certain gathering data. In the process of associative analysis those rules are created. The rules have the following structure: „if  $x$  then  $y$ ”, where  $x$  and  $y$  are logical formulas in the conjunction form.

For the medical diagnosis problem dealing with the decision table we are interested in finding the answer for the following question: *On the basis of the data what symptoms are characteristic for a disease?* The answer of this question could be very useful in practice. Before presenting the methods we give in the next section an overview of medical diagnosis systems, which are used in practice. The method presented in this work should be useful in recommendation process for physicians to facilitate their decision making process in medical diagnosis.

## 2 Related Works

Currently there is plenty of software to run small and average medical posts, for example: “Physician”, “Consulting Room”, “Assistant for Windows” and many others. Although these programs are simple enough to help in medical post, they do not contain supporting medical diagnosis systems. From the available materials it follows that a lot of such systems have been in a beginning phase of realization and still need much time to be finished.

On the other hand one can find huge, complex medical systems. For example, “Help” has been improved since 1980 and appropriated in the hospitals in Utah.

Instead of full hospital administration and patient monitoring it has an additional advantage: supporting medical diagnosis system. This module is constructed in an interactive way and is activated whenever new data about a patient are saved, or in strictly specific time intervals [7],[11]. Supporting diagnostic system "Dxplain" was invented 10 years ago in Massachusetts General Hospital and is being used by thousands of users. Database and system are still being improved. Database of this system (Bayesian Networks) contains descriptions of above 2000 diseases, including: symptoms, etiologic, pathologic, prognosis and references to literature. System delivers the list of possible diseases related to more than 5000 symptoms and clinical tests. This system is a kind of electronic medical handbooks, therefore it is being used not only in decision supporting but also as a knowledge source by the students [7],[8].

There also exist systems based on a database limited to one discipline. For example: system GIDEON supports infectious decision making [9]; system TxDENT is designed for dentists [23]; and system Isabel concentrates on pediatric diseases. It is popular that the set of patient symptoms is compared with data from literature (medical books, articles) and afterwards the disease is diagnosed [12].

Data mining methods have been proved to be very useful in practical applications in many fields. These methods find application in business, telecommunication, etc. The opportunities connected with these methods are enormous, from market analysis and prediction of customers' behaviour to saving human life. Due to data mining in medicine the time needed for diagnosis is deducted, the precision and authority of physician's diagnosis grew up incredibly. It was possible with the support of intellectual knowledge and experience with information obtained from work of other people.

According to information included for example in [16] data mining methods were applied successfully in supporting medical decision making [13] disciplines: oncological diagnostic and prevention [2], liver treatment [17], psychology [19] and also genealogy [20]. Supporting medical diagnosis in aforementioned disciplines is possible due to the development of programs with automatic data analysers. Such significant problem analysis is possible due to the application data mining methods. According to [18] these methods can be differentiated in three groups. The first group consists of inductive learning of symbolic rules, for example: inductive rules [4], decision trees [21] and logical programs [16]; the second group contains statistics methods such as  $k$ -nearest neighbour; the third group gathers algorithms based on artificial neural networks [22].

To analyse long-term medical data it is Apriori algorithm and choice of representative method which are selected. The choice was made after careful analysis data mining methods and other deduction methods. Adaptation of these methods is an innovative solution in case of medical databases.

Algorithm Apriori is very well known [1],[3],[10]. It begins from looking for all possible formulas with the small number of descriptors then joining them into more complex structures. Algorithm Apriori is generally used in market basket analysis. In this article its usability in medical database analysis is also shown.

The other way of medical knowledge discovering is the choice of representative method presented in [6]. Using it in medicine is also innovative solution. The method relies on choosing of ordered coverings. It is much faster than previous one and assures support for medical decision making as well. In this paper the thorough analysis and comparison is presented. On its basis the conclusion about similarities and differences of both methods were presented.

### 3 Solutions of the Problem

#### 3.1 Using Data Mining Methods

It is natural that for the problem formulated in Introduction one should use data mining methods, which enable to find the association rules being the answer of the question. The Apriori algorithm finds all frequent item sets and then convert them into the association rule set. To show this algorithm a few auxiliary definitions are quoted.

**Definition 1.** By the support  $f((s_1,1) \wedge (s_2,1) \wedge \dots \wedge (s_i,1) \wedge (d_k,1))$  of formula  $(s_1,1) \wedge (s_2,1) \wedge \dots \wedge (s_i,1) \wedge (d_k,1)$  in the decision table we call the number of tuples where values referring to attributes  $s_1, s_2, \dots, s_i$  and  $d_k$  are simultaneously equal 1 for  $1 \leq i \leq I$  and  $1 \leq k \leq K$ .

**Definition 2.** The confidence of an association rule

$$(s_1,1) \wedge (s_2,1) \wedge \dots \wedge (s_i,1) \Rightarrow (d_k,1)$$

is equal:

$$\text{Con}((s_1,1) \wedge (s_2,1) \wedge \dots \wedge (s_i,1) \Rightarrow (d_k,1)) = \frac{f((s_1,1) \wedge (s_2,1) \wedge \dots \wedge (s_i,1) \wedge (d_k,1))}{f((s_1,1) \wedge (s_2,1) \wedge \dots \wedge (s_i,1))}.$$

The Apriori algorithm is presented as follows:

**Given:** Decision table  $T$ , threshold  $f_{min}$ ,  $Con_{min}$  for the support and the confidence values, respectively.

**Results:** All rules  $R_k$  that have support and confidence greater than  $f_{min}$  and  $Con_{min}$  respectively

BEGIN

1. Let  $j:=1$  and find frequent item set  $L_1$  of 1-itemsets  $s$  such that  $f((s,1)) > f_{min}$ ;
2. Let  $j:=j+1$ , choose from  $L_{j-1}$  these pairs which have exactly  $j-2$  common elements, add all these pairs to candidate set  $C_j$ ;
3. Delete from  $C_j$  the  $j$ -itemsets for which subset of size  $j-1$  is not frequent (does not appear in  $L_{j-1}$ );
4. Choose from  $C_j$  frequent item sets and add them to set  $L_j$ ;
5. If  $L_j \neq \emptyset$  and  $j \neq J$  go to Step 2;
6. Create from  $L = \bigcup_{j \geq 1} L_j$  association rule set such that:  

$$\text{Con}(X \Rightarrow (d_k,1)) > Con_{min}, \text{ for } X \subset L;$$
7. If  $j < K$  then GOTO 2.

END.

As the result of this algorithm we obtain sets  $L_1, L_2, \dots, L_K$  of itemsets.

#### 3.2 Using Consensus Theory

In this section we present an algorithm relying on consensus theory, which enables determining representations of sets. In this approach we restrict the decision table  $T$  to

only one disease  $d \in D$ . The restriction is based on removing from  $T$  these columns referring to other diseases than  $d$ , and these rows in which the value referring to  $d$  is equal 0. In the result we obtain table  $T(d)$ . Thus table  $T(d)$  contains only data referring to the patients for whom disease  $d$  has been diagnosed. We mention the question: *What symptoms are characteristic for disease  $d$ ?*

To answer this question we formulate the following representation choice problem:

Let  $S_i \subseteq S$  ( $i=1,2,\dots,N$ ) be the set of these symptoms, which have been observed for patient  $p_i$  referring to disease  $d$ . In fact on the basis of table  $T$  we have:

$$S_i = \{s \in S: \rho(p_i, s) = 1 \wedge \rho(p_i, d) = 1\}.$$

One should determine a set  $S^* \subseteq S$  such that  $S^*$  best represents the sets  $S_1, S_2, \dots, S_N$ .

For solving this problem first we define the distance function between sets of symptoms as follows:

$$d(S_i, S_j) = (1/N) \cdot \text{card}(S_i \div S_j)$$

for  $i, j = 1, 2, \dots, N$ . This function has been proved to be a metric. In work [6] it has been shown that set  $S^*$  should satisfy the following condition:

$$\sum_{i=1}^N d(S^*, S_i) = \min_{S' \subseteq S} \sum_{i=1}^N d(S', S_i).$$

To find set  $S^*$  satisfying this condition we use the following problem [6]: By  $\eta(s)$  we denote the number of occurrences of symptom  $s$  in sets  $S_1, S_2, \dots, S_N$ .

**Theorem 1.** *For each symptom  $s \in S$  the following dependencies are true:*

- a) *If  $\eta(s) > N/2$  then  $s$  should occur in  $S^*$ .*
- b) *If  $\eta(s) < N/2$  then  $s$  should not occur in  $S^*$*
- c) *If  $\eta(s) = N/2$  then  $s$  may or may not occur in  $S^*$*

Form this theorem it is easy to determine set  $S^*$ . The algorithm for determining set  $S^*$  is presented as follows:

**Given:** Sets  $S_1, S_2, \dots, S_N$ .

**Result:** Representative  $S^*$  of given sets.

BEGIN

1. For each  $s \in S$  calculate number  $\eta(s)$  being the number of occurrences of symptom  $s$  in sets  $S_1, S_2, \dots, S_N$ ;
2. Set  $S^* := \emptyset$ ;
3. For each  $s \in S$  do
  - Begin
  - If  $\eta(s) > N/2$  then set  $S^* := S^* \cup \{s\}$ ;
  - If  $\eta(s) = N/2$  then set  $S^* := S^* \cup \{s\}$  or do not change  $S$ ;
  - End;

END.

The above algorithm may in fact determine not only one, but a set of representations. Denote this set by symbol  $\mathcal{S}$ . A representation  $S^*$  may be interpreted as an association rule:



$$r^* = (s_1, 1) \wedge (s_2, 1) \wedge \dots \wedge (s_M, 1) \Rightarrow (d, 1)$$

where  $S^* = \{s_1, s_2, \dots, s_M\}$ .

In the next section we present some results of the analysis of the two algorithms.

## 4 Comparison of Algorithms

Algorithms described in Section 3 give different solutions, however, several cases for which it is possible to get a similar set of rules can be pointed out. Assume that the Apriori algorithm is run for table  $T(d)$  and as the result we obtain sets of itemsets  $L_1, L_2, \dots, L_K$ .

We now present some results of analysis of the two algorithms worked out in Section 3.

**Theorem 2.** *The following dependences are true:*

- If  $f_{\min} = N/2$  then set of representative of ordered covering is equal to 1-itemsets, that is  $L_1 = S$ .
- If  $f_{\min} < N/2$  then  $S \subset L_1$ .
- If  $f_{\min} > N/2$  then  $L_i \subset S$  for  $i = 1, 2, \dots, K$ .

**Theorem 3.** *Let's introduce value  $\alpha = \sum_{i=1}^N \delta(S, S_i)$ , where*

$$\delta(S, S_i) = \begin{cases} 0, & \text{for } S_i \subseteq S \\ 1, & \text{otherwise} \end{cases}$$

*Then set of representation is a frequent item set of support  $f_{\min}$  if  $\alpha > f_{\min}$ .*

Theorem 2 shows the relationships between sets  $S$  and  $L_i$  ( $i = 1, 2, \dots, K$ ) referring to value  $f_{\min}$ . From these relationships it follows certain coherence of the algorithms. Theorem 3 presents the situation when set  $S$  may become a set of itemsets. These properties show the coherence between the results given by the data mining method and the representation choice method. This in turn shows the usefulness of the second method. The proofs of these theorems are given in report [1]. In view of the small computation complexity of the second method we can state that some improvement has been done.

## 5 Conclusions and Further Work

In this paper two methods for discovering knowledge from medical data are presented. The Apriori algorithm and the method of representation choice are chosen to analyze the medical data collected over the period of many years. Description of algorithms and theorems are shown. Next, the comparison of these algorithms is analysed. It turned out that the sets of rules received by using these algorithms are different in many cases, but there exist several cases when both methods give similar solutions. The considerations are carried in this work for zero-one data. In future the use of algorithms for other discrete data, implementation and tests with real medical data are planned.

Proposed solutions should make data analysis and generation of complicated decision rules easier. It is possible to reduce the time of medical diagnosing and increase its precision. This way we could improve medical help for each patient. It is a huge advantage, which explains why problems of discovering knowledge from medical data are so important.

## References

1. Aftarczuk, K., Kozierekiewicz, A.: The method of supporting medical diagnosis based on consensus theory. Report of Institute of Information Science & Engineering, Wrocław University of Technology. Series PRE No. 1 (2006).
2. Bratko, I., Kononenko, I.: Learning diagnostic rules from incomplete and noisy data. In: Phelps B, editor. *AI Methods in Statistics*. London: Gower Technical Press (1987) 142-153
3. Child, Ch., Stathis, K.: The Apriori Stochastic Dependency Detection (ASDD) Algorithm for Learning Stochastic Logic Rules. In J. Dix, J. Leite, and P. Torroni (eds), *Proceedings of the 4th International Workshop on Computation* (2004) 201-216
4. Clark, P., Niblett, T.: The CN2 induction algorithm. *Mach Learn* 3(4) (1989) 261–83.
5. Coiera, E.: *The Guide to Health Informatics* (2nd Edition). Arnold, London (2003).
6. Daniłowicz, Cz., Nguyen, N. T.: *Methods of choice of representation of ordered covering*. Wrocław University of Technology Press, Wrocław (1992).
7. Duch, W., Adamczak, R., Grąbczewski, K., Jankowski, N., Żal, G.: Medical diagnosis support using neural and machine learning methods”. In *Proceedings of Conference EANN’98, Gibraltar* (1998) 292-295.
8. Kononenko, I., Kukar, M.: Machine learning for medical diagnosis. In: *Proceedings Workshop on Computer-Aided Data Analysis in Medicine*; Bled, Slovenia (1995) 9-31.
9. Kusiak, A., Kern, J. A., Kernstine, K. H., Tseng, B. T. L.: Autonomous Decision-Making: A Data Mining Approach. *IEEE Transactions on Information Technology in Biomedicine* 4(4) (2000) 274-284.
10. Lavrac, N.: Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine* 16 (1999) 3-23.
11. Lavrac, N., Dzeroski, S.: *Inductive Logic Programming: Techniques and Applications*. Chichester, Ellis Horwood (1994).
12. Lesmo, L., Saitta, L., Torasso, P.: Learning of fuzzy production rules for medical diagnosis. In: Gupta M, Sanchez E (eds), *Approximate Reasoning in Decision Analysis*. Amsterdam: North-Holland (1982) 249-260.
13. Michie, D., Spiegelhalter, D. J., Taylor, C. C.: *Machine learning; neural and statistical classification*. Chichester: Ellis Horwood (1994) 50-83
14. Muggleton, S.: *Inductive Acquisition of Expert Knowledge*. Addison-Wesley (1990).
15. Nunez, M.: Decision tree induction using domain knowledge. In: Wielinga B (Ed), *Current Trends in Knowledge Acquisition*. Amsterdam, IOS Press (1990).
16. Quinlan, J.: Induction of decision trees. *Mach Learn* 1 (1986) 81–106.
17. Rumelhart, D., McClelland J.: *Parallel Distributed Processing, vol. 1: Foundations*. Cambridge, MA: MIT Press (1986),
18. [http://www.openclinical.org/aisp\\_txdent.html](http://www.openclinical.org/aisp_txdent.html)
19. [http://www.openclinical.org/aisp\\_dxplain.html](http://www.openclinical.org/aisp_dxplain.html)
20. [http://www.openclinical.org/aisp\\_gideon.html](http://www.openclinical.org/aisp_gideon.html)
21. [http://www.openclinical.org/aisp\\_help.html](http://www.openclinical.org/aisp_help.html)
22. <http://www.isabelhealthcare.com>
23. [http://www.openclinical.org/aisp\\_isabel.html](http://www.openclinical.org/aisp_isabel.html)

# Construction of School Temperature Measurement System with Sensor Network

Ayahiko Niimi, Masaaki Wada, Kei Ito, and Osamu Konishi

Department of Media Architecture, Future University-Hakodate  
116-2 Kamedanakano-cho, Hakodate 041-8655, Japan  
{niimi, wada, kei, okonishi}@fun.ac.jp

**Abstract.** We propose the sensor network system using the microcomputer board that can connect to the Internet. This proposed system can acquire information from the sensor of the microcomputer group arranged on the network, and can view collected information on Web browser. In this paper, it is shown to be able to construct easily the microcomputer's sensor network which is combined microcomputer modules (Micro Cube) and the database server and the Web application server. The system that measured the room temperature in school campus was constructed, it has run for four months, and the effectiveness is verified.

## 1 Introduction

The digital measurements of the temperature and humidity, etc. become possible, and connecting the system that acquired the measured data on the network becomes possible. However, there are many measurement systems which are rich systems that are used sensors on PC or which are cheap microcomputer systems that need to construct a special network for the sensor network. We propose the sensor network system using the microcomputer board that can connect to the Internet. This proposed system can acquire information from the sensor of the microcomputer group arranged on the network, and can view collected information on Web browser.

In this paper, it is shown to be able to construct easily the microcomputer's sensor network which is combined microcomputer modules (Micro Cube) and the database server and the Web application server. The system that measured the room temperature in school campus was constructed, it has run for four months, and the effectiveness is verified.

The research of ecopic [1,2] is similar to this research. The ecopic research has aimed the construction of ecopic of the weather observing system that can make it easily by user. Our proposed system has aim of improving the extendibility by using Micro Cube that is generality module.

Chapter 2 describes the sensor module using Micro Cube. Chapter 3 describes the composition of the sensor network. We describe the installation of Micro Cube, the server and the client and the technique of collecting and viewing data. In chapter 4, we describe construction of our proposed system, and discuss about the problem when constructing sensor network. Section 5 describes conclusion and enhancing in a future.

## 2 Sensor Module

In this section, we describe proposed sensor module.

### 2.1 Outline of Micro Cube

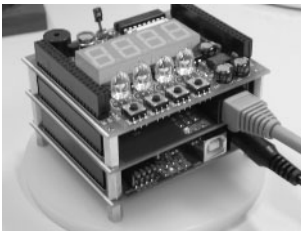
The Micro Cube is a board computer and is composed of several stackable boards [3,4]. Fig. 1 is a photo showing one of the combinations of stacked Micro Cube. The specifications of the CPU and extension boards are summarized in Table 1. It has a CPU board with a RENESAS H8 CPU and a TCP/IP Protocol stack. Stackable boards can vary as follows: Ethernet LAN board, compact flash board, PCMCIA board, serial board (RS232C and RS422) and so on (some boards shown in Fig. 2-3). Since the different combinations of stackable boards make a seamless connection with the sensors, users can structure an ad hoc sensor network very easily. To get sensor information through the Internet, HTTP is also employed so that user can get data via a standard Web browser.

### 2.2 Instrumentation of the Present System

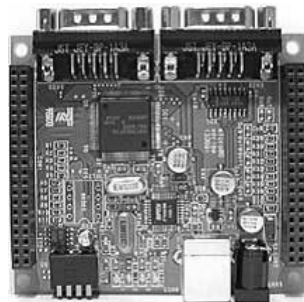
The Micro Cube used in the system to get the information of room condition is composed of the H8/3069 CPU board, LAN board, and special sensor board.

**Table 1.** Lineup of the CPU and extension boards of Micro Cube

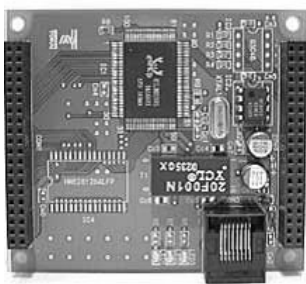
Board Name	Features
H8/3048BV	CPU board for battery operation
H8/3069	Same for general purpose (see Fig. 2)
H8s/2638	Same for Controller Area Network (CAN)
LAN	Extension board for Ethernet Connection (see Fig. 3)
CF	Same for Compact Flash slot
IDE	Same for storage devices
ADIO	Same for analog/digital IO
COM4	Same for 4-port serial interface
RF	Same for wireless communication
PCMCIA	Same for PCMCIA slot



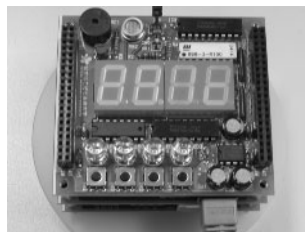
**Fig. 1.** Photo of a Stacked Micro Cube



**Fig. 2.** CPU board



**Fig. 3.** LAN board



**Fig. 4.** Sensor board

The special sensor board is utilized the board used of the programming practice class in Future University-Hakodate. (The sensor board is shown in Fig. 4) Future University-Hakodate has the programming practice class with the microcomputer and assembler language as “Media Architecture Practice II”. The special board for Micro Cube was designed for its practice class. The push switch, the thermally sensitive resistor (temperature sensor), and CdS sensor (optical sensor) were attached on this board as an input. Moreover, four digits seven-segments LED and four two-color LED were attached as an output. Because an accurate temperature measurement using the thermally sensitive resistor is difficult, a digital sensor is added in this board for our experiment. Humidity can be also measured in this digital sensor. Only the temperature data is acquired this experiment though some sensors are attached on the board. The exchange and the addition of the sensor can be easily done by exchanging the sensor boards.

To confirm the measurement data easily, the measured temperature was displayed in seven-segments LED. Moreover, data can be got by HTTP though the network. When only one sensor module runs, the user can display a present temperature when the user accesses it using Web browser.

### 3 Network Configuration

The sensor network was constructed by using the microcomputer that explained in Chapter 2. Fig. 5 shows the composition of the constructed sensor network system.

The used software is shown below (see table 2). The data store part is implemented by Perl, and the data display part is implemented by JSP.

The flow of the collection of data and the display stored data is shown in Fig. 6. Fig. 6 shows the following steps.

1. Data storage
  - (a) The Perl script accesses to URL of Micro Cube.
  - (b) Micro Cube returns the measurement result by HTML format.
  - (c) HTML is parsed, and necessary data is preserved in the database.

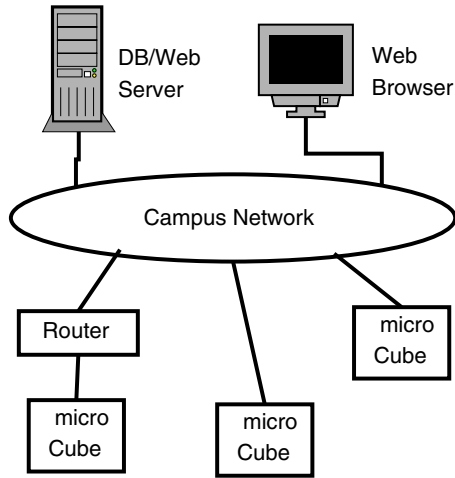


Fig. 5. Network Configuration

Table 2. Software used in the web database server

System	Software
OS	Red Hat Linux release 9
HTTP	Apache 2.0.40
Database	PostgreSQL 7.3.2
Software codes	Tomcat 5.0.28 and Perl 5.8.0

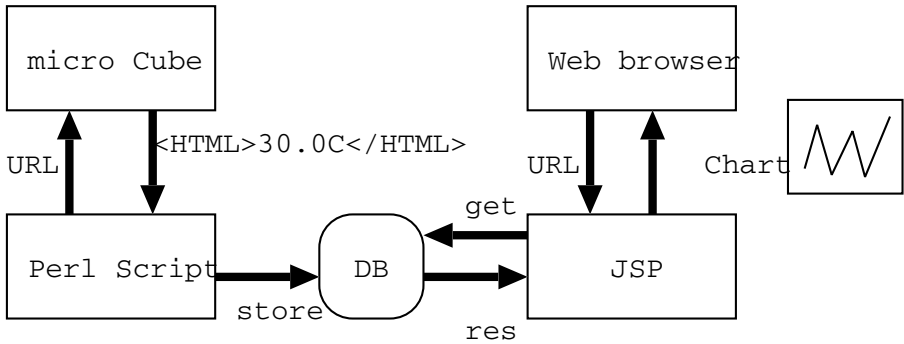


Fig. 6. Flowchart of Proposed System

2. Data browse

- (a) URL of the server is opened from a Web browser.
- (b) JSP accesses the database.
- (c) Necessary data is acquired from the database.
- (d) The result is processed to the graph and displayed it on a browser.

The micro cube arranged in school is connected with campus network (LAN). The data of each sensor module is acquired with the server set up on the campus network at regular intervals, and stores in the database. In this experiment, data is acquired from the sensor module every ten minutes. The acquired data is processed with the Web application server set up on the same server, and can be displayed from Web browser of PC on the campus network.

At first, Micro Cube connected to campus network by arranging it in the router because the router had not been exceeded in LAN of the micro cube. Afterwards, connecting Micro Cube to the campus network even if we modified the program, and the router is not set up became possible so that the router was exceeded.

## 4 Experimental Results

The system that explained in Chapter 3 was actually constructed. The system is constructed in December, 2005, and it is running at March, 2006. Because a lot of modules were able to be reused, the time that had constructed to development was about one week.

Some data display examples are shown as follows. (See Fig. 7, 8) The displayed data can be switched to all or a part of room. The displayed range can be switched to a day, a week, a month, or all. Fig. 7 shows all data in one chart, and Fig. 8 shows one data in one chart.

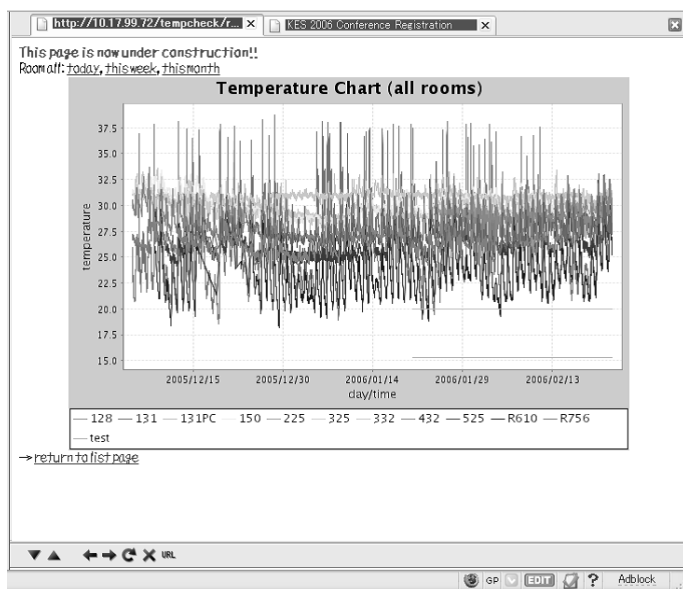
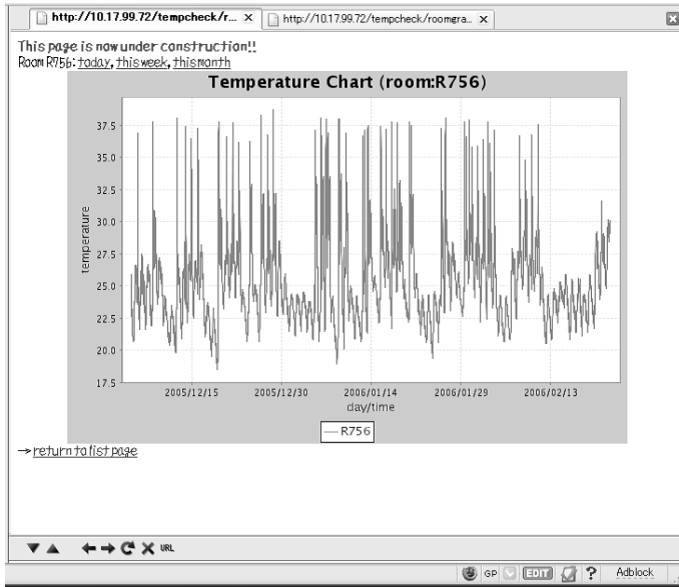


Fig. 7. All Data in One Chart



**Fig. 8.** Selected Data in One Chart

At first, Micro Cube connected to campus network by arranging it in the router because the router had not been exceeded in LAN of the micro cube. However, data might not be able to be acquired normally when some router's passing. Then, connecting Micro Cube to the campus network even if we modified the program, and the router is not set up became possible so that the router was exceeded.

It took time to divide the problem when data was not able to be acquired normally because Micro Cube connected to the campus network. Then, the problem was considered in cooperation with SE that resided in school. For the exceeding router problem, we constructed some environment, such as the dummy server with the PC-UNIX server, passing the router setting, bypassing the router setting, and it undertook the resolution of a problem while capturing the packet that flowed in the network.

Because data began to collect, temperature data is scheduled to be analyzed in the future.

## 5 Conclusion

We proposed the sensor network system using the microcomputer board that can connect to the Internet. This proposed system can acquire information from the sensor of the microcomputer group arranged on the network, and can view collected information on Web browser. In this paper, it was shown to be able to construct easily the microcomputer's sensor network which is combined microcomputer modules (Micro Cube) and the database server and the Web



application server. The system that measured the room temperature in school campus was constructed, it has run for four months, and the effectiveness was verified. Because data began to collect, temperature data is scheduled to be analyzed in the future. In the analysis of data, it is thought that it is possible to refer to a technique of the multiagent base [5,6] and an analytical technique of the analysis of the fixed point observation data [7,8].

The system that expands acquired more different type of sensor information will be constructed, and then, obtained data is scheduled to be analyzed in the future. In addition, we want to attach the IR I/O module on the microcomputer board, and to do the research for the ubiquitous computing of the indoor environment controlling.

## References

1. Toda, M., Akita, J., Kimura, K.: Construction of Ubiquitous Weather Observation System. Proc. of The 7th World Multiconference on Systemics, Cybernetics and Informatics (SCI2003), Vol.1 pp.192–197 (2003)
2. Akita, J., Toda, M., Kimura, K.: Handy Meteorological Observation System for Hands-on Study. Proc. of The 8th World Multiconference on Systemics, Cybernetics and Informatics (SCI2004), Vol.X pp.155–159 (2004)
3. Wada, M., Hatanaka, K., Kimura, N., Amagai, K.: Applying an Information Technology in Fishing Industry I. Acquisition of a Three Dimensional Topography. Journal of Japan Institute of Navigation, vol. 112, pp.189–198 (2004) (in Japanese)
4. Hatanaka, K., Wada, M., Kotaki, M.: Instrumentation for the Measurement of Shallow Seabed Topography by a Fishing Echo Sounder. Oceans 2005 MTS/IEEE Conference Proceedings, Ocean Instrumentation I, pp.1–6 (2005)
5. Niimi, A., Konishi, O.: Data Mining for Distributed Databases with Multiagents. KES'2003, Proceedings, Part II, Lecture Notes in Artificial Intelligence 2774, Springer, pp.1412–1418 (2003)
6. Niimi, A., Konishi, O.: Extension of Multiagent Data Mining for Distributed Databases. KES'2004, Proceedings, Part III, Lecture Notes in Artificial Intelligence 3215, Springer, pp.780–787 (2004)
7. Teiten2000 Project, <http://www.teiten2000.jp/> (in Japanese)
8. Toda, M., Osanai, A., Niimi, A., Akita, J., Kimura, K., Konishi, O.: Information Access Method to Meteorological Data for Educational Application. World Conference on Education Multimedia, Hypermedia and Telecommunications (ED-MEDIA2005), Association for the Advancement of Computing in Education, Montre'al, Canada, Proceedings, pp.4404–4408 (2005)

# Land Cover Classification from MODIS Satellite Data Using Probabilistically Optimal Ensemble of Artificial Neural Networks

Kenneth J. Mackin<sup>1</sup>, Eiji Nunohiro<sup>1</sup>, Masanori Ohshiro<sup>2</sup>, and Kazuko Yamasaki<sup>2</sup>

<sup>1</sup> Department of Information Systems, Tokyo University of Information Sciences  
1200-2 Yatoh-cho, Wakaba-ku, Chiba, Japan  
{mackin, nunohiro}@rsch.tuis.ac.jp

<sup>2</sup> Department of Environmental Information, Tokyo University of Information Sciences  
1200-2 Yatoh-cho, Wakaba-ku, Chiba, Japan  
{ohshiro, yamasaki}@rsch.tuis.ac.jp

**Abstract.** Terra and Aqua, 2 satellites launched by the NASA-centered international Earth Observing System project, house MODIS (Moderate Resolution Imaging Spectroradiometer) sensors. Moderate resolution remote sensing allows the quantifying of land surface type and extent, which can be used to monitor changes in land cover and land use for extended periods of time. In this paper, we propose applying a probabilistically optimal ensemble technique, based on fault masking among individual classifier for N-version programming. We create an optimal ensemble of artificial neural networks and use the majority voting result to predict land surface cover from MODIS data. We show that an optimal ensemble of neural networks greatly improves the classification error rate of land cover type.

## 1 Introduction

With the increased interest in monitoring the global ecological changes, the demand for satellite remote sensing has increased. NASA-centered international Earth Observing System project has launched many satellites to monitor the earth for scientific purposes. Two such satellites, Terra and Aqua house MODIS (Moderate Resolution Imaging Spectroradiometer) sensors. Moderate resolution remote sensing allows the quantifying of land surface type and extent, which can be used to monitor changes in land cover and land use for extended periods of time.

Past research in intelligent classification of remote sensor data using neural networks[1][2][3] has given promising results. Kushardono et al. [3] compared different neural network structures in order to use several different bandwidth data as input for a target classification cell.

For this research, we investigate the effect of applying probabilistically optimal ensemble of neural networks for the classification of land cover type from MODIS satellite remote sensing data. Neural networks, as with other training based classifiers, inherently have a risk that when classifying an untrained data set, the classifying error rate may be much worse than the training result. In order to overcome this risk, we applied a probabilistically optimal ensemble technique proposed by Imamura et al. [4]

to N-version programming of neural networks. Our purpose is to research the validity of using a training based classifier method for land cover type, compared against the results of previous statistical methods. We applied the proposed method to MODIS data for classifying land cover for Chiba prefecture, Japan to test the validity of the method.

## 2 MODIS Data

With the increased interest in monitoring the global ecological changes, the demand for satellite remote sensing has increased. NASA-centered international Earth Observing System project has launched many satellites to monitor the earth for scientific purposes, including Terra and Aqua. A key instrument aboard the Terra and Aqua satellites is MODIS (Moderate Resolution Imaging Spectroradiometer). Terra's orbit around the Earth is timed so that it passes from north to south across the equator in the morning, while Aqua passes south to north over the equator in the afternoon. Terra MODIS and Aqua MODIS enable the viewing of the entire Earth's surface every 1 to 2 days. MODIS captures data in 36 spectral bands, or groups of wavelengths. Moderate resolution remote sensing allows the quantifying of land surface type and extent, which can be used to monitor changes in land cover and land use for extended periods of time. This data is used to monitor and understand global dynamics and processes occurring on the land, oceans, and lower atmosphere.



**Fig. 1.** Terra and Aqua satellites (c)NASA

Monitoring of land cover and land use is an important element of global monitoring. Moderate resolution remote sensing enables the quantifying of land surface characteristics such as land cover type and extent, snow cover extent, surface temperature, leaf area index, and fire occurrence. Satellite measurements of leaf area, leaf duration and net primary productivity provide important inputs to capture or model ecosystem processes. High quality, consistent and well-calibrated satellite measurements are needed to detect and monitor changes and trends in these variables.

For this paper, we used MODIS data collected at Tokyo University of Information Sciences, Japan. Tokyo University of Information Sciences receives satellite MODIS data over eastern Asia, and provides this data for open research use, as part of the research output of the Japanese government funded Frontier project.

Table1 describes the 36 spectral bands of MODIS sensor data.

**Table 1.** MODIS sensor band specifications

Primary Use	Band	Bandwidth	Spectral Radiance	Spatial Resolution
Land/Cloud/Aerosols Boundaries	1	620 - 670 nm	21.8	250 m
	2	841 - 876 nm	24.7	250 m
Land/Cloud/Aerosols Properties	3	459 - 479 nm	35.3	500 m
	4	545 - 565 nm	29	500 m
	5	1230 - 1250 nm	5.4	500 m
	6	1628 - 1652 nm	7.3	500 m
	7	2105 - 2155 nm	1	500 m
Ocean Color/ Phytoplankton/ Biogeochemistry	8	405 - 420 nm	44.9	1000 m
	9	438 - 448 nm	41.9	1000 m
	10	483 - 493 nm	32.1	1000 m
	11	526 - 536 nm	27.9	1000 m
	12	546 - 556 nm	21	1000 m
	13	662 - 672 nm	9.5	1000 m
	14	673 - 683 nm	8.7	1000 m
	15	743 - 753 nm	10.2	1000 m
Atmospheric Water Vapor	16	862 - 877 nm	6.2	1000 m
	17	890 - 920 nm	10	1000 m
	18	931 - 941 nm	3.6	1000 m
Surface/Cloud Temperature	19	915 - 965 nm	15	1000 m
	20	3.660 - 3.840 $\mu\text{m}$	0.45(300K)	1000 m
	21	3.929 - 3.989 $\mu\text{m}$	2.38(335K)	1000 m
	22	3.929 - 3.989 $\mu\text{m}$	0.67(300K)	1000 m
Atmospheric Temperature	23	4.020 - 4.080 $\mu\text{m}$	0.79(300K)	1000 m
	24	4.433 - 4.498 $\mu\text{m}$	0.17(250K)	1000 m
Cirrus Clouds Water Vapor	25	4.482 - 4.549 $\mu\text{m}$	0.59(275K)	1000 m
	26	1.360 - 1.390 $\mu\text{m}$	6	1000 m
	27	6.535 - 6.895 $\mu\text{m}$	1.16(240K)	1000 m
Cloud Properties	28	7.175 - 7.475 $\mu\text{m}$	2.18(250K)	1000 m
	29	8.400 - 8.700 $\mu\text{m}$	9.58(300K)	1000 m
Ozone	30	9.580 - 9.880 $\mu\text{m}$	3.69(250K)	1000 m
Surface/Cloud Temperature	31	10.780 - 11.280 $\mu\text{m}$	9.55(300K)	1000 m
	32	11.770 - 12.270 $\mu\text{m}$	8.94(300K)	1000 m
Cloud Top Altitude	33	13.185 - 13.485 $\mu\text{m}$	4.52(260K)	1000 m
	34	13.485 - 13.785 $\mu\text{m}$	3.76(250K)	1000 m
	35	13.785 - 14.085 $\mu\text{m}$	3.11(240K)	1000 m
	36	14.085 - 14.385 $\mu\text{m}$	2.08(220K)	1000 m

### 3 Land Cover Classification with Neural Networks

Artificial neural networks (ANN) can be characterized by its "black box" approach to learn and classify complex data patterns. For this research, we propose applying 3 layer network structure (1 input layer, 1 hidden layer, 1 output layer) for the training of land cover type classification, using the neural network to learn the complex relationship between MODIS sensor data. For this paper, we will propose methods applying N-version programming of neural networks to construct the land cover classifier.

First we describe the basic artificial neural network for land cover classification. We considered the 3 layer artificial neural network (1 input layer, 1 hidden layer, 1 output layer) as the basic training classifier. We use a sigmoid function for the synapse function of the neuron, with back propagation (BP) training of the MODIS sensor data. The number of hidden neurons was decided by results of preliminary experiments of the neural network.

For the network training we used the database of collected MODIS sensor data, and applied BP training based on the difference between classified land cover and land-truth data provided by the Japanese Ministry of the Environment.

As for the neural network input, we considered the possibility that the large number of input nodes increases the problem domain and complicates the classification, causing an adverse affect on the network training efficiency. With this assumption, we decided to minimize the number of input nodes in order to first achieve a workable learning curve and classification accuracy. It has been previously shown that bands 1 and 2 (visible red and infra-red) can be used to classify land cover type, and bands 1 and 2 also have the best spatial resolution (250m) among the MODIS bands. For these reasons, for this research we use only bands 1 and 2 as input to the classifiers. For output classes, we use the same 5 major classifications used by Kushardono et al.[3]. The 5 major classifications are paddy, trees, urban, water, and other.

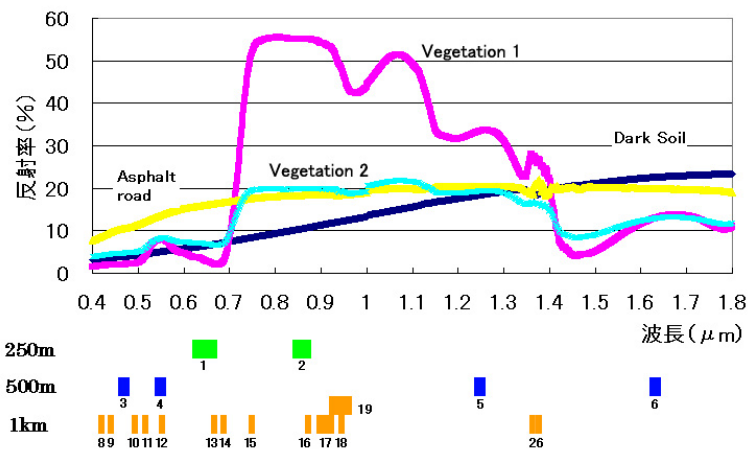


Fig. 2. Example spectral readings for different land cover types

## 4 Training of Single Neural Network

We trained the above described neural network using BP and evaluated the classification accuracy. A standard sigmoid function was used as the neuron's base synapse function. The number of neurons used in each layer was 3 input neurons (band 1 input, band 2 input, and 1 fixed input), 6 hidden layer neurons, and 1 output neuron.

For the training data, 4 MODIS data images of the same location (Chiba prefecture), 1 image for each season of spring, summer, autumn, and winter, were used. For the untrained data used to plot the training curve of network accuracy, similarly, 4 MODIS data images of the same location (Chiba prefecture), 1 image for each season of spring, summer, autumn, and winter, were used.

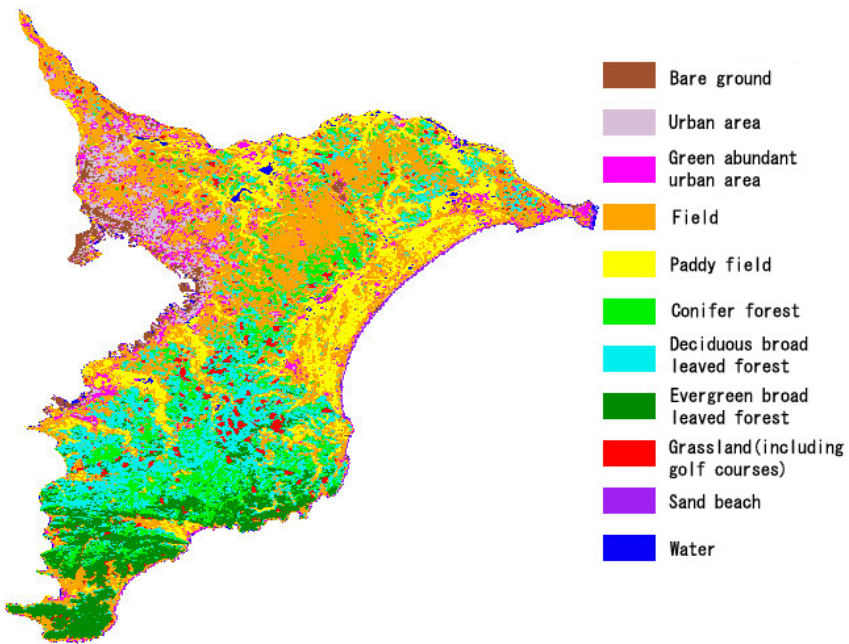


Fig. 3. Land cover data for Chiba prefecture, Japan

## 5 Probabilistically Optimal Ensemble of Neural Networks

Neural networks, as with other training based classifiers, inherently have a risk that when classifying an untrained data set, the classifying error rate may be much larger than the training result. In order to overcome this risk, we applied a probabilistically optimal ensemble technique proposed by Imamura et al. [4] to N-version programming of neural networks.

Fault masking in N-version programming assumes that the individual members give completely independent results. If certain members give similar output, then correct fault masking will not occur. In a probabilistically optimal ensemble, the members of the ensemble are chosen so that the members are correctly independent of each other. This is realized by selecting members so that the measures error rate of the ensemble comes closest to the expected error rate of the ensemble. If the members are correctly independent of each other, then proper fault masking should allow the measured error rate to become very close to the expected error rate.

The expected failure rate  $f$  of the probabilistically optimal ensemble can be calculated by the following equation [4]

$$f = \sum_{k=m}^n \binom{n}{k} (1-p)^{n-k} p^k \tag{1}$$

where  $p$  is the failure rate of each individual,  $n$  is the size of the ensemble,  $m$  is the minimum number of faulty outputs for an ensemble to fail. We assume the same failure rate  $p$  for individuals for simplicity.

In the case where the ensemble has 3 members, majority vote (2 votes) for output, and  $p = 0.19$ , then  $f = 0.086$ .

For our research we trained 6 artificial neural networks using different initial weights, and the same training set. Using the same training set, we compared the measured ensemble failure rate for ensemble size 3, for all combinations of neural networks. Using individual failure rate  $p = 0.19$ , the expected ensemble error rate was  $f = 0.086$ .

Table 2 shows the results of the measured ensemble failure rate. The ensemble with the closest ensemble failure rate ( 0.085 ) was selected as the optimal ensemble.

The selected optimal ensemble was evaluated using untrained data. The resulting ensemble error rate was  $f = 0.087$ , indeed very close the training ensemble error rate, and vastly improved over the error rate 0.19 for the single neural network.

**Table 2.** Measured ensemble failure rate

ANN no.	ensemble failure rate	ANN no.	ensemble failure rate
0,1,2	0.11	1,2,3	0.13
0,1,3	0.115	1,2,4	0.135
0,1,4	0.095	1,2,5	0.115
0,1,5	0.085	1,3,4	0.14
0,2,3	0.115	1,3,5	0.13
0,2,4	0.12	1,4,5	0.115
0,2,5	0.1	2,3,4	0.145
0,3,4	0.13	2,3,5	0.135
0,3,5	0.12	2,4,5	0.125

## 6 Conclusion

In this research we applied a probabilistically optimal ensemble technique [4] to create an N-version programming classifier system using 3-layer artificial neural networks to classify land cover from MODIS satellite data. We were able to confirm that by using an optimal ensemble of neural networks, the classification error rate can be greatly reduced.

For future works, we will consider methods to improve classification accuracy of the individual neural network, including the increase in the types of sensor input data, reevaluation of neural network structure, combining fuzzy rules to treat input data, as well as effect of using different base synapse functions for neurons.

## References

1. Y. Shkvarko, J. Montiel, L. Rizo, J. Salas, "Neural Network-Based Signal Processing for Enhancing the Multi-Sensor Remote Sensing Imagery", 14th International Conference on Electronics, Communications and Computers, pp. 168-172, 2004
2. F. Roli, S.B. Serpico, L. Bruzzone, "Classification of Multisensor Remote-Sensing Images by Multiple Structured Neural Networks", 13th International Conference on Pattern Recognition (ICPR'96), Volume 4, pp.180-184, 1996
3. D. Kushardono, K. Fukue, H. Shimoda, T. Sakata, "A Study on Neural Network Landcover Classification Models with the aid of Co-occurrence Matrix for Multiband Images", Journal of The Remote Sensing Society of Japan, Vol.16, No.1, pp.36-49, 1996 (in Japanese)
4. Kosuke Imamura, Kris Smith, "A Probabilistically Optimal Ensemble Technique for Training Based Classifiers", Proceedings of Joint 2nd International Conference on Soft Computing and Intelligent Systems and 5th International Symposium on Advanced Intelligent Systems, Japan, 2004



# Structure-Based Categorization of Programs to Enable Awareness About Programming Skills

Kei Kato and Toyohide Watanabe

Department of Systems and Social Informatics,  
Graduate School of Information Science, Nagoya University  
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan  
{kkato, watanabe}@watanabe.ss.is.nagoya-u.ac.jp

**Abstract.** Programming is a kind of design activity with various skills which include not only knowledge about programming languages, etc. but also abilities for problem-analysis/solving. Expert programmers can acquire new skills easily interrelated to already-improved experiences and also adapt the skills to solve new problems. However, it is difficult for programming learners to grow up their own usable skills successively. In this paper, we focus on the relations among programming skills. In order to make learners be aware of the relations among their own skills, structural similarities among learners' programs are found out, and then these programs are categorized into their structural features based on the similarities. In our approach, the categories are expressed by a tree structure, and the similarities among programs are managed on the basis of the categories generated from learners' input programs.

## 1 Introduction

A program is a set of instructions to make computers perform a certain solution of a problem. In order to code a program, a programmer stepwise understands the property of a problem and solves it through a trial and error process. Therefore, programming is considered to be a design activity to code programs with various skills about not only how to analyze/solve problems but also how to transform solutions of the problems into programs. Of course, these skills include programming knowledge about languages, data structures, algorithms, etc. In order to code programs correctly and efficiently, programmers have to acquire and handle these high-level skills.

The objective in this research is to support programming learners who improve their programming skills. A research on cognitive theories about programming learning has shown evidences that well-experienced programmers store and retrieve old experiences on problem-solving which can be applied to a new problem and adapted to solve the problem[1]. In other words, expert programmers reproduce various relations among their own skills based on their already-experienced knowledge, and can use these relations effectively even in coding programs newly. However, programming learners cannot grow up their own skills unlike experts. This is because the primary programming learners cannot find out the relations

between knowledge heuristically or effectively or cannot infer the conceptual meanings among knowledge.

We consider, therefore, that relations (e.g., similarity) among learner's programs are important factors for the learner to grow up his/her skills. In our approach, learner's programs are positioned into other programs and are grouped with programs which have the same feature. That is, the learner's programs are categorized based on the feature. Our categorization process is first to find out the particular features from learner's programs, second to categorize the programs based on the features and finally to present the categorization result to the learner.

## 2 Framework

### 2.1 Knowledge Handling in Programming

In order to support learners who grow up their skills successively, it is important to discuss how experts handle and interrelate their own skills continuously. Knowledge handling model is one of knowledge management models, which explains personal activities in learning and designs learning support functions in the personal activities[2,3]. This model consists of three distinguished knowledge states: inside of in-world, outside of in-world and out-world, and four individual knowledge manipulation functions defined in accordance with the states: acquisition, refinement, reproduction and presentation. Experts can consciously/unconsciously interrelate their own skills based on their experiences with these functions.

In the relations which experts reproduce among their skills, the concept of similarity plays an important role in applying their skills to new problems and inferring the conceptual meanings among their skills by analogy. In above-mentioned functions, the categorization is defined as one kind of refinement functions and is deeply related to the similarity. This is because the categorization is to put things into a particular group according to their peculiar features: as a result, skills which belong to a particular category have the same/similar features.

### 2.2 Programming Skills Reflected in Programs

One's programs are considered to be his/her own explicit knowledge. That is, one's programming skills (e.g., skills about how to design control structures) adapted to code programs are reflected in the programs.

If we point out the transition between the knowledge states in the knowledge handling model, the skills belong to inside of in-world, while the programs belong to outside of in-world as explicit products derived from the skills(Fig. 1(a)). Broken arrows between skills and programs in Fig. 1(a) indicate correspondences between one's programs and skills used to code the programs.

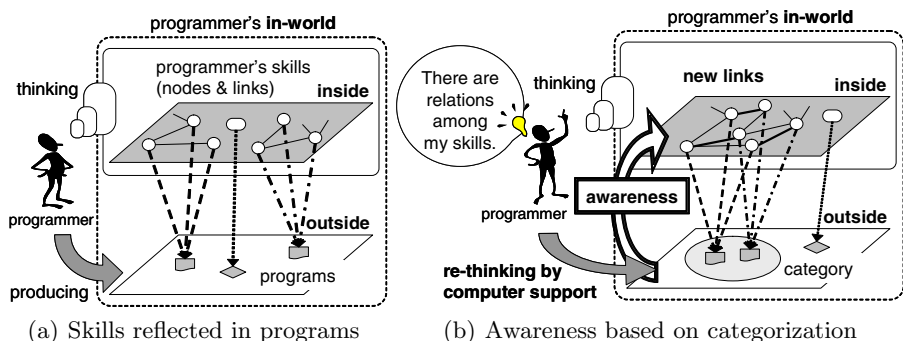


Fig. 1. Conceptual framework

### 2.3 Approach

Based on the previous discussions, we find out the similarities among learner's programs and then categorize the programs based on the similarities. Our basic idea is to enable the awareness about programming skills with the similarities among learner's own skills by presenting the similarity-based categorization result to the learner. If the learner recognizes the correspondences between programs and skills, the categorization based on the program similarity may make the learner be indirectly aware of the similarities among his/her own skills (Fig. 1(b)).

## 3 Categorization Strategy

### 3.1 Categorization Viewpoint

If the similarities are found out based on the well-extracted features of programs, it is possible to suggest the similarities among the learner's own skills. Therefore, it is important to discuss which features our categorization is based on.

In this paper, the control structure is looked upon as a categorization viewpoint, and learner's programs are categorized based on the similarity among structural features of programs. This is because the control structures are very fundamental and important factors of programs. Moreover, in procedural programming, programs are designed as processes to solve problems. Therefore, the learner's skills to solve problems are reflected in the control structures.

### 3.2 Categorization Tree

In our categorization method, the structural similarities among programs are managed by a tree structure, which we call a categorization tree (Fig. 2). The tree structure reflects a hierarchical structure among concepts and has high affinity with human thinking processes. Therefore, the tree structure helps the learner to grasp the relations among his/her programs intuitively.

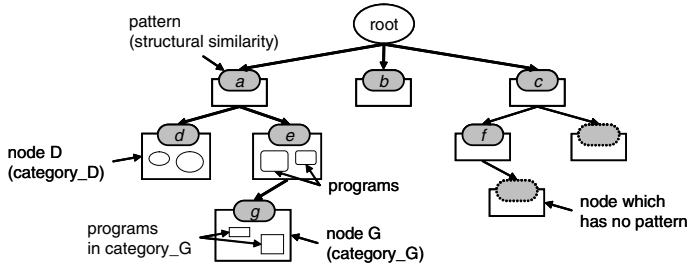


Fig. 2. Categorization tree

Our tree structure represents the constructive relations of already-positioned programs, and the node in the tree means the particular position of programs. Grouping programs is represented by positioning two or more programs to the same node, and the structural similarity among these programs is defined as the feature of the node. Of course, a node does not have such a feature when only one program is positioned to the node. In the rest of this paper, the feature of a node is called a pattern.

In this tree, a category is expressed as a node in the tree and is characterized by patterns of sequenced nodes on the path from the root to the node. In Fig. 2, category\_G is characterized by patterns “*aeg*” and has programs positioned only to the node G. Category\_D and category\_G have a pattern “*a*” in common and are respectively distinguished by a pattern “*d*” and patterns “*eg*”.

## 4 Categorization Process

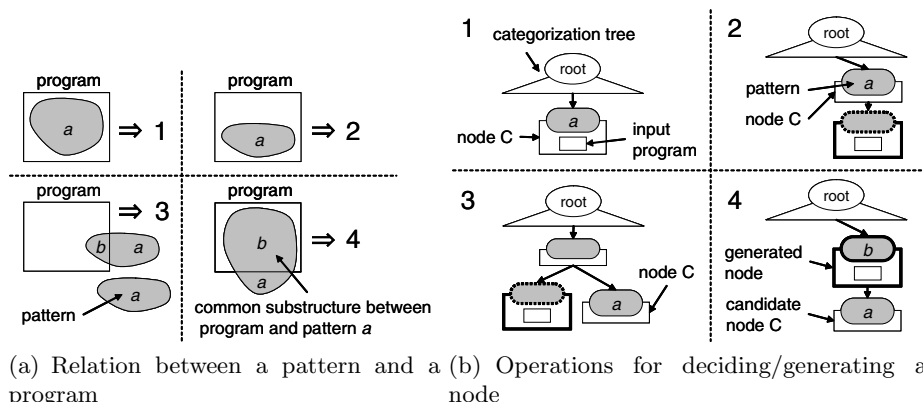
### 4.1 Representation of Programs

In our method, the structural features of programs are represented based on PAD[4], represented as nodes and links. In structured programming paradigm, the control structures are designed by combining three types of structures: sequence, selection, and iteration. In addition to these structures, hierarchical structures of programs should be considered to categorize programs based on the similarity among the structural features. PAD represents these characteristics of the control structures successfully.

### 4.2 Categorization Process

Our categorization process is as follows:

- step1** Transform an input program into PAD.
- step2** Put children of the root to a list of nodes *N*.
- step3** if *N* = null then generate a new node as a child. call adjustment\_2.  
return categorization tree.



**Fig. 3.** Rules for positioning programs to nodes

**step4** Select a node which has the largest common substructure between the pattern of the node and the program.

Here, **if** a node does not have a pattern **then** take a common substructure between a program positioned to the node and the input program as a temporary pattern.

**step5** Put the node selected in **step4** to a candidate node C.

**step6** Decide a relation between the pattern of the node C and the program, and then execute one of the following operations according to the relation.

These relations and operations are described in detail in the next subsection.

**Operation1:** Put the program to the node C. **call** adjustment\_1. **return** categorization tree.

**Operation2:** Put children of the node C to N. **goto** step3.

**Operation3:** Generate a new node as a sibling of the node C. **call** adjustment\_2. **return** categorization tree.

**Operation4:** Generate a new node as a parent of the node C. **call** adjustment\_2. **return** categorization tree.

Procedures adjustment\_1 and adjustment\_2 are called in the process to position the programs more adequately. If programs positioned to the candidate node C have a larger sub-structure than the pattern of the node C in common, adjustment\_1 generates a child of the node C newly and repositions these programs to the child. Adjustment\_2 is applied to shift programs to an already-generated child when the programs include the pattern of the child.

### 4.3 Rules for Positioning Programs

Fig. 3(a) and Fig. 3(b) conceptually represent the relations and the operations in **step6** of our categorization process.

The relations shown in Fig. 3(a) are decided according to how large a structure which a program and a pattern have in common is. Four relations are defined as follows:

**Relation1:** A pattern “*a*” is included in a program and accounts for  $\alpha\%$  or more ratio of the program.

**Relation2:** A pattern “*a*” is included in a program and accounts for  $\alpha\%$  or less of the program.

**Relation3:** A pattern “*a*” is not included in a program, and a common sub-structure “*b*” accounts for  $\alpha\%$  or less of the pattern “*a*” and the program.

**Relation4:** A pattern “*a*” is not included in a program, and a common sub-structure “*b*” accounts for  $\alpha\%$  or more ratio of the pattern “*a*” and the program.

Here,  $\alpha$  is a threshold and is set by the learner before the categorization process is executed.

In Fig. 3(b), the node generated in the operation 4 has pattern “*b*”. In contrast, patterns in the operations 2 and 3 are represented by broken lines, and this means these nodes have no patterns. This is because these nodes have only one program and therefore it is not clear what a feature these nodes have at this point.

## 5 Prototype System

### 5.1 Functions

A main function which the system provides is to construct a categorization tree according to learner’s input programs. The categorization result is represented in the main interface (Fig. 4(a)) and includes not only the constructed tree but also information about individual nodes.

In addition, several functions are provided to support learner’s activities more effectually: to help the learner to reconfirm correspondences between his/her programs and skills. These functions help the learner with more information

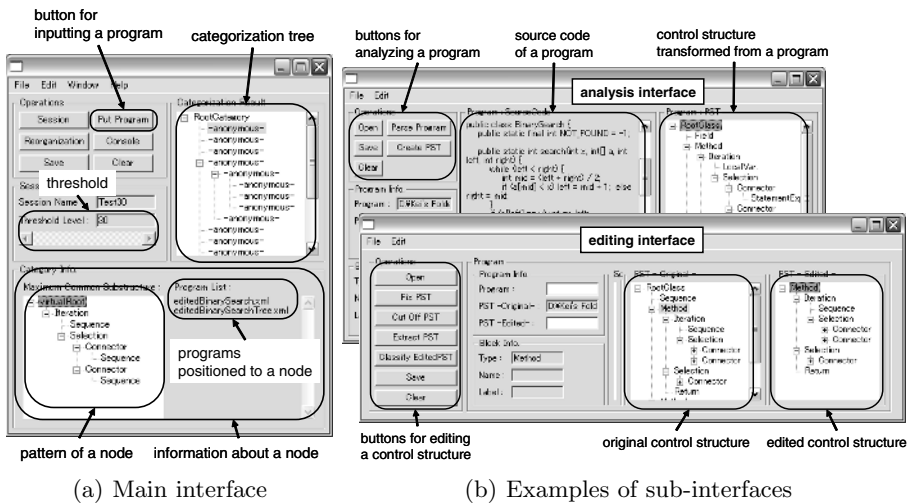


Fig. 4. Interfaces of prototype system

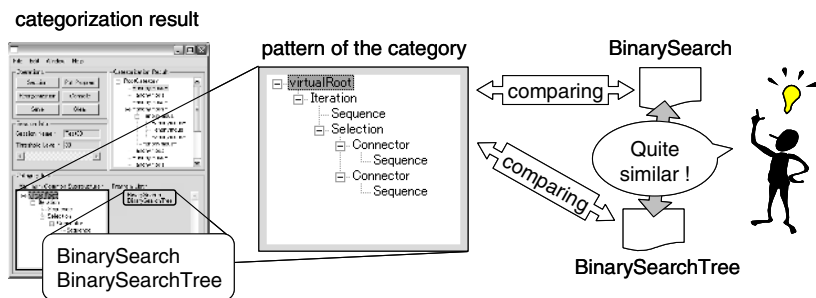


Fig. 5. Execution Example

about his/her programs. The learner can analyze a correspondence between a program and its control structure in the analysis interface and can extract any parts of control structures to categorize them in the editing interface (Fig. 4(b)).

## 5.2 Execution Examples

An execution example shown in Fig. 5 represents how our system supports the learner. The left side in Fig. 5 shows the categorization result. In this example, two programs “BinarySearch” and “BinarySearchTree” are categorized into the same category. The pattern of the category is the structural similarity among these programs and is expressed based on PAD. Comparing the pattern with each program is expected to make a learner be aware that the control structures of these programs can be designed in a quite similar way (the right side in Fig. 5).

# 6 Evaluation and Consideration

## 6.1 Structure of Categorization Tree

In order to analyze the characteristics of our categorization tree, several programs were put in the prototype system. The result shows that the tree structure varies greatly with orders in which the programs were put in even if a set of input programs were the same. This is because the input program is preferentially positioned to a node which has the largest similarity among the program (see **step4** in Section4.2).

Various structures of trees do not conflict with each other but rather result from categorizing the same set of programs based on variant similarities. Therefore, this structural diversity of trees can provide learners with not pre-fixed/stereotyped similarities but various ones according to the input programs. From a viewpoint of awareness about programming skills, this is an advantage of our method.

## 6.2 Awareness

We investigated whether learners can find out significant similarities among their programming skills. In this experiment, four examinees in our laboratory were asked to categorize their programs with our prototype system.

Based on the categorization results, they understood the conceptual meanings of the similarities in accordance with the semantics of programs. During the experiment, all of examinees frequently tried to reconfirm the semantics and intentions which they want to achieve with the programs by using the analysis interface.

This result suggests that our method makes examinees be aware of significant similarities among their skills and that the correspondences between programs and reflected skills play important roles in enabling the awareness. In fact, although the examinees could not find out significant similarities from the categorization results in some case, this is frequently because the structural similarities which the system found out did not correspond to the examinees' intentions.

## 7 Conclusion

In this paper, we focused on the similarity among programs and proposed a categorization method for programs. Our basic idea is to make learners be indirectly aware of similarities among their programming skills by categorizing their own programs. Experimental results suggest that our support enables to make up the awareness about programming skills.

For our future work, we must perform further experiments to verify the effectiveness of our support more precisely. For example, we have to investigate what the differences are observed between the awareness about programming skills when the learners use our system or when they do not. Currently, we focus only on the control structures of programs. In order to manage further similarity, we would like to consider the method for categorizing the programs from different viewpoints.

## Acknowledgements

This work was partly supported by the 21<sup>st</sup> Century COE (Center of Excellence) Program for 2002, a project titled Intelligent Media Integration for Social Information Infrastructure, proposed by Nagoya University.

## References

1. W. L. Johnson and E. Soloway: Proust: Knowledge-based program understanding, Proc. of ICSE '84, (1984) pp.369–380.
2. T. Watanabe: A Knowledge Handling Model to Design Learning Activity, Proc. of E-Learn '05, (2005) pp.1566–1571.
3. T. Watanabe: Framework of Computer Supported Learning/Education System, Based on Knowledge Handling Model, Proc. of E-Learn '05, (2005) pp.2483–2488.
4. Y. Hutamura, T. Kawai, H. Horikoshi and M. Tsutsumi: Design and Implementation of Programs by Problem Analysis Diagram (PAD), Journal of IPSJ, Vol.21, No.4, (1980) pp.259–267 (in Japanese).



# On-Screen Note Pad for Creative Activities

Norikazu Iwamura, Kazuo Misue, and Jiro Tanaka

Department of Computer Science,  
Graduate School of Systems and Information Engineering, University of Tsukuba,  
1-1-1 Tennoudai, Tsukuba, 305-8573 Japan  
{iwamura, misue, jiro}@iplab.cs.tsukuba.ac.jp

**Abstract.** We propose a system that enables users to write memos directly on a computer screen by hand. The system helps users save the screen images along with a user's handwritten strokes and in searching for them. We describe some of the important functions of the system and the effective retrieval keys for searching for saved images. We explain that the implementation of our prototype system consists of two subsystems: note-taking and browsing systems. An example of the usage of the system is also described.

## 1 Introduction

Currently, various types of researches are being conducted on computers. We are often working with intellectually creating ideas, and there are a lot of situations where we unconsciously hit upon some new ideas. For example, when researchers are browsing the web, they might hit upon an idea for their new research, or when a designer is looking at pictures on a PC, the idea for a new design might by chance be realized. However, when the idea presents itself, it might not be able to be put immediately into practice. In many cases, after some time is put into it, it will be reflected. Therefore, taking notes is generally done as a way of saving ideas.

When we are using a computer, using a pen and paper is currently the easiest way to take notes. This is the widely used means of saving ideas, because users can take notes quickly by hand. If there are materials that have been printed out on paper, it is possible to freely write on them, and the relation to the content of the paper can be written at the same time. However, if neither paper nor pen is available, it is impossible to take notes. It takes time to print out the paper and use it. Therefore, we often temporarily take notes on the computer.

It is possible to save notes as text data using word-processing software (note pad and word pad, etc.) to save information on the computer. However, because word-processing software can only save text data, the user's idea cannot be properly reflected using this method. Moreover, when the user starts the application to write notes, a new window opens. This new window hides the original information that brought about the user's idea because it uses a different application, making it possible for the user to forget their original idea.

Our purpose is to support "writing memos". We are developing a system where users can take notes on the computer screen, just like they would on paper. This system would help users save their ideas and helps in the reproduction of those ideas. We have

also considered some of the functions that would be necessary to achieve this purpose, and have developed a prototype system with these functions.

## 2 Overview of on-Screen note pad

We aim to make a note pad system that gives users the feeling that they are writing on paper, and supports users in using saved notes with the advantage of a computer system.

### 2.1 Natural Interface Like Paper

Paper is advantageous to users because it can be easily and simply used. We propose a function where a user can use our system like they use paper.

#### **Starting Without Disturbing Thinking Process**

If users must click an icon to start a system, users must be conscious of the application switch, and it is necessary to display the icon on the screen. The application switching might bring about the disturbance of a user's idea. We adopted a start method that uses either the mouse or a shortcut key. Therefore, users can start the system without disturbing their thought process because the displayed screen doesn't change.

#### **Writing Notes on Computer Screen Using Handwritten Input**

This system adopts a pen interface and the use of handwritten input. With handwritten input, users can freely write characters and construct figures just like on paper. Users often have a lot of vague broken ideas, and saving them by handwritten input is very useful for collecting them.

It is possible to write notes just like on paper. This system uses a displayed screen as the window background, and the user writes strokes on that window. The relationship between the information on a stroke position and the screen image can be used for writing notes.

### 2.2 Advanced Functions Beyond Paper

A computer system has its own advantages that differ from paper, such as using saved notes as data. We will describe some functions that make use of these advantages in this section.

#### **Browsing Notes**

A browsing system is necessary to look over saved notes. When users browse notes, they expect the note that was used the most to probably be the one saved last. When users start up their systems, the memo that was saved last should be the one that is displayed.

When a user wants to read saved notes, they go through a lot of trouble if they have to check their saved notes one by one. This system displays a lot of notes at the same time using a thumbnail format. When using paper, it might be possible for information on paper to be hidden by a large amount of strokes. It is possible for this system to

display only the screen image without the strokes, because it saves the stroke and screen image data separately.

To maintain a user's idea, this system can reproduce strokes according to the time they were written. This system can reproduce the order and speed of strokes similar to the time that the user took the notes, so the user's initial impression can be effectively reproduced. Users can not only browse the saved memo but also edit it, and the saved memo can be expanded. Since the edited notes might have greatly changed the original ones, they are treated as another new note.

### **Searching Notes**

It is necessary to efficiently search the stored notes that users want to retrieve to effectively exploit the notes. Therefore, some retrieval functions should be prepared. The keys to the retrieval functions should be their effectiveness at searching for the notes and being user-friendly. We use two kinds of retrieval keys. We will describe these retrieval keys in more detail in the next section.

## **2.3 Retrieval Keys**

We took into consideration the time and stroke information for the retrieval keys. The stroke information has the advantage in that the stroke impressions remain because the users write by hand. Therefore, it is advantageous for the users to use the stroke information as one of the retrieval keys. However, it is difficult to obtain a highly accurate retrieval using the stroke information, because the strokes that the user writes as the retrieval keys are vaguely shaped. On the other hand, the time information has the advantage that retrieval is easy from a technical point of view because it is exact numerical information. However, the load needed to remember the accurate time might be too large for users. We thought that the retrieval could work effectively by using both of these advantages together.

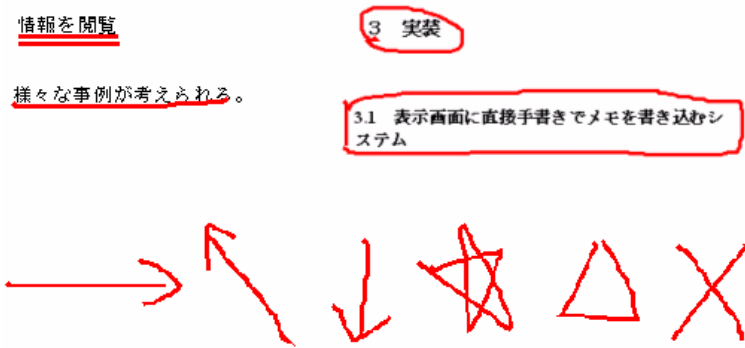
### **Time information**

One of the simplest types of time information that can be used as a retrieval key is the time that a memo was stored. We can use a retrieval method that uses the length of note-taking time by counting the time from when the user began taking notes to when they finished taking notes. A retrieval method based on a user's memory might also be possible considering the length of time needed to take the notes. For example, users can use their memory of consideration how worried the user was when they were writing notes, or whether they took the notes concisely or not. However, a users' memory might not be accurate enough when searching for notes using time information. Therefore, attempting exact matching using time information is not useful to users, we should allow some of the ambiguities found in matching using time information keys.

### **Stroke information**

The properties of written strokes are also effective as retrieval keys. We can obtain the amount of information that was written by examining the number and length of the written strokes. We can judge how intensively the user wrote during a unit of time by using the time and stroke information.

The shape of a stroke can also be used as a retrieval key. When users take notes, it can be expected that the annotation is often applied because users can use the screen image as a window background. To use such information, strokes are analyzed and used as retrieval keys. Figure 1 shows some examples. The shapes of strokes, like the examples shown in Figure 1, are counted by an analyzing system, and users can retrieve notes according to the numbers. However, there is vagueness because strokes are written by handwritten input. The shape of stroke like strokes shown in Figure 1 might not be written. When users retrieve information, some analytical results that have a high-level of similarity are also shown. If the similarity level is higher, the retrieval hit rate is higher.



**Fig. 1.** Examples of shapes used as retrieval keys (Underlines, inclusions, arrows, and marks, etc.)

### 3 Prototype Implementation

We have implemented a prototype system called “Leafletnote” to achieve the functions that are proposed in Section 2. The system consists of two subsystems: a note-taking subsystem and a browsing subsystem.

#### 3.1 Note-Taking System

With the note-taking system, users can write strokes on the displayed screen by hand, and save the strokes with the screen image.

Figure 2 shows a screen image of the note-taking system. When a user starts the system, they can freely write strokes on the screen by hand. When they are finished taking notes, they just tap the screen. Then the system saves the memo in the save folder, and the system shuts down.

Figure 3 shows the explanation of making a starting window. When this system is started, the system captures the screen image on the display at that time, and shows the image data on the window. So, the user can operate the window with the background maintained.

Users can write strokes by hand on this window. Therefore, we have achieved the ability to write directly on the display.

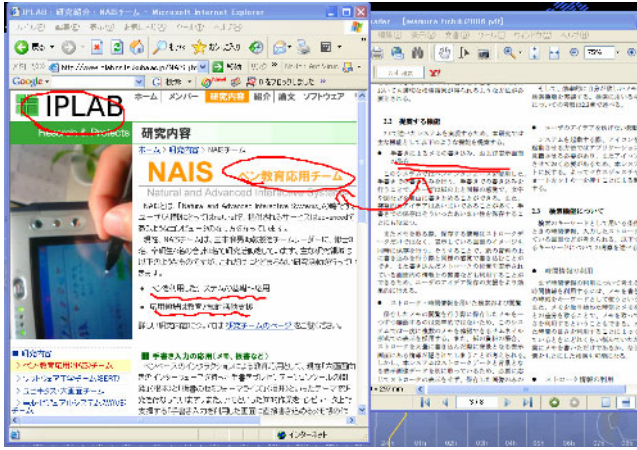


Fig. 2. Example of Note-taking system image

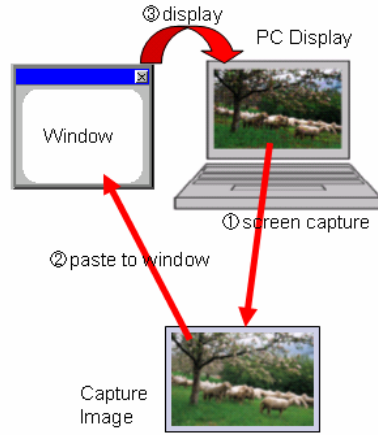


Fig. 3. Explanation of making starting window

### 3.2 Browsing System

The browsing system helps the users retrieve notes stored with the note-taking system and helps to remind them of their ideas.

Figure 4 shows an image of a browsing system. When a user starts this system, the note that was most recently saved is displayed in the largest frame. If the user clicks one of the thumbnails on the left side of the window, the note that corresponds to the thumbnail is displayed in the large section on the right side of the window.

The stroke and image data are saved separately. So, it is possible to display only the screen image without the strokes. This function enables users to see the original screen images at any time. The time and coordinate data are also saved with the stroke data.

This data can be used to replay the writing of the strokes. This function helps users to remember the situation in which they wrote the strokes. They can use these functions by selecting items from the menu.

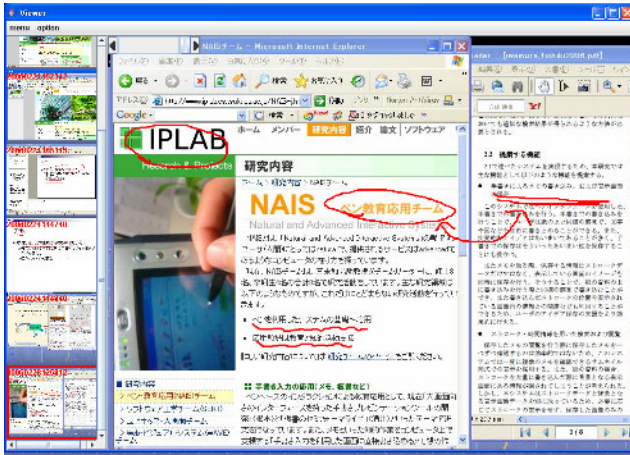


Fig. 4. Example of browsing system image

### 3.3 Implementation Environment

We used the JAVA language for implementing the system. We used the making support tool kit SATIN [1] for the pen-based application. The processing and analysis of the stroke data, and the drawing canvas were used from the SATIN library.

## 4 Example of System Application

In this section we will describe a concrete scene of the use of this system. We assume a situation where users were performing research on their areas of interest on their computers. The following examples are imaginable:

1. When the users find research papers related to their area of interest on the Internet, they read the papers on their computer screens.
2. When they find important information in the papers and hit on some ideas, they start the note-taking system by pushing a shortcut key. This operation does not disturb their activities. They write notes, make marks, or underline items on the interesting area of the display screen. When they finish writing notes, they tap the screen and the system saves the notes with the screen image.
3. While they are reading the research papers, if they want to know the details concerning the research topics, they open the related web pages and study the detailed information. By using this system, they can write notes while displaying both the research paper and the web page.
4. They look at the saved note pad later. They can look over the content of the web pages and research papers that they saw before, because the screen image and

written strokes are saved in the notes, making it is easier to remember their idea. They can review their strokes by using the reproduction function, and reconfirm the image when writing a note. If they hit on more ideas, they can add their ideas to the note, and the content of the note is expanded.

5. When they want to look over a note that was saved in the past, they use a retrieval function. At that time, they can find the notes by using some vague information (stroke, time, and so on), even if their memory is unclear.

## 5 Related Works

Researches that aim to support the thinking process or for taking notes by hand have been conducted in the past.

ScreenCrayons [2] is common to our research as it targets the saving of written text by hand. This system provides a function that is able to write the annotation in various applications, and to save the part in which the annotation was written as an image. However, the saved object is only a part of the image where the annotation was written. We aim to support the user's intellectual creation by using information on the background. Electronic Cocktail Napkin [3] is a system that can freely take handwritten notes. However, it differs from our system in that the displayed image cannot be used as a background image. InkDesktop [4] and PenPlus pro [5] are systems that can write in the displayed screen using handwritten pen input, but they don't support how to use the saved notes. SmartCalendar/SmartWrite [6] is a system that aims to take notes using handwritten input. It has capture screen and retrieval functions using time information, but it does not have a retrieval function using the stroke information.

Kamicopi [7] is a system that aims to save data and reuse it. It takes data from web pages and e-mail and so on, and freely reuses it. However, users cannot use handwritten input in this system. Microsoft OneNote [8] has a note pad function that has been enhanced. This software can be used with data from various media and applications, and memos can be written by hand. However, there is the problem of hidden background images when the new window opens. Another problem is the operation might be too complex because the system provides various functions.

## 6 Conclusion and Future Work

We have proposed and implemented a note pad system that was able to write notes directly on a computer screen by hand. Users can save the screen image and handwritten strokes with this note-taking system. They can use a saved note and reproduce saved strokes using a browsing system.

There are some future works that we are planning. We plan to implement the retrieval function and keyboard input to enhance our prototype system. We will consider some natural interfaces, for example, whether the "tap" method to save notes is appropriate or not.

Users who will actually use the system will be used to evaluate the effectiveness of the system, and we will investigate the frequency of use and the users' satisfaction ratings (about input, inspection, and retrieval phase).

## References

1. J. Hong and J. A. Landay, 2000. SATIN: A Toolkit for Informal Ink-based Applications. In Proceedings of User Interface Software and Technology: UIST 2000, pp.63-72.
2. Dan R. Olsen Jr., Trent Taufer, and Jerry Alan Fails. ScreenCrayons: Annotating Anything. In Proceedings of User Interface Software and Technology: UIST 2004 pp.165-174.
3. M. D. Gross, and E. Do. Demonstrating the Electronic Cocktail Napkin: a paper-like interface for early design, Conference Companion, ACM Conference on Human Factors in Computing: CHI 96, pp.5-6.
4. Microsoft: Microsoft Windows XP Tablet PC Edition <http://www.microsoft.com/japan/windowsxp/tabletpc/downloads/enhancementpack/default.mspix>
5. PlusSoft: PenPlus pro <http://www.plussoft.co.jp/penplus/pro/>
6. Kaoru Misaki and Naoki Kato. SmartWrite: Paperlike pen-based memo tool pursued brevity, In Proceedings of Workshop on Interactive Systems and software 2005 (WISS 2005), pp. 37 - 42. Japan Society for Software Science and Technology (in Japanese)
7. kamilabo.jp: Kamicopi <http://www.kamilabo.jp/copi/index.html>
8. Microsoft: OneNote <http://www.microsoft.com/japan/office/onenote/prodinfo/default.mspix>



# IdeaCrepe: Creativity Support Tool with History Layers

Nagayoshi Nakazono, Kazuo Misue, and Jiro Tanaka

Department of Computer Science,  
Graduate School of Systems and Information Engineering, University of Tsukuba  
1-1-1 Tennoudai, Tsukuba, Ibaraki, 305-8573, Japan  
{zono, misue, jiro}@iplab.cs.tsukuba.ac.jp

**Abstract.** We developed a creativity support tool called IdeaCrepe that manages the history of ideas using a layered structure. Handling ideas with a creativity support tool that is able to record history is very useful for understanding these ideas, the evolution of the ideas over time, and the relationships between the ideas. Looking back through the history of creative activity may contribute to the further development of ideas. In this research, we introduce a concept of “layered structure” based on a time series to manage the history. Using this structure, we can look back at past states of ideas, and examine their development. These features can manage the creation and modification of ideas.

## 1 Introduction

We developed a creativity support tool called “IdeaCrepe” that manages the history of ideas using a layered structure. This tool has a history management function with which we can see the evolution of states of ideas easily. Recalling when ideas were created or what the previous states were is difficult, because successive ideas are created intermittently and frequently. Moreover, generally, we cannot check previous states when we want to because they are overwritten by more recent developments.

We developed a creativity support tool that enables us to look back at the history of our creative process. We think intuitively that recalling previous trends, failures, and successes will make our future activities better. We introduce a recall function and then examine the tool.

## 2 History Management in Creativity Support Tool

### 2.1 Methods for Creative Activities

There are currently any propositions of methods that can support our creative activities. Some of those methods use network diagrams which are composed of nodes (ideas) and edges (relationship lines). We selected a network diagram method for our creativity support tool.

## 2.2 Looking Back Makes New Ideas

During many creative activities, we frequently compile observations about one theme. In activities developed in a time series, one important factor in deciding the guideline of future activities is looking back at the history of the creative process. One way of doing this is by checking previous states when ideas were created.

The general method of saving work is by overwriting it. The computer always saves the most recent snapshots of data. If you want to check previous state, you must save each snapshot without overwriting the data, and then compare multiple snapshots.

This problem is also suited to creativity support tools. We developed a creativity support tool called IdeaCrepe that implements a framework of data processing to manage the history of the creative process. We focused on a creativity support tool that uses a network diagram.

## 3 Layered History Structure

We employ a layered structure [7] as a method to manage the history of creative activity. Network data of creativity is managed separately as several layers, not as a single diagram, and each one depending on the update time. We display the data and color it in a flexible manner. This approach helps users to read large amounts of information from network diagrams.

### 3.1 History Management of Network Using Layered Structure

In IdeaCrepe, ideas produced through a creative activity are expressed as a network diagram using a concept of layered structure as shown in Fig. 1. The structure is similar to layers of transparencies like those used which an overhead projector. Changes from each update are saved on each layer.

### 3.2 Application to Creativity Support

Our concept is that each layer has only one event. When an event is made, a new layer is added. The operations piling layers represent the development of

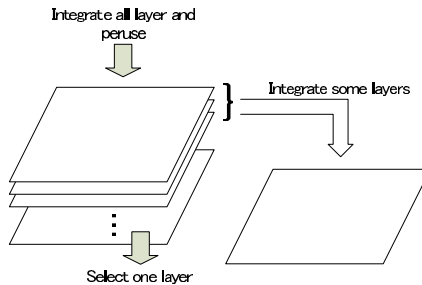


Fig. 1. Conceptual image of layered structure

knowledge during creative activity. Due to this method, the order in which events were created is clear, so we can manage the history of the creative activities more certainly. As the history is recorded in layers, we call this method a “layered history structure.”

In a layered history structure, events are organized according to a time series and users can look at the optional layers easily. This merit causes that users can see the order in which their ideas created. With a layered history structure, users can grasp how much their ideas have created over a period of time, unlike existing methods that save snapshots. With the layered history structure, users can see the evolution of the creative process in more detail than is possible with existing methods.

## 4 IdeaCrepe – Creativity Support Tool

We developed a knowledge creativity support tool called “IdeaCrepe” that has a layered history structure. This tool manages ideas by using network diagrams in the same way as some existing tools.

### 4.1 Operations with Tool

Users perform various operations using IdeaCrepe. The tool provides the following operations:

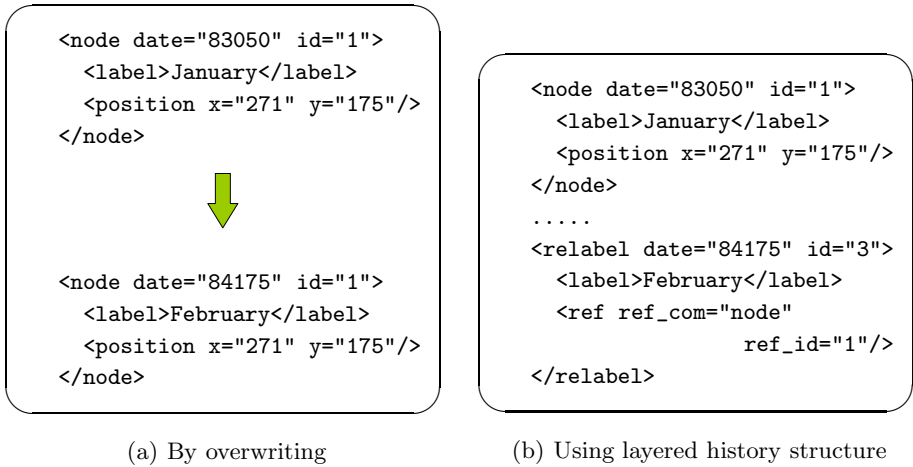
- Create node: creates a node for each individual idea
- Create edge: creates an edge (a relationship line) to link ideas
- Change label: changes the label of a node or an edge
- Move node: moves the position of a node
- delete node: deletes a node or an edge

We call these operations “events,” and call nodes and edges “objects.”

### 4.2 Recording Ideas Using Layered History Structure

IdeaCrepe has a large advantage in that it saves all previous information which we can see by moving, deleting, or changing labels. To create those operations, we described events of IdeaCrepe in an original XML document format. In an XML document, created elements (events) are not changed or deleted. Instead, another new element is added, which is how events are recorded in IdeaCrepe.

For example, imagine a situation where we want to change the label of a node from “January” to “February.” When we create a node, an element is added to the XML document. Then, if we change the label and overwrite it, the element is changed as shown in Fig. 2(a), and the information of past label “January” is lost. IdeaCrepe does not overwrite the information; instead, a `relabel` element is added. We update the data by piling a new layer that has a `relabel` element on top of the related layers (see Fig. 2(b)).



**Fig. 2.** Save by overwriting vs. using layered history structure

### 4.3 Interface of IdeaCrepe

Screenshots of IdeaCrepe are shown in Figs. 3(a) and 3(b). IdeaCrepe has two modes. Each mode implements a different interface. Fig. 3(a) is a screenshot of the idea processor mode. Operating buttons are on the left side of the window and a canvas on the right side acts as an interface of this mode. Users do creative activities by arranging objects on the canvas. Basic functions such as file operation, and mode changing are called up easily by clicking the operating buttons.

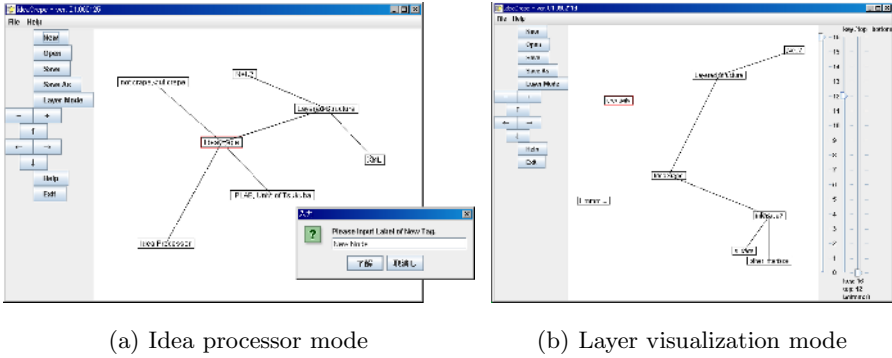
Fig. 3(b) is a screenshot of the history layer visualization mode of IdeaCrepe. The interface of this mode is similar to the other interface; however, the right side has three sliders. Users select a key layer which users pay attention to by using the leftmost slider. The other two sliders select the range of visible layers (top and bottom).

For example, when users set all sliders at the oldest layer, only the oldest layer is visible. From this state, if users move the slider up, which indicates the top layer, they can look back at the changes in creative activities along the time series. When users want to check activities that occurred at a certain time, they only have to set the bottom slider at the beginning point, and the top slider at the ending point. Selecting the visible time period by operating the sliders enables users to look back at the evolution of their creative activities.

### 4.4 User Operations

We suppose that a mouse and a keyboard are used together to operate IdeaCrepe. Users of IdeaCrepe handle ideas by combining the following operations.

**Create Node.** When a user comes up with an idea, he or she clicks any point on the canvas. A dialog box will appear, and the user can create a new node to record an idea by inputting a label in the dialog box. This node is created at the position clicked on the canvas.

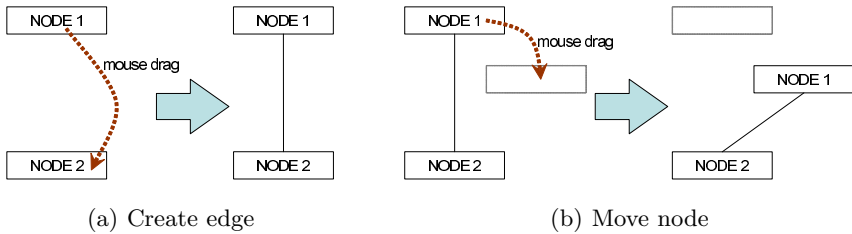


(a) Idea processor mode

(b) Layer visualization mode

**Fig. 3.** Screenshots of IdeaCrep

**Move Node and Create Edge.** In creative activities using pen and paper, we can pick up a piece of paper or card with our fingers and move it to a new location. When we want to create a relationship between two cards, we can draw a line between them with a pen. Those activities are similar to moving a node and creating an edge, both of which can be done using a mouse – dragging. The difference between those operations is whether there is already another card at the dragged position, as in Fig. 4(a), or not, as in Fig. 4(b).



(a) Create edge

(b) Move node

**Fig. 4.** Move node and create edge by dragging mouse

**Change Label and Delete Objects.** When you want to change a label, you have to select an object to be edited. With IdeaCrep, you can edit the label by double-clicking the node or the edge. To delete operation, users select an object to be deleted. However, users must be aware which operation is occurring: creating or deleting. Therefore, to avoid confusion, we designed the delete operation of IdeaCrep to operate with a right click (not a left click).

**Switching to History Layer Visualization Mode.** IdeaCrep consists of three modules as shown in Fig. 5(a): idea processor module on which we operate knowledge creativity, file I/O module, and the history layer visualization module. By clicking the Layer Mode button on the left side of the window, IdeaCrep switches to the history layer visualization mode (see Fig. 3(b)). Users may want as many operations in this mode as in the idea processor mode. Therefore, we

implemented a way for users to change views clearly. The history layer visualization mode changes the layers using sliders.

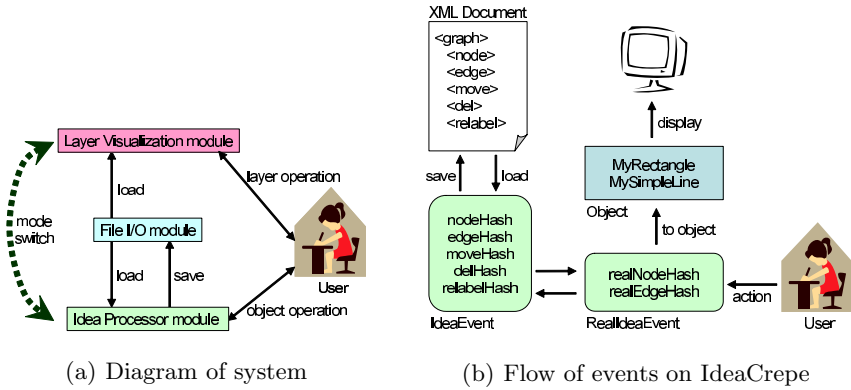


Fig. 5. System images of IdeaCrepe

#### 4.5 Visualize Ideas Using Layered Structure

By saving ideas in a XML document using a layered structure, we can modify the ideas in the previously saved data. However, an object will be overwritten by multiple events because this tool adds new events instead of changing existing ones. To avoid this, IdeaCrepe handles data in the way shown in Fig. 5(b).

After loading an XML document, IdeaCrepe records data in an instance of IdeaEvent class and stores it in a hash table. Events with move, delete, and change labels are stored among events as overwrite nodes or edges that correspond to and can be stored in a hash table. By converting data that is stored in this hash table to the Graphics objects of Java, the data is visualized.

### 5 Evaluation and Study

#### 5.1 Testing Impressions of IdeaCrepe

We performed a creative activity by using IdeaCrepe, and tested how using this tool feels. There is still no strong evidence what effects are produced by looking back at the history of our creative activities. However, a test user said that looking back at past activities assisted him in guided future creative processes. In our future work, we should observe the circumstances of activities quantitatively, such as the number of layers and objects created, and evaluate how effective looking back at the history is.

#### 5.2 How to Improve Layered History Structure

In the present implementation of IdeaCrepe, sliders operation only express layered structure. For that reason, Users have difficulty relating sliders to the

concept of layers. We must add something that can remind users of the layers at a glance. Alternatively, we can replace sliders with another operation interface. For example, if users could operate the concept image of layers directly, we think that they could understand and operate the layers intuitively.

Users who look back at a process may want to continue a new branch of creativity from a certain point in the past. The piling of layers has a tree structure, so we should examine a suitable way to look at and operate such layers.

### 5.3 Study of Input Devices

IdeaCrepe was designed with the assumption that users would use a mouse and keyboard for input. However, some users operate their equipment with pens (styluses) on real world-oriented designs. Therefore, some researchers have tried to make use of merit of handwriting in creativity support [6]. Many designers like to design freely using pen and paper during the early stage of creativity [3]. We think that IdeaCrepe will have to support handwriting input using a pen as well as a mouse and keyboard. Moreover, an interface with a three-dimensional drawing board [4] may make the piling of layers more realistic.

## 6 Existing Tools and Related Work

The field of creativity support has a long history. Stefik et al. researched computer support by computer in collaboration with some people [8].

Many researchers have been developing creativity support tools. Misue et al. developed an interactive support system called D-ABDUCTOR which grasps the processes of the KJ method [9] as part of graphical thought extension [5].

Some concept mapping tools [1] have a function to playback processes, such as the tool created by Funaoi et al. [2]. This playback function is, however, uses snapshots. This function can only play back selected snapshots along the time series. By comparison, IdeaCrepe can show users previous states of ideas and enable them to see the changes (difference). Users are able to look back at the evolution of their creative activities with greater flexibility and variety than is possible with other existing tools.

## 7 Conclusions

We applied a layered structured network to our creativity support tool and created IdeaCrepe. This is an all-purpose tool that can operate various kinds of creative activities and that can save the histories of events in a layered structure. We can recall processes of past creative processes by looking at the history layers.

In our future work, we will modify the view of the layers to create a network diagram that users can operate layers intuitively and directly. And also, we will introduce a new operating system to replace the sliders. In addition, IdeaCrepe will draw out the creative activities of users naturally by supporting their input intuitively.

There is still no strong evidence about what effects can be produced by looking back at the history of our knowledge creative processes. In future work, we have to evaluate quantitatively how effective looking back is.

## References

1. Hanson, E.: A Survey of Concept Mapping Tools.  
<http://datalab.cs.pdx.edu/sidework/pub/survey.of.concept.maps/>
2. Funaoi, H., Yamaguchi, E., Inagaki, S.: Collaborative Concept Mapping Software to Reconstruct Learning Processes. in *Proceedings of the International Conference on Computers in Education (ICCE'02)* (2002) 306–310
3. Landay, J. A., Myers, B. A.: Sketching Interfaces: Toward More Human Interface Design. in *IEEE Computer*, Vol.34, No.3 (2001) 56–64
4. Lapedes, P., Sharlin, E., Sousa, M. C., Streit, L.: The 3D Tractus: A Three-Dimensional Drawing Board. in *First IEEE International Workshop on Horizontal Interactive Human-Computer Systems (TableTop2006)* (2006) 169–176
5. Misue, K., Sugiyama, K.: How Does D-ABDUCTOR Support Human Thinking Processes? in *Proceedings of CG International '94 – Insight through computer graphics* (1994) 10–21
6. Misue, K., Tanaka, J.: A Handwriting Tool to Support Creative Activities. in *Proceedings of 9th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES2005)* (2005) 423–429
7. Nakazono, N., Misue, K., Tanaka, J.: NeL2: Network Drawing Tool for Handling Layered Structured Network Diagram. in *Proceedings of Asia Pacific Symposium on Information Visualization 2006 (APVIS2006)* (2006) 109–115
8. Stefik, M., Foster, G., Bobrow, D. G., Kahn, K., Lanning, S., Suchman, L.: Beyond the Chalkboard: Computer Support for Collaboration and Problem Solving in Meetings. in *Communications of the ACM*, Vol.30, No.1 (1987) 32–47
9. Kawakita, J.: The KJ-method. Chuokoron-sha (1967) (in Japanese)



# Personalized Voice Navigation System for Creative Working Environment

Kaoru Tanaka and Susumu Kunifuji

School of Knowledge Science, Japan Advance Institute of Science and Technology  
1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan  
{tanaka-k, kuni}@jaist.ac.jp  
<http://css.jaist.ac.jp/>

**Abstract.** The development of mobile devices and information network environments has been changing working environments and the way people communicate and collaborate. For creative collaboration, awareness about communities is important. We propose a personalized voice navigation system to support awareness of related research and researchers. We explain a voice navigation system based on pervasive technology that is designed to heighten awareness of research related to a user's interest and his movement in the working environment.

## 1 Introduction

The development of mobile devices and information network environments has made the working environment border-less for knowledge workers. The sight of a person using a PDA or a mobile PC hooked into a wireless network environment to communicate as if sitting in front of a desk, and collecting information by reading two-dimensional pattern code printed on a poster is now commonplace. This sight suggests that the development of mobile devices and information networks has brought about changes in how people collaborate. These changes will continue with the development of pervasive computing [7] technologies. Pervasive devices are expected to become more miniaturized and have more advanced functions. Pervasive systems will be developed based on users' activities in everyday situations.

Many groupware and know-who systems have been proposed to support collaborative work. However, we think people's daily working environments should be taken into account while developing pervasive computing technologies. For example, there are posters in front of university laboratories that introduce the lab's research. However, people reading the posters would not know about research and researchers related to that laboratory's work as they would if they were browsing the web. We think that there is an opportunity to support knowledge workers by making them aware of the implicit relevance of their work to the work of other people who may have insight into common problems. Therefore, it is important to remind relevant information by focusing on objects located in workspace. In other words, it is possible to support communication by applying pervasive technologies to real-world objects.

The most useful method for users to receive related information is an issue in itself. In particular, it is important to introduce pervasive technology in a way that would not interfere with activities of daily work. Therefore, using voice information services have considerable potential. However, the contents of voice information systems would have to be uniform, and the detailed information that the individual wants would have to be supported.

In this paper, we propose a system that enables personalized voice navigation through everyday interactions with real world objects such as informational posters in the working environment. To generate and provide the personalized voice navigation content about related laboratories and research based on the user's interests, we use keywords input previously by the user and browsing information as contained in research posters posted outside offices. Our system will provide the user with information about research and researchers doing work similar to their own, which will improve communication and creative collaboration.

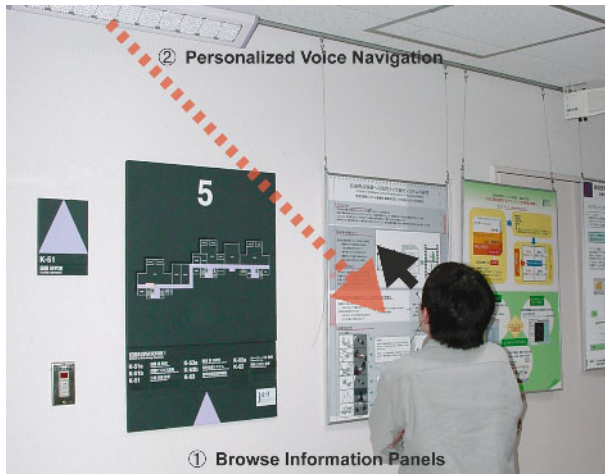
## **2 Personalized Voice Navigation System for a Creative Working Environment**

### **2.1 Introduction**

Collaboration between researchers depends on smooth communication which in turn depends on awareness of common or similar interest [5]. Our goal is to establish a communication support environment by enabling knowledge workers to interact with pervasive computing embedded in real-world objects in their working environment. To support awareness of relevant work, we focus on interests of the user and implied relations behind each real world object and offer recommendations for potentially useful exhibits in a voice navigation format.

### **2.2 Scenario**

The intended users of our system are master's degree students at the School of Knowledge Science at the Japan Advanced Institute of Science and Technology (JAIST). Our school has many department and research topics of each laboratories straddle multiple areas. One researcher's interest or problem may have unexpected links to another's research. With our system, the user personalizes his or her system registration page by inputting his or her research interest keywords. Then, using an optical ID badge that indicates his or her location a linked earphone to listen to navigation contents, the user is guided to posters or guide panels containing information relevant to the user's interests. The more the user browse posters, the more personalized for the user the voice navigation will become. Figure 1 shows an example of someone using our system.



**Fig. 1.** Example of use

### 2.3 Device

For daily use, navigation device should be compact and non-intrusive. Our system uses Aimulet [2], a portable audio device customized for our school. Figure 2 shows Aimulet's transmitter (top-left) and portable device (bottom-left). Aimulet translates infra-red radiation emanated from the exclusive transmitter to audio. The device is easy to take anywhere and requires no electrical power source.

## 3 Prototype System

### 3.1 Introduction

There are two approaches to providing voice navigation based on the user's interest. One is using natural language processing with a text-to-speech machine. In this case, voice navigation for the user should be in plain text format, and should be shifted into sound format by the text-to-speech machine. The other is using pre-generated sound files. In this case, all conceivable contents of the system must be written in advance, and played based on pre-defined playing rules. The former approach has procedures that initially generate the guide contents as text format, so the voice navigation can be more flexible than with the latter approach. The natural language processing approach commonly requires more processing time than the pre-generated sound approach because of the need to translate text format to voice. Also, pre-generated sound provides natural voice contents. However, the pre-generated sound approach cannot provide the same level of detail as the natural voice approach.

With these considerations in mind, we invented a template method that combines the merits of both approaches to generate a personalized voice navigation.



**Fig. 2.** System devices - Aimulet transmitter (upper-left), optical ID reader embedded in ceiling (upper-right), Aimulet receiver(lower-left), and optical ID badge (lower-right)

This approach is that our system use templates for navigation with some blank slots, and when requiring navigation, fill slots with voice content like "Kunifuji Laboratory" recorded in advance. This approach has the advantage of being easy to extend and modify as needed.

### 3.2 System Outline

To work out our approach, two external systems are needed, one to present voice navigation content, the other to track the user's location in the building. Therefore, we designed our system, based on a service oriented approach, to provide seamless communication between subsystems and to accommodate functional extension with the addition of sensors in the future.

Our prototype consists of four main components, Personal Profile Manager, Recommender, Template Engine and Voice Navigation Manager. Personal Profile Manager manages the user's information (ID number of the infrared ray tag and some keywords of interest to user). Recommender determines voice navigation content using information about keywords of user interest, the user's location in buildings, and features of laboratories or researcher's work as extracted from articles. Template Engine chooses suitable voice parts and fills slots in the templates based on Recommender's recommendations. Voice Navigation Manager selects relevant navigation contents from the database and sends them to the Aimulet audio transmitter.

### 3.3 Process Flow

Our system creates real-time voice navigation content when the user is near an information panel, and transmits it through the Aimulet transmitter in that

location. To pinpoint the location of the user, our system communicates through a location awareness system [3]. The system infers from a user’s behavior, such as lingering over certain posters or floor maps, what his or her interests or immediate concerns are.

This method makes it possible to provide beneficial information to the user about research related to his or her own without the user having to request information by inputting search keywords. The user simply has an earphone in one ear. To address the problem of the difference between a user’s actual movements and the lagged information sent to the navigation system, we set up a threshold of only a few seconds in the prototype. Figure 3 shows our system’s process flow.

The location awareness system transmits the ID number on the infrared ray tag and the user’s location in a building to Personal Profile Manager if the threshold is exceeded. After the ID number is received, Personal Profile Manager checks Personal Profile Database, and retrieves the user’s interest keywords, which were input in advance. Then, Personal Profile Manager sends these data to Recommender. Subsequently, Recommender checks the characteristic database of articles and recommendations based on the degree of similarity. In addition, Recommender decides on the applicable template and voice parts considering the user’s location at the time and transmits these results to Template Engine. Having received these results from Recommender, Templates Engine chooses the pertinent template and generates the navigation play list. After that, Voice Navigation Manager takes pertinent voice parts from the voice parts database and transmits them to Aimulet transmitter.

### 3.4 Recommender

Our prototype makes recommendations about which posters displayed in front of laboratories a user may want to read. To arrive at recommendations, we apply natural language processing methods to articles about each poster. First, terms from the articles are identified as morphemes using Chasen [6], a Japanese

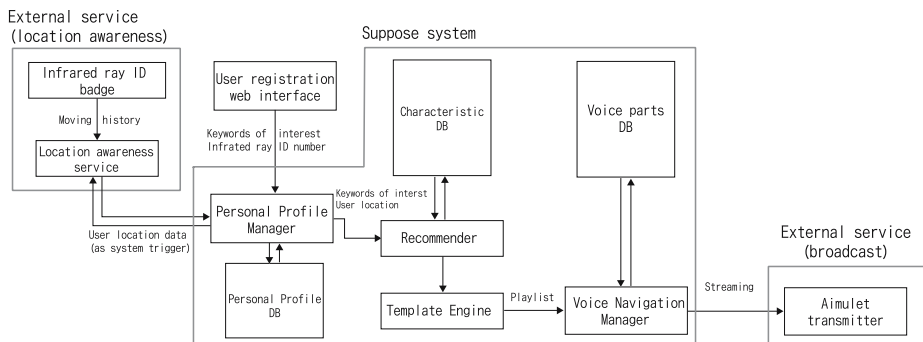


Fig. 3. System outline and process flow

language morphological analysis library. Then Chasen extracts domain-specific terms based on a TFIDF algorithm that identifies valuable characteristics of commonly used words in documents. Subsequently, Chasen calculates the degrees of similarity between documents based on a vector space model. In our prototype, we use the Automatic Domain Terminology Extraction System "termex" and "termmi"<sup>1</sup> developed at the Nakagawa laboratory of the University of Tokyo and the Mori laboratory of Yokohama National University.

### 3.5 Navigation Sample

Our prototype creates navigation content by binding parts belonging to four categories: laboratories, researchers, research keywords, and laboratory locations. Here are sample navigation templates<sup>2</sup> with navigation parts slots.

1. {researcher name} researches {keyword} in {laboratory}.
2. {laboratory} researches {keyword}.
3. {laboratory} is in {location}.

In this sample, terms framed by brackets are Template Slots, which are filled with the navigation content resulting from calculation of recommendations. In these sample templates, the selection rule is simple: when the user's keywords of interest are highly relevant to the topic of the poster, Recommender selects 1. The method for representing location has two styles: absolute and relative (based on the user's location at the time).

## 4 Related Work

Our system recommends information about related researchers and laboratories or research topics based on the interests of each user in the working environment to support communication based on similar research interests.

Streitz proposed Roomware [1], which treats the environment as an extension of groupware with embedded computers in objects such as tables, chairs, and walls in meeting rooms. The central aim of this project is to support users in the process of collaboration. Roomware is different from our system whose purpose is to support opportunities for communication.

The Exhibition Visitor Support System at the Aichi Expo 2005 Global House based on Kurumatami's work [4] uses devices and provides navigation content about exhibits that are similar to our system's. This project has as its central aim supporting users by providing navigation to avoid congestion based on traffic line history of users in Global House. This project is different from ours in purpose and in that it is not personalized based on user interests.

<sup>1</sup> <http://gensen.dl.itc.u-tokyo.ac.jp/index.html>

<sup>2</sup> We described these sample in English for explanation. Our voice navigation system is operated in Japanese in fact.

## 5 Conclusion

We described a system to increase awareness among colleagues of one another's research in a pervasive technology environment. In particular, we explained a voice navigation system designed to make researchers aware of related research information through voice navigation content based on each user's interest and actions in the real world. While we did not evaluate our personalized voice navigation system in detail, it was enthusiastically received by test users. We will report the results of our evaluation at the conference.

## Future Work

In this prototype, navigation contents are recommended only for related laboratories and researchers and their locations. We plan to investigate similar navigation that might be valuable for users of our system. We will also design more navigation templates based on the results of our pilot study. After that, we will construct a more personalized navigation system by adopting new features.

## Acknowledgment

Our research is partly supported by the fund from the second knowledge creation support systems of Center for Knowledge Science of JAIST. I would like to acknowledge to Dr. Koiti Hashida of National Institute of Advanced Industrial Science and Technology for advising the research and development of the above systems.

## References

1. Norbert A. Streitz, Jorg Geisler, and Torsten Holmer: "Roomware for collaborative buildings; Integrated design of architectural spaces and information spaces", *Cooperative Buildings (LNCS 1370)*, Springer, 4-21 (1998)
2. Hideo Itoh, Takeshi Akiyama, Yoshiyuki Nakamura, Takuichi Nishimura, Yoshinobu Yamamoto, Takehiko Hidaka, and Hideyuki Nakashima: "Spatial Optical Interconnection Technique with Low Power Consumption for a Location-Based Information Service Environment", *Optical Memory & Neural Networks (Information Optics)*, Vol. 11, No. 3, pp. 155-158 (2002).
3. Toshiyuki Hirata, Susumu Kunifuji: "Information Sharing System Based on Location in Consideration of Privacy for Knowledge Creation", *Proc. of the 8th International Conference on Knowledge-Based Intelligent Information & Engineering System (LNCS 3213)*, pp. 322-329 (2004)
4. Koichi Kurumatani: "Mass User Support by Social Coordination among Citizens in a Real Environment", in *Multi-Agent for Mass User Support (LNCS 3012)*, pp. 1-17, Springer, 2004.
5. P. Dourish, S. Bly: "Portholes: Supporting Awareness in a Distributed Work Group", in *Conference proceedings on Human factors in computing systems*, pp. 541-547, ACM Press, 1992.

6. Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano: "Japanese morphological analysis system Chasen version 2.0 manual", Technical Report, NAIST, Ikoma (1999)
7. Weiser, M.: "The Computer for the 21st Century", *Scientific American*, 1991, 265 (3), pp. 94-104.



# An Editing and Displaying System of Olfactory Information for the Home Video

Dong Wook Kim<sup>1</sup>, Kazushi Nishimoto<sup>2</sup>, and Susumu Kunifuji<sup>1</sup>

<sup>1</sup> School of Knowledge Science, <sup>2</sup> Center of Knowledge Science  
Japan Advance Institute of Science and Technology  
1-1 Asahidai, Nomi-City, Ishikawa 923-1292, Japan  
{dongwook, Knishi, kuni}@jaist.ac.jp  
<http://css.jaist.ac.jp/>

**Abstract.** Olfactory displays developed in the past emphasized the display of multiple fragrances using one device, and, hence, problems arose, such as huge device size, high cost, troublesome operation and complicated maintenance. Therefore, a home environment to edit and play back images and sound with fragrances has not been achieved to date. In this work, we proposed and developed the “FragrantMemories” system, which can be implemented very easily at home and is convenient to use. Users can edit FragrantMemories and play back images and sounds with fragrances.

## 1 Introduction

Human beings acquire information of the external world through their sense of sight, hearing, smell, taste and touch. Moreover, the growth and popularization of various media such as TV and computers have made it easy to acquire information from remote areas. However, these media are based on audiovisual (AV) information interaction and do not provide any information about the fragrance and atmosphere of a particular place. Recently, several different studies have been conducted, and a variety of systems and displays that transmit smell has been proposed [1][2]. For example, the ambient atmosphere in movie theaters has been tested successfully [3]. Different Web sites provide services that transmit several different fragrances with a click; an aroma generator, which is connected to the PC beforehand, transfers a fragrance when specific links are clicked [4][5]. Olfactory displays developed in the past emphasized the display of multiple fragrances using one device; hence, problems arose such as huge device size, high cost, troublesome operation and complicated maintenance. Therefore, a home environment to edit and play back images and sound contents with fragrances has not yet been achieved. In the present work we proposed and developed the “FragrantMemories” system, which can be implemented very easily at home and is convenient to use. Users can edit FragrantMemories and play back images and sounds with fragrances. Moreover, we conducted an evaluation experiment using FragrantMemories to verify the necessity of smell broadcasting on TV. Generally, FragrantMemories can be operated either

automatically or manually, and so we also verified a suitable process to manipulate this system.

## 2 Structure of FragrantMemories

Light has three primary colors, taste has its own primary flavor, but unfortunately smell does not have any primary odor. Hence, developing an olfactory display which can transmit multiple smells is a rigid task. For the realization of an olfactory display, an appropriate amount of mixtures of different fragrances have to be prepared. For the above reasons, a huge and complex system was previously unavoidable. Therefore, in the present work, we focused on only a particular image and sound that employ a limited fragrance (the smell of grapefruit). In general, however, we used the one-sheet principle, in which one piece of sheet is used at a time: when audiovisual contents change, an ‘‘Aroma Sheet’’ is replaced. Consequently, we are able to design small systems and simplify the problems of maintenance. The structure of the FragrantMemories system can be divided into three different elements. The first one is the olfactory display controller (Aroma Module Controller), a device that controls the aroma signals which help to generate a fragrance. The second one is the olfactory display (Aroma Module), which transmits the aroma signals to control the release. The third and the inner part of the Aroma Module is the non-liquid aromatic Sheet (Aroma Sheet), which works as the source of fragrance. The structure of FragrantMemories is presented in Figure 1.

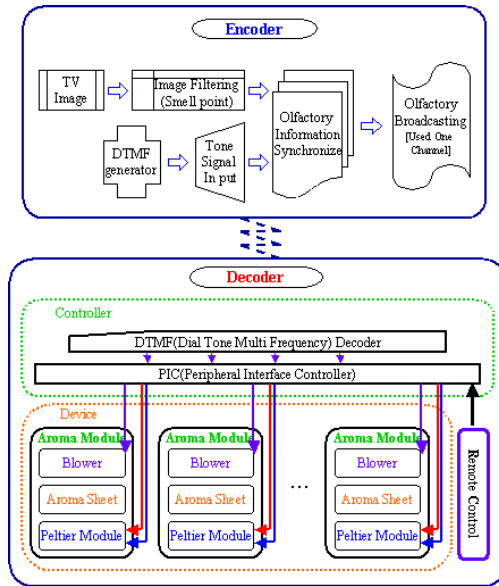


Fig. 1. Structure of FragrantMemories

## 2.1 Olfactory Display Controller(Aroma Module Controller)

The signal that controls the olfactory display is synchronized with the audiovisual content, which is produced beforehand. In detail, an aromatic information (signal) is embedded as sound data in one of the unused sound channels. Hence, there is no need to reconstruct existing AV equipment, and the olfactory display can be added as set-up box. Moreover, because the synchronized information is merely sound information, this information can be played back and recorded very easily, even in a home environment. This results in very inexpensive and convenient olfactory information transmitted along with audiovisual content, and it can be easily implemented in the home. We can even produce home videos with embedded aroma signals and enjoy the fragrance. The olfactory display controller uses DTMF (Dual-Tone Multi-Frequency) signals [6]. Different signals can be used for different fragrances; therefore, signals 1 to 9 would correspond to nine different fragrant sources. For example, at the place where the 2nd signal is recorded, the smell of the corresponding 2nd Aroma Sheet is released. The signal numbering "0" suspends the release of all fragrances. The signals used to release and suspend smells can be transmitted by using one channel among 5 different channels such as in the 5.1 multi-channel system. For the present experiment, we used the channel of a stereo as a signal to transmit smell. We can record these signals on an empty sound track. At the time of playback, the DTMF signals recorded on the sound tracks do not give any output from the speaker as sound data, but instead they give input to the olfactory display controller. The olfactory display controller decodes the DTMF signals and directs the olfactory display to the specified numbers. Figure 2 shows the synchronizing of the olfactory information and Figure 3 shows the prototype of the Aroma Module Controller.

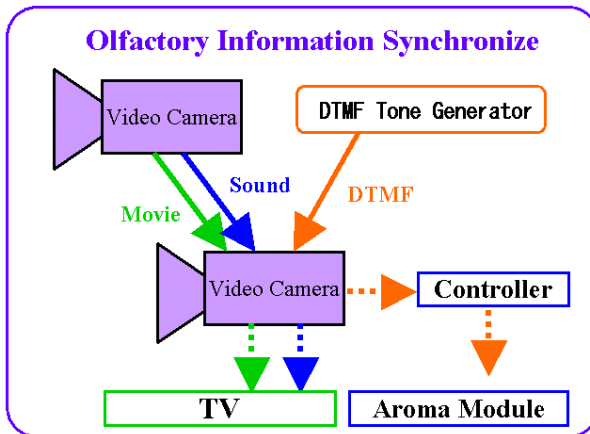


Fig. 2. Synchronizing Olfactory Information

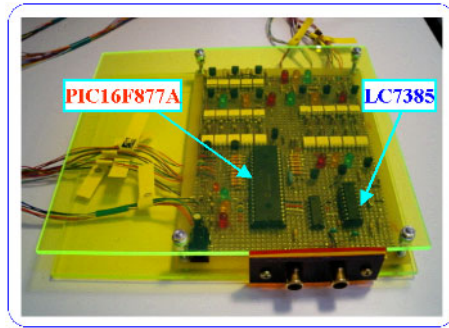


Fig. 3. Aroma Module Controller

### 2.2 Olfactory Display(Aroma Module)

The olfactory display (Aroma Module) consists of three different components: the Peltier module, which helps to direct the action of the reversible conversion between electric energy and heat energy; the Blower, which has advantage to resists high air flow; and the Physical Valve, which works as an open-and-close switch between the Peltier module and blower. Figure 4 shows the structure of the olfactory display. The Peltier module helps to heat and cool the Aroma Sheet according to the assigned signal. The diffused fragrance spreads and reaches users through the blower. We used a very small blower that was very quiet. In Figure 5, we present the prototype of the Aroma Module.

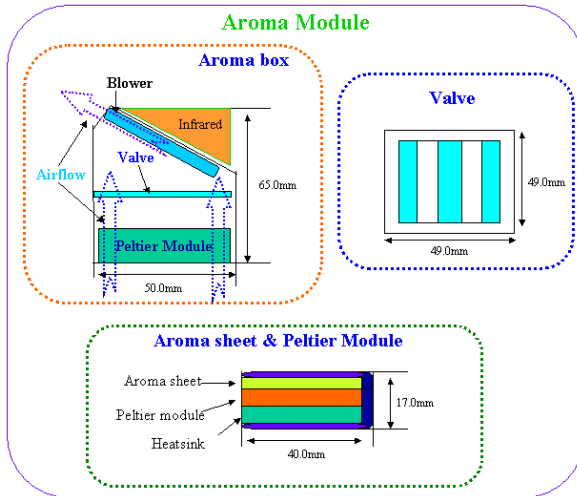


Fig. 4. Structure of Olfactory Display

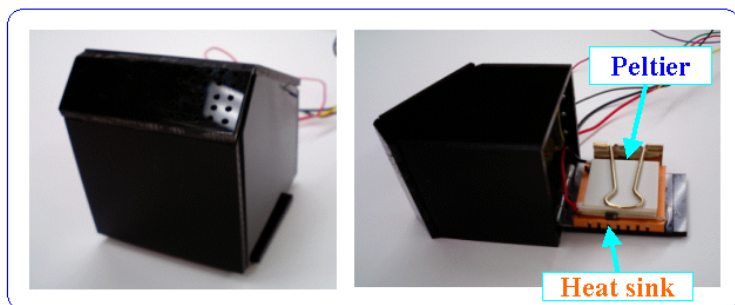


Fig. 5. Prototype of Aroma Module

### 2.3 Non-liquid Aromatic Sheet(Aroma Sheet)

Most previous research soaked liquid aromatic molecules directly in a sponge cartridge as the medium to diffuse fragrance. Economical problems and questions related to troublesome operation and complicated maintenance arose. To eliminate those problems and questions, we adopted a different approach by using non-liquid aromatic components, which led to the development of the Aroma Sheet. The Aroma Sheet has the advantages of mobility and ease of handling. Maintenance is also very convenient. We can bind the Aroma Sheet in a TV guidebook and distribute it to many households in a very effective manner. Moreover, because we generally cannot use any kind of smell for more than six months, as far as the Aroma Sheet in households is concerned, it is just a sheet and there is not much to lose if we are unable to use up the sheet within six months. Controlled release is considered an essential technology in developing the Aroma Sheet. This technology now spans many fields and includes pharmaceutical, food and agricultural applications for pesticides, cosmetics, and household products. For example, the cellulose derivative in medicines helps to control the release of aqueous solutions of substances [7]. We tested natural molecular agars such as konyaku and gelatin, and our results shows that gelatin is suitable for implementation in our research. Regarding the size of the Aroma Sheet, it is natural to assume that, if the size is big it contains more aroma ingredients and hence the fragrance is strong. But for the Aroma Module, a reasonable sheet size is about  $30 \times 30 \times 3$  mm. Figure 6 presents the structure of an Aroma Sheet.

## 3 Editing FragrantMemories

To enjoy synchronized olfactory information, DTMF signals that go to the Aroma Module should be arranged according to the audiovisual contents and they should be embedded as audible signals. As a procedure to edit audiovisual content, the system will send stop signals in all odorless sections. In sections with fragrances, the system sends fragrance-generating signals continuously. This process is recorded beforehand. With this process, even if the contents are re-wound or fast-forwarded, or even started from the middle, the system can stop

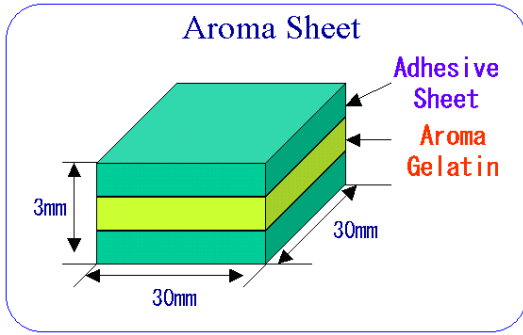


Fig. 6. Structure and figure of Aroma Sheet

or generate the aroma signals accordingly. In this way, by recording DTMF signals according to the content, households can easily edit and play back images and sound with the correct fragrances. In Figure 7 we present an example of recorded DTMF signals.

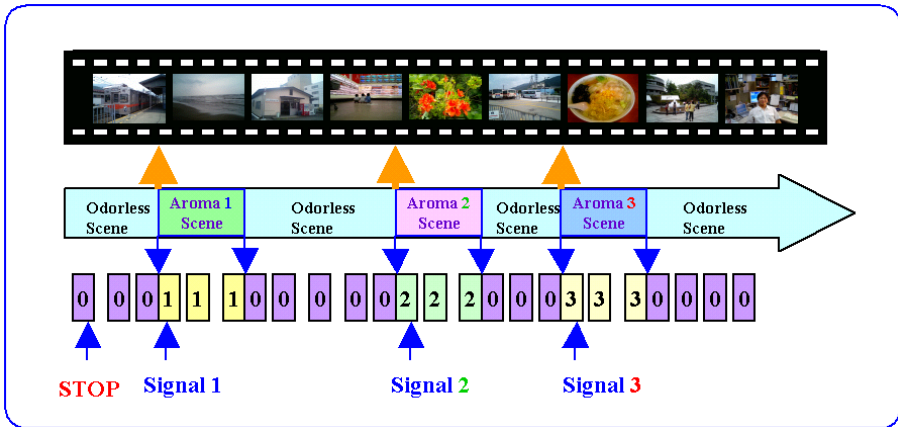
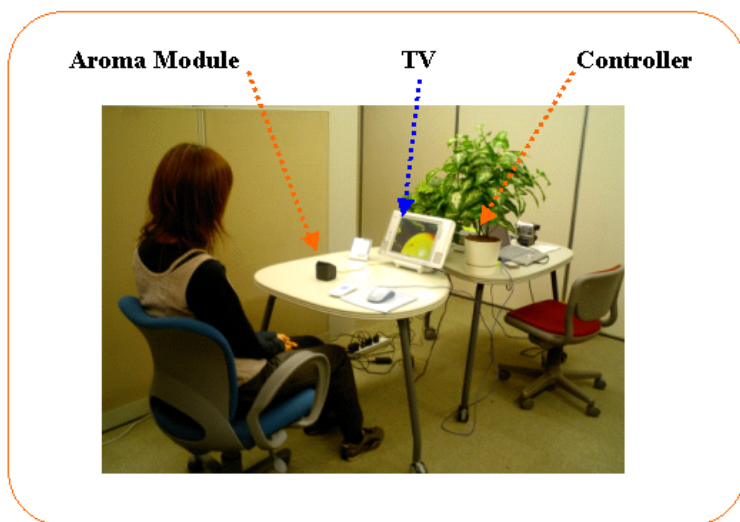


Fig. 7. Recording DTMF signals

## 4 Experiment and Consideration

To evaluate our present work, we conducted a test. Twenty-five individuals were used as participants. One Aroma Module and one Aroma Sheet containing the fragrance of grapefruit were used to test the efficiency of FragrantMemories. We considered a common household as a suitable environment to conduct the FragrantMemories experiment. Figure 8 presents the environment of the experiment.



**Fig. 8.** Environment of the experiment

The sitting environment was arranged beforehand. All participants were asked to sit down and watch three different audiovisual recordings for six minutes. After they finished, they were asked to answer a questionnaire. For the evaluation process, we adopted the 9-grade method usually used by scent manufacturing companies [8]. At first, we evaluated the differences between existing TV broadcasting and audiovisual olfactory broadcasting. FragrantMemories can be operated automatically (automatic diffusion of fragrance in accordance with audiovisual content) or manually (diffusion of fragrance in accordance with a mouse click), and so we could also verify a suitable process to manipulate the system. From the results of the questionnaires we learned that the necessity and demand of audiovisual olfactory broadcasting is rising. But, the manipulation process may differ according to the audiovisual content. For example, automatic manipulation is highly appreciated while watching movies, whereas manual operation is suitable for programs such as quiz shows. In the future, we are planning to conduct several tests to verify the efficiency of FragrantMemories.

## 5 Conclusion and Future Work

Several different studies have been conducted and a variety of systems and displays that transmit smell have been proposed, but suitable systems that can be used in households have not yet been developed. In this work, we proposed and developed the “FragrantMemories” system, which can be implemented very easily at home and is convenient to use. Because the audience of TV broadcasting is larger than that of any other media, we selected TV olfactory broadcasting for our experiment. From the experiment, we verified that, along with existing audiovisual broadcasting, the demand of olfactory broadcasting is also rising. For

the infiltration of FragrantMemories in our daily life, we need to conduct fundamental research to manipulate and control the elements in the Aroma Sheet. Moreover, we need to conduct different tests to verify a suitable distance for using the Aroma Module, and much effort should be put into finding an appropriate size of Aroma Sheet.

## Acknowledgment

Our research is partly supported by the fund from Ministry of Education, Culture, Sports, Science and Technology, Japan, under the name of Cluster for Promotion of Science and Technology in Regional Areas.

## References

1. S. Yokoyama, T. Tanikawak, K. Horota, and M. Hirose *Development of Wearable Olfactory Display(in Japanese)*. Proc.of the Virtual Reality Society of Japan 8th Annual Conference, pp.69-72, 2003
2. Y. Yanagida, H. Noma, N. Tetsutani, and A. Tomono *An Unencumbering, Localized Olfactory Display*. ACM CHI2003 Extended Abstracts, pp.988-989, 2003
3. <http://www.promotool.jp/aroma/index.html>
4. T. Nakamoto, Y. Nakahira, H. Hiramatsu and T. Morizumi *Odor recorder using active odor sensing system*. Sensors and Actuators B, 76(2001) 465.
5. <http://www.tec-tsuji.com/wellness/kaori/index.html>
6. Tho Nguyen and Linga G. Bushnell *Feasibility Study of DTMF Communications for Robots* UWEETR-2004-0013
7. CMC Publisher *Technology of Controlled Release(in Japanese)*. CMC Publishing Co.,Ltd., 2003
8. S. Takagi, T. Shibuya *Science of Smell [Nioi no Kagaku] (in Japanese)*. Asakura Publisher, 1989



# A Divergent-Style Learning Support Tool for English Learners Using a Thesaurus Diagram

Chie Shimodaira<sup>1</sup>, Hiroshi Shimodaira<sup>2</sup>, and Susumu Kunifuji<sup>1</sup>

<sup>1</sup> School of Knowledge Science  
Japan Advanced Institute of Science and Technology  
Asahidai, Nomi, Ishikawa, 923-1292 Japan  
csim@jaist.ac.jp

<sup>2</sup> Centre for Speech Technology Research  
School of Informatics, The University of Edinburgh  
2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom  
h.shimodaira@ed.ac.uk

**Abstract.** This paper proposes an English learning support tool which provides users with divergent information to find the right words and expressions. In contrast to a number of software tools for English translation and composition, the proposed tool is designed to give users not only the right answer to the user's query but also a lot of words and examples which are relevant to the query. Based on the lexical information provided by the lexical database, WordNet, the proposed tool provides users with a thesaurus diagram, in which synonym sets and relation links are presented in multiple windows to help users to choose adequate words and understand similarities and differences between words. Subjective experiments are carried out to evaluate the system.

## 1 Introduction

Language learning with the use of computers has been gaining greater attention due to the increasing demand for English learning as a foreign language (EFL). Some of such examples are, computer assisted language learning (CALL), automatic language translation, and translation supporting tools, many of which are made with free use of latest natural language processing techniques. Moreover, various types of teaching materials, computer software for English learners have been developed, and they are widely available at shops.

In spite of the availability of a large number of software and hardware mentioned above, it is interesting to see that English dictionaries and thesauruses including the electronic ones are still widely used and maybe the most popular. There seems to be some reasons for this: (i) dictionaries are handy, meaning they are portable and they contain richer information than software tools; (ii) software tools have limited uses, i.e. they are designed to specific purposes, whereas dictionaries can be used in various ways. In addition, it is believed that dictionaries help learners to improve their English skills.

Dr. Mochizuki reported his empirical findings through his experience of teaching English [1]:

- Rich vocabulary can not be acquired without consulting dictionaries.
- Teaching English words by means of the learner’s native language does not help a lot.
- Words should be learnt with the contexts.

He also remarked that it would be not until the learner has learnt the pronunciation, spelling, meaning, concept, associated words, grammar, collocation, frequencies and so on that he/she really understands the word.

When we compose documents in English with the aid of dictionaries, we try to choose the right words from numbers of possible candidates and examples in order to reflect the meaning which we want to convey actually. This process of seeking and selecting the right words is not straightforward but rather of trial and error, because it involves a number of backtracking. One of the great benefits with the process is that we can learn not only the target word or phrase in English but also other expressions intentionally or unintentionally in the process.

In the present study, we at first hypothesise that learners improve their language skills through the two types of activities, *convergent-style* and *divergent-style* learning. In the former activity, learners try to narrow down the possible candidates to find the right words and expressions, i.e. the answers, by the aid of various types of resources available. On the other hand, in the latter activity, learners explore various resources not only to have a better idea of the target word they are interested in but also to acquire other words, knowledge and information that are related to the word.

Dictionaries allow learners both types of the activities, whereas most of the learning support tools allow the convergent-style activity only. These tools are designed to show the specific information for users query ignoring other additional information as noises, which might have been useful for learners to improve their skills.

Based on the hypothesis, we predict that learning support tools which enable divergent-style learning and provide users with not only the target word but also other various information relevant to the word would be useful for English learners.

As the first step to develop such a support tool, the present study at first carries out behaviour analysis of English learners to identify functions which are required for the tool.

This paper is organised with four sections. The next section describes the behaviour analysis and system design, and the section three reports experimental evaluations. The final section is devoted to the conclusions.

## 2 Behaviour Analysis and System Design

Behaviour patterns of English learners were collected at an English language school in Scotland. A class of daily English composition was chosen for the analysis. In the class, there were twelve students, who were studying EFL at the school. The class was divided into three groups according to their English skills: A (beginner), B (elementary), and C (intermediate). Each group consists of four students.

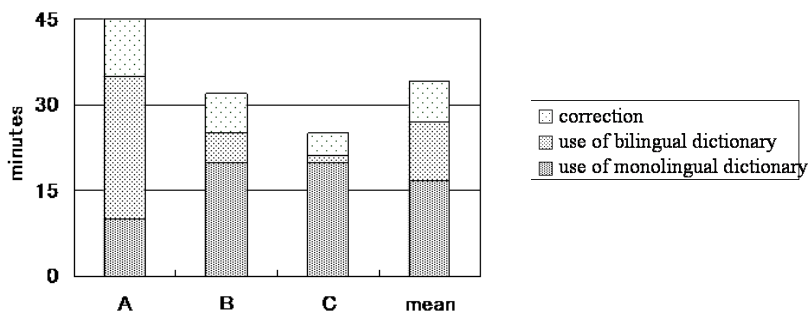


Fig. 1. Behaviour patterns in English compositions

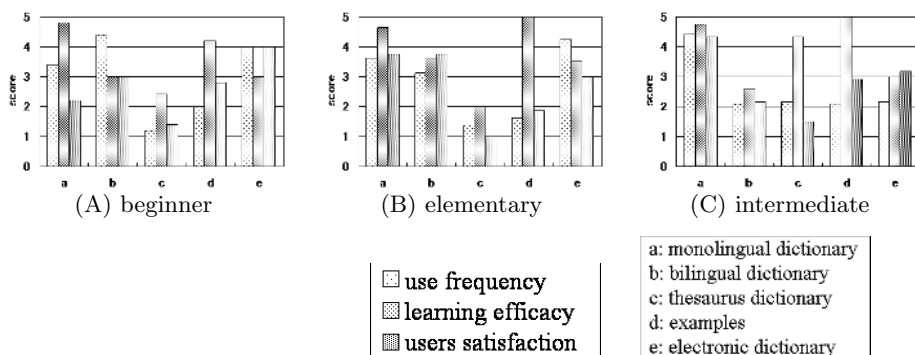


Fig. 2. Subjective evaluation in terms of dictionary types

The students were asked to write a formal letter of about 300 words in length within two and a half hours. During the task, each student can ask a teacher for advice twice. Students are allowed to use any paper dictionaries to complete their task only after they are advised by the teacher twice.

Fig. 1 shows how long each group spent with respect to the three behaviour patterns: (i) correcting their compositions, (ii) consulting bilingual dictionaries (i.e. English - native-language), and (iii) consulting monolingual (English) dictionaries. It can be seen from the figure that monolingual dictionaries are more frequently used than bilingual dictionaries in the higher classes. In addition, further analysis revealed that the students in group C (intermediate) used dictionaries mainly to examine the phrasal verbs, words, and prepositions which they used in the current texts. They also used a thesaurus to find more suitable words.

After the class, we carried out a subjective evaluation on several types of dictionaries in terms of (i) frequency in the use of the dictionary, (ii) usefulness of the dictionary for learning English, and (iii) the degree of satisfaction. Subjects were asked to evaluate each type of dictionary and give a score between 0 (lowest) and 5 (highest). Fig. 2 shows the summary for each group.

It can be found from the figure that the frequencies in the use of a thesaurus and collections of examples increase as the English skills improve. Further survey

which was based on questionnaires to the students revealed that the purposes of using dictionaries change depending on their English skills. Beginners mainly use bilingual dictionaries to find the right words. On the other hand, students of upper classes use dictionaries for the purpose of confirming the words which they already know or finding more suitable words or expressions. This suggests that a divergent style of learning is taking place at this stage.

It should be noted that the users regard a thesaurus potentially useful for language learning, whereas they do not use it that much and they are not much satisfied with it. This is because that a thesaurus does not provide users with detailed information on each word, and users are forced to consult an English dictionary to get further information.

Some of the crucial drawbacks of existing dictionaries can be summarised as follows:

- Users are forced to repeat searching for a word, keeping the search result in mind and backtracking until a suitable expression is found.
- It is not easy for users to compare similar words and find differences between them because it's difficult to see more than one entry at the same time.
- There are not many examples given.

If we could develop a support system which resolves the above problems without losing the merits of paper dictionaries and thesauruses, such a system is expected to be more useful and effective for language learning than the paper ones.

As the first step, the present study focuses on enhancing a thesaurus by solving some of the problems above. To be specific, the following functions are to be implemented.

1. Provides a thesaurus diagram, in which more than one synonym set can be displayed at a same time in multiple windows and collocations are shown as the links between collocated words if existing.
2. Presents examples which are obtained from an internet search engine.
3. Provides a clipboard window to take notes of the search results.

### 3 System Implementation

In order to implement the first function described in the previous section, we've employed the lexical database, WordNet [2], in which English nouns, verbs, adjectives and adverbs are organised into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets.

In the present system, words which belong to a same synonym set are graphically shown in a circular window (*synonym-set window*, here after) by means of a magnetic spring model [3]. Users are allowed to have a look at more than one synonym set at a same time. In such a case, each synonym-set window displays a different synonym set from each other.

Fig. 3 shows a screenshot of the system, where a multi-word query "invade violate infringe law" was given in the query window, which is just under the

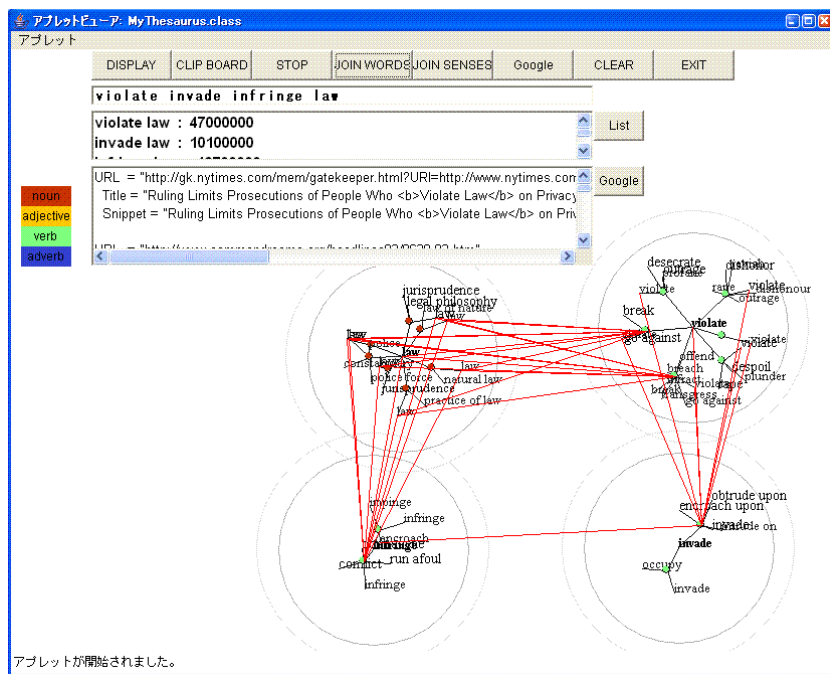


Fig. 3. A screenshot of the proposed system

function buttons located at the the top part of the screen. Since the query consists of four words in this case, there are four synonym-set windows displayed, each of which corresponds to each word in the query. Fig. 4 displays an enlarged view of the synonym-set window for word “law”. In the synonym-set window, each non-terminal (green) node denotes a “sense” of the word, i.e. a meaning of the word. When the word has more than one meaning, more than one sense node appears in the window. By clicking a sense node with the mouse left-button, a small rectangular window appears to show a brief description of the word and examples in that context. The contents of the window can be stored in a “clipboard”, whose window is displayed when the “CLIP BOARD” button in the top menu is pressed.

An user interface for manipulating the windows has been implemented as well: users can move, rotate or delete a diagram and change its size for a better view.

When more than one synonym-set window is displayed, users can check whether some collocations exist between the synonym sets by pushing the “Join Senses” button in the top menu. Collocations are shown by red straight lines in Fig. 3.

The second function, i.e. showing examples retrieved from an internet search engines, was implemented using the Google search engine. The whole-set or subset of the query can be employed as the key words for the internet search. To this end, “List” button is used to select words in the query, and the words selected are shown in the key-word window, which is located just under the query window

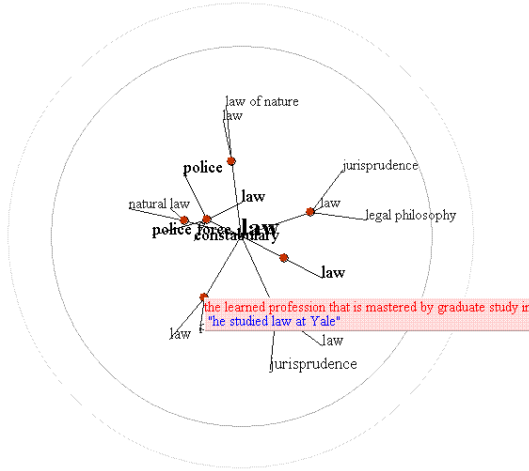


Fig. 4. A screenshot of the synonym-set window for word “law”

(Fig. 3). “Google” button is used to invoke the search engine for the key words, and the search result is shown in the Google search window, which is just under the key-word window. At the same time, the hit count of the search is shown in the key-word window.

## 4 Evaluation and Discussions

### 4.1 Evaluation

A preliminary subjective evaluation by ten EFL students was conducted to assess the prototype system. Fig. 5 shows the proportions of target words and non-target words which were consulted by the users through the experiment, where a target word denotes the one which the user intended to search, whereas non-target words are the others. It can be seen from the figure that the users of the proposed system checked the meanings and examples of not only the target words but also the non-target words. This suggests that divergent style of activity took place.

Another subjective evaluation using the same EFL students was carried out to compare user’s preferences for (i) paper dictionary, (ii) translator (Google), (iii) Visual Thesaurus 3 [4], (iv) Expert System (IdeaFisher) [5], and (v) Thesaurus Diagram (proposed system). As is shown in Fig. 6, when compared in terms of usefulness, the proposed system showed higher score than the other systems and comparable score to paper dictionary.

### 4.2 Related Works

There are two systems which are relevant to the proposed system. One is ‘Visual Thesaurus’ [4] and the other is “Visual Browser” [6].

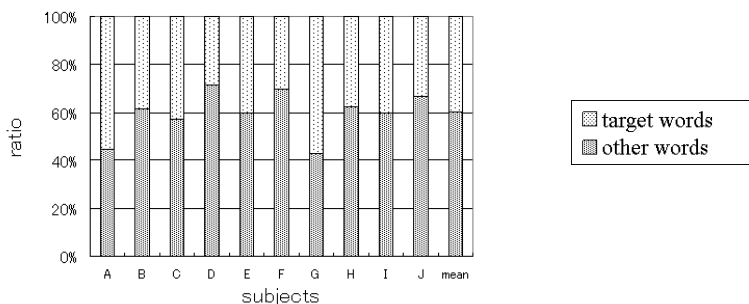


Fig. 5. Proportions of target and non-target words

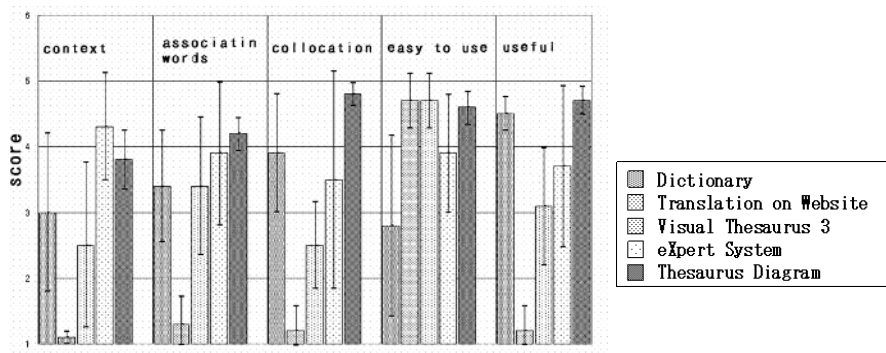


Fig. 6. Subjective evaluation on different systems

The Visual Thesaurus has implemented similar functions with those of 1 and 2 described in section three, however, one of the crucial differences is that Visual Thesaurus accepts only a single-word query at a time while our system accepts a query of multiple words. The multiple word query helps users to check whether the combination of words in a sentence is adequate or not. It also helps users to find more suitable combination of words by showing collocations by lines between words each of which belongs to the synonym set of each word of a user's query.

The Visual Browser is a Java application for visualising the data in RDF (Resource Description Framework) schema [7]. The main difference between this tool and our tool resides in the fact that the former is designed for general applications, whereas the latter is designed for the specific application. As the result, although the Visual Browser can visualise more than one synonym-set of the WordNet, it just displays the synonym-sets which are linked with each other. On the other hand, in our system, it is the user who specifies what synonym-sets to display, and those sets which are not linked with each other can be displayed as well. The user can check whether the two words has some relationships or not. This user initiative interface provides the user with more flexible use of the tool than the Visual Browser.

## 5 Conclusions

The present study aimed to develop a software tool which supports a divergent style of English learning. Subjective evaluation of the prototype system revealed that users used the system not only to find or check the right words but also to find other words which are relevant to the right words. It was also found that the clipboard function was supported by the users since it reduces the user's load of searching the right words.

However, several problems with the system have been pointed out: it is not easy for the user to find the right word if the user does not know the word at all; it is difficult to identify each synonyms in the window when too many synonyms are displayed at the same time; users sometimes get lost when they have followed relevant words too many times. Future works include attempts to resolve these problems and further evaluation of the system.

## References

1. Mochizuki, M., Aizawa, K., Tono, Y.: Manual for English Vocabulary Teaching. Taisyukan (1999) (in Japanese).
2. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* **3** (1990) 235–312
3. Sugiyama, K., Misue, K.: Graph drawing by the magnetic spring model. *Journal of Visual Languages and Computing* **6** (1995) 217–231
4. Visual Thesaurus: <http://www.visualthesaurus.com/>.
5. IdeaFisher: <http://www.ideacenter.com/>.
6. Visual Browser: [http://nlp.fi.muni.cz/projekty/vizualni\\_lexikon/](http://nlp.fi.muni.cz/projekty/vizualni_lexikon/).
7. RDF schema: <http://www.w3.org/TR/rdf-schema/>.



# Full Fuzzy-Logic-Based Vector Control for Permanent Magnet Synchronous Motors

Jae-Sung Yu<sup>1</sup>, Byoung-Kuk Lee<sup>1</sup>, Chung-Yuen Won<sup>1</sup>, and Dong-Wook Yoo<sup>2</sup>

<sup>1</sup> Department of Information & Communication Engineering, SungKyunKwan University, 300, chunchun-dong, Jangan-gu, Suwon, Kyunggi-do, 440-746, Korea  
jaesung75@skku.edu,  
bkleeskk@skku.edu,  
won@yurim.skku.ac.kr  
<http://icc.skku.edu/~won>

<sup>2</sup> Power Electronics Group, Korea Electrotechnology Research Institute, P.O.BOX 20, Changwon, 641-120, Korea  
dwyoo@keri.re.kr

**Abstract.** This paper proposes a full fuzzy-logic-based vector control for a permanent-magnet synchronous motor (PMSM). The high-performance of the proposed fuzzy logic control (FLC)-based PMSM drive are investigated and compared with the conventional proportional-integral (PI) controller at different conditions, such as step change in command speed and load, and etc. In the experimental results, the FLC is employed in the speed and current controller. The comparative experimental results show that the FLC is more robust and, hence, found to be a suitable replacement of the conventional PI controller for the high-performance drive system.

## 1 Introduction

The induction motor has been used for the conventional high-performance drive system because of its simple and rugged construction. However, an induction motor has limitations on compactness and torque due to power factor and efficiency decrease as the number of pole increases. Recently, PMSM is more used in the drive system due to the advantage that the PMSM is smaller and more compact than the induction motor.

In this paper, vector control algorithm is adopted for the high performance control of PMSM. Conventionally, in speed and current controller, PI controller is employed. However, the fixed-gain controllers are very sensitive to parameter variations, load disturbances, and etc. Thus, the controller parameters have to be continually adapted. This problem can be solved by several adaptive control techniques, such as Model Reference Adaptive Control (MRAC) [1], Sliding-Mode Control (SMC) [2], Variable Structure Control (VSC) [3], self-tuning PI controllers [4], and etc. The design of all of the above controllers depends on the exact mathematical model of the system. However, it is often difficult to develop an accurate system mathematical model due to unknown load variation, unknown and unavoidable parameter variations due to saturation, temperature variations, and system disturbances. In order to overcome the above problems, recently, the fuzzy logic controller is employed for motor control purpose [5],[6],[7].

FLC is basically nonlinear and adaptive controller, which gives robust performance for a linear or nonlinear plant with parameter variations [8].

As it mentioned above, even if a plant model is well-known, there may be parameter variation problems, an accurate mathematical model is difficult to find. However the FLC system essentially embeds the experience and intuition of a human plant operator, and sometimes those of a designer and/or researcher of a plant. Recently, the high-speed digital signal processor (DSP) has been widely used for high performance, which makes it possible to implement more advanced control algorithms like FLC [9].

This paper presents a drive system using FLC both in the speed controller and current controller. The complete vector control scheme of drive systems incorporating the FLC has been successfully implemented in real time using digital signal processor (DSP) TMS320VC33. The performances of the proposed drive system have also been compared with the conventional PI controller in the experiment.

## 2 Structure of FLC

Fig. 1 shows the block diagram of the FLC for the PMSM. The FLC is divided into four modules : fuzzifier, knowledge base, fuzzy inference engine and defuzzifier. This chapter explains the structure of speed fuzzy controller. The structure of current controller is similar to speed fuzzy controller.

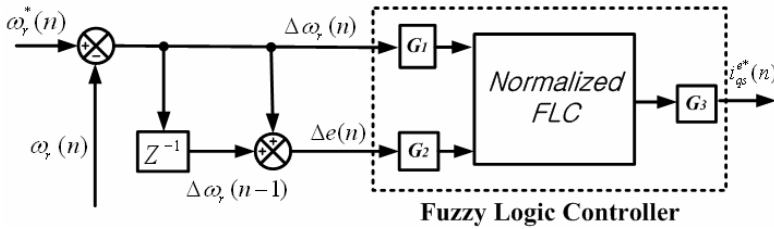


Fig. 1. Block diagram of FLC

### 2.1 Input and Output Variables

To compose the FLC for the PMSM, first of all, input and output variables of the FLC have to be determined. In this paper, in the speed controller, the speed error and the rate of change of the speed error are considered as the input crisp variables, which are defined as,

$$\Delta e(n) = \Delta\omega_r(n) - \Delta\omega_r(n-1) \tag{1}$$

$$\Delta\omega_r(n) = \omega_r^*(n) - \omega_r(n) \tag{2}$$

The output of the FLC is the torque-producing current and is defined as,

$$i_{qs}^*(n) = i_{qs}^*(n-1) + \eta \cdot \Delta i_{qs}^*(n) \tag{3}$$

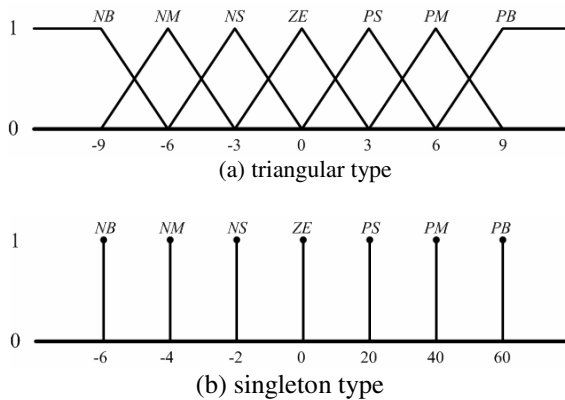
where,  $\Delta i_{qs}^e(n)$  is the inferred change of the torque-producing current by the FLC at the n-th sampling time, and  $\eta$  is the gain factor of the FLC.

In the current controller, the inputs variables are the current error and the rate of change of the current error and the output variable is voltage reference.

### 2.2 Membership Functions

In the FLC, input and output variables are expressed by linguistic variables, which are represented by a membership function. Linguistic variables are defined as fuzzy subsets.

In this paper, seven fuzzy subsets are chosen for input and output variables. : Negative Big (NB), Negative Medium (NM), Negative Small (NS), Zero (ZE), Positive Small (PS), Positive Medium (PM), Positive Big (PB). In order to compose above membership function, the membership function of triangular type is adopted for the error input of FLC and the singleton type is adopted for the output of FLC. Fig. 2 shows the membership function adopted in this paper.



**Fig. 2.** Membership functions. (a) is the input membership function with respect to error and change error of FLC, respectively. (b) is the output membership function of FLC.

### 2.3 Derivation of Control Rules

Control rules are derived by the experience or the knowledge of experts. The fuzzy rules are in the form

$$R_i : \text{IF } \Delta \omega_r \text{ is } A_i \text{ and } \Delta e \text{ is } B_i , \text{ THEN } i_{qs}^* \text{ is } C_i$$

where,  $A_i$  ,  $B_i$  are the fuzzy subset, and  $C_i$  is a fuzzy singleton. The derivation of the fuzzy control is based on the following criteria :

- 1) When the speed of PMSM is far from the reference, the change of torque-producing current must be large so as to bring the speed to the reference.

- 2) When the speed of PMSM is approaching the reference, the small change of torque-producing current is necessary.
- 3) When the speed of PMSM is near the reference and is approaching it rapidly, torque-producing current must be kept constant so as to prevent overshoot.
- 4) When the speed of PMSM reaches the reference and speed is still changing, torque-producing current must be changed a little bit to prevent the output from moving away.
- 5) When the speed of PMSM reaches the reference and the speed is in steady state, torque-producing current remains unchanged.
- 6) When the speed of PMSM is above the reference, the sign of change of torque-producing current must be negative.

According to above criteria, the rule table shown in Table 1 is derived.

**Table 1.** Fuzzy-Rule-Based Matrix in the speed controller

$\Delta i_{qs}^{e*}(n)$		$\Delta \omega_r(n)$						
		NB	NM	NS	ZE	PS	PM	PB
$\Delta e(n)$	NB	-60	-60	-60	-60	-40	-20	0
	NM	-60	-60	-60	-40	-20	0	20
	NS	-60	-60	-40	-20	0	20	40
	ZE	-60	-40	-20	0	20	40	60
	PS	-40	-20	0	20	40	60	60
	PM	-20	0	20	40	60	60	60
	PB	0	20	40	60	60	60	60

Input linguistic variables are converted to output singleton variables through the fuzzy inference engine and the rule base. The inference result of each rule consists of two parts, a weighting factor and the degree of the torque-producing current,  $C_i$ . For example, if  $\Delta \omega_r$  is 35[rad/sec] and  $\Delta e$  is 12[rad/sec<sup>2</sup>], respectively,  $\Delta \omega_r$  belongs to PS, PM and  $\Delta e$  belongs to ZE, PS. Accordingly, following four rules are possible,

- 1) IF  $\Delta \omega_r$  is PS and  $\Delta e$  is ZE, THEN  $\Delta i_{qs}^e$  is 20[A].
- 2) IF  $\Delta \omega_r$  is PS and  $\Delta e$  is PS, THEN  $\Delta i_{qs}^e$  is 40[A].
- 3) IF  $\Delta \omega_r$  is PM and  $\Delta e$  is ZE, THEN  $\Delta i_{qs}^e$  is 40[A].
- 4) IF  $\Delta \omega_r$  is PM and  $\Delta e$  is PS, THEN  $\Delta i_{qs}^e$  is 60[A].

And, by min operation, the weighting factor is obtained.

$$\omega_i = \min\{\mu_e(\Delta \omega_r), \mu_{ce}(\Delta e)\} \tag{4}$$

Finally, the inference results are obtained by following equation.

$$z_i = \omega_i C_i \tag{5}$$

### 2.4 Defuzzification

The inferred results should be converted to the Crisp set. In this paper, the center of gravity defuzzification is used. The output function is given as[11],

$$\Delta i_{qs}^e = \frac{\sum_{i=1}^4 \omega_i C_i}{\sum_{i=1}^4 \omega_i} \tag{6}$$

According to (8),

$$\Delta i_{qs}^e = \frac{20 \times 0.4 + 40 \times 0.6 + 40 \times 0.166 + 60 \times 0.166}{0.4 + 0.6 + 0.166 + 0.166} = 36.486[A] \tag{7}$$

and the change of the torque-producing current at this sampling time is

$$i_{qs}^{e*}(n) = i_{qs}^{e*}(n-1) + \eta \cdot \Delta i_{qs}^e(n) \tag{8}$$

### 3 Drive Systems Using Fuzzy Logic Controller

The proposed method is confirmed by experimental results. Experiments were performed on a motor-generator set with a 13.3- kW PMSM and a 16.8- kW dc motor.

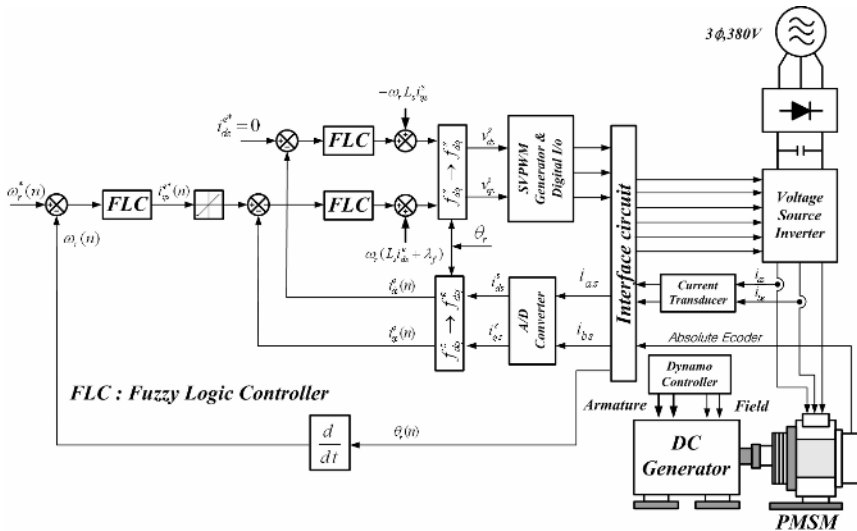


Fig. 3. Control block diagram of PMSM using FLC

Fig. 3 shows a control block diagram of PMSM using FLC in the vector control system. This system has a two control loops; one is for speed controller, the other for current controller. The fuzzy inference of the speed controller make a torque reference through speed error and the fuzzy inference of the current controller make a voltage reference through an output of the speed controller. The SVPWM is applied to generate PWM signals, which will fire the power semiconductor devices of the three-phase inverter to produce the actual voltages to the motor.

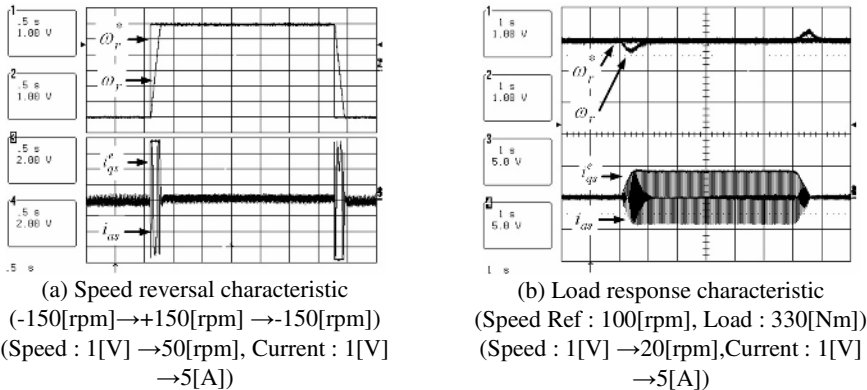
### 4 Experimental Results

Figs. 4 (a) and (b) show the experimental speed and load responses of the drive system using the PI controller, respectively. Figs. 5(a) and (b) show the experimental results with respect to speed and load response of the drive system using the FLC, respectively.

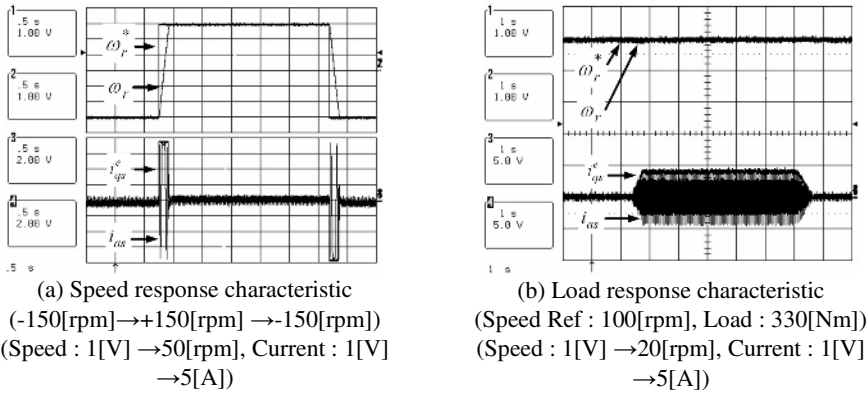
Fig. 4(a) shows the response characteristic when reference is changed from -150[rpm] to 150 [rpm] and back to -150[rpm]. It is spent about 200[ms] to reach to the steady state. As the transient state, PMSM generates a maximum torque. Fig. 4(b) shows the response characteristics of PMSM when reference is 100[rpm] and 330[Nm] (50% of rated torque) is applied in arbitrary time. Load is applied during about 6[sec]. It is confirm that a speed error is about 10[rpm] and 0.7[sec] is spent to reach to the steady state.

The conditions of Figs. 5(a) and (b) are the same as Figs. 4(a) and (b). Fig. 5(a) show that it is spent 200[ms] to reach to the steady state which is the same as result of PI controller.

In Fig. 5(b), there is nearly no speed error. The comparative experimental results of Fig. 4(b) and Fig. 5(b) show that the FLC is more robust to a load torque.



**Fig. 4.** Response characteristics of PI controller in respect of speed reversal and sudden load change at 100[rpm]



**Fig. 5.** Response characteristics of FLC in respect of speed reversal and sudden load change at 100[rpm]

## 5 Conclusion

A FLC-based vector control of PMSM has been presented in this paper. Since exact system parameters are not required in the implementation of the proposed controller, the performance of the drive system is robust, stable, and insensitive to parameters and operating condition variations. In order to verify the superiority of the FLC, a conventional PI-controller-based PMSM drive system has also been experimented.

From the comparative experimental results, it is confirmed that the proposed FLC is more than the PI controller in case of load disturbance.

## References

1. H. Sugimoto and S. Tamai.: Secondary resistance identification of an induction motor applied model reference adaptive system and its characteristics. *IEEE Trans. Industry Applications.*, Vol. IA-23, pp. 296–303, Mar./Apr.1987.
2. C. Y. Won and B. K. Bose.: An induction motor servo system with improved sliding mode control, in *Proc. IEEE IECON'92*, pp. 60–66
3. T. L. Chern and Y. C. Wu.: Design of integral variable structure controller and application to electrohydraulic velocity servo systems, in *Proc. Inst. Elect. Eng.*, Vol. 138, no. 5, pp. 439–444, Sept. 1991.
4. J. C. Hung.: Practical industrial control techniques, in *Proc. IEEE IECON'94*, pp. 7–14.
5. I. Miki, N. Nagai, S. Nishiyama, and T. Yamada.: Vector control of induction motor with fuzzy PI controller, in *IEEE IAS Annu. Rec.*, 1992, pp. 464–471.
6. Y. Tang and L. Xu.: Fuzzy logic application for intelligent control of a variable speed drive, *IEEE Trans. Energy Conversion*, Vol. 9, pp. 679–685, Dec. 1994.
7. E. Cerruto, A. Consoli, A. Raciti, and A. Testa.: Fuzzy adaptive vector control of induction motor drives, *IEEE Trans. Power Electron.*, Vol. 12, pp. 1028–1039, Nov. 1997.
8. Bimal K. Bose.: *Modern Power Electronics and AC Drives.*, Prentice Hall PTR, 2002.
9. Y.M.Lee, J.K.Kang, S,K,Sul.: Acceleration Feedback Control Strategy for Improving Riding Quality of Elevator System, *IEEE IAS Conf. Rec.*, pp.1375-1379, 1999.

10. Wing-Chi So, Chi K, Tse, Yim-Shu Lee.: Development of a Fuzzy Logic Controller for DC/DC Converters : Design, Computer Simulation, and Experimental Evaluation, IEEE Trans on Power Electronics, Vol11, no. 1, pp.24-32, Jan.1996.
11. H.T.Nguyen, M.Sugeno, R.Tong, and R.R.Yager.: Theoretical Aspects of Fuzzy Control. New York:Wiley, 1995



# Artificial Intelligent Application to Service Restoration Considering Load Balancing in Distribution Networks

Sang-Yule Choi<sup>1</sup>, Jae-Sang Cha<sup>2</sup>, and Myong-Chul Shin<sup>3</sup>

<sup>1</sup> Dept. of Electronic Engineering, Induk Institute of Technology,  
San 76 Wolgye-dong, Nowon-gu, Seoul, Korea

<sup>2</sup> Dept. of Media Technology, Seoul National University of Technology,  
Seoul, South Korea

<sup>3</sup> School of Electrical and Computer Engineering, Sungkyunkwan University,  
Suwon 440-746, Korea  
mcshin@ yurim.skku.ac.kr

**Abstract.** Service restoration is an emergency control in distribution control centers to restore out-of-service area as soon as possible when a fault occurs in distribution networks. therefore, it requires fast computation time and high quality solutions for load balancing. In this paper. a load balance index and heuristic guided best-first search are proposed for these problem. The proposed algorithm consists of two parts. One is to set up a decision tree to represent the various switching operations available. Another is to identify the most effective the set of switches using proposed search technique and a feeder load balance index. Test results on the KEPCO's 108bus distribution system show that the performance is efficient and robust.

## 1 Introduction

Electric distribution networks maintain radial structure with normally closed sectionalizing switches along a feeder and normally open interfeeder tie switches for proper protection coordination. For every tie switch closed, another sectionalizing switch is opened. Under feeder faulted conditions, switches are used for fault isolation and service restoration. The resulting feeders must remain radial, without any violations of branches loading and voltage limits. Because of these requirements, the problem of service restoration is a very complicated mixed-integer, non-linear optimization problem. Since there are a numbers of switches in a practical distribution networks, the problem appears to be best solved by heuristic search methods. Heuristic approaches do not guarantee optimal solutions, but they lead to sub-optimal solutions that are technically acceptable. Many heuristic algorithms dealing with feeder restoration have been presented.

Taylor et al. [1] proposed a switch exchange type heuristic method to determine the network configuration for overloads, voltage problem, and for load balancing simultaneously. Its solution scheme sets up a decision tree which represents the various switching operations available, and a best-first search and heuristic rules are used to find feasible switching operations. Wu et al. [2] extended the method proposed by Taylor et al by developing explicit exhaustive method that solves the problem of

overloads, phase current unbalance, service-restoration, and maintenance. This method is to set up a feasible switching options tree which represents possible switching options under constraint of radial structure. Evaluation functions and heuristic rules are used to find feasible switching operations.

In this paper, the authors present a heuristic service restoration algorithm considering load balancing based on an effective exhaustive search method. Its main steps have been implemented in two stages. First stage is to set up a sub-tree that was presented by Wu et al. [2]. Second stage is to identify the most effective the set of switches using proposed search technique called “cyclic best-first search” and a feeder load balance index. This procedure favors solutions with feeder load balancing when feeder faults are restored

Numerical calculations are carried out to show the effectiveness of the proposed algorithm.

## 2 Description of Developed Feeder Load Balance Index

When feeder faults are detected, the loads in the isolated feeder section are energized by transferring these load to adjacent feeders. If adjacent feeders are already overloaded, the load must be transferred to another adjacent feeders. Therefore, when loads are transferred, it must be distributed to adjacent feeders whose actual load are less than their projected loads. In this paper, to distribute loads in proportion to feeder nominal capacities, the authors presents feeder load balance index. This index extents heuristic index proposed by Taylor et al. [1] by considering feeder load balance during service restoration. The whole process is as follows.

$$FL_i = FNC_i \times \frac{\sum_{tek} SL_i}{\sum_{ieU} TAC_i} \tag{1}$$

$$LI_i = FL_i - SL_i$$

$$LI_{sum} = |LI_1| + |LI_2| + |LI_3| + \dots + |LI_i|$$

FLi : Projected load of feeder i (MVA)

LIsun : Feeder load balance index

SLi : Actual load in feeder i (MVA)

TACi : Nominal capacities in transformer i (MVA)

FNCi : Nominal capacities in feeder i (MVA)

U : Set of transformer

K : Set of feeder

During service restoration, the object in distributing feeder loadings with respect to their nominal capacities in the proportional manner is to minimize feeder load balance index.

In this paper, the service restoration considering load balance is to find feasible switch pairs for minimizing feeder load balance index with cyclic best first search.

### 3 Solution Algorithm

The proposed search scheme starts by constructing sub-tree that was suggested in Wu et al. [2] in order to decrease searching space, and finding feasible switching operation with a cyclic best-first search and feeder load balance index.

#### 3.1 Constructing the Sub-tree

Under the constraint of the radial structure in the load transfer process, closing a normally open tie switch should follow the opening of a complementary normally closed sectionalizing switch. Therefore, if n tie switches are closed, then n sectionalizing switches has to be opened.

Fig. 1 shows a sample distribution networks [3] consisting of three feeders with three normally opened tie switches and thirteen normally closed sectionalizing switches.

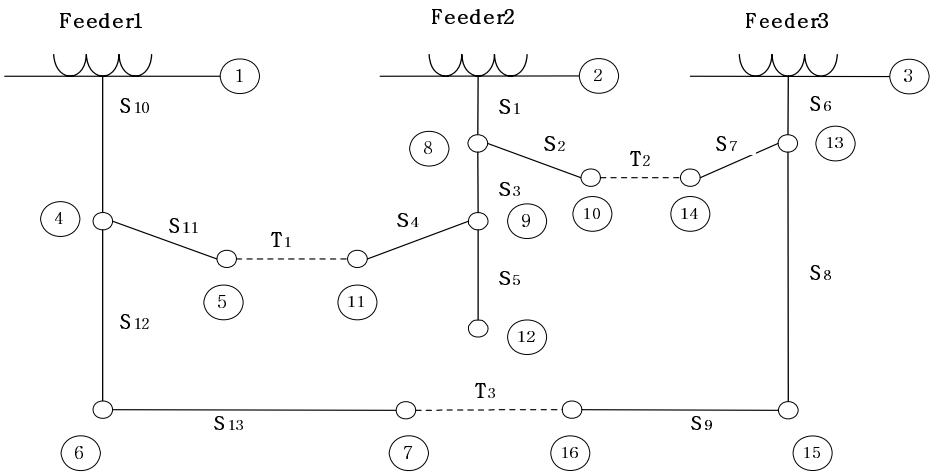


Fig. 1. Three-feeder example system

If feeder section S1 experiencing an fault, then the amount of load on isolated feeder section must be transferred to feeder 1 and/or 3 without creating an overload on either of these feeders. To transfer load at node 11 from feeder 2 to feeder 1, the notation (T1, S4) is used to denote the operation of closing switch T1 and opening switch S4, henceforth.

Feasible (close, open) switching options can be found by searching sectionalizing switches. When each tie switch of the isolated feeder section is closed, a complementary sectionalizing switch to be opened is found by searching from the tie switch, and moving upstream along the faulted feeder to its source, the circuit breaker of the isolated feeder section.

Fig2 shows a searching path for finding feasible switching options when feeder 2 is overloaded.

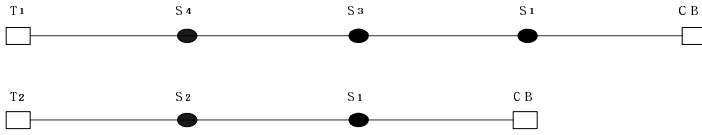


Fig. 2. Main search paths for example system

If the amount of load on isolated feeder section is transferred to only feeder 1, then T1 and either S4, S3 or S1 constitute a switching pair. So feasible switching options are expressed as  $\{(T1, S4), (T1, S3), (T1, S1)\}$ , and one of switching options would be a solution for transferring the isolated feeder section.

Similarly, the amount of load on isolated feeder section may be transferred to feeder 1 and 3 simultaneously by choosing one of following feasible switching options  $\{(T1, S4), (T2, S2)\}$ ,  $\{(T1, S4), (T2, S1)\}$ ,  $\{(T1, S3), (T2, S2)\}$ ,  $\{(T1, S3), (T2, S1)\}$ ,  $\{(T1, S3), (T2, S2)\}$ . But when T1 and T2 are used simultaneously, the switching option  $\{(T1, S1), (T2, S1)\}$  is not a feasible one due to radial structure constraint.

If the results of these feasible options are examined, then the corresponding sub-tree of fig.3 is obtained. In figure.3, both T1 and T2 are tie switches of isolated feeder section and dotted line represents switching options.

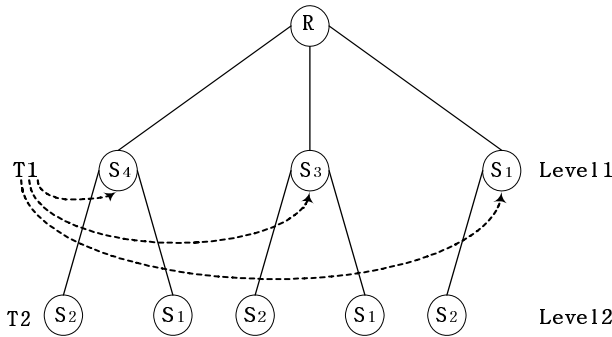


Fig. 3. Sub tree with two backup feeders

### 3.2 Cyclic Best-First Search

An exhaustive search evaluates all feasible switching options of the above sub-tree and it guarantees optimal solution. But this is probably not realizable for large sub-tree because of heavy computation. On the other hand, by using heuristic search, time and effort can be saved by finding reasonable solution promptly. There are usually three heuristic search ways to find optimal(or near-optimal) switching pairs on the above sub-tree : depth-first, breadth-first, best-first. The advantage of the best-first search usually, but not always, yields solution faster than any other heuristic search. But the problem is that it does not always give the optimal solution: unexplored path would

have given an optimal solution. To overcome this defect, the new methodology (so called cyclic best-first search) is presented in this paper. This methodology is based on best-first search. But, by using cyclic methodology, it can usually find more effective solution than best-first search.

As an example for a best-first search, consider fig. 4, where node ① is the start node and node ⑫ is a goal node. Node ① is expanded into its children node ②,③,④,⑤. Since the losses of node ② is less than other nodes, node ② is chosen for expansion. This is continued until a goal node has been found.

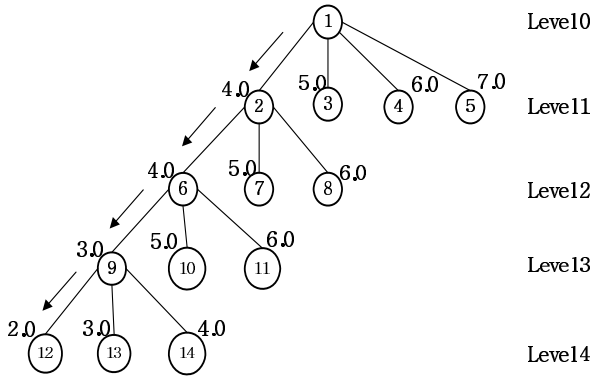


Fig. 4. The first Step of best-first search

In the end, nodes {①,②,⑥,⑨,⑫} are found by best-first search. By always expanding the most likely node, it is possible to get to a goal node or a solution quickly. But this procedure achieves the trade-off between optimality and computational speed. It is possible that unexplored path would have yielded a solution. Therefore, optimality is sacrificed for the sake of increased speed in best-first search. But, in cyclic best-first search, with circulatory reevaluating the unexplored nodes and path, more effective solution for feeder load balance could be found. Although the search space of cyclic best-first search is slightly larger than that of best-first search, the computation difference is negligible due to using heuristic based sub-tree. The cyclic best-first search process is as follows:

First step: nodes are selected by using best-first search.

Second step: Constructing the reversed sub-tree and a search is proceed by using best-first search.

Reversed sub-tree is constructed by reversing levels of sub-tree that was previously constructed. As an example, consider fig. 5, the level-4 of the sub-tree in first step becomes the level-0 of the reversed sub-tree in second step, and the level-3 in first step becomes the level-1 in second step. After reversed sub tree is constructed, a best-first search is used to select near-optimal nodes in a reversed sub-tree.

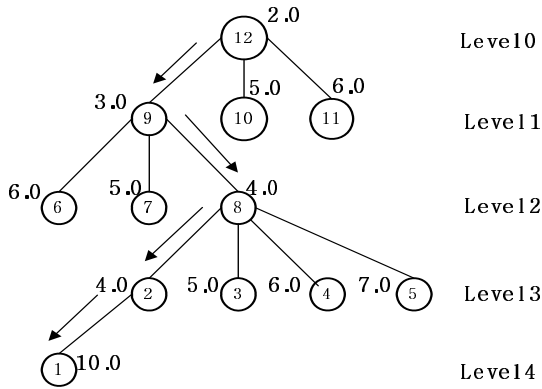


Fig. 5. The second step of best-first search

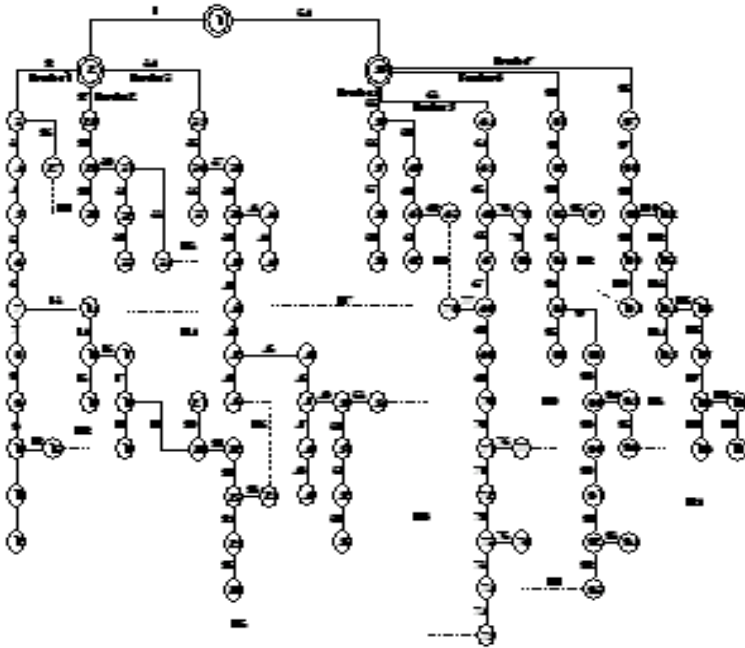


Fig. 6. Initial configuration of 108 bus system

In the second step of best-first search, nodes in each level are evaluated on condition that nodes in the lower levels are already chosen by the first step of best-first search. As an example, node ⑨ is selected on condition that nodes ⑥, ②, ① are already determined from first step. Similarly, node ⑧ in level-2 is selected on condition that nodes ②, ① are already determined from first step:

In the first step of best-first search, the nodes in level-2 are evaluated on condition that node ⑨ in level-3 and node ⑫ in level-4 are not selected by expansion. On the contrary, in second step, nodes ⑫ ⑨ was already selected before evaluating nodes in level-2, and nodes ②,① was also determined from first step. Due to using near-optimal solution from first step, more effective solution can be found in second step. After the second step of best first search, a new nodes {⑫,⑨,⑧,②,①} are selected.

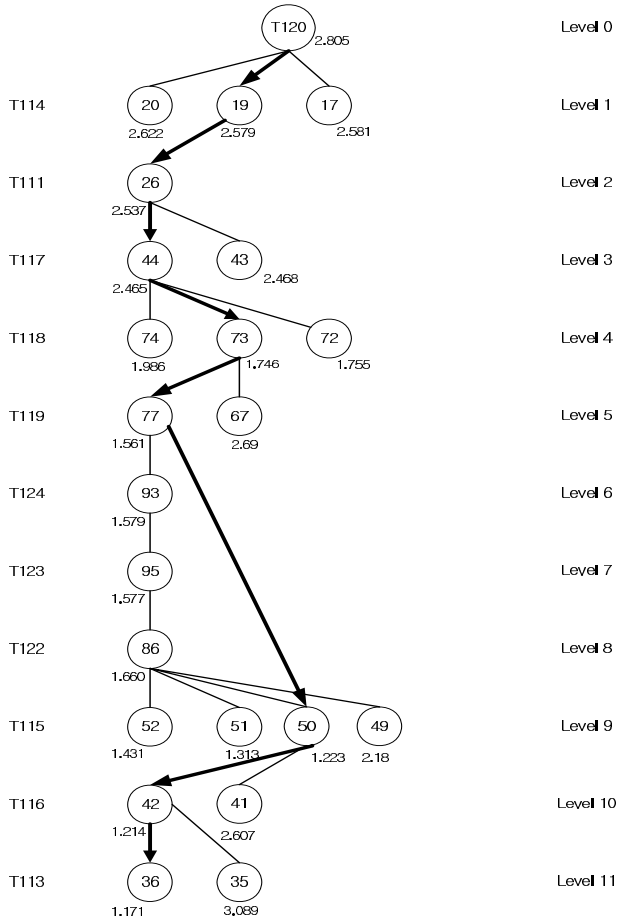


Fig. 7. First step of best-first search for restoration on fault line section 80

### 4 Test Result

The distribution network for KEPCO 108 bus system is used to demonstrate the validity and effectiveness of the proposed algorithm. The network consisting of two feeders with 108 busbars and 14 tie switches as shown in fig.6. The total load are 72.27[MW], 32,78[MVAR] . Table 1 shows initial feeder loadings

**Table 1.** Feeder loadings for 108 bus system

Feeder loadings [MVA]						
Feeder 1	Feeder 2	Feeder 3	Feeder 4	Feeder 5	Feeder 6	Feeder 7
14.47	5.17	13.04	8.38	14.34	9.88	13.80

When feeder fault is detected on section 80, The first step of best-first search for initial sub-tree is shown in fig. 7. Initial sub-tree level is defined by T114, T111, T117, T118, T119, T124, T123, T122, T115, T116, T118 sequentially due to the different voltage across.

In fig. 7, the selected (close, open) switching pair for level 1 is (T114, 19) and feeder balance index is 2.579 for the switching operation. In the process of checking nodes of each level, if checked nodes would increases index then the rest of unchecked nodes are ignored and searching proceeds to next level. By pruning of the most unlikely nodes, this procedure makes it possible to get a solution much faster even if it deep down in the tree. After the first step of best-first search, selected (close, open) switching pairs are {(T114,19), (T111,26), (T117,44), (T118,73), (T119,77), (T115,50), (T116,42), (T113,36)}. This solution seems feasible but it is only locally optimal, because the first step of best-first search dose not examines all the possible nodes. Therefore, it is possible that unexplored path would have presented more feasible solution. Thus, to find more feasible solution, reversed sub-tree is constructed by reversing the level of sub-tree that was constructed in first step. The second step of best-first search is executed in fig 8.

After the second step of best first search, switching pairs{(T114,19), (T111,26), (T117,44), (T118,72), (T119,77), (T115,50), (T116,42), (T113,36)} are selected to minimize feeder balance index.

Comparing feeder loadings before service restoration with those of after service restoration is presented as below table 2 and 3.

**Table 2.** Feeder loadings before service restoration when T 120 is closed to energize isolated section

Feeder loadings [MVA]						
Feeder 1	Feeder 2	Feeder 3	Feeder 4	Feeder 5	Feeder 6	Feeder 7
14.47	5.17	13.04	8.38	<b>26.06</b>	0	13.80

**Table 3.** Feeder loadings after service restoration when T 120 is closed to energize isolated section

Feeder loadings [MVA]						
Feeder 1	Feeder 2	Feeder 3	Feeder 4	Feeder 5	Feeder 6	Feeder 7
13.17	11.29	13.04	11.70	<b>17.91</b>	0	13.80



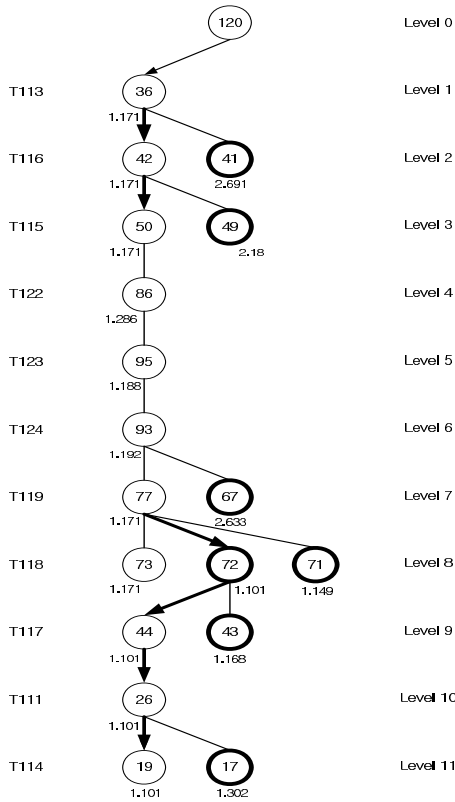


Fig. 8. The second step of cyclic best-first search

The above comparison indicates that loadings on feeder 5 is decreased after service restoration, therefore loadings on isolated feeder section after fault are distributed in proportion to adjacent feeder nominal capacities.

### 5 Conclusion

In this paper, a new heuristic algorithm and feeder load balance index was presented for service restoration considering feeder load balance in distribution networks. The proposed search algorithm adopts the concept of sub-tree proposed by reference [2] and utilizes cyclic best-first search and feeder load balance index developed by the authors. Cyclic best-first search is using best-first search that gets a solution much faster even if it lies deep down in the tree. And, by using revered sub tree, it compensates best-first search for not obtaining the best solution every time. Feeder load balance index is presented in order to distributing feeder loadings with respect to their nominal capacities in the proportional manner.

Test results on the KEPCO's 108 bus distribution system show that the performance is efficient and robust.

## **Acknowledgment**

This work was financially supported by MOCIE through EIRC program.

## **References**

1. T. Taylor, D. Lubkeman : Implementation of heuristic search strategies for distribution feeder reconfiguration. IEEE Trans. on Power Delivery, Vol. 5, No. 1, January (1990) 239 - 246
2. J. S. Wu, K. L. Tomsovic, C. S. Chen : A heuristic search approach to feeder switching operations for overload, fault, unbalanced flow and maintenance. IEEE Trans. on Power Delivery, Vol. 6, No. 4, October (1991) 1579 - 1585
3. M. E, Baran, F. F. Wu : Network reconfiguration in distribution systems for loss reduction and load balancing. IEEE Trans. on Power Delivery, Vol. PWRD-4, 1989, April (1989) 1401-1407

# Minimum Cost Operation Mode and Minimum Loss Operation Mode of Power System – Operation Mode Selection Based on Voltage Stability

Sang-Joong Lee

Dept of Electrical Engineering,  
Seoul National University of Technology,  
Nowon, Seoul,  
139-743 Korea  
85sjlee@snut.ac.kr

**Abstract.** Two formulae - an optimal P-Q generation formula for minimum system cost and an optimal MW allocation formula for minimum system loss - are described in this paper. The author defines two kinds of power system operation mode - minimum cost operation mode and minimum loss operation mode. The system can be operated on either minimum cost operation mode or minimum loss operation mode.

The system is normally being operated on minimum cost operation mode. The system can be shifted to minimum loss operation mode when voltage instability of the system is concerned. The operation mode can be selected by assessment of the system voltage stability, which may be achieved through application of the artificial intelligence techniques.

## 1 Introduction

The economics of the power system is very important. The power system in general is being operated economically by ELD(economic load dispatch). The output of generators and the total system operating cost is optimized while the system satisfies the power balance equation. System security should be considered, however, when the system is heavily loaded. If necessary, the system should be operated in order to maximally increase the system voltage stability.

Two kinds of power system operation mode are defined in this paper. One is the minimum cost operation mode and the other is the minimum loss operation mode. Two formulae - an optimal P-Q generation formula for minimum system cost and an optimal MW allocation formula for minimum system loss - are described in this paper. The loss sensitivities in the formulae are derived using the so-called angle reference transposition. The system can be operated either on the minimum cost operation mode or on the minimum loss operation mode. And the operation mode can be selected considering the system voltage stability.

## 2 System Operation for Minimum Cost Through Optimal P-Q Generation

A simple four-bus system is depicted. In Fig. 1, two load buses 3 and 4 are fed by two generators 1 and 2. Line parameters are given in Table 1. Specified bus data and the power-flow solutions for base case are shown in Table 2 and 3[1].

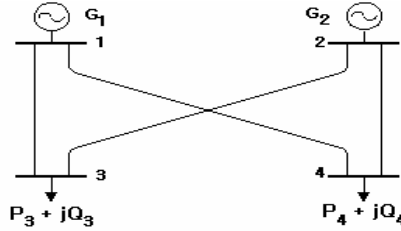


Fig. 1. Single line diagram of four-bus system

Table 1. Line parameters of four-bus system(p.u.)

from	to	R	X	Shunt Y
1	4	.00744	.0372	0.0775
1	3	.01008	.0504	0.1025
2	3	.00744	.0372	0.0775
2	4	.01272	.0636	0.1275

Table 2. Specified bus data(p.u.)

Bus	P	Q	V	Angle(rad)
1			1.0	0
2	3.0		1.0	
3	-2.20	-1.65		
4	-2.80	-1.85		

Table 3. Base-case power-flow(p.u.)

Bus	P	Q	V	Angle(rad)
1	2.097	2.050	1.0	0
2	3.0	1.569	1.0	.033
3	-2.20	-1.65	.954	-.023
4	-2.80	-1.85	.940	-.049

Transposing the angle reference on bus 4, the  $P_{loss}$  sensitivities for all generators in Fig 1 including the slack bus can be obtained as follows[2,3]:

$$\begin{bmatrix} \frac{\partial P_{loss}}{\partial P_1} \\ \frac{\partial P_{loss}}{\partial P_2} \\ \frac{\partial P_{loss}}{\partial P_3} \\ \frac{\partial P_{loss}}{\partial Q_1} \\ \frac{\partial P_{loss}}{\partial Q_2} \\ \frac{\partial P_{loss}}{\partial Q_3} \\ \frac{\partial P_{loss}}{\partial Q_4} \end{bmatrix} = [J^T]^{-1} \begin{bmatrix} \frac{\partial P_{loss}}{\partial \theta_1} \\ \frac{\partial P_{loss}}{\partial \theta_2} \\ \frac{\partial P_{loss}}{\partial \theta_3} \\ \frac{\partial P_{loss}}{\partial V_1} \\ \frac{\partial P_{loss}}{\partial V_2} \\ \frac{\partial P_{loss}}{\partial V_3} \\ \frac{\partial P_{loss}}{\partial V_4} \end{bmatrix} \tag{1}$$

By replacing  $P_{loss}$  in (1) with  $Q_{loss}$ , the  $Q_{loss}$  sensitivities can be easily obtained in a similar manner.

$$\begin{bmatrix} \frac{\partial Q_{loss}}{\partial P_1} \\ \frac{\partial Q_{loss}}{\partial P_2} \\ \frac{\partial Q_{loss}}{\partial P_3} \\ \frac{\partial Q_{loss}}{\partial Q_1} \\ \frac{\partial Q_{loss}}{\partial Q_2} \\ \frac{\partial Q_{loss}}{\partial Q_3} \\ \frac{\partial Q_{loss}}{\partial Q_4} \end{bmatrix} = [J^T]^{-1} \begin{bmatrix} \frac{\partial Q_{loss}}{\partial \theta_1} \\ \frac{\partial Q_{loss}}{\partial \theta_2} \\ \frac{\partial Q_{loss}}{\partial \theta_3} \\ \frac{\partial Q_{loss}}{\partial V_1} \\ \frac{\partial Q_{loss}}{\partial V_2} \\ \frac{\partial Q_{loss}}{\partial V_3} \\ \frac{\partial Q_{loss}}{\partial V_4} \end{bmatrix} \tag{2}$$

The problem of optimal P-Q generations for minimizing the total system operating cost can be formulated as follows:

Minimize  $Cost = \sum_{i=1}^{NG} f_i(P_{Gi}) = \sum_{i=1}^{NG} (a_i P_{Gi}^2 + b_i P_{Gi} + c_i)$

subject to  $\sum_{i=1}^{NG} P_{Gi} - P_D - P_{loss} = 0$  , (3)

$\sum_{i=1}^{NG} Q_{Gi} - Q_D - Q_{loss} = 0$  ,

where

- $Cost$  : total system operating cost(\$/hour)
- $N_G$  : number of generating units,
- $f_i$  : operating cost for i-th generator,
- $P_D, Q_D$  : total MW/Mvar load,
- $P_{Gi}, Q_{Gi}$  : MW/Mvar output of i-th generator,
- $P_G, Q_G$  : vector of generator MW/Mvar outputs.
- $P_{loss}, Q_{loss}$  : active and reactive power loss
- $a_i, b_i, c_i$  : coefficients of operating cost function for  $i^{th}$  generator

The optimality conditions of (3) can be obtained by the method of Lagrange multipliers[4]:

$$\begin{aligned} \mu_P (1 - \frac{\partial P_{loss}}{\partial P_G}) - \mu_Q \frac{\partial Q_{loss}}{\partial P_G} &= 2aP_G + b \\ \mu_P \frac{\partial P_{loss}}{\partial Q_G} - \mu_Q (1 - \frac{\partial Q_{loss}}{\partial Q_G}) &= 0 \end{aligned} \tag{4}$$

Substituting the loss sensitivities of (1) and (2) into (4), the optimal P-Q generation allocation for minimizing the operating cost can be obtained.

A simulation has been performed for Fig. 1 and the following two cases are compared with each other in table 4.

- Case I: Cost calculation by conventional ELD computation, in which the magnitudes of generator voltage  $V_1$  and  $V_2$  are given 1.0 p.u.[1]
- Case II: Cost calculation based on (4) using the loss sensitivities derived by (1) and (2)

We assumed in Case I and II, the voltage magnitude of the load bus 4 is fixed at 0.94 p.u. No other operating constraints are assumed for convenience.

Cost functions  $f_i$  for generators are assumed as follows[1]:

$$\begin{aligned} f_1 &= .0040 P_{G1}^2 + 8.0 P_{G1} + 240, \\ f_2 &= .0048 P_{G2}^2 + 6.4 P_{G2} + 120. \end{aligned} \tag{5}$$

**Table 4.** Comparison of operating cost and P-Q generations

	Case I	Case II
Total cost (\$/hr)	<b>4563.76</b>	<b>4563.13</b>
$P_{G1} / Q_{G1}$ (p.u.)	1.964 / 2.075	1.956 / 1.803
$P_{G2} / Q_{G2}$ (p.u.)	3.135 / 1.555	3.142 / 1.824
$V_1$ (p.u.)	1.0	.9954
$V_2$ (p.u.)	1.0	1.0077
$V_3$ (p.u.)	.953	.9565
$V_4$ (p.u.)	.940	

Case II of Table 4, equation (4) yields  $V_1=0.9954$  and  $V_2=1.0077$  to propose lower operating cost than Case I, which shows that equation (4) with the loss sensitivities derived by (1) and (2) can yield superior results to the conventional ELD.

### 3 System Operation for Minimum Loss Through Optimal MW Allocation

Equation (6) is the classical formulation for minimizing the system total operating cost[1,2].

$$\text{Minimize } \sum_{i=1}^{NG} f_i(P_{Gi})$$

subject to 
$$\sum_{i=1}^{NG} P_{Gi} - P_D - P_{loss} = 0 \tag{6}$$

The solution of above minimization problem can be expressed as follows:

$$\frac{df_i}{dP_{Gi}} \frac{1}{1 - \frac{\partial P_{loss}}{\partial P_{Gi}}} = \lambda \tag{7}$$

where  $\lambda$  is the Lagrangian multiplier introduced for solving (6). In similar way to the classical ELD formulation, let us substitute the system loss for the objective function of (6).

Minimize  $P_{Loss}$   
 subject to 
$$\sum_{i=1}^{NG} P_{Gi} - P_D - P_{loss} = 0 \tag{8}$$

The optimality condition of (8) can be obtained using the principle of Lagrangian multipliers as follows:

$$\frac{\partial P_{loss}}{\partial P_{G1}} = \frac{\partial P_{loss}}{\partial P_{G2}} = \dots = \frac{\partial P_{loss}}{\partial P_{G_{NG}}} = \frac{\mu}{1 + \mu} \tag{9}$$

where  $\mu$  is the Lagrangian multiplier introduced for solving (8). Equation (9) implies that the system loss is minimized when all generators are being operated with equal loss sensitivities. The loss sensitivities for all generators can be derived by (1)[5].

Simulation has been performed for Fig. 1. Specified bus data and the power-flow solutions for base case are shown in Table 5 and 6.

**Table 5.** Specified bus data(p.u.)

Bus	P	Q	V	angle(deg)
1			1.0	0
2	3.18		1.0	
3	-2.20	-1.3634		
4	-2.80	-1.7352		

**Table 6.** Base-case power-flow (p.u.)

Bus	P	Q	V	angle(deg)
1	1.913152	1.87224	1.0	0
2	3.18	1.32543	1.0	2.43995
3	-2.20	-1.3634	.96051	-1.0793
4	-2.80	-1.7352	.94304	-2.6265
remark:		$P_{loss} = .09335$		

The following two cases are compared in Table 7.

- Case I: Loss calculation based on (9) using the loss sensitivities obtained by the conventional B-matrix method.

- Case II: Loss calculation based on (9) with the loss sensitivities derived by (1) and (2)

The B- matrix used[1] is expressed as:

$$B = \begin{bmatrix} 8.3831 & -.0494 & .3750 \\ -.0494 & 5.9635 & .1949 \\ .3750 & .1949 & .0901 \end{bmatrix} \times 10^{-3} \tag{10}$$

**Table 7.** Comparison of generation allocation and system loss

	Case I	Case II
$P_{loss}$	<b>.09008</b>	<b>.08567</b>
$P_{G1}$	2.1063	2.7488
$P_{G2}$	2.9838	2.3369
$\partial P_{loss} / \partial P_{G1}$	<b>.03577</b>	<b>.02035</b>
$\partial P_{loss} / \partial P_{G2}$	<b>.03577</b>	<b>.02035</b>

In table 7, the generation allocation by (9) with the loss sensitivities from (1) demonstrates an improved system loss in comparison to the conventional B-coefficient method.

The B-matrix used above should be corrected to conform to the new operating condition when shifts of outputs among generators occur if more accurate results are required[1]. However, the optimal generations can be directly calculated by using the loss sensitivities derived by (1) because these loss sensitivities can reflect the current system operating conditions.

#### 4 System Operation Mode Selection by Voltage Stability Assessment

Here, the author defines two kinds of power system operation mode. One is the minimum cost operation mode and the other is the minimum loss operation mode. Due to the importance of the power system economics, the system in general is being operated on minimum cost operation mode. The output of generators and the total system operating cost can be better optimized using (4) with the loss sensitivities derived by (1) and (2).

System security should be considered, however, when the system is heavily loaded. If necessary, the system should be operated to maximally increase the system voltage stability rather than minimize the operation cost. Reduction of the system loss is very important in order to improve the system voltage stability.

The power system loss is minimized when all generators are being operated with equal loss sensitivities as described in (9). Optimal MW allocation for minimum system loss can be obtained by solving (9) with the loss sensitivities derived by (1). The



system can be shifted to minimum loss operation mode when voltage instability of the system is concerned. The voltage stability, however, is very vague and difficult to be clearly quantified. Application of artificial intelligence techniques may be a solution for assessment of the system voltage stability and selection of the system operation mode.

## 5 Conclusion

Two kinds of power system operation mode - minimum cost operation mode and minimum loss operation mode - are defined in this paper. Two formulae - an optimal P-Q allocation formula for minimum cost operation mode and an optimal MW allocation formula for minimum loss operation mode - are introduced. The loss sensitivities in the formulae are derived using the so-called angle reference transposition. Example calculations have been demonstrated for a sample system.

The system is normally being operated on minimum cost operation mode and can be shifted to minimum loss operation mode when voltage instability of the system is concerned. Application of artificial intelligence techniques may be a solution for assessment of the system voltage stability and selection of the system operation mode.

## References

1. John J. Grainger, William D. Stevenson, Jr., "Power System Analysis", Mcgraw Hill Inc., 1994. pp. 548-560
2. S.J.Lee, K. Kim, "Re-construction of Jacobian Matrix by Angle Reference Transposition and Application to New Penalty Factor Calculation", IEEE Power Engineering Review, vol.22, no.2, Feb 2002, pp. 47-50
3. S.J. Lee, S.D. Yang, "Derivation of P-Q Loss Sensitivities by Angle Reference Transposition and An Application to Optimal P-Q Generation for Minimum Cost", IEEE Trans on Power Sys, vol.21, no1, Feb 2006, pp.428-430
4. J.H.Kim et al, Power System Engineering, Chungmoon-gak, Seoul, 1998, pp.137-140, 332-333
5. S.J.Lee, "Calculation of Optimal Generation for System Loss Minimization Using Loss Sensitivities Derived by Angle Reference Transposition", IEEE Trans on Power Sys, vol.18, no3, Aug 2003, pp.1216-17

# Optimal Voltage and Reactive Power Control of Local Area Using Genetic Algorithm

Hak-Man Kim<sup>1</sup>, Jong-Yul Kim<sup>2</sup>, Chang-Dae Yoon<sup>3</sup>,  
Myong-Chul Shin<sup>4,\*</sup>, and Tae-Kyoo Oh<sup>5</sup>

<sup>1</sup> Korea Electrotechnology Research Institute  
Uiwang-city, Gyeonggi-do, S. Korea  
hmkim@keri.re.kr

<sup>2</sup> Korea Electrotechnology Research Institute  
Changwon-city, Gyeongnam-do, S. Korea  
jykim@keri.re.kr

<sup>3</sup> Sungkyunkwan University .  
Suwon-city, Gyeonggi-do, S. Korea  
phasors@skku.edu

<sup>4</sup> Sungkyunkwan University  
Suwon-city, Gyeonggi-do, S. Korea  
mshin@yurim.skku.ac.kr<sup>1</sup>

<sup>5</sup> Korea Electrotechnology Research Institute  
Uiwang-city, Gyeonggi-do, S. Korea  
tkoh@keri.re.kr

**Abstract.** To improve the voltage profile and to reduce system losses, many switched shunt capacitors are used in a power system. It is necessary to coordinate them for operating the system effectively. In this study, a genetic algorithm (GA) is used to find optimal coordination of switched shunt capacitors within a local area of a power system. To verify the effectiveness of the proposed method, a simulation is performed of KEPCO (Korea Electric Power Corporation) power system.

## 1 Introduction

In power system planning and operation, voltage and reactive power control are very important. The voltage deviation and system losses can be reduced through control of reactive power sources. There are many types of reactive power sources, such as switched shunt capacitors, synchronizing condensers, Static Var Compensators (SVC), Static Synchronous Compensators (STATCOM), etc.

Switched shunt capacitors are the most popular reactive power source. Hundreds of switched shunt capacitor banks exist within a local area of a power system. For effective control of voltage and reactive power, an optimal decision on the number of inserted switched shunt capacitor banks is needed. To find optimal coordination of the switched shunt capacitor banks, various approaches, such as sensitivity analysis, simulated annealing, expert system and neural network, have been studied [1-4].

---

\* Corresponding author.

In this paper, we suggest a genetic algorithm (GA) approach to control voltage and reactive power using switched shunt capacitors. The approach is tested on a portion of the 239 generator, 743 bus KEPCO (Korea Electric Power Corporation) power system. The simulation is performed using the industry standard PSS/E power system analysis software from PTI [5-6]. The results are compared with those obtained using the sensitivity method.

## 2 Mathematical Model of the Optimal Decision on the Number of Inserted Switched Shunt Capacitor Banks

One bus has several switched shunt capacitor banks. The number of banks is inserted discretely for voltage and reactive power control. The objective function for the optimal decision is established to maintain the bus voltage target range and to reduce transmission losses after the occurrence of a contingency. The objective function is expressed by equation (1).

$$\min F(X) = \sum_j V_{d,j}(X) + Loss(X) \tag{1}$$

$$\text{s.t. } SC_i^{\min} \leq SC_i \leq SC_i^{\max} \tag{2}$$

$$V_{\min 1} \leq V_j \leq V_{\max 1}, \quad V_{d,j} = 0 \tag{3}$$

$$V_{\min 1} < V_j \leq V_{\max 2}, \quad V_{d,j} = (V_j - V_{\max 1}) \tag{4}$$

$$V_{\min 2} \leq V_j < V_{\min 1}, \quad V_{d,j} = (V_{\min 1} - V_j) \tag{5}$$

$$V_j < V_{\min 2} \text{ or } V_j \leq V_{\max 2}, \quad V_{d,j} = a \tag{6}$$

where

$V_{d,j}(X)$  = jth bus voltage deviation when sw. shunt capacitor (X) is inserted

$Loss(X)$  = transmission loss when sw. shunt capacitor (X) is inserted [MW]

$X = [SC_i]$ ; the vector of  $SC_i$

$SC_i$  = quantity of inserted sw. shunt capacitance of the ith bus [MVar]

$V_j$  = jth bus voltage [PU]

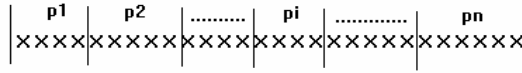
$V_{\min 1}, V_{\max 1}$  = bus voltage target range [PU]

$V_{\min 2}, V_{\max 2}$  = allowable bus voltage range [PU]

$a$  = penalty factor depending on bus voltage deviation

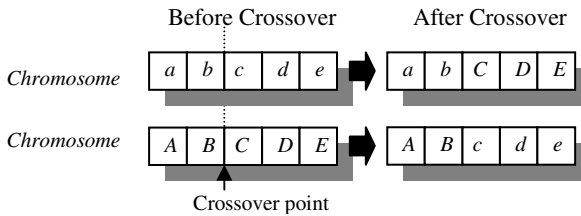
## 3 Genetic Algorithm

GA does not use real parameters but uses chromosomes composed of string coded genotypes (genes), such as is shown in Fig. 1.

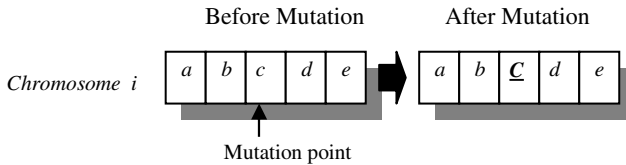


**Fig. 1.** A chromosome composed of parameters

The basic operators of a GA are reproduction, crossover and mutation. Reproduction is a process in which individual strings are copied according to their objective function (fitness function) values. After reproduction, crossover and mutation operators alter chromosomes in the new population. Under the crossover operator, two structures in the new population exchange portions of their genes. After a crossover operator, a mutation operator may be applied to explore other patterns. Fig. 2 and Fig. 3 show the operation process of crossover and mutation, respectively [7,8].



**Fig. 2.** Crossover operation



**Fig. 3.** Mutation operation

GAs explore the optimal solution with the following procedure:

- Step 1: Set the initial population of  $N$  chromosomes randomly.  
( $N$  = the number of chromosomes)
- Step 2: Evaluate the fitness of each chromosome using the fitness function.  
If the termination condition is satisfied then go Step 5.
- Step 3: Reproduce offspring.
- Step 4: Alter chromosomes with crossover and mutation operators. Go to Step 2.
- Step 5: End.

#### 4 A Genetic Approach to Voltage and Reactive Power Control

In voltage and reactive power control, optimal variables are the amounts of inserted switched shunt capacitance, that is, the number of inserted switched shunt capacitor

banks. For applying a genetic algorithm, chromosomes are composed of switched shunt capacitor banks as shown in Fig. 4.

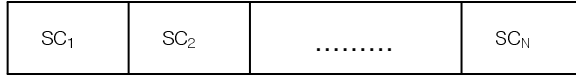


Fig. 4. A chromosome composed of switched shunt capacitor banks

The fitness function can be formulated as given in equation (7) based on maintaining the bus voltage target range and reducing transmission losses after the occurrence of a contingency.

$$Fitness\ Function = \frac{\beta}{W_1 * \sum_j V_{d,j}(X) + W_2 * Loss(X)} \tag{7}$$

where

$W_1, W_2$  = weight values

$\beta$  = arbitrary constant

In this paper, elitism [8] is adopted in the genetic search. The genetic search procedure is as shown in Fig. 5.

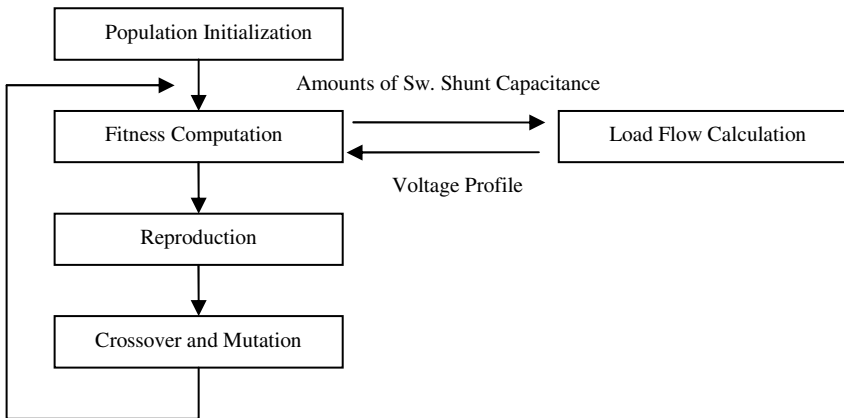
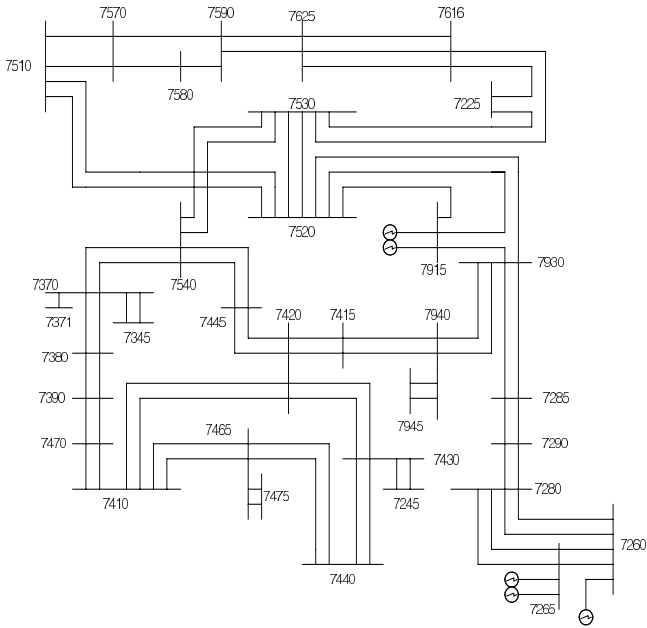


Fig. 5. The proposed genetic search procedure

### 5 Case Study

The genetic approach is tested on a portion of the 239 generator, 743 bus KEPCO power system. Fig. 6 shows the single-line diagram of the test area.



**Fig. 6.** Test area of KEPCO power system

235 switched shunt capacitor banks are placed on 29 buses. The optimal and genetic parameters are as follows:

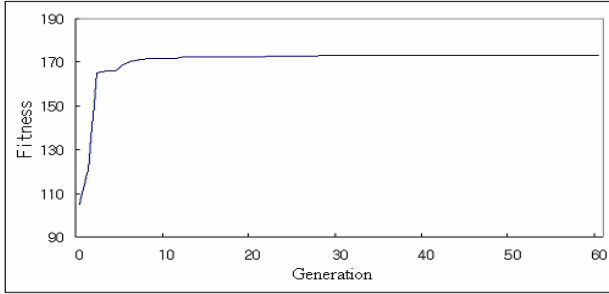
- $V_{\max 1} = 1.03$
- $V_{\min 2} = 0.925$
- $V_{\max 2} = 1.05$
- $a = 100$
- No. of population = 800
- No. of generations = 60
- Mutation probability = 0.05
- Crossover probability = 0.85
- $\beta = 100,000$
- $W_1 = 10$
- $W_2 = 5$

If line 7150-7600 is removed after the occurrence of a contingency, buses 7400 and 7500 violate the allowable voltage range. Table 1 shows voltage changes of those buses.

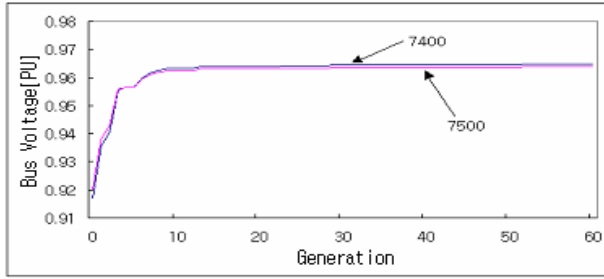
**Table 1.** Voltage changes of bus 7400 and 7500 if line 7150-7600 trips

Bus	Pre-contingency voltage [PU]	Post-contingency voltage [PU]
7400	0.9836	0.9171
7500	0.9865	0.9203

Fig. 7 shows the variation of fitness values and Fig. 8 shows the voltages for buses 7400 and 7500 according to the genetic search. Both fitness and voltage converge rapidly.



**Fig. 7.** Variation of fitness values



**Fig. 8.** Variation of bus 7400 and 7500 voltages

The proposed genetic approach results are compared with the sensitivity method as shown in Table 2. Clearly, the results of the proposed genetic approach are better than those of the sensitivity method.

**Table 2.** Comparison of results

Comparison items		Sensitivity method	Proposed genetic method
Voltage deviation ( $\sum_j V_{d,j}(X)$ )		91.27	1.63
Active transmission losses [MW]		118.36	112.13
Reactive transmission losses [Mvar]		1885.75	1794.24
Bus voltage [pu]	7400	0.926	0.965
	7500	0.928	0.964

## 6 Conclusion

To improve voltage profile and to reduce system losses, many switched shunt capacitors are used in power system. In this paper, we suggest a genetic approach to control voltage and reactive power using switched shunt capacitors.

Key contributions of this paper are as follows:

- 1) The objective function for optimal decision is established to maintain the bus voltage target range and to reduce transmission losses after the occurrence of a contingency.
- 2) For applying a genetic algorithm, chromosomes are composed of switched shunt capacitor banks and the fitness function is formulated based on both maintaining the bus voltage target range and reducing transmission losses after the occurrence of a contingency.

The genetic approach is tested on a portion of the 239 generator, 743 bus KEPCO power system. By comparison with the sensitivity method, the results of the suggested genetic search show excellent performance.

## Acknowledgments

This work was financially supported by MOCIE through the EIRC Program.

## References

1. H. Kobayashi, et al., "Diakoptic Approach to Sensitivity Analysis in Large-scale Systems", IEEE/PES Summer Meeting, Los Angeles, Paper No. A78531-6, July (1978)
2. R.H. Liang and Y.S. Wang, "Main Transformer ULTC and Capacitors Scheduling by Simulated Annealing Approach", *Electrical Power & Energy Systems*, Vol. 5, No. 2, May (2001) 531-538
3. T.L. Le and M. Negnevitsky, "Expert System Application for Voltage and VAR Control in Power Transmission and Distribution Systems", *IEEE Trans. On Power Delivery*, Vol. 12, No. 3, July (1997) 1392-1397
4. N.I. Santoso and O.T. Tan, "Neural-net Based Real-time Control of Capacitors Installed on Distribution Systems", *IEEE Trans. On Power System*, Vol. 5, No. 1, Jan. (1990) 266-272
5. Power Technologies Inc., PSS/E-24 Power System Simulator Program Operation Manual & Application Guide, Dec. (1995)
6. Power Technologies Inc., PSS/E-24 IPLAN Ver. 11.0, Dec. (1995)
7. D.E. Goldberg: *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison Wesley (1989)
8. Z. Michalewicz: *Genetic Algorithms + Data Structures = Evolution Programs*. 3rd Ed. Springer-Verlag, Berlin Heidelberg New York (1996)



# Equivalent Electric Circuit Modeling of Differential Structures in PCB with Genetic Algorithm

Jong Kang Park, Yong Ki Byun, and Jong Tae Kim

School of Information and Communication Engineering, Sungkyunkwan University,  
South Korea  
jtkim@skku.ac.kr

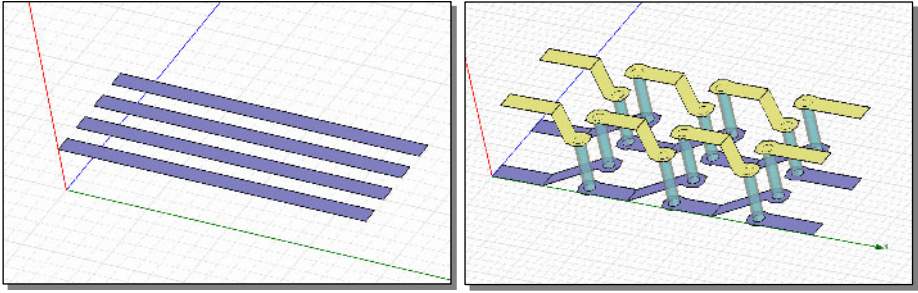
**Abstract.** This paper introduces an equivalent circuit extraction technique for differential structures using a hybrid genetic algorithm. The procedure searches for the proper parameters of lumped circuit elements to fit the scattering parameters which can be obtained from actual measurement or 3D electro-magnetic simulation. The genetic operators include the probabilistic arithmetical crossover, non-uniform mutation and local optimizer techniques. These operators cause the solutions to converge faster and be more precise. The experimental results show that the circuit modeling of parallel differential pairs with the genetic approach has excellent fitting ability.

## 1 Introduction

Signal integrity is becoming more and more important as both the physical geometry and signal-level decrease. Today's numerous off-chip interfaces require data rates of more than 1 Gbps and robust communication. Differential signaling techniques such as the LVDS(low-voltage differential signaling) or GLVDS(ground low-voltage differential signaling) standards, define the key features required in current high-frequency PCB (printed circuit board) designs. These techniques provide high-speed transmission/reception, less EMI(electro-magnetic interference) radiation, invulnerability to common-mode noise and lower power consumption.

Fig. 1 shows some of the possible physical structures of differential pairs. By using the equivalent circuit models of these transmission lines and the I/O interfaces of active devices, we can observe the time-domain signal behavior and evaluate the signal integrity in their geometry(e.g., the spacing between the conductors, dielectric constant, thickness of the conductor and substrate, etc.). To model accurate equivalent circuits accurately from their physical implementation, we need to fit them to the scattering parameters, or so called s-parameters, which are defined by the ratio of the reflected voltage to the induced voltage at the ports.

This paper presents a technique which can be used for the equivalent circuit modeling of differential structures with a GA(genetic algorithm). Basically, equivalent circuits are pre-defined and parameterized with lumped circuit elements including the resistance, self/mutual inductance, and capacitance. In this procedure a search is made for the proper parameter values for these lumped elements to fit into the s-parameters which can be obtained from actual measurements or 3D electro-magnetic(EM)



**Fig. 1.** Examples of differential structures : straight line(left) and twisted differential line(right)

simulations in CAD software. To accomplish this, the arithmetical crossover, non-uniform mutation and local searcher techniques are used to make the solutions converge faster to near-optimum and be more precise.

## 2 Related Works

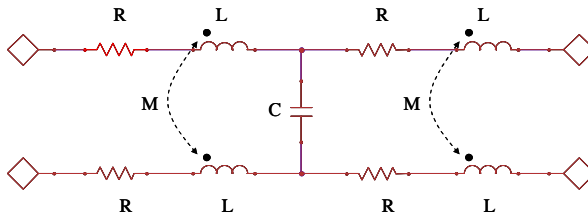
Interconnect modeling at both on- and off-chip is regarded as an important issue for signal integrity. As stated above, we can simulate signal attenuation, crosstalk noise and EMI radiation using interconnect modeling. Zhang et. al.[1] proposed the effective selection of VLSI(very large scale integration) on-chip interconnects through multi-objective optimization involving the crosstalk and performance of their structures. In [2], Azadpour et. al. introduced a fast extraction method for inductance in high-speed clock trees.

Actually, equivalent circuit extraction using a GA is not a new approach. There have been several studies in which a GA was used to search for the parameter values of lumped elements. In 1998, Werner et. al.[3] proposed that a GA could be used to extract the equivalent circuit models of right-angled bands in microstrip lines, parallel-plate capacitors and planar inductors. Their results also show that it is possible for a GA to model the impedance discontinuity in microstrip lines. In [4], the SPICE model of ACF(anisotropic conductive film) flip-chip interconnections was obtained using a GA. Cao et. al.[5] employed another evolutionary approach to optimize electro-chemical impedance of lumped elements.

## 3 Equivalent Circuit Extraction

Single-ended transmission lines such as microstrips and stripline structures have a single signal trace and its return path(ground plane). Differential signaling, however, transmits and receives voltage difference between two signals, and typically uses the ungrounded model as shown in Fig. 2.

A lumped circuit with specific values for the passive device parameters has to conform to the characteristics of the  $s$ -parameters obtained using a VNA(vector network analyzer) or a 3D-EM simulation. In Fig. 2,  $R$  denotes the conductor loss,  $C$  the mutual capacitance between the two coupled lines, and  $L$  and  $M$  correspond to the self- and



**Fig. 2.** Equivalent lumped circuit model for a differential pair

mutual-inductance, respectively. Changing the RLCM values does not cause the  $s$ -parameters to exhibit linear characteristics.

Thus it is an intricate task to search for a precise value of the lumped elements intuitively. Some commercial tools can compute the  $s$ -parameters for given 3D geometrical and physical structures accurately, but such analytical methods typically require long computation times which means that it is difficult to use them directly for a fully transient analysis with IBIS(input/output buffer information specification) of the active devices, and they are not portable to other development systems.

Fig. 3 shows the equivalent circuit extraction flow for differential lines, referred to in [6]. We do not fit this equivalent circuit to the measured or EM-modeled single-ended  $s$ -parameters but, rather use the two-port differential-mode  $s$ -parameters, which can be calculated using the mixed-mode  $s$ -parameter theory[7]. In this paper, various experiments including the calculation of the differential-mode  $s$ -parameters were conducted with Agilent's ADS(Advanced Design System) command line simulator. Since the  $s$ -parameters include the reactance, the fitting process iteratively compares each real and imaginary component at every frequency step. The corresponding objective functions are discussed in a later section.

## 4 The Genetic Algorithm for Modeling a Differential Line

This section describes the problem formulation and evolutionary procedure with genetic operators to extract a lumped circuit model of a differential pair. As in other previous studies[3][4], it is natural that facile encoding and feasibility tests be available for lumped RLC elements. However, careful selection of the genetic operator is required for faster convergence and higher accuracy.

### 4.1 Objective Function, Constraints and Encoding

Let us assume that the two-port differential-mode  $s$ -parameters, either measured or simulated, are represented as magnitude/angle(MA) or real and imaginary components, there exists a sampled frequency bandwidth of interest. Then, in the case where  $N=2$  ports, the objective[3] for minimizing the least squares error(LSE) and its constraints can be written as follows,

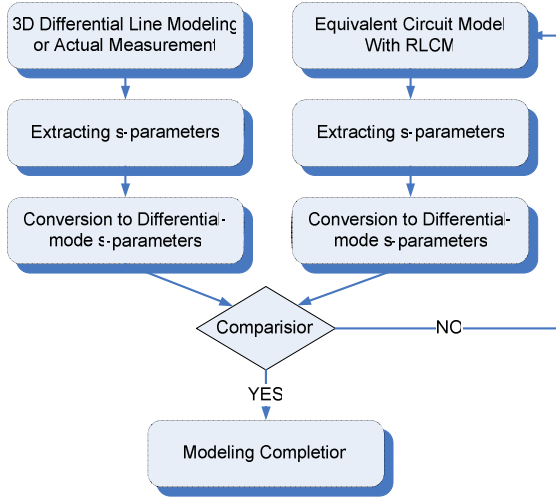


Fig. 3. Equivalent circuit fitting procedure for differential lines

$$\text{Minimize } F = \sum_{f=f_s}^{f_e} \sum_{i=1}^N \sum_{j=1}^N |S_{ij}^m - S_{ij}^a|^2 \tag{1}$$

$$\text{s.t. } R^b \leq R \leq R^u, L^b \leq L \leq L^u, C^b \leq C \leq C^u, 0 \leq M \leq L \tag{2}$$

where  $f_s$  and  $f_e$  are the start- and stop-frequencies in the frequency sweep, respectively, and  $S_{ij}^m$  and  $S_{ij}^a$  denote the s-parameters of the measured and GA generated equivalent circuits, respectively. As shown in Eq. (2), each chromosome can easily be encoded by a positive real-number as  $\{R, L, C, M\}$  in the pre-defined range. Note that both self-inductances are the same in a differential pair, and the corresponding mutual inductance cannot be more than  $\sqrt{L^2} = L$ . Those chromosomes outside of the ranges given in Eq. (2) are regarded as infeasible solutions in this paper.

### 4.2 The Evolutionary Extraction Procedure and Its Genetic Operators

Fig. 4 shows the block diagram for the lumped model extraction procedure using the hybrid GA. First, the initial P population is determined by randomly generated real numbers satisfying Eq. (2). Then, each chromosome is evaluated using Eq. (1) and the HPEESOF-SIM command-line simulator. Both arithmetical crossover and non-uniform mutation are performed to yield offspring  $O(t)$  at the current  $t$ -th generation, where  $p_c$ , and  $p_m$  denote the probabilities of crossover and mutation, respectively. The repair function moves each gene of the offspring that violates the constraints to the boundary of a feasible solution. Hill-climbing searches for the local optimum  $\{R', L', C', M'\}$  are performed, so as to allow for faster convergence to the solution near the global optimum. After evaluating  $O(t)$ ,  $(\mu+\lambda)$  deterministic sampling selects the best P chromosomes for the next generation. The details of the genetic operators and the local searcher are,

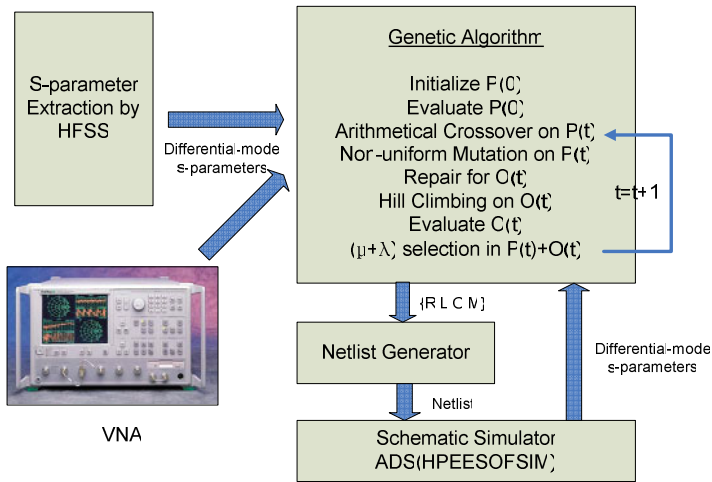


Fig. 4. Equivalent circuit generation flow using hybrid genetic algorithm

- Arithmetical crossover – Select two chromosomes from  $P$  with  $p_c$ . Let  $r$  be a random variable in the range of  $[0,1]$ . Then, the corresponding offspring is calculated as,

$$x = rx_1 + (1 - r)x_2 \tag{3}$$

where  $x$ ,  $x_1$  and  $x_2$  denote each gene ( $R, L, C, M$ ) of the offspring and its two parents.

- Non-uniform mutation[8] – Select genes from a chromosome with  $p_m$ . Determine each gene of the offspring through either one of the following equations.

$$\begin{aligned} x' &= x + \Delta(t, x^u - x) \\ x' &= x - \Delta(t, x - x^b) \end{aligned} \tag{4}$$

In this paper, we select either of the forms of Eq. (4) with equal probability. The function  $\Delta(t,y)$  is given as follows,

$$\Delta(t, y) = yr(1 - t/T)^k \tag{5}$$

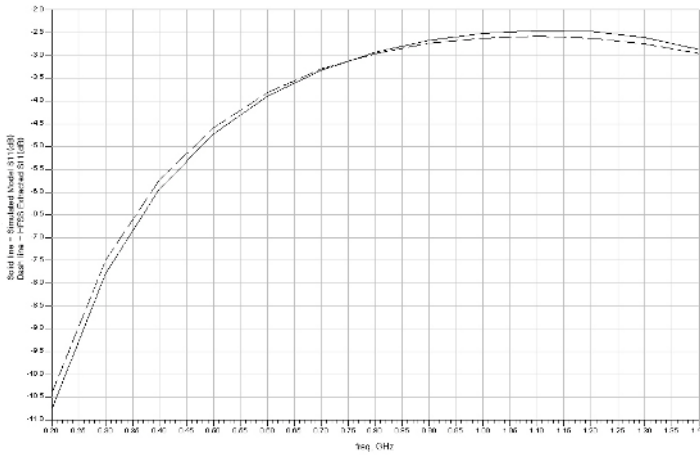
where  $r$  is a random number in the range of  $[0,1]$ ,  $T$  the maximum generation number, and  $k$  the degree of non-uniformity.

- Hill climbing – The local search determines the direction of  $R$  to the local optimum with the pre-defined step size. Then, whenever no improvement is achieved through hill climbing, the target gene is changed from the left to right in the chromosome.

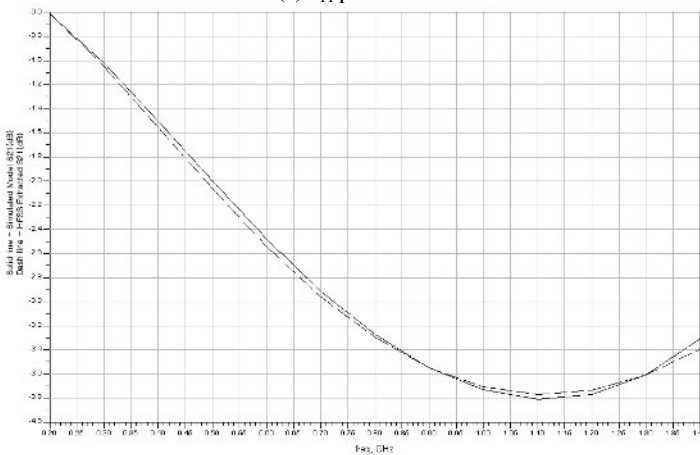
In s-parameter fitting, arithmetical crossover and hill climbing has a tendency to converge the chromosomes to near optimal solutions. Both crossover and non-uniform mutation provide the possibility of leaping over a local optimum, but the latter is particularly effective in improving the accuracy in later generations, due to the use of Eq. (5). This feature makes the local move size used in hill climbing be not too small.

### 5 Experimental Results

In the experiment, we simulated two straight conductors with different widths and spacings with  $\sigma=5 \times 10^7$  siemens/m on a 59mil-thickness FR4 substrate. The calculated differential-mode s-parameters( $s_{11}$ - $s_{22}$ ) were sampled between  $f_s=0.2$ GHz and  $f_e=1.4$ GHz with a 0.1GHz step size. The genetic parameters were chosen as,  $T=30$ ,  $P=100$ ,  $p_c=0.2$ ,  $p_m=0.05$ ,  $R^u=100\Omega$ ,  $R^b=0.01\Omega$ ,  $L^u=100$ nH,  $L^b=0.01$ ,  $C^u=1500$ fF,  $C^b=0.0$ fF and  $k=2.0$ . The fitting results of a differential pair with a length( $l$ ) of 40mm, a width( $w$ ) of 1mm, and a spacing( $d$ ) of 1mm are shown in Fig. 5. We can see that the s-parameters of the extracted lumped circuit are very similar to those of the 3D EM result. Table 1 also lists the extracted lumped model parameters obtained using GAs with different geometrical configurations. In each case we obtained accurate models before the 20-th generation.



(a)  $s_{11}$  parameter



(b)  $s_{21}$  parameter

**Fig. 5.** s-parameter fitting results

**Table 1.** Extracted equivalent circuits in different physical configuration

$l(\text{mm})$	$w(\text{mm})$	$s(\text{mm})$	$R(\Omega)$	$L(\text{nH})$	$C(\text{fF})$	$M(\text{fH})$
40.0	1.0	1.0	0.70	6.80	906.20	0.14
40.0	0.5	0.5	1.50	7.50	856.10	0.00
40.0	1.0	0.5	0.81	5.72	1050.15	0.01
40.0	1.0	2.0	0.90	8.00	819.23	0.01

## 6 Conclusion

This paper presents an evolutionary approach to the extraction of equivalent lumped circuit models with differential structures from actual measurements or EM-analyzed results. The local optimizer and genetic operators used in this paper allow the models to be more precise and more easily be obtained. We verified that our GA has excellent fitting ability with different physical configurations. Adding flexibility to the process of defining the templates of the lumped circuit enables the GA to accurately extract the various shapes of differential lines, such as bent/twisted off-chip structures.

## Acknowledgement

This work was supported by grant No. RTI04-03-04 from the Regional Technology Innovation Program of the Ministry of Commerce, Industry and Energy(MOCIE) .

## References

1. Q. Zhang, and J. Liou, J. McMacken, J. Thomson, and P. Layman "Development of Robust Interconnect Model Based on Design of Experiments and Multiobjective Optimization," IEEE Trans. on Electron Devices, Vol. 48, pp.1885-1891, 2001.
2. M. A., Azadpour and T. S. Kalkur, "VLSI Interconnect Modeling at Multi-GHz Frequencies Incorporating Inductance," Southwest Symposium on Mixed-Signal Design, pp.54-59, 2003.
3. P. L. Werner, R. Mittra and D. H. Werner, "Extraction of equivalent circuits for microstrip components and discontinuities using the genetic algorithm," IEEE Microwave and Wireless Components Letters, Vol.8, pp.333-335, 1998.
4. W. Ryu, M. Yim, S. Ahn, J. Lee, W. Kim, K. Paik and J. Kim, "High-frequency SPICE model of anisotropic conductive film flip-chip interconnections based on a genetic algorithm," IEEE Trans. on Component and Packaging, Vol.23, pp. 542-545, 2000.
5. H. Cao, J. Yu and L. Kang, "An Evolutionary Approach for Modeling the Equivalent Circuit for Electrochemical Impedance Spectroscopy," Proc. of 2003 Congress on Evolutionary Computation, Vol.3, pp.1819-1825, 2003.
6. D. G. Kam, H. Lee and J. Kim, "Twisted differential line structure on high-speed printed circuit boards to reduce crosstalk and radiated emission," IEEE Trans. on Advanced Packaging, Vol. 27, pp.590-596, 2004.
7. D. Bockelman and W. Eisenstadt, "Combined differential and common mode scattering parameters : theory and simulation," IEEE Trans. Microwave Theory Tech., vol. 43, pp. 1530-1539, 1995.
8. M. Gen and R. Cheng, "Genetic Algorithms and Engineering Optimization," John Wiley & Sons, Inc., 2000.

# Rule-Based Expert System for Designing DC-DC Converters

Seok Min Yoon and Jong Tae Kim

School of Information and Communication Engineering, Sungkyunkwan University,  
Korea  
jtkim@skku.ac.kr

**Abstract.** The design of DC-DC converters mainly relies on the experience of the designer who is usually aided by simulation programs such as SPICE. Since it is very difficult to design highly efficient and reliable converters, even for the experienced designer, we developed a design automation tool for isolated DC-DC converters based on a rule-based expert system. We define a set of rules specifying the decisions that need to be made for selecting the design topology and circuit elements. The rule-based system was implemented and verified in JESS.

## 1 Introduction

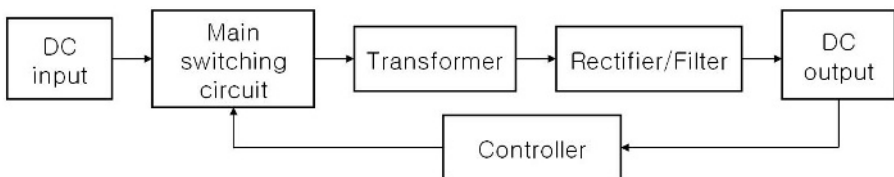
Each electronic circuit has its own power requirement, and therefore requires a power supply which is specially designed for it. Power supplies can be classified into linear power supplies, SMPSs (switch mode power supplies) and resonant converters. In contrast to traditional linear power supplies in which a resistor is inserted between input and output ports to control the output voltage, SMPSs control the output voltage by varying the ratio of the turn-on to duty cycle time of the electronic switch. The control provided by the electronic switch of the SMPS makes its power supply more efficient, lighter and smaller than the traditional linear power supply. To design an SMPS, an understanding of switching elements design techniques is necessary, along with the ability to design a transformer and knowledge of control theory. However it is very difficult to design highly efficient and reliable SMPSs, even for a designer who has knowledge of all of the above techniques. This makes the design of SMPSs highly dependent on the designer's design experience or skill, and designers usually verify their designed circuit with a separate circuit analyzing tool. Therefore, to make the design process more efficient, a design automation tool is necessary. We developed such a tool using the forward chaining method of a rule-based system that can produce a suitable design from the specification of the application. The proposed system can select a suitable circuit topology, switching elements and transformer, and determine the values of the inductances and capacitances from a given specification automatically. Consequently, non-experienced designers can design SMPSs like experienced designers do and this results in a reduction in the complexity of the design process.



## 2 Related Works

Designing the DC-DC converter is a core part of the development of an SMPS. Semiconductor components have been put to practical use for designing DC-DC converters since the 1960s, and various kinds of converter design topologies have been developed including buck, boost and buck-boost for basic switching converters, and flyback, forward, push-pull, half-bridge and full-bridge for isolated transformer converters, and resonant converters. Research on SMPSs has been focused on the modeling and analyzing of new topologies. Since Middlebrook[1] began a systematic study on the modeling and analyzing of SMPSs in 1970, there have been many studies on SMPS modeling and analyzing. The state space and the averaging method[2], which consists of the state space method and circuit averaging method, and SCAP (switching converter analysis and measurement program)[3] are representative examples of such research. Computer simulation is essential to the modeling and analyzing of DC-DC converters, since DC-DC converters have a highly nonlinear nature. SPICE is the major tool used in computer simulation and circuit analysis[4]. Nowadays, designing DC-DC converters using computers mainly depends on the designer's design experience of using SPICE as a simulator [5].

The isolated DC-DC converter has the advantages of circuit protection and multiple outputs, due to the separation of the input port from output port, using a transformer. Also, it is very easy to boost or drop down the output voltages. Flyback, forward, push-pull, half-bridge, and full-bridge are the isolated DC-DC converter topologies. Forward, push-pull, half-bridge and full-bridge converters are based on the non-isolated buck converter, whose characteristic feature is that the output voltage is lower than the input voltage. The flyback converter is based on the non-isolated buck-boost converter, which controls the output voltage by varying the ratio of the turn-on time so that it is higher or lower than the input voltage. Table 1 shows that input to output voltage ratio equation for each topology, where  $N_1$ ,  $N_2$ , and  $D$  represent the numbers of primary turns, second turns of transformer and the duty ratio, respectively. Figure 1 shows the basic structure of an isolated transformer DC-DC converter. It consists of 4 parts, which are the high-speed switching element, the transformer used to isolate the output port from the input port as well as boosting or dropping down the output voltage, the diode and capacitor used as a rectifier and filter, respectively, and the controller.



**Fig. 1.** Basic structure of an isolated transformer DC-DC converter

**Table 1.** Input to output voltage ratio equation

Topology	Input to output voltage ratio
Flyback	$V_{OUT} = \frac{N_2}{N_1} \frac{D}{1-D} V_{IN}$
Forward	$V_{OUT} = \frac{N_2}{N_1} D V_{IN}$
Push-Pull	$V_{OUT} = 2 \frac{N_2}{N_1} D V_{IN}$
Half-bridge	$V_{OUT} = \frac{N_2}{N_1} D V_{IN}$
Full-bridge	$V_{OUT} = 2 \frac{N_2}{N_1} D V_{IN}$

### 3 Rule-Based Expert System for Designing an Isolated DC-DC Converter

Until now, the design and verification of DC-DC converters have been conducted by the designer with computer simulation programs such as SPICE[5]. It is the objective of the proposed rule-based system for non-experienced designers to be able to design a DC-DC converter which is comparable to that produced by experienced designers. This design automation tool automatically selects a suitable topology and decides the values of the inductance and capacitance. This tool can reduce the development time, facilitate redesign, and improve the accuracy of the design for the sake of cost-cutting and competitiveness.

An overview of the Design Automation System for isolated DC-DC converters is shown in Fig. 2. The rule-based design automation system consists of a rule base, an inference engine, and a component library.

#### 3.1 Design Automation Process

The design automation process can be divided into the following steps. First, the selection of the DC-DC converter topology based on the design specification given by the user. The flyback and forward converter are selected when the power requirement is lower than 150W. If the power requirement is between 150W to 500W, then the push-pull and half-bridge converters are selected. The full-bridge converter is used for power requirements higher than 500W[6]. Next, the selection of appropriate elements such as MOSFETs, diodes, capacitors, inductors, and a transformer for the selected topology. The flyback converter has the characteristic that it operates more stably in discrete mode than in continuous mode. For this reason, we let the design automation tool decides the values of the inductances for the transformer which are used to make flyback converter always operate in continuous mode. All other topologies are set to

operate in continuous mode. If semiconductor elements exist for the selected topology, then suitable elements are suggested. Otherwise, the process returns to the first step in order to allow the user to choose another topology and start again. Once the topology has been recommended, the SPICE simulator performs verification with the suggested elements. If none of the topologies have suitable elements for the DC-DC converter, the design automation tool makes no suggestion.

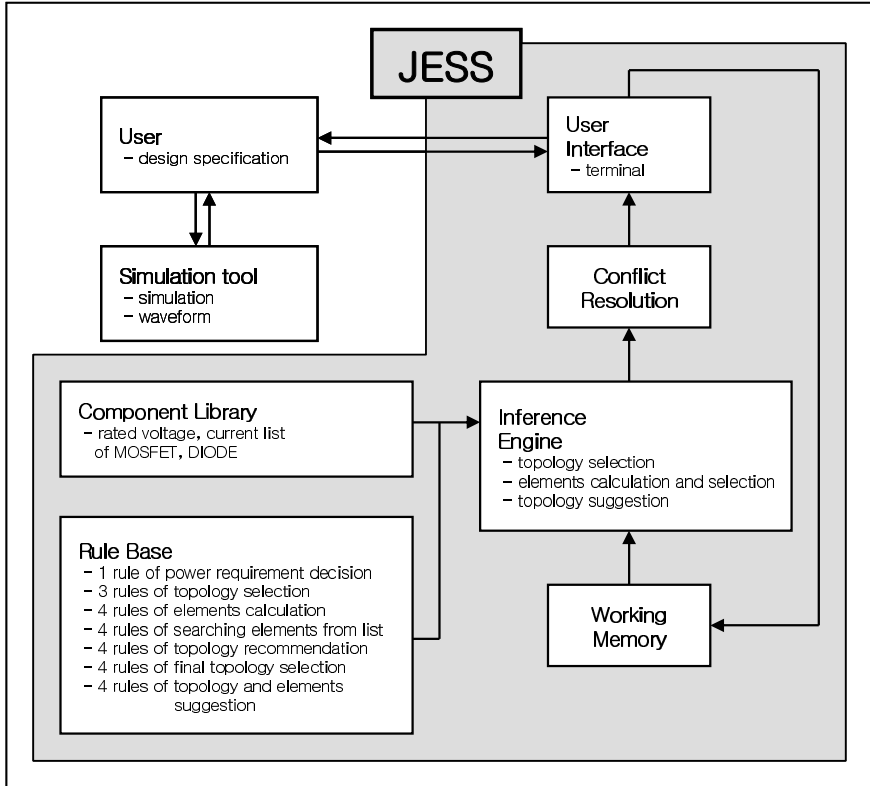


Fig. 2. Design Automation Process

### 3.2 Rule-Based Expert System

In order to develop a DC-DC converter design automation tool which is intelligent and efficient, the use of a rule-based approach is necessary. We developed 7 different kinds of rules for the design automation process. These are rules for the power requirement decision, topology selection, element calculation, searching of elements from a list, topology recommendation, final topology selection, and suggestion for the topology and elements. Some of the rule sets are shown below.

RULE 1: IF all specifications have been entered,  
THEN calculate the power requirement.

- RULE 2: IF the power requirement is lower than 150,  
 THEN the selected topologies are Flyback, Forward, Half-bridge, and Full-bridge.
- RULE 3: IF the power requirement is lower than 500,  
 THEN the selected topologies are Half-bridge and Full-bridge.
- RULE 4: IF the power requirement is higher than 500,  
 THEN the selected topology is Full-bridge.
- RULE 5: IF the topology is selected,  
 THEN calculate the elements for each selected topology  
 and search for suitable elements from the list.
- RULE 6: IF the topology is selected and suitable elements exist,  
 THEN recommend the selected topology.
- RULE 7: IF Flyback is the recommended topology,  
 THEN the final-topology is Flyback.
- RULE 8: IF Forward is the recommended topology and  
 Flyback is not recommended,  
 THEN the final-topology is Forward.
- RULE 9: IF Half-bridge is the recommended topology and  
 Flyback and Forward is not recommended,  
 THEN the final-topology is Half-bridge.
- RULE 10: IF Full-bridge is the recommended topology and  
 Flyback and Forward and Half-bridge is not recommended,  
 THEN the final-topology is Full-bridge.
- RULE 11: IF the final topology is determined,  
 THEN suggest the final topology and elements and exit.
- RULE 12: IF the recommended topology doesn't exist,  
 THEN exit without any suggestion.

We would like to explain the operation of the rule-based system using a real example. In this example, the specification for the converter is as follows: input voltage 50V, output voltage 10V, output current 1~2A, output ripple voltage 50mV, and switching frequency 33KHz. The user enters the design specification, and then the system asserts the power requirement to be 20W (RULE1). Because a power requirement of 20W is stipulated, the flyback, forward, half-bridge and full-bridge topologies are selected (RULE2). Then, the elements for each selected topology are calculated, and a search is made for suitable elements which are referred to as the calculated elements on the reference list (RULE5). The selected topology is recommended if suitable elements are available for it (RULE6). According to the recommended topology, the determination procedure for the final topology is performed (RULES 8~10). When the final topology has been determined, the design automation process is terminated after suggesting suitable elements for it (RULE 11).

The DC-DC converter design automation tool was implemented and verified with JESS (JAVA Expert System Shell) which supports the development of expert systems. JESS provides the inference engine and conflict resolution part of the tool. The rule set is stored in the rule base part, and the working memory contains the specification entered by means of the user interface. When the rule stored in the rule base is matched with the factor in working memory, JESS selects the rule to

fire by means of its stored conflict resolution strategy. The results can be observed through the user interface. Figure 3 shows the rule description for the expert system developed in JESS.

```
(defrule select-flyback
  (recommended-topology ?top&:(eq ?top flyback))
  ?final <- (final-topology ?fin&:(not (eq ?fin flyback)))
  =>
  (retract ?final)
  (assert (final-topology ?top))
)

(defrule select-forward
  (recommended-topology ?top&:(eq ?top forward))
  ?final <- (final-topology ?fin&:(not (or (eq ?fin flyback) (eq ?fin forward))))
  =>
  (retract ?final)
  (assert (final-topology ?top))
)

(defrule select-half-bridge
  (recommended-topology ?top&:(eq ?top half-bridge))
  ?final <- (final-topology ?fin&:(not (or (eq ?fin flyback) (eq ?fin forward) (eq ?fin half-bridge))))
  =>
  (retract ?final)
  (assert (final-topology ?top))
)

(defrule select-full-bridge
  (recommended-topology ?top&:(eq ?top full-bridge))
  ?final <- (final-topology ?fin&:(not (or (eq ?fin flyback) (eq ?fin forward) (eq ?fin half-bridge) (eq ?fin full-bridge))))
  =>
  (retract ?final)
  (assert (final-topology ?top))
)
```

Fig. 3. Rule Description in JESS

Table 2. Specification

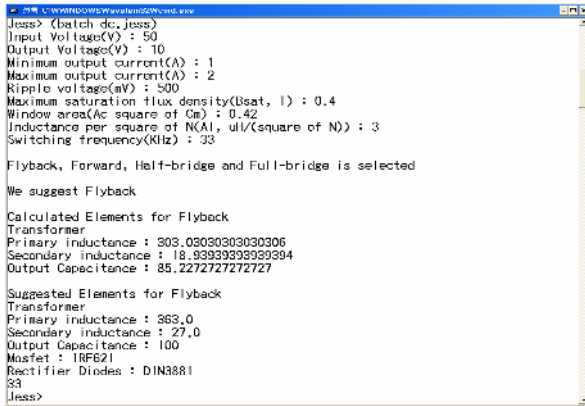
Specification	Case1	Case2	Case3	Case4
Input Voltage	50V	100V	100V	200V
Output Voltage	10V	48V	20V	100V
Output Current	1~2A	1.6~2A	1~10A	1~10A
Switching Frequency	33KHz	125KHz	60KHz	100KHz
Output Voltage Ripple	500mV	480mV	1V	1V

## 4 Experimental Results

In order to verify its performance, some specifications are entered into the implemented design automation tool, as shown in Table 2. Table 3 shows the outputs of our expert system in JESS, which gives the suggested circuit topologies, switching elements and transformer, and determines the values of the inductances and capacitances. PSPICE simulation shows that the design automation tool suggests converters which have an efficiency of more than 90%. We can see that the design automation tool suggests a converter which is very similar to that which would be proposed by an experienced designer. A snapshot for the execution of the expert system is shown in Fig. 4.

**Table 3.** Results

	Case1	Case2	Case3	Case4
Power Requirement	20W	76.8W	200W	1000W
Selected Topology	Flyback	Forward	Half-bridge	Full-bridge
Output Voltage	9.1V	47.1V	18.9V	93.1V
Transformer	324uH 25uH	1728.0uH 2523uH	1875uH 588uH	10800uH 17328uH
Inductance	.	75uH	27uH	75uH
Capacitance	100uF	100uF	100uF	22uF
Mosfet	IRF621	IRF621	IRF631	IRF641
Diode	D1N3881	D1N3881	D1N3891	D1N3892



**Fig. 4.** Execution of the Rule-Based Expert System in JESS

## 5 Conclusion

In this paper, we present a design automation tool for DC-DC converters which form a core part of an SMPS. We developed this tool using the forward chaining method of a rule-based system that can produce a suitable design from the given specification. The proposed system can select a suitable circuit topology, switching elements and transformer, and determine the values of the inductances and capacitances from the

given specification automatically. The experiment and verification were performed in JESS. We confirmed that the design automation tool is able to design a high efficiency converter just as an experienced designer would do. We hope that the proposed system will enable non-experienced designers to design SMPSS which are as efficient as those developed by experienced designers, while reducing the complexity of the design process.

## References

1. R. D. Middlebrook and S. Cuk, "Advances in Switched-Mode Power Conversion", Vols. 1, 2, and 3, Teslaco, Pasadena, CA, 1983.
2. R. D. Middlebrook and S. Cuk, "A General Unified Approach to Modeling Switching - Converter Power Stages", IEEE Power Electronics Specialists Conference, 1976 Record, pp. 18-34.
3. F. Barzegar, S. Cuk, and R. D. Middlebrook, "Using Small Computers to Model and Measure Magnitude and Phase of Regular Transfer Functions and Loop Gain", Powercon 8, The English International Solid-State Power Electronics Conference, Dallas, TX, April 1981.
4. V. G. Bello, "Computer-Aided Analysis of Switching Regulators Using SPICE2", IEEE Power Electronics Specialists Conference, 1980 Record, 80CH1529-7, pp. 3-11.
5. Yim-Shu Lee, "Computer-Aided Analysis and Design of Switched Mode Power Supplies", Marcel Dekker, Inc., 1993
6. Marty Brown, "Power Supply Cookbook", Motorola, 1994.

# DNA-Based Evolutionary Algorithm for Cable Trench Problem

Don Jyh-Fu Jeng<sup>1</sup>, Ikno Kim<sup>2</sup>, and Junzo Watada<sup>3</sup>

Graduate School of Information, Production and Systems  
Waseda University

2-7 Hibikino, Wakamatsu, Kitakyushu, Fukuoka 808-0135, Japan

<sup>1</sup> don\_jeng@yahoo.com, <sup>2</sup> oktoberkim@hotmail.com,

<sup>3</sup> junzow@osb.att.ne.jp

**Abstract.** An evolutionary DNA computing algorithm is proposed for solving a cable trench problem in this paper. The cable trench problem is a combination of the shortest path and minimum spanning tree problems, which makes it difficult to be solved by conventional computing method. DNA computing is applied to overcome the limitation, where fixed-length DNA strands are used in representing numerical values and the weights are varied by melting temperatures. Biochemical techniques in terms of DNA thermodynamic properties are used for effective local search of the optimal solution.

## 1 Introduction

Ever since scientists discovered that conventional silicon-based computers have an upper limit in speed, they have been searching for alternative media with which to solve computational problems. That search has led them among other places, to deoxyribonucleic acid (DNA). Scientists have found the new material which has the potential to be the next generation of microprocessors, the DNA, and the technique is known as DNA computing. DNA computing began in earnest with Adleman [1] solved a Hamiltonian path problem in 1994 [3]. Since then, the implementation of computation using DNA has attracted high attention in the fields of computer science and biochemistry.

In the last decade, many evolutionary algorithms have proposed in terms of DNA computing. The massive parallelism of the DNA computing provides an opportunity to solve a host of difficult problems, especially the NP (nondeterministic polynomial time) problems. NP problems are a class of mathematical problems which have most likely exponential complexity, for which no efficient solution has been found yet [2].

The cable trench problem (CTP), classified as network routing problem, is a combination of two problems universally discussed in operations research and management science, the shortest path and minimum spanning tree problems, which is difficult to be solved by conventional computer. CTP is a well-known NP-complete problem, in which there is a tradeoff between the fixed cost associated with constructing the network and a variable cost associated with operating it [12].



Seldom researches in DNA computing study the representation of numerical value in DNA strands. Nevertheless, many practical applications involve edge-weighted graph such as CTP. There exist previous works to represent the numerical data with DNA, but the results are not satisfactory yet [5]. Lee et al. [5] recently proposed the temperature gradient method to overcome the drawbacks of previous works. In this paper, we expand the capability of DNA computing to solve numerical optimization problems in the example of CTP, specifically, to solve an instance consisting of six vertices and eight edges of four differing path lengths.

The remainder of this paper is organized as follows: In section 2, we briefly review the CTP. The DNA computing is then introduced in section 3. The DNA encoding scheme and molecular algorithm for solving the CTP is described in section 4. Result and analysis are given in section 5. Section 6 recaps our conclusions.

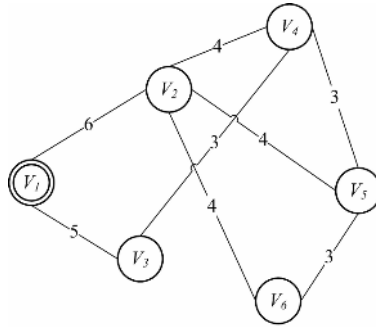
## 2 Cable Trench Problem

In a manufacturing plant design, for example, the connection between buildings and a computer room is an application of CTP. The computer room and all other buildings are the vertices to be connected. The only allowable routes between two buildings for digging trenches and laying cables define the edges of the graph. A trench may carry more than one cable once it is dug and the trench cost is proportional to the total trench distance. The cable cost is proportional to the total length of the cable required [15].

Conceptually, the CTP is a combination of the minimum spanning tree problem and the shortest path problem. Given a spanning tree  $T$  of graph  $G$ , denote the total length of the spanning tree by  $l(T)$  and the sum of the path lengths  $p_i$  from node 0 to each node  $i$  in  $T$  by  $s(T)$ . The CTP examines the tradeoff between  $s(T)$  and  $l(T)$ . The weighted sum version of the CTP is to find  $T$  such that  $\alpha l(T) + (1-\alpha)s(T)$  is minimized.

The difficulty of the CTP comes from the complexity of solving the weighted sum version depends heavily on the value of  $\alpha$ . If  $\alpha \rightarrow 1$ , the solution to this problem is a minimum spanning tree. On the contrary, if  $\alpha \rightarrow 0$ , then the problem becomes a shortest paths tree. Solving this problem, however, for values of  $\alpha$  arbitrarily close to one, i.e., finding among all minimum spanning trees the spanning tree  $T$  that minimizes  $s(T)$ , is an NP-hard optimization problem, whereas the problem of finding the shortest paths tree that minimizes  $l(T)$  can be solved in polynomial time. A proof of this fact is contained in [15]. The cases  $\alpha = 0$  and  $\alpha = 1$  are of course both solvable in polynomial time, and hence, the ideal point can be computed in polynomial time. For general  $\alpha$ , it is easily shown that this problem is NP-hard. Hence, the weighted sum version of the CTP exhibits the very interesting property that the difficulty of a particular instance depends heavily on the actual weight.

A CTP as shown in Fig. 1, for example, giving a graph  $G$  with vertices  $V = \{1, 2, 3, 4, 5, 6\}$  and edges  $E = \{(1, 2), (1, 3), (2, 4), (2, 5), (2, 6), (3, 4), (4, 5), (5, 6)\}$  with edge lengths 6, 5, 4, 4, 4, 3, 3, and 3, respectively, is an instance to be solved by DNA computing in this paper.



**Fig. 1.** A 6-vertex, 8-edge, 4-weight cable trench problem. A graph  $G$  with vertices  $V=\{1, 2, 3, 4, 5, 6\}$  and edges  $E=\{(1, 2), (1, 3), (2, 4), (2, 5), (2, 6), (3, 4), (4, 5), (5, 6)\}$  with edge lengths 6, 5, 4, 4, 4, 3, 3, and 3, respectively.

### 3 DNA Computing

Since Adleman’s pioneering accomplishment in 1994 [1], DNA computing has been applied to various fields of research including combinatorial optimization [10], massive parallel computing [7], Boolean circuit development [11], nanotechnology [18], very large scale database [13], etc. DNA plays the role of memory in nature, which carries the genetic information of an organism to be copied into the next generation of the species. DNA computing is a form of computing which uses DNA and molecular biology, instead of the traditional silicon-based computer technologies.

We use DNA molecules as information storage media, often DNA sequences of about 8-20 base pairs are used to represent bits, and numerous methods have been developed to manipulate and evaluate them. DNA is a convenient choice, as it is both self-complementary (A to T, and G to C), allowing single-stranded DNA to match its own Watson-Crick complement, and can easily be copied. The techniques of molecular biology can be used for manipulating DNA, including restriction enzyme digestion, ligation (by DNA ligase), sequencing (by dideoxy method), amplification (by polymerase chain reaction, PCR), gel electrophoresis, fluorescent labeling, and so on, which all giving DNA a surplus over alternative computational media [4].

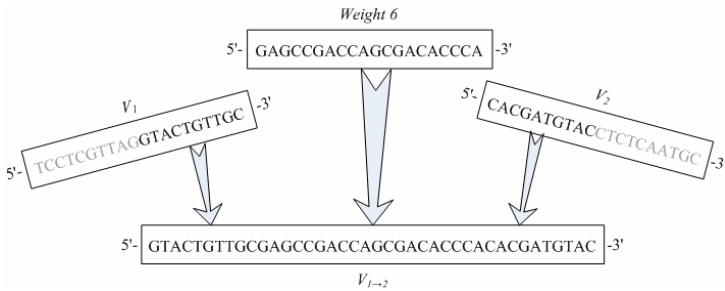
DNA computing has many advantages over conventional computing methods that utilize solid-state semiconductors. Although DNA computing performs individual operations slowly, it can execute billions of operations simultaneously. This fact is contrasted with the silicon-based computers where individual operations are very fast; however, the operations are executed sequentially. The massive parallelism of DNA computing comes from the huge number of molecules which chemically interact in a small volume [8]. DNA also provides a huge storage capacity since they encode information on the molecular scale [5].

## 4 Molecular Algorithm for Solving CTP

### 4.1 Encoding Scheme

The concept of encoding scheme is to design the DNA sequences that have heavier weights with higher melting temperature than those lighter weights. The DNA sequences that represent vertices are fixed length, and weights are distinguished by melting temperatures of the given DNA strands.

As shown in Table 1, firstly, each vertex is randomly associated with a 20-mer sequence of DNA, denoted  $V_i$ , which has a similar melting temperature due to vertex sequences should contribute equally to the thermal stability of paths. Weight sequences are designed to have different melting temperatures depending on its weight. The lighter in weight, the lower the melting temperature is. In other words, the more economical path has a lower melting temperature. The edge between vertices are generated by the beginning vertex, weight sequence, and ending vertex. For each edge  $i \rightarrow j$  in the graph, an oligonucleotide  $V_{i \rightarrow j}$  is created that is the 3' 10-mer of  $V_i$  followed by the weight sequence of path length, and then the 5' 10-mer of  $V_j$ . The edge from vertex 1 to vertex 2 in Fig. 1, for example, the 3' 10-mer of  $V_1$ : 'GTACTGTTGC' followed by the weight sequence of 6: 'GAGCCGACCAGCGACACCCA', then the 5' 10-mer of  $V_2$ : 'CACGATGTAC', as illustrated in Fig. 2.



**Fig. 2.** Example of the encoding scheme for edge from  $V_1$  to  $V_2$ . The 3' 10-mer of  $V_1$ : 'GTACTGTTGC' followed by the weight sequence of 6: 'GAGCCGACCAGCGACACCCA', then the 5' 10-mer of  $V_2$ : 'CACGATGTAC'.

The key factor of our encoding scheme is the melting temperature of a DNA strand. DNA amplification and detection techniques often depend on oligonucleotide melting temperature ( $T_m$ ). The  $T_m$  of a DNA duplex, defined as the temperature where one-half of the nucleotides are paired and one-half are unpaired [16]. The  $T_m$  indicates the transition from double helical to random coil formation and is related to the DNA GC (guanine-cytosine) base content [9]. Usually expressed as a percentage, it is the proportion of GC-base pairs in the DNA molecule or genome sequence being investigated. GC-pairs in the DNA are connected with three hydrogen bonds instead of two in the AT-pairs. This makes the GC-pair stronger and more resistant to denaturation by high temperatures [17]. The most accurate prediction of  $T_m$  for oligonucleotide DNA uses

the thermodynamic nearest-neighbor (N-N) model [14], which calculates the  $T_m$  by the enthalpy and entropy of all N-N pairs. In this research,  $T_m$  is calculated by N-N model with  $1\mu\text{M}$  oligonucleotide concentration and 50 mM salt concentration as shown in Table 1.

**Table 1.** DNA sequences associated with each vertex and weight for the 6-vertex, 8-edge, 4-weight cable trench problem. The melting temperature ( $T_m$ ) to each DNA sequence is calculated by nearest-neighbor model with  $1\mu\text{M}$  oligonucleotide concentration and 50 mM salt concentration.

	DNA Sequence (5'→3')	$T_m$ (°C)	GC Content
Vertex			
$V_1$	TCCTCGTTAGGTACTGTTGC	46.02	50%
$V_2$	CACGATGTACCTCTCAATGC	45.59	50%
$V_3$	TTCATGGTGCCTTGACGTGAG	47.59	50%
$V_4$	GCGGTTCTAAATTCGGTCAC	46.51	50%
$V_5$	ATTGGACCCAGATGCAAAGG	46.92	50%
$V_6$	GTTAGACCTCGCGTTGCTAT	46.97	50%
Weight			
3	ATGTTGGAGTTATTGCCAAC	43.18	40%
4	GCGTAACGTTACCGTTGTGT	48.48	50%
5	TGCTGGGGTATGGCCTCACA	51.88	60%
6	GAGCCGACCAGCGACACCCA	55.84	70%

## 4.2 Hybridization and Ligation

For each Watson-Crick complement of  $V_i$  in a CTP graph, and for each edge  $i \rightarrow j$  were mixed in a single ligation reaction. The added amount of an edge is varied according to weight, where as the weight increased, the amount is decreased. The oligonucleotide mixture is heated to  $95^\circ\text{C}$  and cooled to  $20^\circ\text{C}$  at  $1^\circ\text{C}/\text{min}$  for hybridization. The reaction mixture was then subjected to a ligation [5].

## 4.3 PCR Amplification, Gel Electrophoresis, and Affinity Separation

Double-stranded DNAs with sticky ends were generated from hybridization and ligation. To satisfy the condition of a CTP, the route must begin at a specified vertex ( $V_1$  in Fig. 1) and the route must pass by all the vertices ( $V_2$  to  $V_6$  in Fig. 1). PCR can be applied to reach the former requirement, which it is a technique for amplifying DNA that rapidly synthesize many copies of a specific DNA segment by providing specific complementary sequences (primers) and enzymes (DNA polymerases). Unlike the

Hamiltonian path problem Adleman proposed [1], CTP route may end at any vertex that hard to predict. Hence, we let the route begin at  $V_i$  and pass by every vertex, then return back to  $V_i$ . In other words, the CTP route encoded by DNA would make the round trip and back to where it originally started. The DNA strands of  $V_i$  and Watson-Crick complement of  $V_i$  are used as two primers in two successive PCRs to re-produce the routes that both begin and end at  $V_i$ .

To test the latter requirement, agarose gel electrophoresis is applied. Agarose gel electrophoresis is a method to separate DNA strands by size, and to determine the size of the separated strands by comparison to strands of known length. All PCR products are sieved by agarose gel electrophoresis, and the unreasonable lengths are excluded due to the candidate paths should be at least twice of six-vertices long (round trip).

To verify the DNA strands pass by every vertex, the product from the above step is affinity-purified with a biotin-avidin magnetic beads system. The single-stranded DNA is generated from the above step produced double-stranded DNA and then incubated with Watson-Crick complement of  $V_i$  that conjugated to magnetic beads. Only those single-stranded DNA molecules that contained the sequence  $V_i$  annealed to the bound are retained. This process is repeated six times with Watson-Crick complement of six vertices to ensure every vertex has passed by.

#### 4.4 Denaturation Temperature Gradient PCR

The denaturation temperature gradient PCR (DTG-PCR) is a modified PCR method that denaturation temperature changes with cycle [6]. DTG-PCR is a specified PCR protocol that modifies the denaturation temperature profile. If the denaturation temperature is decreased to a certain level in PCR, the DNA strands with denaturation temperatures lower than that temperature will be denatured and amplified. As the denaturation temperature is increased cycle by cycle in PCR, other DNA strands with higher denaturation temperature will also be amplified. However, the economical paths that have lighter weights will be amplified more and will occupy the major part of the solution. Hence, they can be detected more easily [5].

#### 4.5 Temperature Gradient Gel Electrophoresis

Temperature gradient gel electrophoresis (TGGE) is an electrophoresis method for separation of nucleic acids like DNA or RNA, which uses the temperature dependent change of conformation for separating molecules. The TGGE is applied in this study for finding the most economical route among other possibilities.

## 5 Result and Discussion

The optimal route to CTP is determined by most economical round trip. Since the temperature difference between weights is doubled, a more accurate result can be benefited by the temperature gradient methods in DTG-PCR and TGGE.

In the step of conventional gel electrophoresis, the DNA strands that are twice of six-vertices long and above are collected as candidates to the affinity separation. The DNA strands that are not passing by every vertex at least once will then be excluded.

In DTG-PCR, the denaturation temperature starts at low temperature in 70°C in the beginning cycles of PCR, which is lower than the melting temperatures of template strands. Then the denaturation temperature was gradually increased until reaches 95°C and maintained at the same temperature for the remaining cycle. After this process, one main band is observed in the gel which might contain three different DNA strands of the possible routes as shown in Table 2. These strands are of the same length which cannot be separated by the conventional gel electrophoresis. From the algorithm design, however, they have distinct behaviors in terms of melting temperature and can be separated by TGGE. TGGE is a gel electrophoresis method that is based on the correlation of the melting characteristic of a DNA strand to its electromigration [5], which the most economical route can be distinguished from other possible routes. The GC content and its  $T_m$  among those possible routes are shown in Table 2, which the DNA strands corresponded to the route ‘1→3→4→2→4→5→6→5→4→3→1’ has the most economical weight. The final sequencing results in finding the optimal solution of  $\{(1, 3), (2, 4), (3, 4), (4, 5), (5, 6)\}$  to our CTP.

**Table 2.** Possible routes resulting from DTG-PCR, and their distinct behavior in melting temperature and GC content

Route	$T_m$ (°C)	GC Content
1→2→6→2→5→2→1→3→4→3→1	82.15	52.1 %
1→2→6→2→1→3→4→5→4→3→1	82.09	51.5 %
1→3→4→2→4→5→6→5→4→3→1	80.57	47.89 %

## 6 Concluding Remarks

In this paper, an evolutionary DNA computing algorithm is proposed in solving the CTP. The melting temperature plays the key concept in numerical representation of weight. The quantitative expression of real numbers is implemented by fixed-length DNA strands. The proposed algorithm successfully finds the optimal solution of the 6-vertex, 8-edge, 4-weight cable trench problem.

In the previous investigation of DNA computing, none of any research has engaged in both the shortest path and the minimum spanning tree problem. The combination of these two objectives becomes a more realistic application; however, due to its characteristic of NP-completeness, the conventional computer should eventually encounter its limit in solving such problem. With DNA molecules’ enormous parallelism, DNA computing provides an opportunity to overcome the limitation of a conventional computer. The proposed algorithm would superior then the conventional computing method as the problem size increases.

## References

1. Adleman, L.M.: Molecular Computation of Solutions to Combinatorial Problems. *Science* 266 (1994) 1021-1024
2. Garey, M.R., Johnson, D.S.: *Computers and Intractability: A Guide to the Theory of NP-completeness*. W. H. Freeman and Company (1979)
3. Jeng, D.J.-F., Watada, J., Kim, I.: Solving a Real Time Scheduling Problem Based on DNA Computing. In: *Proceedings of International Conference on Intelligent Technologies and Applied Statistics*, Taipei, Taiwan. (2005) 317-322
4. Jeng, D.J.-F., Wu, J.-Y.: Recent Development of DNA Computing. In: *Proceedings of the International Conference on Recent trends in Information Systems*, National Engineering College, Kovilpatti, Tamil Nadu, India. (2006) 579-584
5. Lee, J.Y., Shin, S.-Y., Park, T.H., Zhang, B.-T.: Solving Traveling Salesman Problems with DNA Molecules Encoding Numerical Values. *BioSystems* 78 (2004) 39-47
6. Lee, J.Y., Zhang, B.-T., Park, T.H.: Effectiveness of denaturation temperature gradient-polymerase chain reaction for biased DNA algorithms. In: *Preliminary Proceedings of the Ninth International Meeting on DNA Based Computers* (2003) 208
7. Liu, Q., Wang, L., Frutos, A.G., Condon, A.E., Corn, R.M., Smith, L. M.: DNA computing on surface. *Nature* 403 (2000) 175-179
8. Maley, C.C.: DNA computation: theory, practice, and prospects. *Evolutionary Computation* 6(3) (1998) 201-229
9. Marmur, J., Doty, P.: Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. *Journal of Molecular Biology* 5 (1962) 109-118
10. Ouyang, Q., Kaplan, P.D., Liu, S., Libchaber, A.: DNA solution of the maximal clique problem. *Science* 278 (1997) 446-449
11. Owenson, G.G., Amos, M., Hodgson, D.A., Gibbons, A.: DNA-based logic. *Soft Computing* 5(2) (2001) 102-105
12. Ralphs, T.K., Saltzman, M., Wiecek, M.: An Improved Algorithm for Biobjective Integer Programming. To appear in *Annals of Operations Research*
13. Reif, J.H., LaBean, T.H., Pirrug, M., Rana, V.S., Guo, B., Kingsford, C., Wickham, G.S.: Experimental construction of a very large scale DNA database with associative search capability. In: *the 7th International Workshop on DNA-Based Computers* (2001) 241-250
14. SantaLucia Jr., J.: A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. In: *Proceedings of the National Academy of Sciences of the United States of America* 95 (1998) 1460-1465
15. Vasko, F. J., Barbieri, R.S., Rieckts, B.Q., Reitmeyer, K.L., Stott Jr., K.L.: The cable trench problem: combining the shortest path and minimum spanning tree Problems. *Computer & Operations Research* 29 (2002) 441-458
16. Wetmur, J.G.: DNA probes: applications of the principles of nucleic acid hybridization. *Critical Reviews in Biochemistry and Molecular Biology* 26(3) (1991) 227-259
17. Wikipedia, GC-content, retrieved December 27, 2005, Web site: [http://en.wikipedia.org/wiki/GC\\_content](http://en.wikipedia.org/wiki/GC_content).
18. Winfree, E., Lin, F., Wenzler, L.A., Seeman, N.C.: Design and self-assembly of two-dimensional DNA crystals. *Nature* 394(6693) (1998) 539-545

# The Influences of R&D Expenditures on Knowledge-Based Economy in Taiwan

Lily Lin

Department of International Trade, China University of Technology  
No. 56, Sec. 3, Shinglung Rd., Taipei, Taiwan 116, R.O.C.  
lily@cute.edu.tw

**Abstract.** This article adopts Multivariate Time Series Approach to analyze the impacts of research and development (R&D) on overall economy. By using Johansen-Juselius Cointegration Methodology and Vector Autoregression (VAR) approach, we analyze the causality between R&D and economic contribution, such as employment, investment and consumer's expenditures. The results suggest that in the long term, public funded R&D expenditures has stronger long run dynamics of the relationship to investment and employment than private funded R&D does. Moreover, the study also proves that a four-year lagged effect of R&D expenditure existing to employment and investment.

## 1 Introduction

As the tide of globalization comes with the era of knowledge-based economy, the industries in Taiwan are eager to enhance their competitiveness through research and development (R&D) and technology innovations on the base of comparative advantages of existing technology and capital. The purposes of invigorating technology should be not only the sustainable economic growth and enhancement of people's welfare but also contributions to the sustainable development of the world. In addition to the enhancement of industry competitiveness, the sustainable growth is one of the primary goals of expenditures for research and development of the government. Therefore, in this study, we would like to know the effects of R&D expenditures of both public and private sectors on economic growth in Taiwan.

In the substantial economic growth rate, the contributing factors are divided into three types: labor force, capital and technological improvement-dominated total factor productivity. For promoting industry upgrade, Taiwan should strengthen technology-intensive industries. The governmental R&D funds should not only inject into the R&D of basic technologies, talent cultivation and improvement of research environment but also positively and effectively exercise available governmental resources to spur input from private sectors into R&D and bring along industry upgrades.

However, there is no significant growth in R&D expenditures in the government in current stage. The study aims to explore the influence of reduction of government budget for technology development on overall national economic growth and industry competitiveness. Citing the effects of R&D expenditures in public sectors and private



sectors on employment, investment and consumer's expenditures, the study is planning to adopt Multivariate Time Series Approach to analyze the effects of research and development in two sectors and the effects of governmental research and development expenditures on overall economy.

## 2 Literature Review

The study of the effects of R&D on output and productivity shows that R&D plays a statistically and economically significant role in explaining productivity growth. However, not all R&D is created equal. Basic research has different effects from applied research and from development, and privately funded R&D has different effects from publicly funded R&D. Our focus here is on the differences between the effects of public and private R&D. According to previous studies, all of the positive effects of R&D on output and productivity are traced to privately funded R&D. However, our study has different results.

Perhaps the main conclusion that emerges from the R&D-driven endogenous-growth literature is that public policies can have long-run growth effects by influencing the incentives firms have to innovate. Romer [8], Segerstrom, Anant, and Dinopoulos [9], and Aghion and Howitt [1] all find that R&D subsidies encourage firms to devote more resources to R&D activities and as a result increase the long-run rate of economic growth.

In this study, we analyze the effects of private and public R&D on economic growth performance in the context of a multivariate time-series model. The choice of this methodological approach, which represents our point of departure from the literature, is based on a casualty concept. This article contributes to the economic growth debate by discussing the casual relationship between two sectors funded R&D and economic growth in Taiwan in a VAR econometric context by using both Granger causality and impulse response analysis. This is the first attempt to use Multivariate Time Series Approach to test the hypothesis that public sector R&D importance leads economic growth in Taiwan. The article is designed as follows: Section 2 describes the data and its sources. In Section 3, we introduce the model and present the empirical results. Finally, Section 4 is devoted to the concluding remarks.

## 3 Preliminary Data Analysis

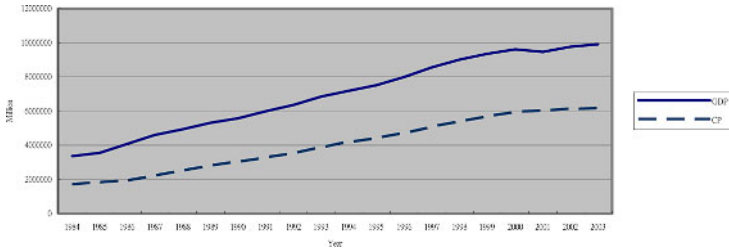
### 3.1 The Data

Our analysis focuses on the relationships among GDP, consumption, investment, R&D expenditures of two sectors, and employment in Taiwan. All annual data in this study for Taiwan consists of the period 1984-2003, therefore, the data set contains 20 yearly observations. All data are measured in billions of 1990 dollars as well as employment is measured in 10,000 full-time equivalent workers. The description and source for all variables is shown as Table 1.

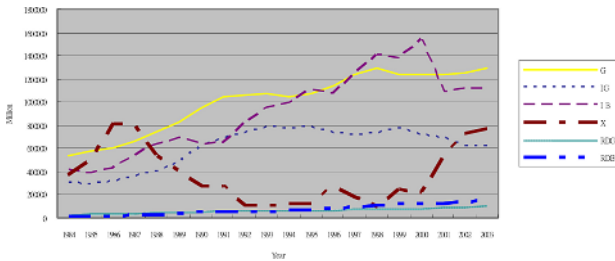
**Table 1.** Reference for Data

Variable	Description	Source
GDP	Real GDP, (1990=100), 1984-2003	Directorate-General of Budget, Accountin and Statistics, Executive Yuan, R.O.C.
CP	Private Final Consumption Expenditure (1990=100), 1984-2003	Same as above
G	Government Final Consumption Expenditure, (1990=100), 1984-2003	Same as above
$I_G, I_B$	Public Sector, Private Enterprises Investment (1990=100), 1984-2003	Same as above
X	Net Exports (1990=100), 1984-2003	Same as above
$RD_G, RD_B$	Public, Private sector R&D Expenditures (1990=100), 1984-2003	"Indicators of Science and Technolog Republic of China" published Nation Science Council
E	Employment (Unit: 10,000 persons), 1984-2003	Directorate-General of Budget, Accountin and Statistics, Executive Yuan.

Fig. 1, 2 and 3 indicate that, although variable  $RD_G$  and X (particularly in X) do not show obvious trend during the period 1984-2003, other variables have more stable as a up going trend.



**Fig. 1.** Plots of data for Taiwan (GDP, CP)



**Fig. 2.** Plots of data for Taiwan (G,  $I_G, I_B, X, RD_G, RD_B$ )

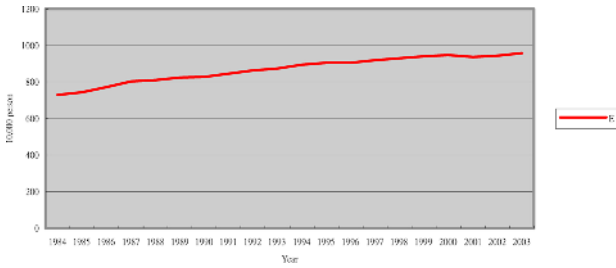


Fig. 3. Plots of data for Taiwan (E)

Table 2. Testing the Null Hypothesis of a Unit Root

Variable	Deterministic Components	Optimal Lag (AIC)	Lag	Test Statistic	Critical Value		
					10%	5%	1%
GDP	Constant and trend	1		-3.685	-3.287	-3.691	-4.572
CP	Constant and trend	2		-4.770	-3.298	-3.710	-4.616
G	Constant and trend	1		-3.867	-3.298	-3.710	-4.616
I <sub>G</sub>	Constant	2		-4.923	-2.667	-3.052	-3.887
I <sub>B</sub>	Constant and trend	1		-4.868	-3.287	-3.691	-4.572
X	Constant	1		-2.972	-2.661	-3.040	-3.857
RD <sub>G</sub>	Constant and trend	2		-5.121	-3.298	-3.710	-4.616
RD <sub>B</sub>	Constant and trend	1		-3.930	-3.287	-3.691	-4.572
E	Constant and trend	1		-3.771	-3.298	-3.710	-4.616

Note: AIC= Akaike information criterion.

## 4 Empirical Analysis

Our analysis consists of three steps. The first step is checking stationary of the series, the second step is testing for cointegration, the third step we examine the impact of R&D on economic growth in Taiwan by using a Vector Autoregression (VAR) approach. Impulse response analysis and Granger causality test are applied to examine interrelationships between variables in the VAR system and, therefore, differs from the more usual approach.

### 4.1 Variables Stationary Checking

We first test for stationary. To decide the order of integration of these variables, we start by testing the null hypothesis of unit root. Traditional econometric techniques using time-series data implicitly assume that the data used in estimation are stationary. If stationary is violated, this could lead to spurious results. In analyzing the time-series data properties, the Augmented Dickey-Fuller (ADF) (Dickey and Fuller [4], [5]) unit root test is most commonly applied. The results of the augmented Dickey-Fuller (ADF) *t* test are shown in Table 2. The optimal lag structure was chosen using Akaike information criterion (AIC). A deterministic component was considered if statistically significant. In no case can we find evidence against the null

hypothesis that the series contain unit roots in levels. However, we reject the null hypothesis for first or second differences of all variables. Accordingly, the empirical evidence is that all variables are stationary in first or second differences.

**Table 3.** Johansen Cointegration Test

Variable	Optimal Lag (AIC)	Trace Statistic	5 % Critical Value	1 % Critical Value
GDP**	1	29.795	3.76	6.65
CP**	2	27.701	3.76	6.65
G**	1	16.311	3.76	6.65
I <sub>G</sub>	2	2.560	3.76	6.65
I <sub>B</sub> **	1	27.781	3.76	6.65
X**	1	42.748	3.76	6.65
RD <sub>G</sub>	2	0.358	3.76	6.65
RD <sub>B</sub> *	1	5.857	3.76	6.65
E*	1	6.223	3.76	6.65

Note: The level data variables have lineal trends except I<sub>G</sub> and X but the cointegrating equations have only intercepts (constant).

\*(\*\*) denotes rejection of the hypothesis significant at the 5%(1%) level with the trace test

## 4.2 Cointegration Analysis

Cointegration analysis, on the other hand, provides an estimation of the relationship between the variables. We employ the Johansen-Juselius [7] multivariate cointegration methodology to determine the number of cointegrating variables.

To investigate the possible existence of cointegration, we applied the ADF *t* test to the residuals obtained from regressing each variable on the remaining eight. This provides us with night different tests of the null hypothesis of no cointegration. The results are shown in Table 3. In all but two cases (I<sub>G</sub> and RD<sub>G</sub>), the value of trace statistic is smaller than the 5% critical value. This means that the null hypothesis of a unit root in residuals cannot be rejected and that, therefore, there is no evidence for the existence of cointegration among the variables under consideration. In the remaining case, denoted by “\*, \*\*” the null can be rejected.

Through empirical analysis, the trace statistic finds that there are seven cointegrating vectors at 1% or 5% significance level when we assume linear trend in the data except I<sub>G</sub> and X and allow for an intercept and a trend term in the cointegrating relationship. Therefore, we conclude that there is a cointegrating relationship among GDP, consumption, private enterprises investment, private sector R&D expenditures and employment in Taiwan.

## 4.3 VAR Estimation

Next, we use a VAR modeling framework to capture the dynamics of the relationship between R&D and other economic factors while avoiding the pitfalls of endogeneity and integration of the variables. Moreover, we will utilize VAR estimation in

investigating the R&D-economic nexus by using the innovation accounting technique (impulse response function and variance decomposition) to investigate causality.

Juselius [7] procedure has superior properties than the Engle-Granger [6] two step procedure. The maximum likelihood methodology suggested by Juselius is based on the following VAR model:

$$Y_t = \mu + A_1 Y_{t-1} + A_2 Y_{t-2} + \dots + A_p Y_{t-p} + B X_t + \varepsilon_t \tag{1}$$

where  $Y$  is a  $(n \times 1)$  vector of endogenous variables,  $X$  is a  $(k \times 1)$  vector of exogenous variables,  $\mu$  is a  $(n \times 1)$  vector of constant terms,  $A_1, A_2, \dots, A_p$  are  $(n \times n)$  and  $B$  are  $(n \times k)$  coefficient matrices and  $\varepsilon$  is a  $(n \times 1)$  innovation vector of error terms with zero mean and constant variance. The reparameterization of (1) can be written as:

$$\Delta Y_t = \mu + \Gamma Y_{t-p} + \sum_{i=1}^{p-1} \Gamma_i \Delta Y_{t-i} + B X_t + \varepsilon_t \tag{2}$$

where  $\Gamma_i = -I + A_1 + \dots + A_i$  ( $i = 1, 2, \dots, p-1$ ), and  $\Gamma = -I + A_1 + \dots + A_p$ .

The rank of the matrix  $\Gamma$ , the matrix determining the long-run relationships between variables, is equal to the number of independent cointegrating vectors denoted by  $r$ . If  $r = 0$ , then the elements of  $Y$  are nonstationary, and (2) is a usual VAR in first differences. Instead, if the rank( $\Gamma$ ) is  $n$  and  $r = n$ , then the elements of  $Y$  are stationary.  $\Gamma Y_{t-p}$  is error-correction factor, if  $r = 1$ . For other cases,  $1 < r < n$ , there are multiple cointegrating vectors. The number of distinct cointegrating vectors can be obtained by checking the significance of characteristic roots of  $\Gamma$ . The empirical result is explained as Table 4.

According to Table 4, there is an important finding that public funded R&D expenditures has stronger long run dynamics of the relationship to investment and employment than private funded R&D does.

#### 4.4 Granger Causality Test

We also conducted a Granger causality test to verify the finding. Engel and Granger (1987) suggests that if cointegration exists between two variables in the long run, then, there must be either unidirectional or bi-directional Granger-causality between these variables, although this may not be uncovered using a finite sample. The result of Granger Causality Test is shown as Table 5. It indicates that R&D expenditures cannot have a direct effect on GDP, but both private enterprises investment and employment are directly affected by public sector R&D expenditures in lag four periods ( years ) test. It is interesting to note that the Ganger causality from public sector R&D expenditures to private enterprises investment and employment is more strongly obvious than the causality from private sector R&D expenditure to both. Therefore, plus the result of impulse response test, it clearly suggests that the increase of government R&D expenditures may increase investment and employment to result in economic growth in Taiwan.

Given the test of running for the appropriate lag length, we found the most appropriate lag length to be 4 (please refer to Table 5) for explanation the causality

between R&D expenditures and other factors. Therefore, we conclude that there is a four-year lagged effect of public funded R&D expenditures in Taiwan.

**Table 4.** Vector Autoregression Estimates

	I <sub>G</sub>	I <sub>B</sub>	E	RD <sub>G</sub>	RD <sub>B</sub>
I <sub>G</sub> (-1)	0.751421 [2.5053]**	-0.076300 [-0.09064]	4.18E-05 [ 0.73904]	0.030450 [ 1.34415]	-0.053123 [-1.7432]*
IB(-1)	0.107197 [1.43141]*	0.135658 [ 0.64541]	-1.74E-05 [-1.23241]	-0.017938 [-3.17]**	-0.000195 [-0.02569]
E(-1)	230.7232 [ 0.17423]	4350.933 [ 1.17062]	0.559679 [ 2.23882]**	-11.73462 [-0.11732]	-72.02475 [-0.53530]
E(-2)	686.7781 [ 0.60284]	-7865.634 [-2.4599]**	-0.200265 [-0.93121]	90.62393 [ 1.05321]	120.2182 [ 1.03860]
RDG(-1)	14.19094 [ 2.26510]**	27.39397 [ 1.55791]*	0.001120 [ 0.94733]	0.367045 [ 0.77568]	0.214287 [ 0.33664]
RDG(-2)	-7.759108 [-1.69684]*	-15.35496 [-1.19644]	4.71E-05 [ 0.05454]	-0.164446 [-0.47615]	0.355008 [ 0.76412]
RDB(-1)	4.594972 [ 1.22044]	-12.07563 [-1.14276]	-0.000794 [-1.11711]	0.485435 [ 1.70707]*	0.282990 [ 0.73977]
RDB(-2)	-8.230999 [-2.1746]**	-9.887955 [-0.93077]	4.66E-05 [ 0.06527]	-0.028813 [-0.10079]	-0.018037 [-0.04690]
GDP	-0.054115 [-1.05051]	0.471551 [ 3.2615]**	2.37E-05 [ 2.4334]**	-0.001071 [-0.27516]	0.012934 [2.4711]**
Adj. R-squared	0.979045	0.963923	0.992871	0.992613	0.996671
F-statistic	73.20739	42.29236	216.2503	208.6778	463.7051

t-statistics in [ ], \*, \*\*, \*\*\* denote significant at the 10%, 5%, 1% level

**Table 5.** Granger Causality Test

Lags: 2		Lags: 3		Lags: 4		Lags: 5	
Variable	P-value	Variable	P-value	Variable	P-value	Variable	P-value
I <sub>B</sub> =>GDP	0.01**	X=>CP	0.03*	X=>CP	0.02*	I <sub>B</sub> =>GDP	0.05*
X=>CP	0.01**	GDP=>I <sub>G</sub>	0.05*	I <sub>B</sub> =>GDP	0.00**	X=>GDP	0.02*
E=>G	0.02**	I <sub>B</sub> =>GDP	0.00**	X=>GDP	0.02*	X=>G	0.02*
X=> I <sub>G</sub>	0.03**			I <sub>B</sub> =>G	0.05**	I <sub>B</sub> =>I <sub>G</sub>	0.01**
				X=>G	0.04*	RD <sub>B</sub> => I <sub>G</sub>	0.02*
				RD <sub>G</sub> =>I <sub>B</sub>	0.05*	I <sub>B</sub> =>RD <sub>G</sub>	0.02*
				RD <sub>G</sub> =>E	0.04*	X=>RD <sub>G</sub>	0.02*
						RD <sub>G</sub> =>E	0.00**

Note: => indicates the direction of causality. \* significant at 5%; \*\* significant at 1%.

## 5 Conclusion

In this article, we examine the causal relationship between R&D expenditures and economic growth in Taiwan using annual time-series data for the 1984-2003. Based

on our study, the results as vector autoregression estimates and Granger causality test all show the relation existing between public funded R&D and private enterprises investment.

Though the results indicate a non-direct causality relation running from R&D expenditures to GDP, there is causality relation existing among R&D expenditures, private enterprises investment and employment. Based on the study results, in the long term public funded R&D expenditures has stronger dynamics positive relationship to investment and employment than private funded R&D does. This suggests that Taiwan Government should increase government budget of R&D expenditures for overall national economic growth and industry competitiveness. Moreover, the study also proves that there's a four-year lagged effect of R&D expenditure on employment and investment.

Perhaps the shortcoming of our study is that it is inherently incapable of examining lagged relationships and, therefore, is inappropriate for test. Also, the median estimates from VECM are, in general, more accurate than VAR. Differences between VAR and VECM can be both economically and statistically significant.

## References

1. Aghion, P. and Howitt, P.: A Model of Growth Through Creative Destruction. *Econometrica* Vol. 60, (1992) 323-351
2. Archibald, R.B., and M.P. Alfredo: Effects of Public and Private R&D on Private-Sector Performance in the United States. *Public Finance Review* Vol. 31(4), (2003) 429-451
3. Benat, B.O., and R.P. Andres. From R&D to Innovation and Economic Growth in the EU. *Growth and Change* Vol. 35(4), (2004) 434-456
4. Dickey, D., and W. Fuller: Distribution of the Estimators For Autoregressive Time Series With A Unit Root. *Journal of the American Statistical Association* Vol. 74, (1979) 427-431
5. Dickey, D.A., and W.A. Fuller: Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* Vol. 49(4), (1981) 1057-1072
6. Engle, R.F., and C.W.J. Granger: Co-integration and error correction: representation, estimation and testing. *Econometrica* Vol. 55(2), (1987) 251-276
7. Johansen, S., and K. Juselius: Maximum Likelihood Estimation and Inference on Cointegration – With Applications To the Demand For Money. *Oxford Bulletin of Economics and Statistics* Vol. 52(2), (1990) 169-210
8. Romer, P.: Endogenous Technological Change.” *Journal of Political Economic* Vol. 98, (1990) S71-S102
9. Segerstrom, Anant, and Dinopoulos: A Schumpeterian Model of the Product Life Cycle. *American Economic Review* Vol. 80, (1990) 1077-1092

# A Process Schedule Analyzing Model Based on Grid Environment

Huey-Ming Lee, Tsang-Yean Lee, and Mu-Hsiu Hsu

Department of Information Management, Chinese Culture University  
55, Hwa-Kung Road, Yang-Ming-San, Taipei (11114), Taiwan  
{hmlee, tylee}@faculty.pccu.edu.tw, hsu\_clay@yahoo.com.tw

**Abstract.** Grid computing architecture was defined to be a complete physical layer. The functions of information systems based on grid architecture are resources sharing, collaborative processing, etc. Resources are used by processes. System performance is calculated from resources usages. Process scheduling is more important when jobs are not uniformly distributed in all grid nodes. In this paper, we proposed a process schedule analyzing model based on grid computing architecture. This model can make all grid nodes be loading-balance. When the load of the node is heavy, it can select the other grid nodes to execute its job by the verification of supervisor grid node and then the job is transferred. Via implementing this model, we can have the best system performance.

## 1 Introduction

The term “Grid” was coined in the mid 1990s to denote a proposed distributed computing infrastructure for advanced science and engineering [8]. In grid environment, users may access the computational resources at many sites [6, 10]. The functions of information systems based on grid computing architectures are resources (e.g., CPUs, storages, etc.) sharing, collaborative processing, reliable and secure connection, etc. However, each resource of coordinate nodes in the grid environment, (e.g., CPU loading, memory rate of utilization, etc.) changes dynamically, how to optimize these resources usage is an important issue.

Many researchers have addressed the resource allocation problem. Lee and Yang [11] proposed a distributed backup agent which can assist user to backup the data files in different nodes storages, update replicas synchronously, and access the replicas if necessary. Lee *et al.* [12] proposed a dynamic supervising model which can utilize the grid resources, e.g., CPU, storages, etc., more flexible and optimal. Lee *et al.* [13] proposed a dynamic analyzing resources model which can receive the information about CPU usage, number of running jobs of each grid resource node to achieve load-balancing and make the plans and allocations of the resources of collaborated nodes optimize. Condor provides a general resource selection mechanism based on the ClassAds language [14, 18], which allows users to describe arbitrary resource requests and resource owner to describe their resources. Raman *et al.* [17] developed and implemented the classified advertisement (classad) matchmaking framework



which was designed to solve real problems encounter in the deployment of Condor. The matchmaker is used to match user requests with appropriate resources. Liu *et al.* [15] presented an extended set matching matchmaking algorithm that supports one-to-many matching of set-extended ClassAds with resources. Foster *et al.* [8] presented the grid resource allocation and management (GRAM). GRAM simplifies the use of remote systems by providing a single standard interface for requesting and using remote system resources for the execution of “jobs”. The most common use of GRAM is remote job submission and control. GRAM is designed to provide a single common protocol and API for requesting and using remote system resources, by provide a uniform, flexible interface to local job scheduling systems.

In the scheduling criteria, Silberschatz *et al.* [20] proposed that it includes CPU utilization, throughput, turnaround time, waiting time and response time; Stallings [21] and Tanenbaum [22] proposed that it not only includes the above five criteria but also deadlines, predictability, fairness, enforcing priority, balancing resources.

In the communication, Crowley [3] proposed that it has three modes of communication, saying procedure call within a process, message between processes and remote procedure calls. Nutt [16] proposed that remote files and applications can be read and written information from remote machines as if the file was stored on local storage devices.

From the above statements, we know that if we want to have good performance in the system based on grid environment, we should have load-balancing of all grid nodes in the system. When a new job is entered to be processed and the load of this node is heavy, we should schedule some jobs to the other nodes to process. Through the data base of grid information, we can find the best suitable grid nodes to process these jobs.

In this study, we proposed a process schedule analyzing model. Via the implementing this model, it can decide which node(s) may accept some jobs to process, also, we can have that the resource allocations are more effective and the best system performance.

## 2 Framework of the Proposed Model

In this section, we presented the framework of the proposed process schedule analyzing model based on grid computing architecture. We divided grid nodes as three kinds, saying, supervisor grid node ( $S_0$ ), backup supervisor grid node ( $B_1$ ) and execute grid nodes ( $X_i$ ).

The functions of supervisor grid node are as follows:

- (1) builds process grid node capacity information of all grid nodes;
- (2) collects grid node information;
- (3) processes the new job;
- (4) transfers job to other grid node;
- (5) verifies transfer job from requested grid node;
- (6) receives job transfer from other grid node;
- (7) writes grid node information and total account information to log file.

The functions of the execute grid node are as follows:

- (1) processes new job;
- (2) sends grid node information to the supervisor and receives other grid node information from supervisor;
- (3) requests to do job transfer through supervisor's verification;
- (4) do job transfer to the selected grid node and receive the complete job from the selected grid node;
- (5) receives job transfer from other grid node.

The functions of backup supervisor grid node are as follows:

- (1) does the functions of execute grid node normally;
- (2) does the function of supervisor grid node, when supervisor is inactive.

We presented the supervisor process schedule analyzing module (SPSAM) on the supervisor grid node, execute process schedule analyzing module (EPSAM) on the backup supervisor grid node and execute grid node, as shown in Fig. 1.

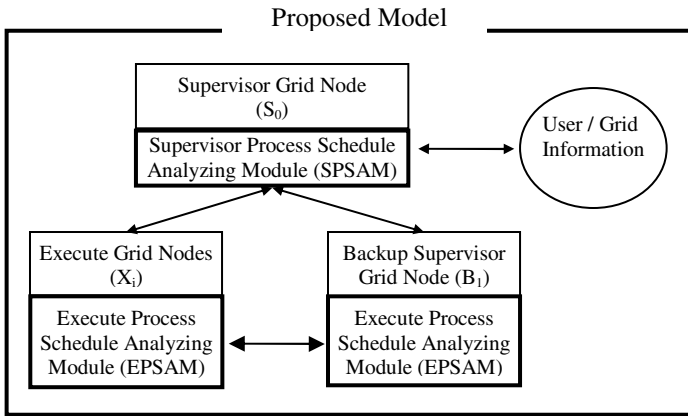


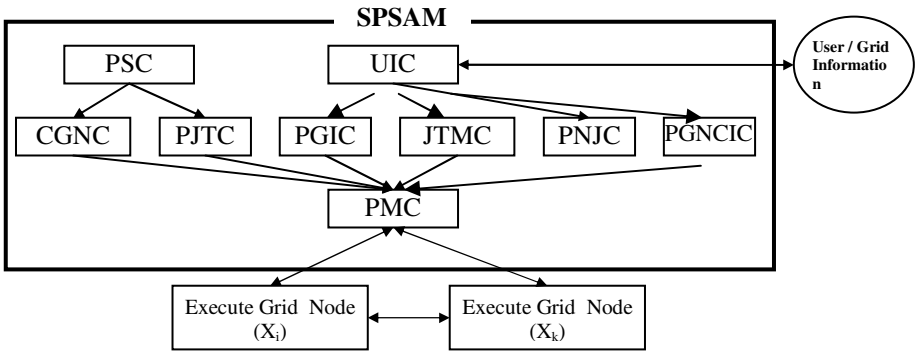
Fig. 1. Architecture of the proposed model

## 2.1 Supervisor Grid Node

We presented the supervisor process schedule analyzing module (SPSAM) on the supervisor grid node. There are nine components in this module, saying, process supervisor component (PSC), user interaction component (UIC), check grid node component (CGNC), process job transfer component (PJTC), process grid information component (PGIC), job transfer message component (JTMC), process new job component (PNJC), process grid node capacity information component (PGNCIC) and process message component (PMC), as shown in Fig. 2.

The functions of these components are as the follows:

- (1) Process supervisor component (PSC): Process supervisor component (PSC) periodically checks available grid nodes and the loading of the system. PSC can call CGNC to check whether the grid node is available, and call PJTC when loading is too heavy and transfer the job to the selected grid node.



**Fig. 2.** Framework of the SPSAM

(2) User interaction component (UIC): User interaction component (UIC) receives the grid node information, as shown in Table 1, and calls PGIC to update the supervisor grid information data base (SGIDB), and send the updated information to all grid nodes to update execute grid information data base (EGIDB). If it receives transfer job messages then calls JTMC to return yes or no. If it receives new job then calls PNJC to update job table (JT).

**Table 1.** Contents of the grid node information

Node-name	Job-count	CPU-usage	Running-count	Completed-count
-----------	-----------	-----------	---------------	-----------------

, where Node-name is name of this node, Job-count is total jobs in this node, CPU-usage is total CPU usage in this period, Running-count is total running jobs, Completed-count is total completed jobs in this period.

(3) Check grid node component (CGNC): If the supervisor node doesn't receive messages any more, the check grid node component (CGNC) will send messages, as shown in Table 2, to check whether this grid node is inactive or not, also, it writes total information of the system to log file, as shown in Table 3.

**Table 2.** Grid node inactive message from supervisor node

Grid node	Inactive? Yes/no
-----------	------------------

**Table 3.** Log file 1

Clock-time	Process-count	CPU-usage	Running-count	Completed-count
------------	---------------	-----------	---------------	-----------------

(4) Process job transfer component (PJTC): When the loading of supervisor node is too heavy, process job transfer component (PJTC) will select the highest transfer order from job table, as shown in Table 4, and find the best grid node from supervisor grid information data base (SGIDB), as shown in

Table 5, and supervisor grid node capacity information (SGNCI), as shown in Table 9, to do job transfer and send this message to the selected node.

**Table 4.** Job table (JT)

Job -no	Time -begin	CPU -used	Memory -size	File- size	Transfer -to-node	Transfer -from	Transfer -order
------------	----------------	--------------	-----------------	---------------	----------------------	-------------------	--------------------

where Job-no denotes job order number, Time-begin is time of job entered, CPU-used is total CPU used, Memory-size is the required memory used, File-size is total size of used files, Transfer-to-node is node to be transferred, Transfer-from is job accepted to process from this node, Transfer-order is the order of job transfer in this grid node.

**Table 5.** Supervisor or Execute grid information data base (SGIDB, EGIDB)

Node-name	Job-count	CPU-usage	Running-count	Completed-count
-----------	-----------	-----------	---------------	-----------------

where the fields in Table 5 are the same as Table 1, each node has one entry.

- (5) Process grid information component (PGIC): Process grid information component (PGIC) can receive the information of each grid node and update supervisor grid information data base (SGIDB). This data base keeps the current usage of each grid node. The grid information is also written to log file, as shown in Table 6, for analyzing later and sending to all grid nodes to update execute grid information data base (EGIDB).

**Table 6.** Log file 2

Clock -time	Node -name	Job -count	CPU -usage	Running -count	Completed -count
----------------	---------------	---------------	---------------	-------------------	---------------------

- (6) Job transfer message component (JTMC): Job transfer message component (JTMC) can receive job transfer message, as shown in Table 7, from the request grid node, and check supervisor grid information data base (SGIDB) to see whether the selected node can accept the job transfer or not, then return messages, as shown in Table 8, to the request node.

**Table 7.** Transfer Job message

Job-no	Transfer-node	Select-node	Memory-size	File-size
--------	---------------	-------------	-------------	-----------

**Table 8.** Job Transfer return message from supervisor

Job-no	Transfer-node	Select-node	Yes / No
--------	---------------	-------------	----------

- (7) Process new job component (PNJC): If a new job is entered, the process new job component (PNJC) computes memory used and data volume, and updates the job table (JT), as shown in Table 4. If the PNJC can't process

immediately in this grid node, it will assign the number of transfer order to the new job.

- (8) Process message component (PMC): Process message component (PMC) sends messages to the grid nodes.
- (9) Process grid node capacity information component (PGNCIC): When the grid node is inserted, deleted or changed, process grid node capacity information component (PGNCIC) updates supervisor grid node capacity information (SGNCI), as shown in Table 9, in supervisor grid node and sends these updated information to execute grid node and updates execute grid node capacity information (EGNCI), as shown in Table 10.

**Table 9.** Supervisor grid node capacity information (SGNCI)

Node -name	CPU -speed	CPU -cost	Memory -size	Memory -cost	Disk -capacity	Disk -cost	Network -address	Priority
---------------	---------------	--------------	-----------------	-----------------	-------------------	---------------	---------------------	----------

where CPU cost denotes CPU charge rate, Memory cost is memory charge rate, Disk cost is disk charge rate, Priority is value used to decide supervisor or supervisor backup grid node

**Table 10.** Execute grid node capacity information (EGNCI)

Node -name	CPU -speed	CPU -cost	Memory -size	Memory -cost	Disk -capacity	Disk -cost	Address	Priority	Transfer -priority
---------------	---------------	--------------	-----------------	-----------------	-------------------	---------------	---------	----------	-----------------------

where Transfer-priority is value to decide transferring order of this grid node, set by execute grid node itself

## 2.2 Execute Grid Node

Execute grid node is processing job usually and it sends grid node information, as shown in Table 1, to the supervisor grid node periodically. We presented the execute process schedule analyzing module (EPSAM) on the execute grid node. There are eight components in this module, saying, execute process supervisor component (EPSC), execute user interaction component (EUIC), execute send grid node information component (ESGIC), execute process job transfer message component (EPJTC), process grid information message component (PGIMC), process grid job transfer component (PGJTC), execute new job component (ENJC) and execute process message component (EPMC), as shown Fig. 3.

The functions of these components are as the follows:

- (1) Execute process supervisor component (EPSC): Execute process supervisor component (PSC) periodically sends grid node information to the supervisor grid node and checks the loading of the system. EPSC can call ESGIC to send grid node information, or calls EPJTM to find the best selected grid node to transfer job when loading is too heavy
- (2) Execute user interaction component (EUIC): If execute user interaction component (EUIC) receives the grid node information from the supervisor grid node then it calls PGIMC to update the execute grid information data

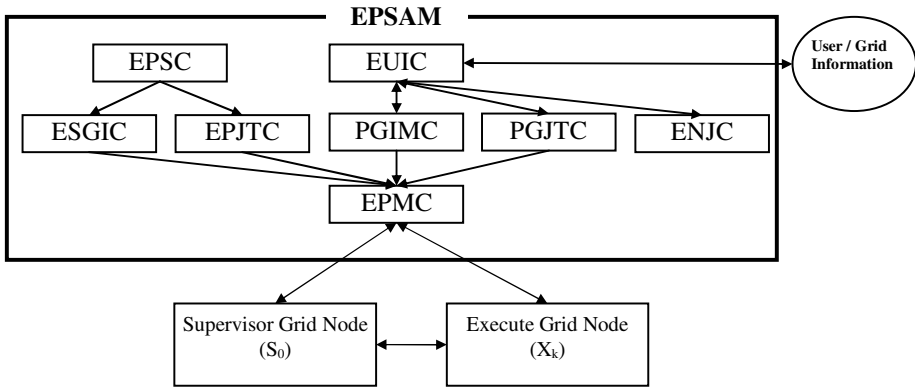


Fig. 3. Framework of EPSAM

base (EGIDB) or execute grid node capacity information (EGNCI). If EUIIC receives grid transfer job messages then it calls PGJTC to handle job transfer or receive job completed from other grid node. If it receives new job then it calls PNJC to update job table (JT).

- (3) Execute send grid information component (ESGIC): Execute send grid information component (ESGIC) sends grid node information, as shown in Table 1, to the supervisor grid node periodically.
- (4) Execute process job transfer component (EPJTC): When the loading of this grid node is too heavy, execute process job transfer component (EPJTC) will select the highest transfer order job from job table, as shown in Table 4, and find the best grid node from execute grid information data base (EGIDB), as shown in Table 5, and execute grid node capacity information (EGNCI), as shown in Table 10, to do job transfer and send this message to supervisor
- (5) Process grid information message component (PGIMC): Process grid information message component (PGIMC) can receive the information of each grid node from supervisor and update execute grid information data base (EGIDB) or execute grid node capacity information (EGNCI). When updating execute grid node capacity information (EGNCI), we set the suitable value (from our view point) to the transfer priority field of each grid node.
- (6) Process grid job transfer component (PGJTC): Process grid job transfer component (PGJTC) will do the following thing:
  - (i) When it receives that the response of job transfer message is positive from supervisor, it transfers job to the selected node. If negative, can not do job transfer.
  - (ii) It receives the return job completed from other grid node.
  - (iii) It receives job transfer from other requested grid node.
  - (iv) All the above, it will update job table (JT).
- (7) Execute new job component (ENJC): If this grid node has a new job to enter, the execute new job component (ENJC) computes memory used and data volume, updates the job table (JT), as shown in Table 4. If the new job

can't process immediately in this grid node, we set the suitable number to transfer order in job table..

- (8) Execute Process message component (EPMC): Execute process message component (PMC) sends grid node information, grid active, job transfer messages to supervisor and job transfer to grid nodes.

### 2.3 Backup Supervisor Grid Node

Backup supervisor grid node works as the execute grid node normally. When it does not receive any message from supervisor for a finite period, it sends supervisor grid node inactive message, as shown in Table 11, to supervisor to make sure supervisor is active or not. If it does not receive in a definite period, it means supervisor is inactive now. It sends supervisor change message, as shown in Table 12, to inform all grid node that backup supervisor is supervisor now until original supervisor grid node is available again. From the priority of grid node capacity information, it can select new backup supervisor grid node.

**Table 11.** Supervisor grid node inactive message from backup supervisor

Grid-node	Inactive ? Yes/No
-----------	-------------------

**Table 12.** Supervisor change message from backup supervisor

Grid-node	Supervisor-grid-node
-----------	----------------------

## 3 Discussion

We made some discussion as follows:

- (1) Each grid node can have the information from grid information data base, and capacity from grid node capacity information in its site.
- (2) The supervisor and backup supervisor grid node work as execute grid node normally. Therefore, via this model, we can have that the system will increase performance.
- (3) Job transfer is handled by the supervisor. The reasons are as follows:
  - (i) to avoid many grid nodes to select the same grid node to accept job transfer at the same time.
  - (ii) to keep the load balance.
- (4) Transfer priority is set by execute grid node based on the relative activity with this node.
- (5) Determination of the job transfer order is based on:
  - (i) file volume: light volume is high priority
  - (ii) the location of execute file stored in the grid
  - (iii) job high priority
  - (iv) CPU computing ability
  - (v) memory size

- (6) Each executing grid node processes the job scheduling itself. When the loading is too heavy, it can select the best grid node to do job transfer.
- (7) In the execute grid node, the grid node information is sent to supervisor periodically. Moreover, if the job status is changed (like job entered, completed or deleted), it will send grid node information to supervisor immediately.
- (8) Log file is written by supervisor when supervisor receives grid node information or supervisor calculates total account information from supervisor grid information data base periodically.
- (9) We can analyze log file according to our requirements, such as
  - (i) Kind: Each, which and all of grid node;
  - (ii) Time: per minute, hour, day, month, year, or range of time;
  - (iii) Content: total of process count, CPU usage, running count and completed. count

## 4 Conclusion

At present, the requirement of high performance computing is increasing at high speed. Grid computing is getting more and more significantly. If jobs and load distributions are not uniform. The load of some grids may be heavy and must transfer job to other grid node to process and increase performance of all system. If every grid can schedule its jobs and decide the selected grid node to transfer job, the result will be best.

In this study, we proposed a process schedule analyzing model based on grid computing architecture. With this model, we not only can optimize to run jobs in the grid, but also can get best performance in all system.

## Acknowledgement

This work was supported in part by the National Science Council, Republic of China, under grant NSC94-2745-M-034-008-URD.

## References

1. Buyya, R.: Economic-based distributed resource management and scheduling for grid computing, Doctor of Philosophy, School of Computer Science and Software Engineering, Monash University, Melbourne (2002).
2. Chun, B., & Culler, D.: Market-based Proportional Resource Sharing for Clusters, University of California at Berkeley, Computer Science Division, Technical Report CSD-1092, (1999).
3. Crowley, Charles : Operating Systems : A Design-Oriented Approach, Richard D. Irwin, a Times Mirror Higher Education Group, Inc., company, 1977.
4. Ferng, J.: *UAGrid* [Online], Available: <http://computing.arizona.edu/uagridinternet2day.pdf> (2004).



5. Ferreira, L., Jacob, B., Slevin, S., Sundararajan, S., Brown, M., Lepesant, J., & Bank, J. : Globus toolkit 3.0 quick start, IBM RedPaper, 23-63 (2003).
6. Foster, I., Kesselman, C.: Globus: A Metacomputing Infrastructure Toolkit, International Journal of Supercomputer Application, Vol. 11, No. 2, (1997), 115-128.
7. Foster, I., Kesselman, C., & Tuecke, S.: The Globus alliance globus toolkit [Online]. Available: <http://www.globus.org/> [1998, May 31].
8. Foster, I., Kesselman, C., & Tuecke, S. : GRAM: Key concept [Online]. Available: <http://www-unix.globus.org/toolkit/docs/3.2/gram/key/index.html> [1998, July 31]
9. Foster, I., Kesselman, C. (ed.): The Grid2: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (2004).
10. 10 Globus: <http://www.globus.org> (2005).
11. Lee, H.-M. and Yang, C.-H.: A Distributed Backup Agent Based on Grid Computing Architecture, Knowledge-Based Intelligent Information & Engineering Systems, LNCS 3682/2005, Springer-Verlag, pp.1252-1257, 2005.
12. Lee, H.-M., Hsu C.-C. and Hsu M.-H. : A Dynamic Supervising Model Based on Grid Environment, Knowledge-Based Intelligent Information & Engineering Systems, LNCS 3682/2005, Springer-Verlag, pp.1258-1264, 2005.
13. Lee, H.-M., Lee, T.-Y., Yang, C.-H. and Hsu, M.-H.: An Optimal Analyzing Resources Model Based on Grid Environment, WSEAS Transactions on Information Science and Applications, Issue 5, Vol. 3, 960-964 (2006)
14. Litzkow, M., Livny, M., and Mutka, M.: Condor-A Hunter of Idle Workstations, in Proceedings of the 8th International Conference of Distributed Computing Systems, 1998.
15. Liu, C., Yang, L., Foster, I., Angulo, D.,: Design and Evaluation of a resource Selection Framework for Grid Applications, Proceedings of IEEE International Symposium on High Performance Distributed Computing, (2002), 63-72.
16. Nutt, Gary : Operating Syswtems, by Pearson Education, Inc., Third Edition (2004).
17. Raman, R., Livny, M. & Solomon, M. : Matchmaking: Distributed Resource Management for High Throughput Computing. In Proceedings of the Seventh IEEE International Symposium on High Performance Distributed Computing, (1998), 140-146, 04, November
18. Raman, R.: ClassAds Programming Tutorial (C++), 2000
19. Research, Development, and Evaluation Commission, Executive Yuan: The Information Outsourcing Promotion of the Government Department, December 15, 2003, Available: [http://web.rdec.gov.tw/cisa/CaseRule\\_Profile.htm?mdy=14](http://web.rdec.gov.tw/cisa/CaseRule_Profile.htm?mdy=14),
20. Silberschatz, Abraham, Galvin, Peter Bear, Gagne, Greg : Operating System Principles, John Wiley & Sons(Asia) Pte Ltd, Seventh Edition (2004).
21. Stallings, William : Operating System :Internals and Design Principles,by Prentice-Hall, Inc. Simmon & Schuster / A Viacom Company, Upper Saddle River, NJ 07458. Third Edition (1998).
22. Tanenbaum, Andrew S. : Modern Operating Systems, 2001 by Prentice-Hall, Inc. Upper Saddle River, New Jersey 07458, Second Edition.
23. Winston, W. L., Operations Research, Duxbury Press, Belmont, California, Third Edition, (1994).
24. Wolski, R., Plank, J., Brevik, J., & Bryan,T.: G-commerce: Market Formulations Controlling Resource Allocation on the Computational Grid. In Proceedings of the 15th International Parallel & Distributed Processing Symposium, (2001) 46.

# Fuzzy Set Theoretical Approach to the RGB Color Triangle

Naotoshi Sugano

Department of Intelligent Information Systems, Tamagawa University  
6-1-1 Tamagawagakuen, Machida, Tokyo 194-8610, Japan  
sugano@eng.tamagawa.ac.jp

**Abstract.** The present study considers a fuzzy color system in which triangular pyramid-like membership functions are constructed on the RGB color triangle. This system can process a fuzzy input to an RGB system and output the center of gravity of three weights associated with respective grades. Triangular pyramid-like membership functions are applied to the RGB color triangle relationship. By treating three membership functions of redness, greenness, and blueness on the RGB color triangle, a target color can be easily obtained as the center of gravity of the output fuzzy set. In the present paper, the differences among fuzzy input, inference output, and chromaticity are described, and the relationship between inference outputs for crisp inputs and inference outputs for fuzzy inputs on the chromaticity diagram are shown for colors that have circular or elliptical vagueness areas.

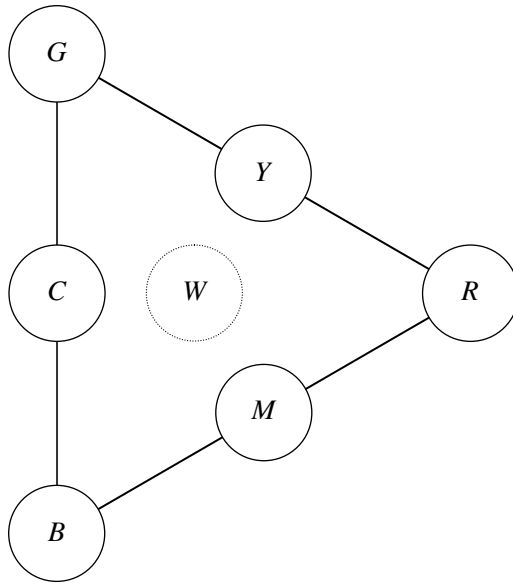
## 1 Introduction

Additive color mixing occurs when two or three beams of differently colored light combine. It has been found that mixing just three additive primary colors, red, green and blue, can produce the majority of colors. In general, a color can be described by certain quantities, called the tristimulus values,  $r$  for the red component,  $g$  for the green component, and  $b$  for the blue component, as follows:

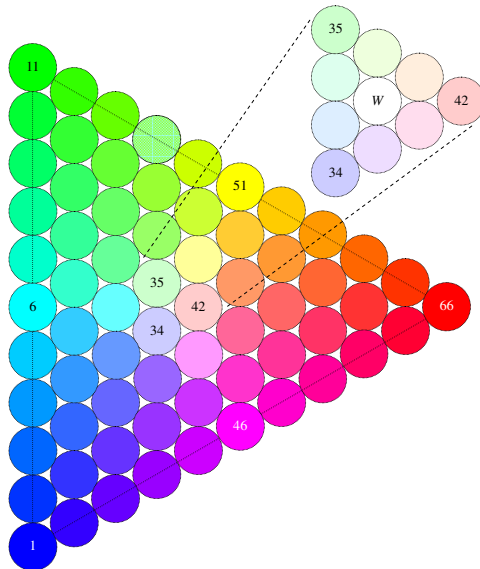
$$\text{color} = r + g + b \quad (1)$$

This is called the RGB color model. This concept allows colors to be represented by a planar diagram. The first step is to draw the red, green and blue components as the vertices of a color triangle, as in Fig. 1. The coordinates in the plane of the color triangle can specify various colors. The location given by the coordinates corresponds to the amounts of  $r$ ,  $g$  and  $b$  that make up the color. The coordinates specifying the center of the color triangle represent the case in which the three primary colors are mixed in equal proportion and indicate the color white. Such representations are called chromaticity diagrams. The diagram represents hue and saturation but not lightness [9]. On the RGB color triangle, the percentages of redness, greenness, and blueness, where the total of the three attributes is equivalent to 100%, specify a color.

In the Natural Color System (NCS), a method similar to the fuzzy set theoretical method for obtaining hue expressions with vagueness has been reported by Sivik [4]. Using the fuzzy set theoretical method, a technique for acquiring tone expressions



**Fig. 1.** A color triangle. A point in the plane of the triangular system represents the hue and saturation of a color.



**Fig. 2.** Sixty-six crisp color inputs and white with six neighboring colors (*detail*) on the RGB color triangle

with vagueness on the NCS color triangle has been investigated by Sugano [5], [7]. In a recent study, the triangular membership functions of achromatic colors and conical membership functions of chromatic colors were used as vagueness [5], [7], which caused a gathering effect toward the center of the NCS tone triangle. In this previous study, fuzzy achromatic colors of triangular membership functions and fuzzy modified achromatic colors of conical membership functions were used on the NCS color triangle in a manner corresponding to the HLS (hue, lightness, and saturation) tone plane consisting of lightness and saturation. The vagueness effects of achromatic colors and modified achromatic colors (e.g., reddish, yellowish, greenish, and bluish achromatic colors) have been clarified [6], [8].

However, a technique for obtaining expressions of the RGB color triangle using the fuzzy set theoretical method has not been reported. In the present study, input fuzzy sets of a triangular pyramid-like shape with a plateau on the RGB triangle and fuzzy inputs of conical membership functions are examined. The RGB color triangle (plane) represents the hue and saturation of a color [9]. The six fundamental colors and white can be represented on the same color triangle (See Fig. 1). Vague colors on the RGB color triangle and chromaticity diagram are clarified. Such a system will help us to determine the average color value as the center of gravity of the attribute information of vague colors. This fuzzy set theoretical approach is useful for vague color information processing, color-naming systems, and similar applications.

## 2 Methods and Results

The present study considers a system of the three primary colors, red, green, and blue (RGB), presented on an RGB color triangle. As Fig. 1 shows, blue, cyan, green, yellow, red, magenta, and white are abbreviated as  $B$ ,  $C$ ,  $G$ ,  $Y$ ,  $R$ ,  $M$ , and  $W$ , respectively. Six fundamental color coordinates, e.g.,  $(r_1, g_1, b_1)$ ,  $(r_6, g_6, b_6)$ ,  $(r_{11}, g_{11}, b_{11})$ , ..., were selected, where  $r_n$ ,  $g_n$ , and  $b_n$  are the red, green, and blue components, respectively, of the  $n^{\text{th}}$  color.

Figure 2 corresponds to the schematic diagram shown in Fig. 1. The color names in Fig. 2 are No.1: blue, No.6: cyan, No.11: green, No.46: magenta, No.51: yellow, and No.66: red. White is surrounded by six neighboring colors, as shown in the detail inset, and these seven colors are surrounded by No.34, No.35, and No.42.

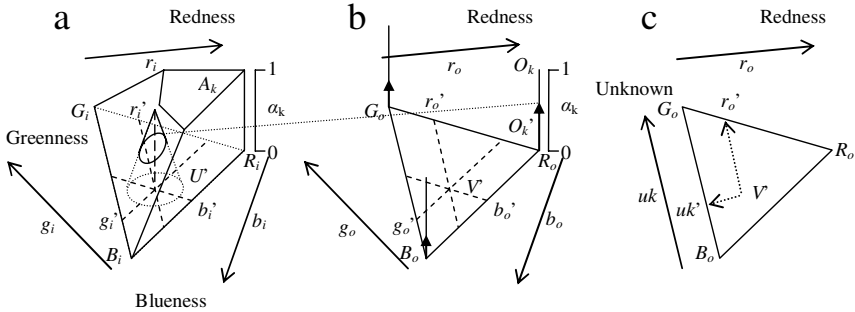
Figure 3 illustrates input fuzzy sets of triangular pyramid-like shape, fuzzy input, and crisp output on the RGB color triangle, and crisp output on the graphical plane. The fuzzy rules are as follows (See Figs. 3 and 5):

$$R^1 : \text{IF } U \text{ is } A_1 \text{ THEN } V \text{ is } O_1 \quad (2)$$

$$R^2 : \text{IF } U \text{ is } A_2 \text{ THEN } V \text{ is } O_2 \quad (3)$$

$$R^3 : \text{IF } U \text{ is } A_3 \text{ THEN } V \text{ is } O_3 \quad (4)$$

where  $k$  is the rule number,  $A_k$  is a fuzzy set of inputs,  $O_k$  is a crisp set of outputs,  $U = (r_i, g_i, b_i)$  are input parameters (variable), and  $V = (r_o, g_o, b_o)$  are output parameters.



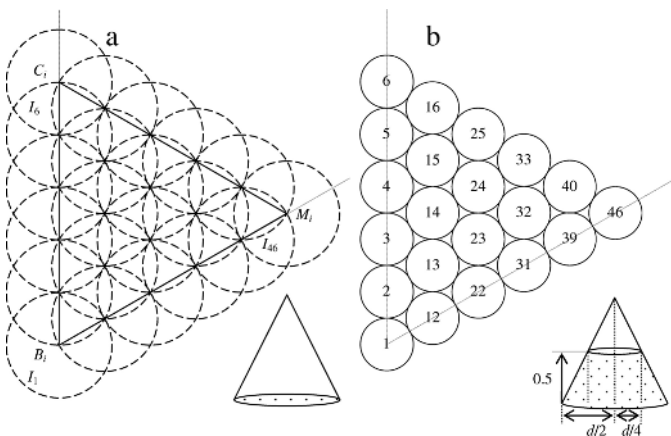
**Fig. 3.** Fuzzy system using the membership function of a triangular pyramid-like shape on the RGB color triangle

Here,  $U$  and  $V$  are fixed to these RGB parameters. A fuzzy set  $A_k$  of inputs shows a triangular pyramid-like shape at corner points  $R_i, G_i$ , and  $B_i$ , and a crisp set  $O_k$  of outputs of rule  $R^k$  is shown at corner points  $R_o, G_o$ , or  $B_o$  (a fuzzy set  $O_k'$  indicated by vertical arrows in Fig. 3b) on the color triangular, and the output is  $O_k$  if the input is  $A_k$ .

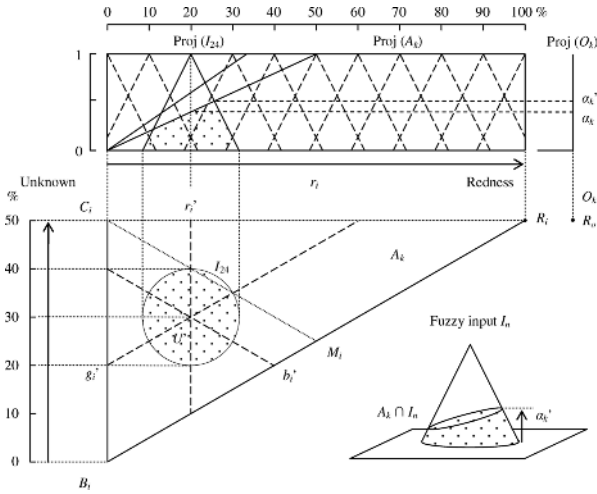
The fuzzy inference method is as follows. Let the inputs be  $r_i = r_i', g_i = g_i'$ , and  $b_i = b_i'$ . 1) The input of rule  $R^k$ , grade  $\alpha_k = A_k(U')$ , where  $k = 1, 2, 3$ . 2) The output of rule  $R^k$ , and the  $\alpha_k$  level-set is shown as a vertical filled allow. 3)  $O_k' = \alpha_k O_k$ , where  $O_k$  is fuzzy set in Fig. 3b. 4) The complete inference result  $O'$  of rules  $R^1, R^2$ , and  $R^3$ .

$$O' = \alpha_1 O_1 \cup \alpha_2 O_2 \cup \alpha_3 O_3 = O_1' \cup O_2' \cup O_3' \tag{5}$$

The output parameter,  $V = (r_o', g_o', b_o')$ , corresponds to the coordinates of the central axis of the membership function of  $O'$ , which is a de-fuzzification. In addition, in Fig. 3c,  $V = (r_o', uk')$  corresponds to a coordinate of the graphical system, where  $uk'$  (on the vertical axis) is calculated from  $g_o'$  and  $b_o'$ .



**Fig. 4.** Fuzzy inputs on part of the RGB color triangle and top areas of 0.5 level-sets indicated by number



**Fig. 5.** Membership functions of triangular pyramid-like shape on the half of the RGB color triangle and one of sixty-six conical fuzzy inputs (*vague colors*)

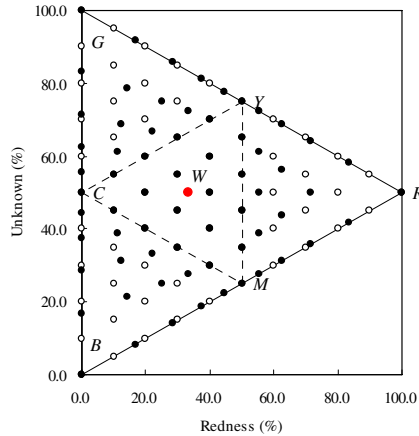
Figure 4a (left) illustrates fuzzy inputs ( $I_1 - I_{46}$ ) on the RGB color triangle as a triangle with color names ( $B$ ,  $C$ , and  $M$ ). The fuzzy inputs are formed by conical membership functions, and the membership functions are made to mutually overlap. The edge of the basal plane (circle) of the conical membership function passes through the centers of the overlapped circles. Figure 4b (right) shows the arrangement of numbers corresponding to the conical membership functions of Fig. 4a, and the numbers are shown inside circles representing the top of the 0.5 level set (bottom-right). The color names are No.1: blue, No.6: cyan, and No.46: magenta.

Figure 5 illustrates half of the RGB color triangle as a base of input fuzzy set  $A_k$  and one of the sixty-six conical fuzzy inputs ( $I_1 - I_{66}$ ) on the RGB color triangle. The triangular membership function  $\text{Proj}(I_{24})$  on the redness axis is one of eleven projections of the sixty-six fuzzy inputs ( $I_1 - I_{66}$ ) by the rays from the lower part, and the triangular membership function  $\text{Proj}(I_n)$  on the unknown axis is not used in the present study.

The intersection of input fuzzy set  $A_k$  for fuzzy input  $I_n$  is  $A_k \cap I_n$ . (See the dotted area at the bottom-right of Fig. 5.) Grade  $\alpha_k' = \text{height}(A_k \cap I_n)$ . If the input is crisp,  $\alpha_k'$  becomes  $\alpha_k$ .  $R_o$  is the new red.  $\text{Proj}(O_k)$  is a projection of an output crisp set at the corner point  $R_o$ . (See Fig. 3b).

What happens if a vague color is input into the RGB system? The system considered in the present study can translate input data  $U$  of a vague color to output data  $V$  of a simple color on the RGB color triangle. The fuzzy input (No.24) on the RGB is made up of the center  $U' = (r_i', g_i', b_i') = (20, 20, 60)$  in % and the diameter  $d = 23.0\%$  of the basal plane (circle) of the cone indicated vagueness.

Figure 6 illustrates the relationship between the unknown value  $uk$  and the redness value  $r_o$  obtained from data  $(r_o', uk')$ . Filled circles indicate outputs for crisp inputs of colors corresponding to Fig. 3c, and open circles indicate crisp inputs of colors. The inference outputs (filled circles) for crisp inputs are not the same as the coordinates



**Fig. 6.** Inference outputs (*filled circles*) for crisp inputs (*open circles*) on the graphical plane. White (*filled circle*) exists in the coordinates (33.3%, 50.0%).

for the inputs (open circles) on the RGB color triangle. However the positions of twenty-one colors (circles) are no change on the inside of YMC triangle and three positions of RGB colors are also no change. The inference outputs (filled circles) for crisp inputs are grouped at the center of the RGB color triangle.

Figure 7 also illustrates the relationship between the unknown value  $uk$  and the redness value  $r_o$ . Filled circles indicate outputs for fuzzy inputs of colors, corresponding to Fig. 3c. The inference outputs (filled circles) for fuzzy inputs are gathered at the center of the RGB color triangle. The circles are clearly different in this case. Vague color inputs to the RGB color triangle (Fig. 3a), the system outputs crisp color on the RGB color triangle (Fig. 3b), and also outputs crisp color on the graphical plane (Fig. 3c).

The chromaticity coordinates are denoted by  $r_o', g_o', b_o'$  and  $x, y, z$ . The transformation from  $R, G,$  and  $B$  to  $X_i, Y_i,$  and  $Z_i$  can be shown as follows [2]:

$$X_i = 2.77 r_o' + 1.75 g_o' + 1.13 b_o' \tag{6}$$

$$Y_i = 1.00 r_o' + 4.59 g_o' + 0.06 b_o' \tag{7}$$

$$Z_i = 0.00 r_o' + 0.06 g_o' + 5.59 b_o' \tag{8}$$

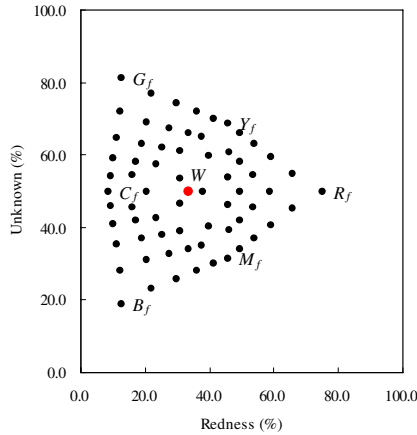
The general definitions of the chromaticities  $x, y, z$  [1], [3] are:

$$x = X_i / (X_i + Y_i + Z_i) \tag{9}$$

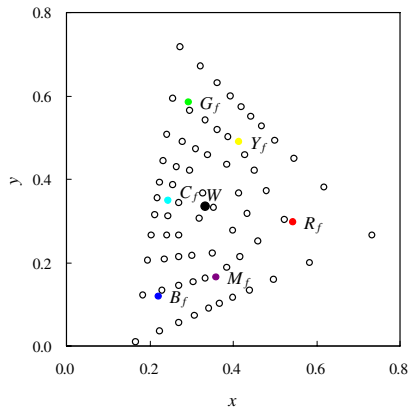
$$y = Y_i / (X_i + Y_i + Z_i) \tag{10}$$

$$z = Z_i / (X_i + Y_i + Z_i) \tag{11}$$

where  $x + y + z = 1$ .



**Fig. 7.** Inference outputs (*filled circles*) for fuzzy inputs on the graphical plane. White (*filled circle*) exists in the coordinates (33.3%, 50.0%).



**Fig. 8.** Transformed inference outputs (*open circles*) for sixty-six crisp inputs and transformed inference outputs (*filled circles*) for six conical fuzzy inputs on the chromaticity diagram. White (*filled circle*) is located at the coordinate (0.33, 0.33).

Figure 8 illustrates the differences between crisp and fuzzy inputs. Only the six fundamental colors (filled circles) show changes in direction from coordinates  $(x, y)$  for crisp input to those for fuzzy input. The direction indicates toward white  $W$  (central filled circles), for example,  $R_f$  lies midway between  $R$  and  $W$ . The output  $(x, y)$  for fuzzy input is dislocated from the center of the conical fuzzy input (vague colors). Crisp input and its inference output for the six fundamental colors do not have the same coordinates. Each output for a fuzzy input is distant from the center of vague colors, although do not examine for sixty colors, except for six fundamental colors.



This implies that vague colors move toward the direction of white, as such vague colors are input to the fuzzy system.

### 3 Conclusions

The present paper proposes a fuzzy system that can extract crisp outputs of the RGB triangle (which is available for use in fuzzy set theory), a graphical system (which is easy to show via graphs), and chromaticity. It is difficult to construct such a fuzzy system on the chromaticity diagram directly, because the membership function of a triangular pyramid-like shape or cone is quite complicated on the nonlinear chromaticity diagram. The system also extracts, in a simple manner, the membership grades from the projection of a conical membership function of a vague color input. Three parameters associated with respective grades indicate vague colors and output the center of gravity as a crisp color value although the RGB triangle does not have a vertical attribute (on the unknown axis).

In the future, this system will help to ensure important color information (e.g. vagueness and color shading) in manufactured goods and art by reducing the confusion between colors that is often experienced by people.

### References

1. Brettel, H., Hardeberg, J. Y., Schmitt, F.: Webcam for interactive multispectral measurements. In: MacDonald, L.W., and Luo, M.R. (eds.): *Colour image science*. John Wiley & Sons, New York (2002)
2. Fukinuki, T.: *Image media technology*. Coronasha Publishing, Tokyo (2002) in Japanese
3. Lee, H.C. *Introduction to color imaging science*. Cambridge University Press, New York (2005)
4. Sivik, L.: Color systems for cognitive research. In: Hardin, C. L. and Maffi, L. (eds.): *Color categories in thought and language*. Cambridge University Press, New York (1997) 163-193
5. Sugano, N.: Natural color system using fuzzy set theoretical approach. *Proc. of the International Workshop on Fuzzy Systems & Innovational Computing, Kitakyushu* (2004) 69-74
6. Sugano, N.: Fuzzy natural color system for achromatic colors. *Proc. of Joint 2nd International Conference on Soft Computing and Intelligent Systems and 5th International Symposium on Advanced Intelligent Systems, WE-1, Yokohama* (2004) 6-11
7. Sugano, N.: Fuzzy natural color system using membership function of triangular pyramid on color triangle. *Biomedical Soft Computing and Human Sciences*, Vol. 10, No. 1 (2004) 1-10
8. Sugano, N.: Fuzzy set theoretical approach to achromatic relevant color on the natural color system. *International Journal of Innovative Computing, Information and Control*, Vol. 2, No. 1 (2006) 193-203
9. Tilley, R. J. D.: *Colour and optical properties of materials, An exploration of the relationship between light, the optical properties of materials and colour*. John Wiley & Sons, New York (1999)

# Spatial Equilibrium Model on Regional Analysis for the Trade Liberalization of Fluid Milk in Taiwan

Lily Lin

Department of International Trade, China University of Technology  
No. 56, Sec. 3, Shinglung Rd., Taipei, Taiwan 116, R.O.C.  
lily@cute.edu.tw

**Abstract.** The purpose of this study is to analyze and solve the equilibrium demand, supply and the prices of fluid milk in each demand area and supply county of Taiwan under fixed import quantities and given income levels. This was accomplished by incorporating a spatial price equilibrium model that was derived for the maximization problem. Then, we modeled problems associated with three simulated fluid milk import policies – including import quotas, tariff, and complete liberalization – to analyze the impact and regional influence of the policies on consumers, producers and factories, respectively, from 2001 through 2006.

## 1 Introduction

Following approval by WTO Ministers meeting Taiwan became a member of the WTO in January 2002. Taiwan is opening up its market for products where imports were previously banned such as fluid milk through a system of tariff quotas. The simulations of this study were made by the assumption of Taiwan becoming a member of the WTO in 2001, although there is one-year lag, the issues, methodology, major findings would be valuable and useful for the future research. In Taiwan, over 90% of raw milk is used to produce fluid milk; therefore, there will be an expected influence on the Taiwanese dairy industry caused by the tendency of trade liberalization to result in the import of less costly fluid milk as opposed to the current policy of restricted importation.

The spatial price equilibrium model applied in this study is static and involves partial equilibrium. It assumes perfect competition and a homogeneous product. It also considers that there are no structural changes in demand and supply in the transition from a starting position to the new equilibrium. That is, prices and quantities are determined along demand and supply functions that remain unchanged in the basic model. It should be noted that we consider four cases of the basic model which correspond to four cases of income level.

Applying the theoretical econometric model, both 14-area supply functions and 18-area demand functions of fluid milk can be used as Lin and Kawaguchi [8], but revised in the year of 2000 level. Consequently, with linear demand and supply functions as well as the unit transportation cost of fluid milk, the maximization

problem to achieve equilibrium becomes a quadratic programming problem. As a result, it is feasible to determine the equilibrium solutions of fluid milk supply, demand, trade flows and prices.

## 2 Theoretical Model and Application

### 2.1 Theoretical Model

The model will be described and started with a simple multi-region, one commodity (fluid milk) world. We assume that there are  $m$  supply areas and each of them supplies one given commodity, and there are  $n$  regions (counties) that demand this commodity. The supply areas from 1 to 14 represent the domestic suppliers and the area of 15 refers to all overseas suppliers. Since the world market price of fluid milk is supposed to be unique for Taiwan, to simplify the model and consider the policy factor, it is assumed that only one overseas supplier exists and import quantity is parametrically fixed for analysis. Thus,  $PS_{15}$  stands for the import price of fluid milk in Taiwan. We omit the superscript  $f$  in this section. The commodity is supposed to be traded freely from any supply area to any consuming region; moreover, traded freely by dealers among consuming regions.

The model presented in this paper uses a price formulation, in which the decision variables are prices (fluid milk demand and supply prices). The Lagrangean [10] multipliers are interpreted as shadow quantities. Alternatively, it can be presented in an equivalent quantity formulation (dual), in which the decision variables are quantities, and the Lagrangean multipliers are interpreted as shadow prices. The price formulation is easier in the presence of a fixed supply. On the other hand, the quantity formulation is easier in the presence of a fixed price. The model is specified as follows.

Maximize

$$F = \sum_{n=1}^{18} \int \{ \alpha_n - \beta_n PD_n \} d PD_n - \sum_{m=1}^{14} \int \{ \gamma_m + \delta_m PS_m \} d PS_m - \gamma_{15} PS_{15} \tag{1}$$

subject to  $PD_n - PS_m \leq t_{mn}$  for all  $m$  and  $n$ , and all variables are non-negative.

$\alpha_n$  : intercept value of the linear demand function in consuming area  $n$

$\beta_n$  : slope coefficient of the linear demand function in consuming area  $n$

$\gamma_m$  : intercept value of the linear supply function in supply area  $m$

$\delta_m$  : slope coefficient of the linear supply function in supply area  $m$

$PD_n$  : real fluid milk market price in  $n$ th area

$PS_m$  : real supply price of fluid milk in  $m$ th area

$SQ_{mn}$ : the shipped fluid milk quantity from  $m$ th producing area to  $n$ th consuming (demand) area

$t_{mn}$ : unit transportation cost of fluid milk from supply area  $m$  to demand area  $n$ .

The Kuhn-Tucker [5] conditions associated with this problem are both necessary and sufficient conditions for an optimal solution, under the assumptions of

differentiability and concavity of the objective function, and in the presence of linear constraints. The Kuhn-Tucker conditions are specified as follows.

$$\begin{aligned}
 PD_n - PS_m + sq_{mn} &= t_{mn}. && \text{for all m and n} && (2) \\
 \alpha_N - \beta_N PD_N - \sum SQ_{MN} &= -v_D. && \text{for all n} && (3) \\
 -\gamma_M - \delta_M PS_M + \sum SQ_{MN} &= -v_{SM}. && \text{for } m=1,2,\dots,14 && (4) \\
 -\gamma_{15} + \sum SQ_{15n} &= -v_{S15}. && \gamma_{15} \text{ is fixed import quantity} && \\
 PD_n v_{Dn} = 0, \quad PS_m v_{Sm} &= 0. && \text{for all m and n} && (5) \\
 SQ_{mn} sq_{mn} &= 0. && \text{for all m and n} && (6)
 \end{aligned}$$

where  $sq_{mn}$  and  $v$  are explained as slack variables, and all variables are non-negative.

Statement (2) indicates that the price difference between demand and supply areas is less than or equal to the unit transportation cost. Whenever trade takes place, the price difference is exactly equal to the unit transportation cost (6). In each region, production has to be greater than or equal to the domestic use plus exports to other regions (4), and consumption has to be less than or equal to domestic production plus imports from other regions (3). Statement (5) implies that if the demand price in area  $n$  is positive, the quantity shipped from all supply areas to market  $n$  equals the demand quantity in market  $n$ . In addition, if the supply price in area  $m$  is positive, the shipped quantity from area  $m$  to all demand areas should be equal to the supply quantity in area  $m$ .

All these statements taken together characterize an equilibrium solution for the traditional spatial equilibrium problem in a perfectly competitive market. The Kuhn-Tucker conditions in the case of the Taiwanese fluid milk market are formulated as Lin and Kawaguchi [8], Table 7. The optimal solution is obtained with the Quadratic Programming Method or other equivalent methods.

## 2.2 Simulation

According to the above quantitative method, we can solve the perfectly competitive spatial equilibrium problem by utilizing estimated demand and supply functions as well as the unit transportation cost of fluid milk. The unit transportation cost is calculated by using the data of the tariff for freight traffic combined with the data of the unit transportation cost of raw milk (Lin and Kawaguchi [6], Table 4). We apply the model to analyze the equilibrium condition under a given import quantity of fluid milk. Since the future income level is uncertain, the simulated models used to analyze the policy effect will be created with four cases that have different levels of national disposable income: 100%, 120%, 140%, and 160% of the 2000 income level. The result is shown in both Table 1-1, 1-2. Although the prices of 18 consuming and 14 producing counties – together with their consuming and producing quantities and trade flows among those counties – are also obtained in the result, but we do not show them in this paper due to the limits of available space.

As long as trade takes place between producing area  $m$  and consuming area  $n$ , the difference between the market price and the supply price is equal to the unit transportation cost. Otherwise, no trade occurs when the price difference is smaller

**Table 1-1.** Simulation result-composition of import price and quantity

IQ	00INC	120%INC	140%INC	160%INC
10	44.718	46.591	48.236	49.718
20	44.490	46.344	47.999	49.490
30	44.261	46.127	47.791	49.291
40	44.033	45.909	47.584	49.092
50	43.805	45.692	47.376	48.893
60	43.577	45.475	47.169	48.695
70	43.348	45.258	46.961	48.496
80	43.109	45.040	46.754	48.297
90	42.872	44.812	46.546	48.098
100	42.644	44.594	46.327	47.898
110	42.379	44.375	46.117	47.685
120	42.116	44.121	45.906	47.483
130	41.844	43.892	45.662	47.282
140	41.571	43.639	45.452	47.062
150	41.341	43.348	45.211	46.845
160	41.112	43.130	44.923	46.616
170	40.846	42.905	44.715	46.417
180	40.608	42.688	44.508	46.140
190	40.379	42.470	44.300	45.941
200	39.902	42.216	44.093	45.743
210	39.494	41.998	43.877	45.544
220	39.252	41.616	43.633	45.345
230	38.991	41.223	43.425	45.146
240	38.734	40.912	43.218	44.911
250	38.506	40.694	42.750	44.704
260	38.277	40.402	42.372	44.501
270	38.049	40.185	42.134	44.142
280	37.821	39.968	41.896	43.827
290	37.588	39.750	41.675	43.421
300	37.360	39.533	41.467	43.223
310	37.132	39.316	41.260	43.024
320	36.901	39.099	41.052	42.825
330	36.673	38.882	40.845	42.612
340	36.445	38.660	40.637	42.414
350	36.217	38.443	40.430	42.215
390	35.301	37.572	39.596	41.420
400	35.067	37.355	39.386	41.221
450	33.909	36.265	38.349	40.222
490	32.996	35.379	37.519	39.427
500	32.767	35.157	37.312	39.228

IQ: Import quantity, unit: 1,000 ton  
The other 4 columns indicate the import price (unit: N.T. \$/kg.) under four different income levels.

**Table 1-2.** Simulation result-composition of import quantity and domestic supply

IQ	00INC	120%INC	140%INC	160%INC
10	499.898	554.348	602.947	646.741
20	493.155	547.934	596.822	640.874
30	486.411	541.516	590.692	635.003
40	479.667	535.098	584.563	629.132
50	472.924	528.679	578.433	623.260
60	466.180	522.261	572.303	617.389
70	459.436	515.843	566.173	611.518
80	452.706	509.424	560.043	605.646
90	445.920	503.021	553.913	599.775
100	439.177	496.602	547.799	593.922
110	432.547	490.208	541.702	588.105
120	425.904	483.911	535.608	582.270
130	419.233	477.488	529.606	576.436
140	412.565	471.179	523.512	570.652
150	405.839	464.923	517.495	564.863
160	399.096	458.508	511.470	559.076
170	392.443	452.053	505.340	553.204
180	385.687	445.634	499.210	547.430
190	378.942	439.216	493.080	541.558
200	372.778	432.900	486.950	535.687
210	366.319	426.482	480.784	529.816
220	359.488	420.410	474.767	523.944
230	352.946	414.464	468.637	518.073
240	346.380	408.197	462.508	512.323
250	339.636	401.776	457.140	506.418
260	332.893	395.673	451.325	500.598
270	326.149	389.255	445.407	495.239
280	319.405	382.837	439.486	489.709
290	312.656	376.418	433.267	484.542
300	305.912	370.000	427.137	478.670
310	299.169	363.582	421.007	472.799
320	292.447	357.163	414.877	466.928
330	285.704	350.745	408.747	460.970
340	278.960	344.323	402.617	455.098
350	272.216	337.904	396.488	449.227
390	245.279	312.256	371.966	425.741
400	238.609	305.837	365.862	419.870
450	205.003	273.786	335.213	390.540
490	178.029	248.307	310.693	367.055
500	171.285	241.900	304.563	361.184

IQ: Import quantity, unit: 1,000 ton  
The other 4 columns indicate the domestic supply (unit: 1000 ton) under four different income levels.

than the unit transportation cost. This kind of price relation is the feature of the perfectly competitive spatial equilibrium.

### 3 Results

Finally, we are going to explore the effects of different import tariffs on the supply and demand in each county in Taiwan. The results of research through simulation include volume of transportation all 15 supply areas (14 domestic areas plus one import area) to 18 consumption areas, 15 equilibrium volume of supply and prices in 15 supply areas and equilibrium consumption and prices in 18 consumption areas under various income levels and different import quantity. The key points of simulation are shown as Table 2 and Table 3. There are a lot of significant differences from area to area in the results.

From Table 2, we could see that the variations of demand of fluid milk in Central Taiwan are smaller than that in other areas under the policy of import restrictions. Even under policy of import quotas, different income levels, the growth of demand of fluid milk in Central Taiwan has never been over than 5 % while the growth of demand in the other areas are more than 40%. It is evidenced that the regional difference existed. Besides, if we adopt a policy of fully free trade, under the income level or 120 % income level in 2000, the growth of demand of fluid milk from customers in Central Taiwan will be greater than any other areas. If we adopt measures of import quotas, the growth of demand from East Taiwan will be the greatest among all areas while if we adopt measures of import tariffs plus tariff equivalence, the growth of demand from South Taiwan will be the greatest. Furthermore, if we simulate a policy of complete liberalization, due to the reduction of prices, the demand of fluid milk from customers in all areas shows significant increase.

Table 3 shows that under a policy of import restrictions, the growth of demand of fluid milk from the customers to the north of Miao-li County (area no. 4) is greater than any other counties in Taiwan while the demand from the customers between Central Taiwan and Chia-yi (area no. 9) is relatively small. But if we adopt a policy of fully free trade, due to the significant increase of imports that may affect the supply and demand in Taiwan, the supply in every county is greatly reduced comparing with it in base period. However, comparing with other counties, the impact of trade liberalization on the counties in Central Taiwan will be smaller than the counties in other areas in which the impact on the counties in North Taiwan and Tai-dong County (area no. 13) will be the greatest.

In the research, the author simulates solutions of spatial equilibrium of supply and demand in four different policies of import tariffs under four different income levels and introduced from the results equilibrium quantity and price as well as equilibrium import quantity and price of fluid milk in individual areas in Taiwan. We find that there are significant regional differences in both supply and demand. We also analyze impacts of different import policies on customers, producers and milk industry.

**Table 2.** Simulation result: fluid milk demand for each region under import policy unit: ton, %

Area	Policy I				Policy II		Policy III			
	(1) basic year 2000	(2) 120% imp=20	(3) 140% imp=20	(4) 160% imp=20	(5) 140% imp=50	(6) 160% imp=120	(7) 2000 imp=220	(8) 120% imp=300	(9) 140% imp=390	(10) 160% imp=480
1	9,113	10,426	11,538	12,566	11,704	13,170	10,242	12,009	13,783	15,538
2	65,174	74,498	82,448	89,788	83,637	94,113	73,207	85,824	98,506	111,052
3	78,725	89,990	99,593	108,458	101,030	113,651	88,408	103,647	118,963	134,112
4	36,749	42,012	46,503	50,652	47,171	53,066	41,058	48,135	55,248	62,285
5	16,845	19,261	21,323	23,227	21,627	24,329	18,728	21,957	25,200	28,408
6	9,355	10,695	11,840	12,899	12,010	13,510	10,400	12,192	13,993	15,775
North	215,960	246,883 14.32%	273,245 26.53%	297,589 37.80%	277,179 28.35%	311,839 44.40%	242,043 12.08%	283,763 31.40%	325,693 50.81%	367,170 70.02%
7	66,043	68,677	69,168	69,017	71,275	76,281	79,870	87,426	94,967	102,108
8	25,328	26,350	26,561	26,520	27,362	29,282	30,479	33,366	36,210	38,930
9	10,500	10,926	11,010	10,997	11,342	12,142	12,637	13,836	15,012	16,143
10	18,244	18,981	19,133	19,104	19,710	21,094	21,955	24,035	26,084	28,044
11	20,305	21,125	21,292	21,260	21,934	23,475	24,436	26,751	29,059	31,244
Mid.	140,420	146,059 4.02%	147,164 4.80%	146,898 4.61%	151,622 7.98%	162,273 15.56%	169,377 20.62%	185,414 32.04%	201,332 43.38%	216,469 54.16%
12	44,172	51,463	57,718	63,518	58,743	67,293	49,817	59,652	69,956	80,351
13	60,912	70,969	79,588	87,591	81,001	92,797	68,697	82,262	96,467	110,806
14	22,401	26,101	29,267	32,215	29,787	34,129	25,264	30,254	35,475	40,751
15	2,608	3,036	3,400	3,738	3,462	3,967	2,980	3,580	4,198	4,823
South	130,093	151,569 16.51%	169,974 30.66%	187,063 43.79%	172,994 32.98%	198,185 52.34%	146,759 12.81%	175,748 35.09%	206,096 58.42%	236,732 81.97%
16	4,621	5,367	6,059	6,721	6,104	6,885	4,881	5,736	6,599	7,451
17	7,350	8,539	9,640	10,694	9,711	10,886	7,696	9,059	10,421	11,755
18	8,193	9,517	10,741	11,909	10,822	12,202	8,732	10,280	11,826	13,349
East	20,165	23,423 16.16%	26,440 31.12%	29,324 45.42%	26,637 32.10%	29,973 48.64%	21,310 5.68%	25,074 24.34%	28,845 43.04%	32,555 61.44%

/: comparing with the basic year

The Policy I: Import Quotas. The policy II: Import Tariff plus Tariff Equivalents from 2001.

The Policy III: Complete Liberalization in 2001.

**Table 3.** Simulation result: fluid milk supply for each region under import policy unit: ton, %

Area	Basic year 2000	Policy I			Policy II		Policy III			
		(1) 120% imp=20	(2) 140% imp=20	(3) 160% imp=20	(4) 140% imp=50	(5) 160% imp=120	(6) 2000 imp=220	(7) 120% imp=300	(8) 140% imp=390	(9) 160% imp=480
1	9,792	10,877	12,161	13,318	11,678	11,770	5,415	5,633	5,681	5,704
		<b>11.08%</b>	<b>24.19%</b>	<b>36.00%</b>	<b>19.26%</b>	<b>20.20%</b>	<b>-44.70%</b>	<b>-42.48%</b>	<b>-41.98%</b>	<b>-41.75%</b>
2	30,182	33,529	37,490	41,060	36,000	36,285	16,678	17,351	17,500	17,572
		<b>11.09%</b>	<b>24.21%</b>	<b>36.04%</b>	<b>19.28%</b>	<b>20.22%</b>	<b>-44.74%</b>	<b>-42.51%</b>	<b>-42.02%</b>	<b>-41.78%</b>
3	11,964	13,318	14,922	16,367	14,319	14,434	6,771	7,044	7,104	7,133
		<b>11.32%</b>	<b>24.72%</b>	<b>36.80%</b>	<b>19.68%</b>	<b>20.65%</b>	<b>-43.40%</b>	<b>-41.13%</b>	<b>-40.62%</b>	<b>-40.38%</b>
4	24,044	26,798	30,058	32,995	28,831	29,066	13,783	14,336	14,459	14,518
		<b>11.45%</b>	<b>25.01%</b>	<b>37.23%</b>	<b>19.91%</b>	<b>20.89%</b>	<b>-42.68%</b>	<b>-40.37%</b>	<b>-39.86%</b>	<b>-39.62%</b>
5	10,666	11,162	11,750	12,279	11,529	11,571	8,920	9,021	9,043	9,054
		<b>4.65%</b>	<b>10.16%</b>	<b>15.12%</b>	<b>8.09%</b>	<b>8.48%</b>	<b>-16.37%</b>	<b>-15.42%</b>	<b>-15.21%</b>	<b>-15.12%</b>
6	84,829	88,773	93,443	97,650	91,686	92,023	70,949	71,752	71,928	72,012
		<b>4.65%</b>	<b>10.15%</b>	<b>15.11%</b>	<b>8.08%</b>	<b>8.48%</b>	<b>-16.36%</b>	<b>-15.42%</b>	<b>-15.21%</b>	<b>-15.11%</b>
7	6,852	7,170	7,548	7,888	7,406	7,433	5,707	5,771	5,792	5,799
		<b>4.65%</b>	<b>10.15%</b>	<b>15.11%</b>	<b>8.08%</b>	<b>8.48%</b>	<b>-16.71%</b>	<b>-15.77%</b>	<b>-15.47%</b>	<b>-15.37%</b>
8	56,400	59,041	62,166	64,983	60,990	61,216	47,296	47,826	47,999	48,055
		<b>4.68%</b>	<b>10.22%</b>	<b>15.22%</b>	<b>8.14%</b>	<b>8.54%</b>	<b>-16.14%</b>	<b>-15.20%</b>	<b>-14.90%</b>	<b>-14.80%</b>
9	31,532	34,364	37,716	40,737	36,455	36,697	21,767	22,336	22,522	22,582
		<b>8.98%</b>	<b>19.61%</b>	<b>29.19%</b>	<b>15.61%</b>	<b>16.38%</b>	<b>-30.97%</b>	<b>-29.16%</b>	<b>-28.57%</b>	<b>-28.38%</b>
10	105,116	114,680	126,001	136,203	121,742	122,558	71,797	74,718	75,146	75,349
		<b>9.10%</b>	<b>19.87%</b>	<b>29.57%</b>	<b>15.82%</b>	<b>16.59%</b>	<b>-31.70%</b>	<b>-28.92%</b>	<b>-28.51%</b>	<b>-28.32%</b>
11	33,119	36,131	39,697	42,910	38,356	38,613	22,625	23,545	23,679	23,744
		<b>9.10%</b>	<b>19.86%</b>	<b>29.56%</b>	<b>15.81%</b>	<b>16.59%</b>	<b>-31.69%</b>	<b>-28.91%</b>	<b>-28.50%</b>	<b>-28.31%</b>
12	87,232	95,729	105,788	114,852	102,004	102,728	57,628	60,224	60,603	60,784
		<b>9.74%</b>	<b>21.27%</b>	<b>31.66%</b>	<b>16.93%</b>	<b>17.76%</b>	-	-	-	-
		<b>6.51%</b>	<b>6.51%</b>	<b>6.51%</b>	<b>6.51%</b>	<b>6.51%</b>	<b>33.94%</b>	<b>30.96%</b>	<b>30.53%</b>	<b>30.32%</b>
13	5,935	6,51	7,199	7,816	6,941	6,991	3,742	3,858	3,884	3,929
		<b>9.75%</b>	<b>21.30%</b>	<b>31.70%</b>	<b>16.95%</b>	<b>17.79%</b>	-	-	-	-
		<b>9.84%</b>	<b>10.88%</b>	<b>10.88%</b>	<b>10.49%</b>	<b>10.88%</b>	<b>36.95%</b>	<b>34.99%</b>	<b>34.56%</b>	<b>33.80%</b>
14	8,974	9	5	11,818	5	6	6,409	6,585	6,624	6,692
		<b>9.75%</b>	<b>21.29%</b>	<b>31.69%</b>	<b>16.95%</b>	<b>21.31%</b>	<b>-28.59%</b>	<b>-26.63%</b>	<b>-26.19%</b>	<b>-25.43%</b>

/: comparing with the basic year



## References

1. Hashimoto, H.: A Spatial Nash Equilibrium Model. In "Spatial Price Equilibrium: Advances in Theory, Computation and Application (Lecture Notes in Economics and Mathematical Systems Vol. 249)". (1985 ) ed. by P.T. Harker, Springer-Verlag, Berlin Heidelberg, pp.20-40
2. Kawaguchi, Tsunemasa: On the Importance of Maintaining Control Power to Check the Linkage of Manufactured and Fluid Milk Price. *Dairyman* Vol. 46-9 (1996) 21-25 (in Japanese)
3. Kawaguchi, Tsunemasa and Nobuhiro Suzuki: An Application of Single-Product "Dual-Structure" Spatial Imperfect Competition Equilibrium Model to the Japanese Milk Market. *Journal of Rural Economics* Vol.66-1 (1994) 22-34 (in Japanese) .
4. Kawaguchi, T., Suzuki, N., and H.M. Kaiser: A Spatial Equilibrium Model for Imperfectly Competitive Milk Markets. *American Journal of Agricultural Economics* Vol.79 (1997) 851-859
5. Kuhn, H. W. and A. W. Tucker: Nonlinear Programming. In "Proceedings of The Second Berkely Symposium on Mathematical Statistics and Probabilities". ed. by J. Neyman, University of California Press, California, pp. 481-492 (1951)
6. Lin, Lily and T. Kawaguchi: Study on Reducing Raw Milk Transportation Cost in Taiwanese Dairy Industry. *Journal of the Faculty of Agriculture, Kyushu Univ.* Vol.43-1-2 (1998) 269-280
7. Lin, Lily and T. Kawaguchi: Increase in Raw Milk Demand and Its Long-Run Effects on Dairy Industrial Organization in Taiwan—An Approach with Spatial Equilibrium Analysis. *Journal of the Faculty of Agriculture, Kyushu Univ.* Vol. 43-3-4 (1999) 509-529
8. Lin, Lily and T. Kawaguchi: Impact of the Trade Liberalization of Fluid Milk on the Taiwanese Dairy Industry. *Journal of the Faculty of Agriculture, Kyushu Univ.* Vol. 44-1-2 (1999) 219-234
9. Nagurney, A.: A Computational Comparison of Spatial Price Equilibrium Methods. *Journal of Regional Science* Vol.27 (1987) 55-76
10. Rockafellar, R.T.: Lagrange Multipliers and Variational Inequalities. In " Variational Inequalities and Complementarity Problems-Theory and Applications". ed. by R.W. Cottle et. al., John Wiley, pp.303-322 (1980)
11. Samuelson, P.A.: Spatial Price Equilibrium and Linear Programming. *The American Economic Review* Vol.42 (1952) 283-303
12. Takayama, T. and G.G. Judge: Spatial Equilibrium and Quadratic Programming. *Journal of Farm Economics* Vol.46 (1964) 67-93
13. Takayama, T. and G.G. Judge: "Spatial and Temporal Price and Allocation Models." North-Holland Publishing Company (1971)

# Analysing the Density of Subgroups in Valued Relationships Based on DNA Computing

Ikno Kim<sup>1</sup>, Don Jyh-Fu Jeng<sup>2</sup>, and Junzo Watada<sup>3</sup>

Graduate School of Information, Production and Systems  
Waseda University

2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan  
<sup>1</sup>octoberkim@akane.waseda.jp, <sup>2</sup>don\_jeng@yahoo.com,  
<sup>3</sup>junzow@osb.att.ne.jp

**Abstract.** One method for enhancing the quality of work life for companies or other organisations is to rearrange employees by detecting and analysing employees' close interpersonal relationships based on business implications. Although human resource managers can use various methods to enhance the quality of work life, one of the most widely used and effective methods is job rotation. In this paper, we select a model of a workplace where employees in a variety of job functions are sharing tasks, information, etc. through close interpersonal relationships, and we suppose a personnel network which contains strong terms of mutual understanding. However, with a huge number of employees it becomes extremely difficult to find the maximum clique of employees for rearrangement, meaning this is NP-hard. Therefore, we employ DNA computing, also known as molecular computation, to this rearranging problem. The goal of this paper is to propose a way to apply DNA computing to this human resource management problem, and to measure its effectiveness in rearranging employees to analyse the density of subgroups in a personnel network with valued relationships.

## 1 Introduction

Job rotation [1], [2], [3] simply moves employees from task to task, and is very important for each employee's career growth, learning and development [4]. Rotation can break the monotony of highly specialised work by calling on different skills, aptitudes and capabilities.

Some subgroups are sure to exist in a personnel network, and to execute job rotation, the most important point is to note all the subgroups [5] in the personnel network for analysing the density of subgroups. The most important reason is that the personnel network might become a huge number of employees composed of all the connected subgroups, and those subgroups are also composed of subgroups, both small and large. Another important reason is that human resource managers hope to find the maximal number of employees who mutually exchange and share their information by close interpersonal relationships at work. In addition, we associate all the subgroups in the personnel network with valued relationships to rearrange employees for the density analysis. However, even if employees' close interpersonal relationships in the huge personnel network are known, it would be quite difficult to find all the cliques and all the components for the analysing the density of subgroups efficiently and

effectively, because the maximum clique from all the subgroups in the huge personnel network must be extremely hard to discern exactly. Furthermore, finding the maximum clique [6] of employees becomes NP-hard, and cannot be solved using present electronic computers in polynomial time unless  $P=NP$ , which is widely believed to be false. To address this, we employ DNA computing to solve these rearranging problems of employees and analyse the density of subgroups for job rotation.

DNA computing has drawn attention in various fields since it was proposed to be able to solve a Hamiltonian path with molecular computation by L. Adleman in 1994 [7]. Until now, the attention of DNA computing is almost entirely from either the computer science fields or the biotechnology fields. However, we propose that DNA computing should be a very useful tool for a variety of management problems. Thus, we have shown that efficient solutions are obtained by DNA computing, and the efficiency of DNA computing is examined by rearranging employees and analysing the density in each subgroup for the redesign of subgroups in valued relationships.

## 2 Analysis of Subgroups

As Fig. 1 shows, a personnel network is mainly composed of various types of subgroups. It is necessary to analyse all the subgroups to accurately understand the personnel network. Therefore, in this section, we describe and analyse subgroups in more detail to provide a better understanding of this rearranging problem.

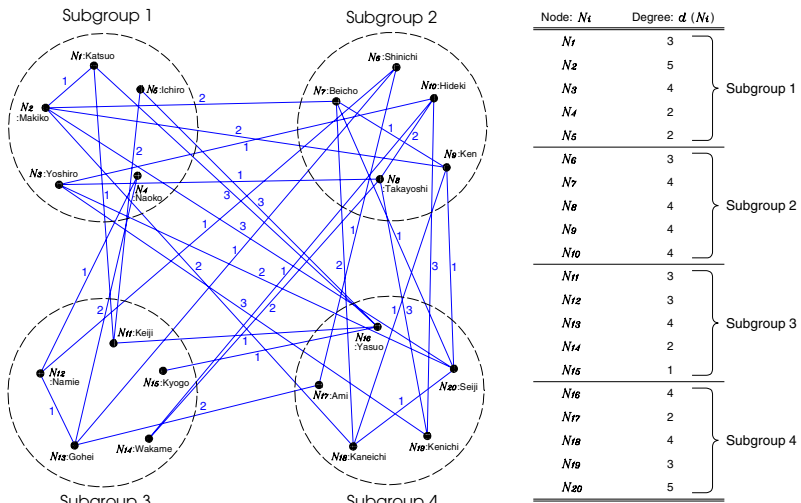


Fig. 1. Personnel network and nodal degrees for employees in valued relationships

### 2.1 Cohesive Subgroup

Cohesion is one of measures of the extent to which individual tasks are needed to execute the same task on closely related to concepts of strong ties among employees

of embedded subgroups or closed social groups. Cohesive subgroups [5] are subsets of employees among whom there are relatively strong, dense, direct, frequent, intense, or positive ties. These are relationships that enable employees to share or exchange their information, create solidarity, or act collectively. Numerous direct contacts among all cohesive subgroup employees, combined with few or null ties to outsiders, dispose a group toward a close interpersonal relationship in business, homogeneity of thought, behaviour, and identity.

## 2.2 Clique

A cohesive subgroup should have a clique or cliques. The clique is a useful starting point for specifying the formal properties. The clique also has well-specified mathematical properties, and captures much of the intuitive notion of cohesive subgroups. In a graph, the clique is a maximal complete subgraph of three or more nodes, and it consists of a subset of nodes, all of which are adjacent to each other, and there are no other nodes that are also adjacent to the employees of the clique.

## 3 Personnel Network with Valued Relationships

Basically, human resource managers rearrange employees and redesign subgroups based on relational quantifications. There are two properties of employees' relationships which are either dichotomous or valued for understanding their measurement. In this section, we select employees in the same firm to create their own personnel network and describe this personnel network with valued relationships.

### 3.1 Model Graph

The network with valued relationships for this model is given in Fig. 1. In the graph, there are  $N = 20$  nodes, and  $E = 33$  edges with values between the pairs of nodes. Even with as few as twenty employees and thirty-three ties, the graph looks very complicated. There are also the four circles that represent present subgroups which do not really look like cohesive subgroups in valued relationships, and the connected lines that represent the relationships among the employees who mutually exchange and share their business information through a close interpersonal relationship in business. Although this personnel network is very intricately connected with four subgroups, it is important for the human resource manager to rearrange employees and redesign better subgroups by finding a subgroup of all the cliques including the maximum clique, and all the components, in order to satisfy job rotation requirements with a new personnel network.

### 3.2 Valued Relationships

It is necessary to measure the relationships between employees to do the best rearrangement. Therefore, relationships are very often valued. Valued relationships usually indicate the strength or intensity of ties between pairs of employees. In a real business situation, personnel network data can be collected by having each employee indicate their degree of co-operation, the content of the business relationships, and

other related measures. In this paper, quantification of scaling is done using a three-level relational rating scale based on the employees' indications. A relational value of 3 indicates a strong tie, 2 indicates a normal tie, and 1 indicates a weak tie in a business situation.

## 4 DNA Computing Approach

DNA computing is recently drawing attention from a number of scientists, engineers and other researchers [8], [9], [10], [11], [12]. The attention is almost always focused on solving NP-completeness problems, combinational problems and difficult mathematical problems. However, we propose a way to solve one of the management problems using DNA computing. Therefore, in this section, we describe how to relate employees with valued relationships to DNA sequences, and how to approach redesigning subgroups based on a DNA experiment.

### 4.1 Algorithm to Find Cliques and Components

We note one of the approaches to cohesive subgroup analysis that is a socio-matrix [5], because this is the most important analyzing procedure to find all the cliques and all the components of the subgroups using DNA computing.

A systematic way for ordering rows and columns of the socio-matrix reveals the subgroup structure of the personnel network. The socio-matrix of the model graph with the rows and columns is composed of employees who have ties and close interpersonal business relationships to each other in valued relationships. The socio-matrix of size  $m \times m$  becomes 20 rows and 20 columns for the model graph. There is a row and column for each node, and the rows and columns are labeled 1, 2, 3, ..., 20.  $x_{ij}$  denotes the value of the tie from employee  $i$  to employee  $j$ , and  $x_{ij}$  records which pairs of nodes are adjacent. If nodes  $N_i$  and  $N_j$  are adjacent, then  $x_{ij} = \{1, 2, 3\}$ , and if nodes  $N_i$  and  $N_j$  are not adjacent, then  $x_{ij} = \{0\}$ .

We designed a new algorithm based on the socio-matrix and the algorithm of the maximal clique problem solution that was proposed by Ouyang et al. [13], the method to find the maximum clique. However, the new algorithm that finds all the 1-cliques, as well as the maximum clique and all the components, because we would like to design more cohesive and efficient subgroups for arranging employees in the personnel network with valued relationships based on the DNA results.

### 4.2 Experiment

In this experiment, the DNA-sequence is designed in the form of double-stranded DNA, and it corresponds to 20 nodes while satisfying the given algorithm above. Each node of the DNA-sequence in a binary number is composed of two sequences that are a position sequence  $E_i$  and a value sequence  $N_i$ . The position sequences are used for connecting each node of the DNA-sequence, and the value sequence is used for distinguishing whether those position sequences contain that node or not. For a twenty-digit binary number in the graph, twenty value sections ( $N_1$  to  $N_{20}$ ) are prepared sandwiched sequentially between twenty-one position sections ( $E_1$  to  $E_{21}$ ). We set  $E_i$  with the length of 10 base pairs (bp),  $N_i$  with the length of 0 bp if the value of 1,

and 6 bp if the value of 0. We repeat selecting the shortest DNA strands, which correspond to all the possible cliques including the maximum clique, and all the components using a gel electrophoresis apparatus.

### 4.3 Experiment Results

The maximum clique is  $\{N_2, N_7, N_9, N_{18}, N_{20}\}$ , which is connected at distance 300 bp by five nodes. The second largest clique is  $\{N_3, N_8, N_{10}, N_{19}\}$ , which is connected at distance 306 bp by four nodes. There are also six possible cliques that are all connected by three nodes, and the three independent lines with a 2-node were found as well. In addition, there are some nodes that are empty spaces which are vertically arranged, so we can understand the possible cliques which are mutually connected together and find out all the nodes are divided into two components that also can be one component and one subgroup.

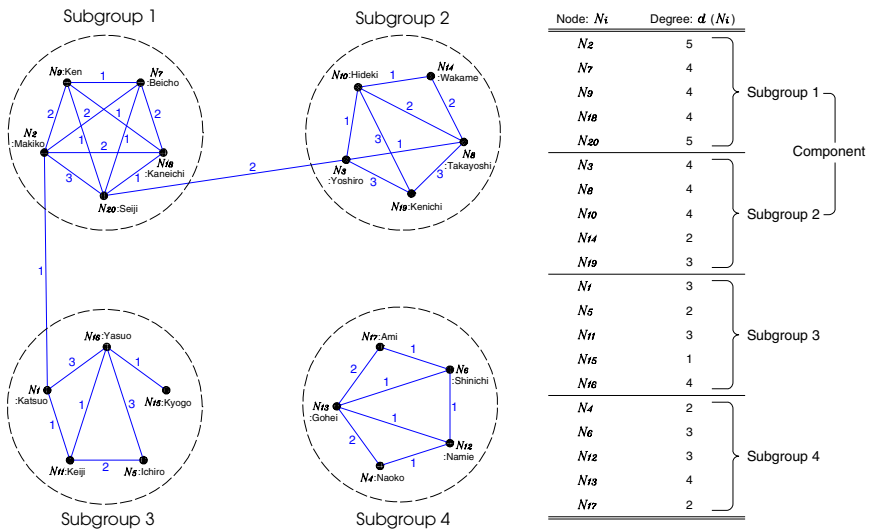


Fig. 2. Redesigned personnel network and nodal degrees for employees in valued relationships

## 5 Density Analysis

Although we found all the cliques including the maximum clique, and also all the components, we hope to analyse the density of subgroups for arranging employees with valued relationships in the large order of cohesive subgroups. Moreover, even though we do not know exactly which cliques connect with which others or how connected subgroups are in the network with valued relationships, we hope to analyse the density of subgroups [14] for rearrangement of employees efficiently based on the DNA results.

### 5.1 Density of Subgroups

First, inclusiveness defines the number of nodes that are included within the various connected parts of the graph. The inclusiveness of a graph is the total number of nodes minus the number of isolated nodes. The most useful measure of inclusiveness for comparing various types of graphs is the number of connected nodes expressed as a proportion of the total number of nodes. Thus, the proportion of the inclusiveness is denoted by  $\Pi$ , and the inclusiveness is calculated as follows:

$$\Pi = \frac{C}{\sum_{i=1}^n N_i} \text{ for } 0 \leq \Pi \leq 1, \tag{1}$$

where  $C$  is the number of connected nodes present.

Second, the density is defined as the number of edges without considering valued relationships in a graph. The density also expressed as a proportion of the maximum possible number of edges. The density without considering valued relationships is denoted by  $\Delta$ , and the density is calculated as follows:

$$\Delta = \frac{L}{n(n-1)/2} \text{ for } 0 \leq \Delta \leq 1, \tag{2}$$

where  $n$  is the number of nodes present, there are  $n(n-1)/2$  possible unordered pairs of nodes, and thus  $n(n-1)/2$  possible edges that could be presented in the graph, and  $L$  is the number of edges present.

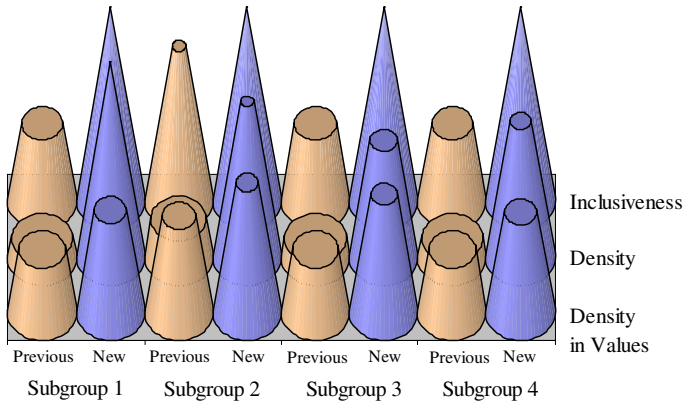
Finally, the density in a valued graph is defined as the number of edges, considering valued relationships in a graph. To generalise the notion of density to a valued graph, one can average the values attached to the edges across all edges. The density in a valued graph is also expressed as a proportion of the maximum possible number of edges. The density considering valued relationships is denoted by  $\delta$ , and the density in a valued graph is calculated as follows:

$$\delta = \sum_{i=1}^n w_i / (\max_i w_i) L \text{ for } 0 \leq \delta \leq 1, \tag{3}$$

where  $w_i$  is the value of that edge. This measures the average strength of edges in the valued graph.

### 5.2 Analysis Results

First, it is known that the new subgroup 1 and subgroup 2 are connected by the relationships between the two employees,  $N_3$  (Yoshiro) and  $N_{20}$  (Seiji). It is also known that the new subgroup 1 and subgroup 3 are connected by the relationships between the two employees,  $N_1$  (Katsuo) and  $N_2$  (Makiko). These four employees are appropriate to be information communicators for each subgroup to exchange their information between subgroups 1 and 2, and subgroups 1 and 3. Subgroup 4 shows that, if the number of degrees is considered,  $N_{13}$  (Gohei) could be a leader for subgroup 4. Furthermore, it can be said that, if the four employees who are  $N_{20}$  (Seiji),  $N_3$  (Yoshiro),  $N_1$  (Katsuo), and  $N_{13}$  (Gohei) are connected, the network could be an integrated personnel network.



**Fig. 3.** Inclusiveness, density, and density in values comparisons in each subgroup

Second, densities in values correspond to each cohesive clique size of employees. Thus, an efficient rearrangement and redesign becomes possible while considering densities in values of each group to the size of the clique of employees.

Third, as Fig. 3 shows, the results of inclusiveness, the density, the density in values, and others are used to compare previous subgroups with new subgroups in 4 subgroups which were redesigned based on DNA computing.

## 6 Concluding Remarks

In this paper, all the employees were rearranged, and the subgroups were redesigned more efficiently and effectively as shown in the density of new subgroups, and other results. A massively parallel computation corresponds to DNA computing that was able to be done to make a new personnel network for an efficient job rotation. In addition, we show various ideas on human resource management based on the results of DNA computing. Furthermore, we can also measure the efficiency of DNA computing in rearranging employees in valued relationships. Thus, we believe that it is better to employ DNA computing rather than other methods when human resource managers redesign of subgroups for rearrangement of employees in any kinds of personnel networks with valued relationships to support job rotation.

## References

1. Miyakawa, T.: Third Edition, Management Information System. Chuo-keizaisha Co., Ltd. (2004) 87-98 in Japanese
2. Umezu, H.: MBA Management of Human Resources and Organisation. Seisansei-Shuppan (2003) 41-60 in Japanese
3. Werther, W. B., Werther, Jr., Keith, D.: Human Resources and Personnel Management. McGraw-Hill, Inc. (1995) 145-157



4. Schwalbe, K.: Fourth Edition, Information Technology Project Management. Thomson Course Technology (2006) 366-384
5. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press (1999) 149-290
6. van Noort, D., Gast, F.-U., McCaskill, J. S.: DNA Computing in Microreactors. Submitted to Special Lectures in DNA Computing, Graduate School of Information, Production and Systems, Waseda University (2005) 1-10
7. Adleman, L.: Molecular Computation of Solutions to Combinatorial Problems, Vol. 266. Science (1994) 1021-1024
8. Rose, J. A., Hagiya, M., Deaton, R. J., Suyama, A.: A DNA-based in vitro Genetic Program, Vol. 28. Journal of Biological Physics (2002) 493-498
9. van Noort, D., Landweber, L. F.: Towards a Re-programmable DNA Computer, DNA9, LNCS 2943. Springer-Verlag Berlin Heidelberg (2004) 190-196
10. Rose, J. A., Deaton, R. J., Suyama, A.: Statistical thermodynamic analysis and design of DNA-based computers, Vol. 3. Natural Computing (2004) 443-459
11. Sakamoto, K., Kiga, D., Komiya, K., Gouzu, H., Yokoyama, S., Ikeda, S., Sugiyama, H., Hagiya, M.: State Transitions by molecules, Vol. 52. Biosystems (1999) 81-91
12. Watada, J., Jeng, D. J.-F., Kim, I.: Application of DNA Computing to Group Control of Elevators. 2005 Anniversary Symposium on the Romanian Society for Fuzzy Systems & A.I., Iasi, Romania (2005) May 5-7
13. Ouyang, Q., Kaplan, P. D., Liu, S., Libacher, A.: DNA Solution of the Maximal Clique Problem, Vol. 278. Science (1997) 446-449
14. Yasuda, Y.: Practice Network Analysis. Shinyosha Co., Ltd. (2004) 55-63 in Japanese

# Structural Learning of Neural Networks for Forecasting Stock Prices

Junzo Watada

Waseda University,  
Graduate School of Information, Production and Systems,  
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, Fukuoka 808-0135 Japan  
watada@waseda.jp

**Abstract.** Generally, a neural network spends much computation time and cost in forecasting the value and movement of a stock. The reason is because a neural network requires exponential time in computation according to the number of units in a hidden layer.

The objective of the paper is to optimally build a neural network through structurally learning. The results enable us to reduce the computational time and cost as well as to understand the structure more easily.

In the paper the method is employed in forecasting the price movement of a stock. The optimization of the network by the structured learning is evaluated based on its real use.

**Keywords:** Forecasting the stock, Structural Learning.

## 1 Introduction

Forecasting methods of stocks can be classified into two groups; fundamental analysis and technical analysis [4].

Fundamentals mean economical circumstance for stocks and include economic macro parameters and such financial indices of individual companies. Analysis depending on such fundamentals is named fundamentals analysis. The forecasting of each stock price can be achieved by surveying the states of individual companies that are invested and provides an economical environment. The technical analysis stands on that all stocks should be a function of a price and time. Therefore, the future price of a stock is analyzed using moving average and stock chart.

In this paper, we build a structural learning of a neural network and employ this model to forecast a stock price. This analysis can be classified into the technical analysis.

Generally, a neural network is built using sufficient number of units. This redundancy makes the neural network enable to evaluate prices using a huge number of data. On the other hand, the redundancy spends much more computation time and building cost of a system. As well, the complex structure bothers a structure from its transparency.

In this paper, it enables us to decrease the computation time that a neural network is previously learned to obtain the optimal structure,

## 1.1 Review

Various researches are pursued on forecasting a stock price by a neural network. Using a hierarchical neural network, learning is executed inputs of forecasted stock prices and dealing amount and output of changing rates of a teaching stock price as well as stock prices. The hierarchical neural network has learning ability to adjust itself to teaching changing rates as well as stock prices. In forecasting of a stock through a neural network, moving rate par the present value is emphasized. As well, the up and down movements of a stock price are forecasted. The forecast precision is defined by the hitting ratio of the real stock movement. The hitting rate is defined as follows:

$$\text{Hitting rate} = \frac{\text{frequency of true hit of up and down movements of a stock price}}{\text{The trial number of stock prices}}$$

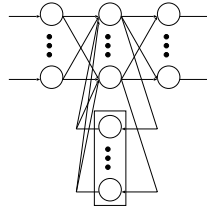
Jason E.Kutsurel [1] reported 90% precision in stock forecasting, the neural network is one of good method.

Nevertheless, there are few researches that a structural learning of a neural network is employed for forecasting. Therefore, it is not clear how much precision a structural obtains.

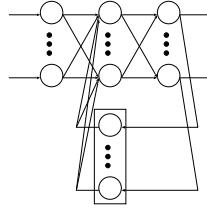
## 2 Neural Network of Stock Forecasting

Stocks are time-series data. General neural network cannot deal with time-series data in the sense of time-series input. It is necessary to input data in a neural network multi-dimensionally or to construct a new structure that can deal with time-series data well. Representative neural networks to treat data in time-series is a concurrent neural network. The recurrent neural network is a hierarchical neural network that has feedback of some output from hidden units to input units in order to have a time-series input, although a general hierarchical neural network has the flow of information from input to output. There are an Elman type of feedback from hidden units to input and a Jordan type of feedback from output [2] [3].

Figure shows that Elman type of a neural network feeds back the values of units in a hidden layer and the Jordan type of neural network the values of units in an output layer. The information from these units is feeded back to the flamed portion. The flamed portion is named a context layer. On the case of the Elman type of a neural network, a unit in a hidden layer is one to one corresponding to a unit in a context layer. Both the numbers in a hidden layer and context layer are the same. On the case of the Jordan type of a neural network, the number of units in the context layer is the same number of units in the output layer. Therefore, the number of units in the context layer is constrained by the number of units in the output layer defined by the considered problem. On the other hand, the number of units feedback in the Jordan type of a neural network has



**Fig. 1.** Elman type of neural network



**Fig. 2.** Jordan type of neural network

no constraint depending on the number of units in the output layer, the number of units in the constraint layer can be freely determined. In this research, the Elman type of a neural network is employed so as to optimize the number of units in the hidden layer by the structural learning.

### 3 Structural Learning

On the case of making a hierarchical neural network learn teaching data through back-propagation method, the numbers of input units and output units are uniquely decided depending on their required numbers of teaching input and output, respectively. On the other hand, the number of hidden units are depending on a learning method as well as the numbers of the input units and output units. There should be the minimal number of the hidden units to obtain the same results. Generally, such a number is not known previously. Conventionally, the number of units in the hidden layer is decided depending on experiences. On this case, if we can decrease the number of units, the computation speed and system cost can be saved. Also, we experience that a general type of back-propagation method has much dependence on the initial setting and is hard to forecast an expected value without convergence even if we select the approximately minimal number of units.

In order to overcome such a problem, there are various trials that a structure of a network is changed recursively and gradually to result in reaching to its optimal structure [5]-[9]. This process is named structural learning. Such trials of structural learning can be roughly classified into two groups. One group is a generating learning method to start from a small structure of a neural network

and expanding recursively and gradually the structure of the neural network up to the optimal one. The other group is an eliminating learning method to start from a sufficiently large structure of a neural network to shrink recursively the structure of the neural network until the optimal one.

### 3.1 Generating Method

The generating method has the following features. On the starting stage, as the structure of a neural network is small, it spends smaller total computation time than the eliminating method. Generally, since the learned portion is frozen without being included in the learning process, it is rather rare to stop at the local minimum. Nevertheless, as the partial optimizations are repeated, an unnatural structure is sometimes created.

### 3.2 Eliminating Method

The eliminating method requires a little bit large computation time because of starting a large network. As the learning can be pursued through the overall check of the network, it has the possibility to find the optimal solution that the generating method cannot reach.

### 3.3 Elimination of Ineffective Units by Goodness Factor

In the paper, we employed the eliminating learning method for structural learning. In this research, a goodness factor proposed by Matsunaga [8] is employed in building the optimal structure of a neural network through the eliminating method to evaluate the effect of each unit in the hidden layer.

The goodness factor defines as the total sum of propagating signals in forward direction as follows:

$$G_i^k = \sum_p \sum_j (w_{i,j}^k x_i^k)^2 \quad (1)$$

where  $x_i^k$  is an output of unit  $i$  of layer  $k$ ,  $w_{i,j}^k$  is the link weight from unit  $i$  of layer  $k$  to unit  $j$  of layer  $k + 1$ ,  $p$  is summation over whole learning patterns. As it is easily understood from the definition, the large value of goodness factor produces the large influence on the whole units of level  $k + 1$  from the unit. Therefore, after calculating goodness factor of all units of layer  $k$ , the unit with the smallest goodness factor is interpreted as the most useless unit in the layer and named a bad unit.

First, the appropriate number of units in a hidden layer is provided. When a back propagation learning gives the result lower than the error rate given, the learning is taken to terminated and a bad unit is eliminated. When the error rate is greater than an arbitrarily defined value, then initialize all link weights of a bad unit and re-learn the teaching patterns. Even if this procedure is executed the result is not improved, then one unit is newly attached on to the hidden layer. The above mentioned procedure is pursued repetitively until the smallest number of units is obtained.

## 4 Structural Learning Method for Forecasting

Generally, the structural learning is to build a neural network with the minimal number of units under the consideration of dependency of initial values. In this research, the structural learning is employed to decide the optimal number of units in order that the stock prices are forecasted with high precision.

The effect of the structural learning in forecasting stock prices by a neural network is verified as follows. First, it is pursued to structurally learn a neural network. Then, depending this procedure, the optimal number of hidden units is determined in this experiment. After then, we compare and verify the effect and efficiency between the neural network with the same number of hidden units and the neural network resulted by the structural learning.

### 4.1 Determination of Input Data

Before the verification let us discuss about input data. In this research we propose to employ plural number of stocks including the focal stock in order that the proposed neural network model can forecast stock prices precisely. Using the plural stocks, the neural network can recognize the tendency of stock movements. Even if focal stock data for forecasting are noisy, the forecasting does not have such an influence.

In order to select plural stocks relating the focal stock, a large number of stocks from various industries are selected and the time-series data of their stock prices are classified based on a SOM ( Self-Organizing Map ). It is necessary to classify the transit of the stock price instead of classifying the stock prices themselves. In order to do so, we normalize the time-series prices of 50 stocks as a pre-processing.

Input data consists of five stocks.

- Hitachi Ltd [code6501]
- Sekisui House Ltd [code1928]
- Saizeriya Co. Ltd.[code7581]
- Squire Enix [code9620]
- TOPIX

It is necessary in time-series forecasting to input time sequence data that are expanded to space data. In this paper, the time-series movement is realized by giving one bundle consisting of continuous five days of each stock and stock index as input to a neural network and also the bundle of continuous five days regarding one day lag data. Therefore, one input consists of 35 patterns.

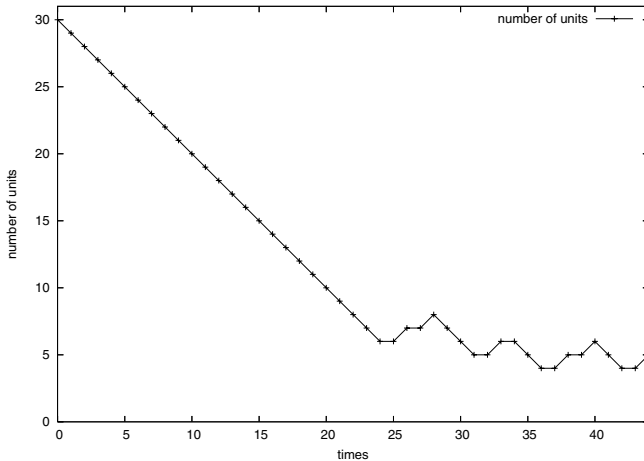
### 4.2 Verification of Structural Learning

In structural leaning of a neural network for forecasting stock price, each parameter of the neural network and learning condition are defined as follows:

- The initial number of units in hidden layer: 30
- The initial linkage weight and threshold of units: randomly given within a value from -1 to 1.

- Learning rate: 0.1
- maximum number of learning repeat F30000
- Error rate at stopping of learning: less than 0.05

The teaching data for learning of a neural network consist of 192 patterns where one pattern has one bundle of 5 days stock prices for each of 5 stocks and stock index out of 197 days from January 4, 2002 to October 22, 2002. Figure 3 shows the transit number of hidden units as the result of structural learning under the above-mentioned conditions.



**Fig. 3.** Transit number of hidden units

In the figure, on the case of 5 hidden units the learning did not succeed and the number of units is increased. On the case of 6 hidden units the learning does not stably succeeded. Therefore, the minimal number of hidden units for learning a neural network should be 6. Nevertheless, there are a few times when the learning did not succeed. Therefore, we employed 8 hidden units for the minimal number of hidden units..

In eliminating a bad hidden unit, as all link weights are maintained from the successful learning, it is not necessary to learn duplicated after elimination. Therefore, as the learning is continuously pursued after learning some length, even if the number of hidden units is 8 and all other parameters and the learning conditions are set to the same, it is not guaranteed to successfully learn the neural network within the given leaning time. So it is necessary to set the maximum learning time larger.

### 4.3 Application to Forecasting Stock Price

In the previous section, we obtained the minimal number of hidden units is 8 and the number of input units is 30. In this section we verify the efficiency and effect

of forecasting stock price by the structural learning. The same learning condition is employed but the maximum learning time is set to 50,000. The verification simulation is executed 20 times about the largest, smallest and mean values of forecasting precision and learning time.

The verification is done using the same normalized learning data of 197 days of each of stocks and stock index from January 4, 2002 to October 22, 2002. The forecast of the price of Hitachi Ltd stock for 40 dealt days from October 23, 2002 to December 28, 2002. When the forecasting of price movement is coincident with the real movement, the forecast is counted as success.

Table 1 shows the result of verification.

**Table 1.** Verification Result of Forecasting Stock Movement

Number of Hidden Units	Structural Learning (8)	Conventional (30)
Forecast Precision (Mean) [%]	65.1	62.7
Forecast Precision (Maximum) [%]	73.5	73.5
Forecast Precision (Minimum) [%]	58.8	55.9
Learning Time (Mean)[sec]	810.6	2289.1
Learning Time (Maximum)[sec]	1025.3	2714.5
Learning Time (Minimum)[sec]	573.5	1849.1

In structural learning of a neural network, the neural network with 30 hidden units takes 1327 seconds as mean learning time. On the other hand, the structurally learned neural network with 8 hidden units can pursue the same results in 413 seconds. So the computational time is shortened into 1/3. Regarding the mean forecasting precision, the neural network with 30 hidden units is 63 %, on the other hand the neural network with 8 hidden units resulted in 67%. Even if the number of hidden units decreases, the forecasting precision was not decreasing but increasing.

Conventionally, it is required to employ a neural network with larger number of hidden units in forecasting data such as stocks using huge number of data. Even if we employed the optimal number of hidden units after structural learning, it is possible to improve the forecasting in the sense of precision and computation time. In conventional forecasting, although the redundant forecasting is employed considering noise it provided too much fitting. Then the resulted precision of forecasting was not expected good.

The computation cost decreased by eliminating redundant hidden units as well. That is, the elimination realize the decreasing of the memory assignment for the computation and also for the link waits. As a result the computational cost was possible.

## 5 Concluding Remarks

The objective of this paper is to remove redundant units from the neural network in forecasting stock prices. The structural learning is employed to decide the op-



timal number of hidden units of the neural network and its appropriateness is verified. As a result, it was shown in structural learning that the neural network can successfully determined and forecast stock prices in the simulation experiment using real data. The simulation showed to increase forecasting precision and shorten the computational time with better forecasting ability.

## References

1. Jason E.Kutsurelis, "Forecasting financial markets using neural networks: An analysis of methods and accuracy", Naval Post Graduate School(1998)
2. J.L.Eلمان, "Finding Structure in Time", Tech.Report CRL-8801, University of California, San Diego(1988)
3. M.I.Jordan, "Serial order: A parallel, distributed processing approach", Tech.Report ICS-8604, University of California, San Diego(1986)
4. Yasushi Hayashi, *Stock Price Is Forecastable - Introduction to Chart Analysis*, Kanki Publishing (2000).
5. Masafumi Hagiwara, "Buck-Propagation with Selection Mechanism," Report of Institute of Electronics, Information, NC89-104(1989)
6. Toshio Asahi, Kenji Murakami, Tsunehiro Sagamihara,"BP Algorithm for Optimality and highly speeding of Neural Network Structural Learning," Report of Institute of Electronics, Information, NC90-64(1991)
7. Masumi Ishikawa, "Recent Topics on Learning Algorithms of Neural Network," Instrument and Control, Vol.30, No.4, pp. 285-290 (1991)
8. Yutaka Matsunaga, Yoshiaki, Nakade, Osamu Nakagawa, Kazuyuki Murase, "Buck Propagation Algorithm to Auto-Elimination of hidden units in Higher Neural Network," Trans of Institute of Electronics, Information, Vol. J74-D-II, No. 8, pp. 1118-1121 (1991)
9. Yutaka Matsunaga, Yoshiaki, Nakade, Osamu Nakagawa, Manabu Kawanaka, "Buck Propagation Algorithm to auto eliminate redundant hidden units from Competition," Trans of Institute of Electronics, Information, Vol. J79-D-II, No. 3, pp. 403-412 (1996)
10. Ippei Murata, Junzo Watada & Jaeseok Choi, "Structural Learning of Neural Network and Its Application to Stock Forecasting," Proceedings, International Conference on Management Engineering (ISME2006), Kitakyushu, JAPAN, March 10-12, 2006, pp. R30/01-R30/06, (2006)
11. Junzo Watada. "Structural Learning o Neural Network and Its Application," Keynote Speech, Proceedings, An International Conference on Intelligent systems (ICIS2005), at Kuala Lumpur, Malaysia on Dec. 1-3, 2005, pp. 4-12, (2005)

# Customer Experience Management Influencing on Human *Kansei* to MOT

Shin'ya Nagasawa

Graduate School of Asia-Pacific Studies  
Waseda University

Sodai-Nishiwaseda Bldg. 7F, 1-21-1 Nishi-waseda, Shinjuku-ku, Tokyo 169-0051, Japan  
nagasawa@waseda.jp

**Abstract.** Recently knowledge obtained from customer experience is effective in producing and designing goods and products, especially ones that influence on *Kansei* and psychological aspects of consumers. If such methods can be summarized combined more organizedly into some theoretical methodology such knowledge will be more powerful. Furthermore, it is also possible to understand hit products and brand goods that have little understanding. The objective of the paper is to explain the relation and the meaning of Customer Experience Management approach, that is manufacturing and fabrication that influencing on the human *Kansei* to Management of Technology and Engineering (MOT).

## 1 Introduction

From the view points of Management of Technology (MOT), the importance of Management of developing new products (Management of New Product Development) can be illustrated as follows:

The production rule in 20th century manufacturing industry is placed stress on “how to make”. The production rule in 21st century is stressed on “what to make”.

Attractive goods and products are most important for all divisions of a corporation. The management of new product development plays a central role in the corporation and is persuade from such multi-direct view as combined viewpoints between technological points and non-technological points (Technology + Non-technology).

The problem of manufacture industries in Japan is that high technological products can not result in high sold ones. This is not coming from technology itself but from management philosophy or methodology, therefore the education and research on management of technology (MOT) should be emphasized.

The Management of New Product Development is to commit in all aspects of developing a new product and to make it sure that developed products should be successfully hit [1].

There are rules and sharable attributes among cases of hit product development.

The author [1]-[3] tried to mine and build rules among successful attributes and features in project management after interviewing directly with product and project managers who succeeded in developing hit products and scrutinizing the gift, talent and ability required for developing hit products. The developing method of new products and hit products is proposed as “seven tools for new product planning”. The tools are widely used and contributes in producing hit products in various fields .

On the other hand, the author recognized and understood through his long experience that products should not have their own functions and quality elements but also something such as design that moves, touches and impresses human *Kansei* (representative word of feeling, taste, emotion, etc.) and psychology. Nevertheless, up to the present, the author could not commit in building the theory and methodology to install such features and attributes in new products, even if he scrutinized and prepared them.

The author concluded that the customer experience can play a central role in a theory and method materializable to install something *Kansei* and psychological in a new product and found that the reasons of hit products and brand products cannot be explained by conventional marketing theory but the experiential marketing and customer experience management can do how to sell so well.

The following will be spent to describe how the philosophy of experiential marketing and customer experience management, which is manufacturing and fabrication that influencing on the *Kansei* connects with, have the relation with Management of Technology (MOT).

## 2 Design of Intangible Event from Tangible Product

It is not always true that highly technical superior and high quality products sell well. Therefore, it is necessary to employ MOT management in developing a new product.

Extremely speaking, MOT is not required if highly technical products and high quality products can sell well. In this sense, MOT discusses other aspects and managements besides technology and quality.

There are many aspects and managements considered in MOT but the most important is about customers. It is frequently said that “a product did not sell well even if the product was good technically.” The product is a technical product but not the goods. Therefore, if customers do not accept highly technological products and high quality products, these are not the goods.

So the most important requirement for “the goods” is to design an intangible event that builds a smooth relation between a tangible product and a user. In this discussion, the stress is placed on the contact point between a product and a user, and an interface or a touch between a product and a user. Furthermore, it is more important to provide some “meeting” to make a user encounter such a product.

The design of meeting between a user and an tangible product does not mean to decide only a shape and styling and form of a product but provides a product such as a user feels well or becomes happy in using it.

A tangible product that was designed under consideration of its users should be sold well and be the goods. The intangible event means to combine a product with the goods such as both sides of a coin. This is to design an intangible event that makes a user meet a tangible product. In other words, the most important is how useful a product designed can be for people who buy this [20].

Designers and producers used to have a strong satisfaction and confidence of designing a high technological product because of its high technology. But customers do not understand its high technology and do not percept whether its technology contributes and has relation with their happiness. This tendency becomes more remarkable as technology becomes higher. The explanation can be obtained from experiential marketing and customer experience management.

### 3 Customer Experience and Strategic Experience Module

“Customer Experience” does not mean individual experiments obtained in the past, but indicates the value of something that impresses and appeals on the *Kansei*, senses and impression of a user’s such as a customer actually directly feels and is impressed on the contact with a company and brand.

“Customer Experience” is not an additional and incidental value but an essential and intrinsic value obtained in the case where a product and service are understood from customer oriented points of view that are provided by a company and a brand. The objective of marketing that creates “Customer Experience” (“Experiential marketing”) is not to provide products and services as a tangible thing to customers, but to take the consume of the customers’ in the context of their life style and to interpret their consumption through putting stress on their senses and feelings in the process.

Bernd H. Schmitt who is a professor of Colombia University promoting the experiential marketing and customer experience management classifies the experience values into five modules illustrated in Table 1. The classification is worth of strategic bases of marketing activities.

**Table 1.** Modules of Strategic Experience Value Provided by Bernd H. Schmitt

Class	Contents of Experience Value
SENSE	Sensitive experience value to appeal on five senses
FEEL	Emotional experience value to appeal on feeling and mood
THINK	Intellectual experience value to appeal on creativity and cognitively
ACT	Behavioral experience value and life style to appeal on physical behaviour
RELATE	Relative experience value to appeal on confirmative group and cultural group.

Note) names used by Bernd H. Schmitt are partly changed in this paper

### 4 Case Study of Customer Experience Creation in Japan

Let us introduce four cases of experience value creation from Nagasawa [21].

#### 4.1 INAX “SATIS”

“SATIS” of INAX shown in Figure 1 is a series of sensational sanitary ceramic products used in a toilet room, a washing room and a powder room. It is a toilet facility without a water tank whose design is impressive on users’ *Kansei*. When it was put in market, it was a very much sensational and very much influent production. As “SATIS” changes the perspective of understanding things such as a toilet room is replaced as a hospital space, the conventional concept of a toilet room is recreated into another customer’s value that can provide a unique value added product.



**Fig. 1.** INAX “SATIS” [21]

Let us analyze a product under a framework of “customer experience” as a concept to create customers’ value whether it gives any impression on customers or what kind of customers’ value it creates. This analysis in Table 2 is also considering competitive advantage.

**Table 2.** Customer experience of INAX “SATIS” as Strategic Experiment Module [21]

Class	Customer Experience included in “SATIS”
SENSE	# Design making customers feel an aesthetic space # High function peeling on customers senses # Toilet space as a new living circumstance
FEEL	# It makes customers percept clean circumstance. # Sense of relief is obtained # ideal toilet space appraisable
THINK	# Tanklessness widened a toilet space. # Sufficient coordination full of imagination # Tankless washing beyond expectation.
ACT	# Toileting behavior is changed according full-automatization # Toilet space providing to guests and friends # Intellect toilet space.
RELATE	# Eco-design depending on social responsibility # Appealing to customers by branding # Building new social categories

#### 4.2 NISSAN “X-TRAIL”

Nissan sells “X-TRAIL” shown in Figure 2 well since November, 2000 when they started its sales. They kept the top position of domestic sales volume in SUV (Sports Utility Vehicle) for 4 years till 2004. They sold totally 140,000 X-TRAILS in 150 countries and kept the top in the class in each of countries and expand sales volume this year. Its reason can be understandable since “X-TRAIL” was developed based on “Seven tools for New Product Planning.”



**Fig. 2.** NISSAN “X-TRAIL“ [21]

Nevertheless, it is not persuasive if we result such a huge sales volume on its function and convenience. It is essential to analyze elements and functions beyond “a hit vehicle”. Therefore, let us analyze the “customer experience” that creates the customer value. On the other hand, we can image Cadillac, Porche, and Pherarii as an luxury vehicle. But NISSAN’s “X-TRAIL” sticks on price 2 million JPY as the waiting customer’s expectation. They take it into consideration that the price can realize customer experience as shown in Table 3.

**Table 3.** Customer Experience of NISSAN “X-TRAIL” as Strategic Experiment Module [21]

Class	Customer Experience of “X-TRAIL”
SENSE	# Design of rectangular
FEEL	# CM of a falling person stimulates on the psychology that a person plays a sport
THINK	# Astonishment and Thinking of washable interior in a car
ACT	# Events of “X-TRAIL JAM” makes people into outdoor sports “X-TRAIL”
RELATE	# Building a flexible Fan Club “X-TRAIL” of outdoor sports

**4.3 Canvas Bug of Small Kyoto Company “Ichizawa Hampu”**

Let us discuss about “Ichizawa Hampu,” a veteran company of canvas product in Kyoto that is famous of producing and selling canvas bugs (Figure 3) all over Japan. According to the interviewing with the fourth owner Shinzaburo Ichizawa, CEO and his wife Emi Ichizawa, Board Director, let us analyze the secrete of Ichizawa Hampu behind its brand from point view of developing power and customer experience. Ichizawa Hampu is a famous Japanese brand talked as “Japanese Louis Vuitton”, and has one shop and catalogue sales with constantly three to four thousand orders.



**Fig. 3.** Canvas Bug of Ichizawa Hampu [21]

**Table 4.** Customer Experience of Ichizawa Hampu’s Canvas Bugs as Strategic Experience Module [21]

Class	Customer Experience owned by “Ichizawa Hampu”
SENSE	# Brand recognition by visual labels # Visual touch of canvas # careful finish of good supple texture
FEEL	# the long life of products creates attachment to products # Labels bring out nostalgia
THINK	# careful craftsman art provides spirit of study
ACT	# life style to take a good care of things
RELATE	# Customers’ Royalty created through re-upholstery service

As shown in Table 4 Ichizawa Hammpu creates customer experience emphasizing on customers of young generation with under consideration of customers life styles, nevertheless it is a veteran company of a traditional industry in Kyoto, symbolized by natural sailcloth adhering craftsman art. The source of customer experience value creation of Ichizawa Hampu is summarized in three points of developing products; 1)

their attitude of careful craftsman’s work, 2) new discovery of traditional natural cloths for sailcloth products, and 3) feedback of customers’ needs. That is, the development power of a company and the customer experience are inextricably linked.

**4.4 Football J1 Team “Albirex Niigata”**

Let us discuss about Football J1 team “Albirex Niigata” that has incomparable big drawing power. It is spectacle and even impressive that every match can gather more than 40 thousand spectators who put on an orange uniform of team as in Figure 4.



**Fig. 4.** Emblem of Football J1 “Albirex Niigata” [21]

Niigata is said unfit to any pro-sports. There is no big company that can be a sponsor and no star player. So nothing fits to pro-sports. From the viewpoints of customer experience we analyzed the phenomena of Albirex Niigata that is evaluated to be a “miracle”. In the analysis we employed the talk at Wa seda Business School and various interviews by media that Mr. Hiroshi Ikeda,CEO of Albirex Niigata gave and illustrated the efforts to create customer experience in Albirex Niigata. From viewpoints of customer experience, it is very effective that Albirex Niigata succeeded in combining between the enthusiastic space of 40 thousand fans and experience to appeal their love of the home town. This success gave us the knowledge that the frame of customer experience is effective not only for developing products but also for sports business and entertainment industry.

**Table 5.** Customer Experience of “Albirex Niigata” as Strategic Experience Module [21]

Class	Experience Value of “Albirex Niigata”
SENSE	# Enthusiastic Experience of 40 thousand fans without encountering it in other places # Visual experience of Orange Team Color
FEEL	# Passionate experience attaching to their own handmade team without any support of big companies and without any star players
THINK	# Experience of turning the negative image of Niigata to its positive image
ACT	# Festival experience once two weeks # Experience of unity of loud cheering by supporters
RELATE	# Experience of unity of supporters in their own home Niigata

**5 Conclusion**

In the conclusion, let us summarize the new relation of customer experience with conventionally functional benefits and the role that the customer experience plays on the basis of analyses of the above-mentioned four cases.

It is effective in analyzing relative relation between functional benefits and customer experience to discuss about what can be provided and what can be created from the view points of creating customers' value.

From viewpoints of functional benefits, for instance, a toilet room provides improvement of functionality and usability of the toilet room as a customer value. Also, the depth and breadth of assortment by providing various options that improve customers' selections by their own tastes and by unifying design that improve image can be understood to create value as same.

From viewpoints of customer experience, for example, the total space including a toilet room creates customer's value nevertheless the toilet room itself creates a customer's value. That is, providing with the space and atmosphere of a toilet room that are different from conventional ones, it creates customer value by appealing customer psychology and *Kansei* to change the recognition of a toilet room totally and radically. This means to create totally new customer value that influences customer' life style.

As mentioned above, on the comparison between customers' values created by functional benefit and customer experience, these values have complimentary relation even though there are some duplicates. The functional benefit of a toilet room provides customers' value by improvement of functional and beneficial aspects but customer experience create customer' value by improving psychological aspects of customers' *Kansei*, that functional benefit cannot provide. In other words, the functional benefit is to give physical and materialistic satisfaction and the customer experience is to provide psychological and *Kansei* satisfaction.

Figure 5 illustrates the relative relation between functional benefit and customer experience. These are interpreted that both have their own field and are in the complimentary relation. It should be noted that supporting the complimentary relation can be understood from MOT approach. That is, when MOT approach realizes an innovative technology (development of direct valve washing and so on) it provides functional benefit (realizing of small space with tanklessness). Customer experience (Image change of a toilet room) influences on customers' mind by inventing innovative technology and creates value to influence *Kansei*. The complementarity between functional benefit and customer experience creates totally different and innovative customers' value (hospitality space of a toilet room).

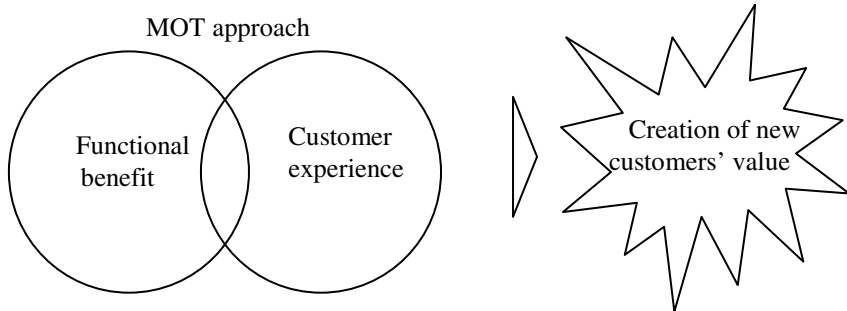


Fig. 5. Image of complimentary relation between functional benefit and customer experience



## References

1. Nagasawa, S. (ed.): Live MOT: Messages of Product Managers, Nikkagiren Shuppansha (2004) in Japanese
2. Nagasawa, S.: An Unpainted Face of the Empire of Luxury Brands: LVMH Moet Hennessy. Louis Vuitton, Nihon Keizai Shimbunsha (2002) in Japanese, and Taiwan Business Weekly (2004) translated in Chinese
3. Iwakura, S., Iwatani, M., Nagasawa, S.: Strategic Design Management in Honda, Nihon Keizai Shimbunsha (2005) in Japanese, and Human Books (2005) translated in Korean
4. Nagasawa, S. (ed.): Value Creation through Customer Experience That Enables to Develop Hit Products: Fabrication That Influencing on Human *Kansei*, Nikkagiren Shuppansha (2005) in Japanese

# Getting Closer to the Consumer: The Digitization of Content Distribution

Peter Anshin and Hisao Shiizuka

Informatics, Kogakuin University Graduate School,  
Nishishinjuku 1-24-2, Shinjuku-ku, Tokyo 163-8677, Japan  
ed05001@ns.kogakuin.ac.jp,  
shiizuka@cc.kogakuin.ac.jp

**Abstract.** Recent advances in digital content distribution has created numerous benefits for both owners and consumers of content, and fundamentally changed the distribution business. This paper presents a new plan of digital distribution platforms that have created numerous benefits for content producers and consumers by decreasing the degree of separation between producers and consumers.

## 1 Introduction

Distribution is the fundamental piece of the lifecycle of content; development, production, distribution and consumption. Recent technical advances have created groundbreaking developments in content distribution, and changed the way that content is consumed as well as produced. This paper supports the view that digital distribution platforms have created numerous benefits for content producers (referred to herein as “producers”) and consumers by decreasing the degree of separation between producers and consumers.

## 2 Distribution and Distributors

Distribution is the fundamental link from producer to consumers. It is the means by which producers’ content is distributed to consumers. The link to producer provided by distributors allows consumers to consume content. Without it, content may be produced but never reach the consumer. Without distribution, producers are not able to expect returns from the exploitation of their content, and therefore less likely to produce content.

In the context of this paper, distributors are those entities who acquire content from content producers, and then sell it to entities who in turn sell it to consumers; they are “middlemen”. Most distributors do not directly sell to consumers (a task undertaken by retailers, who may also act as distributors), although it is conceivable for some distributors to do so. Some distributors are companies who are well connected with both producers and retailers, but who do not have direct ties with consumers, while other distributors are well connected with both producers and consumers. On the other hand, some distributors may also produce content.

Distributors are “shoppers” who shop at various producers for content, and then offer this content to retailers, who then sell the content to consumers. Certainly some

distributors wisely ascertain retailers' tastes before shopping, and later shop to find what they are looking for, going back to the retailer with the content they had hoped to obtain. Still other distributors create their market by convincing retailers regarding the popularity of certain content; then miraculously pulling this exact content from their pockets and offering it to the retailer who is already keen to license it.

It may make more sense to speak of distribution, rather than distributors, as it is this function, and the behavior of those who perform it, that is the subject of this paper, rather than distributors. Nevertheless, since distributors are indeed the actors who are responsible for traditional distribution, they cannot be ignored in the course of this analysis.

Distributors have been pervasively present in the sales of nearly all forms of content. Because of their power, they have had a great influence on content production. Their power comes from their ability to determine the end value that consumers are able to pay for content; that is, by directly influencing the price at which consumers buy content, influencing consumer tastes by the product they distribute, and affecting producers' investment in production.

Distribution has traditionally involved a middleman between the content producer and the consumer. The reason for the middleman is simply that producers didn't have adequate reach to consumers. Why not? Why can't producers adequately distribute their contents to their fans? One reason is the diversity of content delivery platforms, making it impossible for producers to distribute their content to consumers because producers are unable to manage distribution of their content across all of these platforms.

If not for these intermediaries, producers would be forced to distribute content directly to consumers, likely a nearly impossible task for any single producer given the complex and nearly endless supply of non-uniform retail throughout a vast array of locations, representing an infinite array of tastes. This task is certainly all the more daunting given the number and breadth of international markets. The presence of distribution middlemen has enabled producers to specialize in their areas of expertise; production. By providing a means of specialization, resources have been freed up that would otherwise be diverted into distribution, producing waste.

It should be clear that we are talking about content, and therefore licensing from one entity who may hold the copyright for the underlying content, or an entity who has already has a license from the copyright holder that enables it to then license these rights to another entity, a "sublicensee".

Distributors have played an important role in helping content industries to grow by performing a task that would otherwise have to be taken on by content producers. Certainly, some producers do distribute the content they produce, but it might be argued "At what expense?"

## 2.1 Content Delivery Platform

Although "rights" are the subject of the license, the license has also traditionally included a physical embodiment of the content; for example, a compact disk, a 35mm film, or a video cassette. I shall refer to these physical embodiments of the content elements as *content delivery platforms*. *Content delivery platforms* are the means by which content is distributed to consumers. In traditional media, distributors have

distributed content to consumers using several different routes. In some cases, they may sell physical content to retailers from whom consumers purchase content. Take the distribution of music on a Compact Disc (CD) to a music “store”. Distributors contract with producers of the CDs to distribute their CDs to retailers’ stores, where the CDs are housed, and from where consumers then purchase CDs. Another example is a film distributor who sells a 35mm film, another platform, to an exhibitor. The exhibitor then sells (exhibits) the film to the consumer using a projector and theater, yet another *content delivery platform*.

### 3 What Is Wrong with Traditional Distribution?

While traditional distributors have indeed produced many benefits for the content industry, some of which are outlined above, distributors have also arguably produced a good deal of waste. What is their purpose – what do they do? Should they aim to make the widest distribution? Or obtain the best price? Are they acting in the best interest of the seller, consumer, or simply in the best interest of themselves?

Distributors are typically paid a percentage of the price at which they sell content to buyers. Of course there are other compensatory forms, but most distributors aim to distribute as much content, for as high a price as possible so as to maximize their revenue share. Distributors do not necessarily aim to benefit any particular content styles, although they may seek to develop specific artist or content styles if they feel that by doing so, they can reap a financial benefit from such development. Simply put, distributors have become a necessary rather than a desired component of the traditional content business *because they have built the business, and their role in it*. As long as physical distribution has existed, there has been a role for distributors.

#### 3.1 An Example of Dealmaking: The Pre-scale of Film Distribution Right

Middlemen typically negotiate individual content deals. Content “dealmaking” sessions can be as unpredictable as the value of the content under negotiation. This is no surprise, given the fact that the value of the content is indeed more of a subjective than objective analysis. Slowing the deal down is not an issue if, but for the middleman, the deal would not have occurred. But digital distribution has shown that in many cases producers simply desire to distribute their content to as many retailers, and therefore consumers, as possible.

Information that distributors seek to learn typically includes descriptive information regarding the content. Certainly in the case of a “pre-buy”, where rights to distribute the content are purchased before the content is produced, the way in which this information is communicated to the buyer is of paramount importance.

For instance, in the case of a pre-buy of film rights, the producers typically have a script or synopsis of the film, a tentative budget, and know certain elements that will appear in the film. In order to obtain a sale, sellers typically typically make each of these elements seem more attractive to buyers buy puffing. Certainly the line between puffing and misrepresentation is often blurred, and occasionally result in disputes when the content produced is in fact does not meet buyer expectations. This conflict is typically dealt with using contractual measures, but often times, parties do not strictly adhere to their written contracts.

Regardless of the result, the “sale” process for pre-buys of films often involves multiple meetings as well as strategic repositioning by sellers who try to obtain the best deal, so as to maximize their margins. Dealmakers spend an extraordinary amount of time in this process of making the “pitch” and negotiating and closing the deal. The outcome of the deal does not determine whether this time is wasteful. More important is the way in which the deal was negotiated. In many cases, a great deal of time is required, “slow time”, during which there is a good deal of discussion of issues that are of little or no relevance to the transaction.

### 3.2 Waste in Doing Business

Distributors often rely on a network of connections used to both acquire and sell content. These connections are founded on friendships, not in and of themselves harmful or unique to any industry in particular; connections and friendships are a requisite part of distribution in nearly any industry. Still, distributors in the content industry tend to form relationships with buyers and sellers that go far beyond the boundaries of most relationships, and in many cases result in waste because these relationships may limit the number of possible opportunities to distribute content, and the price at which it is distributed.

Even if distributors successfully close a distribution deal, the producer needs to deliver the goods (e.g. videotape, DVD) that embody the content to the retailer. Distributors receive and then deliver these physical goods to retailers. It is indeed true that the distributor takes on risk by being involved with this process, but its involvement also allows it to another “justification” for the distributor to receive its fee to cover its delivery costs. The point here is that, when there is less need to deliver a physical embodiment of the content, there is also indeed less justification for the distributor to receive a fee for this duty. The corollary here is that, delivery costs decline when physical goods needn’t be delivered.

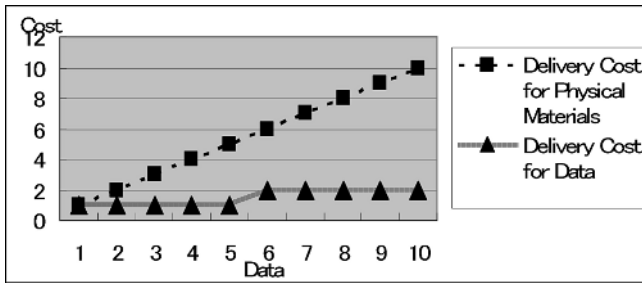


Fig. 1. Delivery cost comparison

### 3.3 Waste in the Payment and Collection Process

Distributors are not only responsible for delivery, but for collecting payments from retailers. Distributors collect these payments, deduct their fee, and pass on the balance to producers. This is obviously a very inefficient process as payments are not directly made to interested parties and rather made “through” parties, creating the likelihood

that payments are unjustifiably delayed, reduced, commingled or otherwise affected in a manner that produces waste.

### **3.4 Prices**

As pointed out above, the distributor deducts its fee from what it receives from retailers. It therefore has an incentive to “markup” the price at which it distributes. Distributors should indeed receive a fee for identifying retailers who maximize returns for producers, entering into deals with these retailers, and servicing these deals. Distributors desire to sell to retailers who will consistently purchase content on predictable terms, so the distributor can focus on obtaining new content and developing new distribution channels rather than negotiating the deal.

As pointed out above, for a given deal for the distribution of a particular content, distributors actual work performed on the deal may include introducing retailers, closing deals, delivering content, collecting payments from the buyers and continually servicing the deal by, for instance, supplying new content into the deal.

It is important to realize that prices at which the distributor sells content may not be the best price for the seller, and rather may simply reflect the distributor’s desire to distribute to a particular buyer because of its relationship with the buyer, or reflect the “padding” of its distribution costs. Prices under this system are not necessarily reflective of actual costs or effort, and may simply represent the biased advantages of a distributor taking advantage of less powerful producers or retailers who, without information, relationships or experience, are in no position to challenge the distributor’s advantageous position.

### **3.5 Conflicts of Interest; Knowing Who the Client Is**

One would expect that distributors’ client is the content producer rather than the buyer. Indeed, if not for the content produced by the producer, there would be no distribution deal. But this is not always the case, and there is no absolute rule dictating who the distributor works for, the content producer, or the retailers.

While this is helpful for concluding a deal where otherwise one might not occur, it may not necessarily bring about the best result for either party. In other words, being partial to neither party helps to more efficiently close the deal, but doesn’t necessarily act in the best interest of one of the parties. This may result in a party having to forego a better deal because it has been “locked” into a deal with terms not as good, that has only happened because a distributor had acted in the interest of both the content producer and the buyer.

### **3.6 Distribution Limited to a Single Media Platform**

Another disadvantage of distribution under the traditional model is the fact that most traditional distributors work only with a single distribution platform. For example, distributors who specialize in distributing theatrical film typically do not also engage in DVD distribution. Retailers of this content differ; the theater owners differ from the DVD rental stores. Each has a platform that is fundamentally different from the other in its terms of its market, pricing and technology. The result is that distributors are forced to specialize in one platform type rather than several. Traditional distributors

don't seem able to expand their expertise beyond this horizon. In an ideal market, distributors might like to distribute to more than one platform to maximize revenues and market breadth.

Distribution is the fundamental link from Producers to consumers. It is the means by which Producers' content is distributed to consumers. The link to Producers provided by distributors has allowed consumers to consume content. Without it, even though content is produced, it may never reach the consumer. As a result, without distribution, Producers are not able to expect returns from their content, and therefore not likely to produce content.

## 4 Why Digital Distribution?

We hear about "Digital distribution" everywhere, but what is it? What is good and bad about it? And what effect is it having on distributors?

Digital distribution is the distribution of content using digital, as opposed to analog, delivery platforms. Content must be digitally formatted to enable distribution on these platforms. Distribution of content digitally, using digital delivery platforms, streamlines the distribution process and in so doing, creates a host of advantages for content producers and consumers. It also reduces the importance of the distributor's job because it shortens the distance between producers and consumers; consumers are able to obtain the content they want, when they want it. Digital distribution is providing a much needed test of the "bads" or waste of traditional distribution, and while the jury is still out, it appears that the use of digital distribution by producers and consumers is on a path to eclipse traditional distribution.

As shown above, under the traditional model, digital distribution shortens the time required for the distribution deal to be made, as well as the time required for the consumers to receive and "consume" content. Pricing on digital distribution deals tends to be more uniform than in traditional distribution, and most deals are not subject to fixed-sum "advance" or "minimum guarantee" payments. Advances are typically required in distribution deals that give the distributor an exclusive right to distribute. Digital distribution deals typically do not call for such deals, as producers distribute to as many digital retailers as possible, much as they would distribute physical product to as many retail stores as possible. Using digital distribution, producers are free to distribute to as many retail stores as they wish without the presence of distributors, who may inflate prices and create other inefficiencies as described above.

Under the physical music distribution model, producers would be forced to ship albums to distributors, who would then ship these to retailers. In contrast, digital distribution enables producers to deliver their product directly to digital service providers, who deliver to consumers. Digital music service providers, such as iTunes, currently receive a "distribution fee" of approximately 30%, leaving producers with much more than they would receive under a traditional distribution scenario. Other service providers pass on even more to producers. Service provider costs consist mainly of formatting, encoding, transmitting and maintaining data ("digital service costs"). In fact, while many service providers have entered the digital distribution business expecting returns that resemble those received by distributors in traditional distribution, these digital service providers are learning a harsh lesson that it is not as easy to justify

distribution fees in the digital world because producers understand the costs, and it is more difficult for service providers to mask these costs.

Digital distribution has lowered the market entry cost because producers are now able to independently digitally distribute their own content. As more producers have entered the market, digital service costs have dropped, and returns for producers have increased.

Another benefit of digital distribution is the amount of content that may be purchased. More content has been made available via digital distribution than with traditional distribution. Moreover, content has been made available in more geographical locations than in traditional distribution. This obviously includes content from different countries around the globe, which can now be purchased from a single terminal, from anywhere in the world.

As content is purchased more rapidly, consumers are also able to sample, and purchase, a greater volume and variety of content than in traditional distribution.

#### **4.1 Portability into More Platforms**

Digital distribution has enabled content to be more easily transferred to a greater variety of delivery platforms previously unavailable under traditional distribution. Or instance, music may be downloaded onto a personal computer, and then transferred to a portable device. This further reduces prices for consumers, because no single platform can have a “monopoly” on the content.

#### **4.2 Not Only 1 Big Music Store, But Thousands of Digital Mom-and-Pop Stores**

While some of the middleman distributors from traditional distribution have sought to shift into digital distribution, digital has allowed anybody to become a retailer using digital distribution technology to open their own digital retail outlets, complete with payment processing and all the other functions found on other services(see reference 1).

#### **4.3 Promotion and Sales**

Distributors have traditionally undertaken to promote and market content to content buyers. In so doing, they have passed on to producers promotion and marketing costs incurred in connection with their selling the content to content buyers. For instance, if we were speaking about the distribution of a film to a theatrical exhibitor, the distributor must incur a “cost” to convince a video retailer that the film will perform well theatrically, and this cost is passed on to the producer.

Digitally distributed content may be promoted using “viral” methods; promoting the content online or using other digital media. Viral and word of mouth promotion and advertising is extremely important to producers digitally distributing their content, as without it, their content may not reach consumers. This very timely subject is worthy of a separate study.

#### **4.4 Consumers**

Consumers benefit from digital distribution in numerous ways. As discussed, digital distribution reduces the content acquisition cost for a variety of reasons. Consumers



certainly benefit as they have more funds leftover to use for other purchases, or to purchase more content, which benefits content owners.

The digitalization of content enables producers to distribute their films using one platform, instead of several. Consumers needn't purchase several platforms in order to enjoy content, obviously reducing costs. Digital distribution also greatly increases the supply of content, as it creates more opportunity for producers' content to reach consumers. It is both easier for producers to offer their content directly to consumers without going through distributors, and easier for consumers to obtain content. As well, digital distribution increases the variation of available content because it enables more content to be more easily distributed.

While at present, not all produced content is available for digital distribution, more and more is rapidly becoming so, through a single or several service locations. It may be argued that the lack of success of some digital music download stores is because not enough digital material has been made available online. Even iTunes (see reference 2), Apple Computer's digital music store, arguably the most comprehensive, has only a fraction of music available in physical form. Certainly, a single store with an extremely comprehensive collection of content would be attractive because it would be a single location from where people could acquire content.

#### 4.5 Better for Producers

Digital distribution is advantageous for content producers because it offers an increased return, clarity, control and freedom from intervention. Producers know who is buying their content, how much they pay, and when they purchase. This information helps producers not only to control risk by managing their content, but also helps them to know who their audience is, and produce content that suits the taste of their audience. In the traditional distribution model, distributors have earned where producers and consumers have not been aware of such earnings. Digital distribution has

**Table 1.** Comparison of Traditional and Digital Distribution

	Traditional Distribution	Digital Distribution
Distributor	Yes	No
Transaction Cost	High	Low
Transaction Time	Long	Shorter
Delivery Time	Long	Short
Delivery Cost	High	Low
Pricing	Advance	No Advance
Exclusive	Yes	No
Service/Format Costs	High	Low
Content Variety/Availability	Less	More
Price to consumer	High	Low
Purchase points	Few	Many
Portability to other delivery platforms	Difficult	Easy
Freedom to consume (when and where)	Less	More
Consumer resale rights/practicality	Difficult	Easy

helped to clarify the obligations of, and meaning and levels of compensation for distributors. Through digital rights management (DRM), producers now have the freedom to better control the uses of their material, rather than relying on distributors to regulate uses, or be subject to distributors' unilateral decisions, and misinformation, regarding the exploitation of their content.

## 5 Conclusion

The net effect of digital distribution is that producers may spend less on distribution, and instead spend this savings on financing production, promotion and marketing. What's next? Is digital the "final" word? Is there another form of next generation distribution? Will digital distribution help producers to more effectively create content that is more tailored to consumer tastes, or to tap into new fan bases? Arguably, with tools such as "Hit Song Science", "Word-of-mouth" and blogcentric communities, it already is. But it is safe to say that while distribution mechanisms may be improved through digitization, content production will, for the time being, remain a task that requires uniquely human creative skills. And shouldn't we keep it that way, as entertainment is entertaining because it is random and unpredictable, while distribution networks are more useful for both producers and consumers if they are more predictable, and less random, not subject to the whims of distributors.

## References

1. [www.burnlounge.com](http://www.burnlounge.com), or [www.indiestore.com](http://www.indiestore.com)
2. [www.itune.com](http://www.itune.com)

# An Action Planning Module Based on Vague Knowledge Extracted from Past Experiences

Grzegorz Popek

Institute of Information Science and Engineering  
Wrocław University of Technology  
50-370 Wrocław, Janiszewskiego 11/17, Poland  
grzegorz.popek@student.pwr.wroc.pl

**Abstract.** In this paper, an agent architecture is presented and a formal model for a preference choice module and planning module based on vague knowledge is described. An agent situated in an environment gathers information about objects placed around it and stores this information inside its own database. According to its preferences, an agent plans its actions in order to change surrounding environment into its preferences. Because of a partial lack of knowledge about an environment, an agent chooses a plan with a highest success ratio due to its previous observations.

## 1 Introduction

A BDI-based [5] agent with a language grounding module [3] and planning module based on a vague knowledge will be described. An agent is placed in an environment and perceives it during an interaction process [3]. He gathers information into a form of base profiles stored in a private profile base. Existence of the base is an important extension of typical BDI architectures which were introductory defined in [5,7,8] and allow modal language grounding process (using operators of knowledge, belief and possibility) implementation. A list of operators is wider than usually proposed solutions [2].

We assume, that an agent performs a cycle consisting of: gathering information about an environment, a comparison of a current world state with a preferred states, computation of a plan leading to one of the preferred states, plan realization, results verification.

In proposed architecture an agent is equipped with an action planning module. This module uses a collection of action's operators ([1]) and refers not only to the current perception of an environment, but also to the knowledge gathered in a private profile base, which allows an agent to come up with a possible image of unperceived facts.

## 2 Model of Reality

### 2.1 Basic Concepts

Let us assume, that a set  $O$  of objects situated in the world  $W$  (see also [4]) and a set of atomic unary relations  $P_k \subseteq O$  ( $k=1,2,\dots,K$ ) is given and that a state of the world is not constant and is observed by an agent in certain moments.

To represent these moments we use unique time pointers  $t_i$  describing moments of the agent's perception. A set of all time pointers is denoted by  $T=\{t_i; i=0,1,2,\dots\}$ .

In every moment  $t \in T$  the world  $W$  is in some certain state. For a description of an irrespective state of the world  $W$  at the moment  $t \in T$  we use a *WorldProfile*( $t$ ).

**Definition 1. *WorldProfile*( $t$ ).** A state of the world  $W$  at the moment  $t \in T$  is represented as a relational system ([4])

$$\text{WorldProfile}(t)=\langle O, P_1(t), P_2(t), \dots, P_K(t) \rangle,$$

where  $O$  describes a set of all objects existing in the world  $W$  and for every  $k \in \{1, \dots, K\}$  an inclusion  $P_k(t) \subseteq O$  holds and for every object  $o \in O$   $o \in P_k(t)$  ( $k=1, \dots, K$ ) holds if and only if at the moment  $t$  object  $o$  has a property  $P_k$ . The fact, that  $o$  has the property  $P_k$  corresponds to the objective state of the world  $W$ .

**Definition 2. *Profile Space*  $\mathfrak{P}$  and *Complete Space*  $\mathfrak{P}'$ .** Let a set of objects  $O$  and a number of properties  $K$  be given. A set  $\mathfrak{P}=\left\{ \langle O, P_1^+, P_1^-, \dots, P_K^+, P_K^- \rangle : \forall i=1, \dots, K \quad P_i^+, P_i^- \subseteq O, \quad P_i^+ \cap P_i^- = \emptyset \right\}$  is a *Profile Space* and elements of  $\mathfrak{P}$  are profiles. A set  $\mathfrak{P}'=\left\{ \langle O, P_1^+, P_1^-, \dots, P_K^+, P_K^- \rangle : \forall i=1, \dots, K \quad P_i^+, P_i^- \subseteq O \right\}$  is a *Complete Space* and of course,  $\mathfrak{P} \subset \mathfrak{P}'$ . Also, for  $P_1=\langle O, Q_1^+, Q_1^-, \dots, Q_K^+, Q_K^- \rangle$ ,  $P_2=\langle O, R_1^+, R_1^-, \dots, R_K^+, R_K^- \rangle$ ,  $P_1, P_2 \in \mathfrak{P}$ , let the following operations be introduced:

- $\setminus: \mathfrak{P} \times \mathfrak{P} \rightarrow \mathfrak{P}$  and  $P_1 \setminus P_2 \stackrel{\text{def}}{=} \langle O, Q_1^+ \setminus R_1^+, Q_1^- \setminus R_1^-, \dots, Q_K^+ \setminus R_K^+, Q_K^- \setminus R_K^- \rangle$ .
- $\cup: \mathfrak{P} \times \mathfrak{P} \rightarrow \mathfrak{P}'$  and  $P_1 \cup P_2 \stackrel{\text{def}}{=} \langle O, Q_1^+ \cup R_1^+, Q_1^- \cup R_1^-, \dots, Q_K^+ \cup R_K^+, Q_K^- \cup R_K^- \rangle$ .

**Definition 3. *Consistency of Profiles.*** Let a *Profile Space*  $\mathfrak{P}$  and  $P_1, P_2 \in \mathfrak{P}$  be given. We say, that  $P_1$  and  $P_2$  are consistent (see also consistency in [4]) if and only if for every  $k=1, \dots, K$   $P_{k'}^+ \cap P_{k''}^-(t) = \emptyset$  and  $P_{k'}^- \cap P_{k''}^+(t) = \emptyset$ .

**Theorem 1.** Let  $P_1, P_2 \in \mathfrak{P}$ .  $P_1 \cup P_2 \in \mathfrak{P}$  if and only if  $P_1$  and  $P_2$  are consistent.

A proof of the theorem follows directly from *Definitions 2, 3*.

**Definition 4. *EncProfile*( $t$ ).** For the state of  $W$  at the moment  $t$  described by *WorldProfile*( $t$ ), the empirically verified knowledge of this state developed by an agent is given as an encapsulated  $t$ -world profile *EncProfile*( $t$ ) ([4]):

$$\text{EncProfile}(t)=\langle O, P_1^+(t), P_1^-(t), \dots, P_K^+(t), P_K^-(t) \rangle \in \mathfrak{P}'.$$

$o \in P_k^+(t)$  ( $o \in P_k^-(t)$ ) denotes the fact, that  $o$  was recognized by an agent as exhibiting (not exhibiting) the atomic feature  $P_k$  and a related representation of this fact has been constructed in a relevant encapsulated database.

The fact, that  $o \notin P_k^+(t)$  and  $o \notin P_k^-(t)$ , where  $o \in O$ , denotes that  $o$  has not been recognized as either exhibiting or not exhibiting the atomic feature  $P_k$ .

**Definition 5. ProfileBase( $t$ ).** Empirically verified knowledge of the states of  $W$  until moment  $t_L$  and in moment  $t_L$  is stored in an encapsulated temporal base  $ProfileBase(t_L)=\{EncProfile(t_l): l=1,\dots,L\}$ .

## 2.2 Preferences

As we have written above, an agent holds information about current and previous results of an observation of the world  $W$  in a base of profiles  $ProfileBase(t)$ . Another information, that is stored, is information about this agent's preferences about a state of affairs.

**Definition 6. WorldPref – Preferred States.** A constant set of agent's preferences are gathered in  $WorldPref$ , given as:

$$WorldPref=\{Pref_i: i=1,\dots,I\},$$

where  $Pref_i \in \mathcal{P}$  and will be referred to as preferred states. The set  $WorldPref$  will be referred to as a set of preferred states.

We may there define a fact, that an agent's perception of the world  $W$  is adequate to a set of preferences  $WorldPref$  included within.

**Definition 7. Adequacy of the EncProfile( $t$ ) to Preferences.** Let a set  $WorldPref=\{Pref_i: i=1,\dots,I\}$  and an encapsulated  $t$ -world profile  $EncProfile(t)$  be given. The perception of the world  $W$  at the moment  $t$ , described by an  $EncProfile(t)$  is adequate to preferences of an agent described by the set  $WorldPref$  if and only if there exist such an index  $i$  that  $Pref_i$  and  $EncProfile(t)$  are consistent (see Definition 3).

## 3 An Architecture Outline

An agent starts from an observation of an environment and stores its results as an encapsulated profile  $EncProfile(t)$  in his private profile base.

According to the perceived state, an agent chooses a preferred state from the set  $WorldPref$ . It is done using a given similarity function reflecting commonsense world state similarity.

Based on the chosen state and its private action set describing its capabilities of changing the environment, an agent comes up with a plan.

Computing the planning process only for one of preferred states greatly reduces the complexity, although it is needed to point a fact, that use of a specific, defined similarity function might not reflect the abilities of an agent in shaping the world. Still, it is a part of a mentioned above commonsense strategy, when one chooses an objective due to the observations and then tries to plan a way to reach it.

Constructed plans are checked on their success rates (due to a fact of only partial agent's knowledge about the state of an environment) and one with a highest success ratio is chosen.

After the choice, an agent performs an action from the plan and checks its effects. If plan is still applicable and will result in a chosen preferred state, an agent continues with a plan. Otherwise, a new plan is computed. If an agent reaches a state that is compatible with at least one of its preferred states, it takes an empty action.

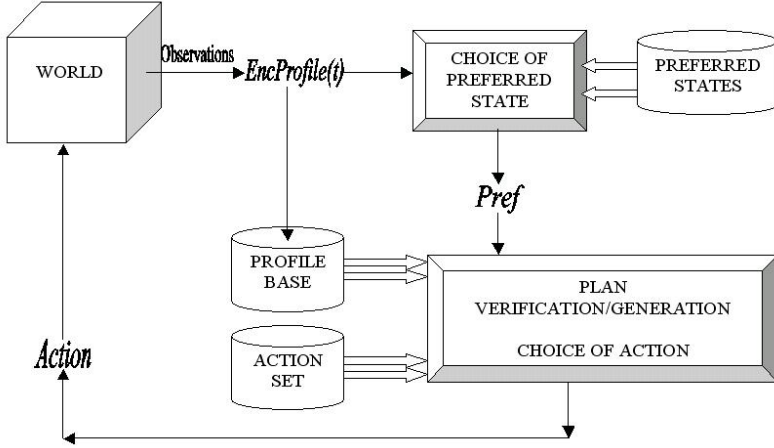


Fig. 1. An architecture model with a prechoice of a preferred state

**Definition 8. Comparison Measure.** Let an  $EncProfile(t)$  and a preferred state  $Pref_i = \langle O, P_{i,1}^+, P_{i,1}^-, \dots, P_{i,K}^+, P_{i,K}^- \rangle$  from the set of preferences  $WorldPref = \{Pref_i: i=1, \dots, I\}$  be given. Comparison measure  $comp$  between  $EncProfile(t)$  and  $Pref_i$  is given as follows:

$$comp(EncProfile(t), Pref_i) = \frac{1}{2K} \sum_{k=1}^K \left( \frac{card(P_{i,k}^+ \cap P_k^+(t))}{card(P_{i,k}^+)} + \alpha \frac{card(P_{i,k}^+ \setminus (P_k^+(t) \cup P_k^-(t)))}{card(P_{i,k}^+)} + \frac{card(P_{i,k}^- \cap P_k^-(t))}{card(P_{i,k}^-)} + \alpha \frac{card(P_{i,k}^- \setminus (P_k^+(t) \cup P_k^-(t)))}{card(P_{i,k}^-)} \right),$$

where  $\alpha$  belongs to an interval  $[0,1]$  and represents an attitude that is paid to unperceived objects' properties.

This measure was introduced with the following requirements:

- If an object  $o$  described in  $Pref_i$  as exhibiting a property  $P_k$  is perceived as exhibiting this property, the measure increases.
- If an object  $o$  described in  $Pref_i$  as not exhibiting a property  $P_k$  is perceived as not exhibiting this property, the measure increases.
- If an object  $o$  described in  $Pref_i$  as not exhibiting a property  $P_k$  is perceived as exhibiting this property, the measure decreases.
- If an object  $o$  described in  $Pref_i$  as exhibiting a property  $P_k$  is perceived as not exhibiting this property, the measure decreases.
- If an object  $o$  described in  $Pref_i$  as exhibiting or not exhibiting a property  $P_k$  is unperceived as exhibiting or not exhibiting this property, the measure increases or decreases according to an agent's attitude.

A parameter  $\alpha$  used in Def. 8 reflects an agent's attitude.  $\alpha=1$  reflects optimistic attitude and unperceived property is considered the same as that given in the preferred

state description.  $\alpha=0$  reflects pessimistic attitude and unperceived property is considered being opposite to the given in a preferred state description. Values of  $\alpha$  between 0 and 1 reflect intermediate agent's attitude to unperceived property.

## 4 Planning

### 4.1 Requirements Description

To describe a way, in which an agent affects the world, an action concept is given.

**Definition 9. Set of Actions, Model of Action.** Let a set  $A=\{a_j: j=1, \dots, J\}$  be given. We call  $A$  a set of actions. Every action  $a_{j(j=1, \dots, J)}$  from  $A$  has a form of a couple  $\langle Prec_j, Post_j \rangle$ , where  $Prec_j = \langle O, P_{Pre,1}^+, P_{Pre,1}^-, \dots, P_{Pre,K}^+, P_{Pre,K}^- \rangle \in \mathfrak{P}$  describes an action's preconditions and  $Post_j = \langle O, P_{Post,1}^+, P_{Post,1}^-, \dots, P_{Post,K}^+, P_{Post,K}^- \rangle \in \mathfrak{P}$  describes postconditions of undertaking the action. Also, for every  $k=1, \dots, K$ , an equality  $P_{Pre,k}^+ \cup P_{Pre,k}^- = P_{Post,k}^+ \cup P_{Post,k}^-$  holds.

An action  $a = \langle \langle O, \emptyset, \emptyset, \dots, \emptyset, \emptyset \rangle, \langle O, \emptyset, \emptyset, \dots, \emptyset, \emptyset \rangle \rangle$  is an empty action and will be also referred to as a WAIT action.

**Definition 10. Change of the world under an action proceeding.** Let a world profile  $WP(t_i) = \langle O, P_1(t_i), P_2(t_i), \dots, P_K(t_i) \rangle$  and an action with description  $Prec = \langle O, P_{Pre,1}^+, P_{Pre,1}^-, \dots, P_{Pre,K}^+, P_{Pre,K}^- \rangle$ ,  $Post = \langle O, P_{Post,1}^+, P_{Post,1}^-, \dots, P_{Post,K}^+, P_{Post,K}^- \rangle$  be given. If for every  $k=1, \dots, K$  both inclusion  $P_{Pre,k}^+ \subseteq P_k(t_i)$  and equality  $P_{Pre,k}^- \cap P_k(t_i) = \emptyset$  hold, then the world will evolve into a state perceived next as  $WP(t_{i+1}) = \langle O, P_1(t_{i+1}), P_2(t_{i+1}), \dots, P_K(t_{i+1}) \rangle$ , where  $P_k(t_{i+1}) = (P_k(t_i) \setminus P_{Pre,k}^+) \cup P_{Post,k}^+$  for  $k=1, \dots, K$ .

### 4.2 Action's Preconditions Verification

An agent's knowledge about effects of undertaken actions corresponds to the real world, which is perceived only partially and results of these observations (at the current moment) are stored in a current profile. An agent, in order to check if it is able to undertake an action  $a$ , compares  $Prec(a)$  with current profile  $EncProfile(t_L)$ .

**Definition 11. Similarity function.** Let a  $ProfileBase(t_L) = \{EncProfile(t_l): l=1, \dots, L\}$  and preconditions  $Prec(a)$  be given. A similarity function  $sim$  is given as follows:

$$sim(Prec(a), EncProfile(t_L)) = \prod_{k=1}^K \{ [1 - sign(card((P_{Pre,k}^+ \cap P_k^-(t)) \cup (P_{Pre,k}^- \cap P_k^+(t))))] \cdot unc^+(k) \cdot unc^-(k) \},$$

where:

$$unc^+(k) = \prod_{o \in (P_{Pre,k}^+ \setminus P_k^+(t))} \frac{1 + card\{EncProfile(t_l): 1 \leq l \leq L, o \in P_k^+(t_l)\}}{1 + card\{EncProfile(t_l): 1 \leq l \leq L, o \in P_k^+(t_l) \cup P_k^-(t_l)\}},$$

$$unc^-(k) = \prod_{o \in (P_{Pre,k}^- \setminus P_k^-(t))} \frac{1 + card\{EncProfile(t_l): 1 \leq l \leq L, o \in P_k^-(t_l)\}}{1 + card\{EncProfile(t_l): 1 \leq l \leq L, o \in P_k^+(t_l) \cup P_k^-(t_l)\}}.$$

If an agent recognized properties of objects noted in  $Prec(a)$  and for every  $k=1, \dots, K$  both inclusions  $P_{Pre,k}^+ \supseteq P_k^+(t)$ ,  $P_{Pre,k}^- \supseteq P_k^-(t)$  hold, the comparison function gives a result of 1.

If during a comparison of  $Prec(a)$  and  $EncProfile(t_L)$  an agent observed such property  $k$ , that at least one of intersections  $P_{Pre,k}^+ \cap P_k^-(t)$ ,  $P_{Pre,k}^- \cap P_k^+(t)$  was nonempty, the comparison function gives a result of 0.

When neither contradiction is found, nor both inclusions mentioned above hold, the comparison function gives result between 0 and 1, according to an agent's knowledge about world  $W$ . For every object and every property from requirements  $Prec(a)$ , if this object wasn't perceived neither exhibiting nor non exhibiting this property, a knowledge about this object and this property is extracted from  $ProfileBase(t_L)$ .

### 4.3 Plan Preconditions and Evaluation

The most important problem during a choice of a plan is an agent's vague knowledge. Due to this knowledge, an agent has to compute a success rate for every plan.

**Definition 12. Plan. Preconditions of a Plan.** Let a set of actions  $A = \{a_j: j=1, \dots, J\}$  and actions' descriptions  $\langle Prec_j, Post_j \rangle$  be given. We define a plan as follows:

- For every action  $a \in A$ ,  $(a)$  is a plan.
- If  $(a_1, a_2, \dots, a_s)$  is a plan and  $a_{s+1} \in A$ , then  $(a_1, a_2, \dots, a_{s+1})$  is a plan if and only if  $Post(a_1, a_2, \dots, a_s)$  and  $Prec_{s+1}$  are consistent.

where:

- For an action  $a \in A$  described by  $\langle Prec_a, Post_a \rangle$ ,  $Prec(a) = Prec_a$ ,  $Post(a) = Post_a$ .
- $Prec(a_1, a_2, \dots, a_{s+1}) = Prec(a_1, a_2, \dots, a_s) \cup [Prec_{s+1} \setminus Post(a_1, a_2, \dots, a_s)]$ .
- $Post(a_1, a_2, \dots, a_{s+1}) = Post_{s+1} \cup [Post(a_1, a_2, \dots, a_s) \setminus Prec_{s+1}]$ .

A SubWorldProfile  $Prec(a_1, a_2, \dots, a_s)$  is called preconditions of a plan  $(a_1, a_2, \dots, a_s)$ .

An agent grades every plan  $(a_1, a_2, \dots, a_s)$ , such that for every  $k=1, \dots, K$  equalities  $O_k^+ \cap P_k^-(t) = \emptyset$  and  $O_k^- \cap P_k^+(t) = \emptyset$  hold, in order to find, which one has a highest possibility of success due to the agent's perception of the world encapsulated in a  $BaseProfile(t)$  and a total knowledge encapsulated in a  $ProfileBase(t)$ .



**Definition 13. Success Rate of a Plan.** Let a *ProfileBase*( $t$ ), a plan  $(a_1, a_2, \dots, a_S)$  and a set of actions  $A$  be given. A Success rate of a plan  $(a_1, a_2, \dots, a_S)$  is given as follows:

$$succ((a_1, a_2, \dots, a_S), Enc\ Profile(t_L)) = sim(Pr ec(a_1, a_2, \dots, a_S), Enc\ Profile(t_L)).$$

When a success rate of a plan is equal to 1, the plan is considered as trustworthy plan. With a decrease of the success rate, the plan becomes less trustworthy, and with a success rate of 0 plan becomes completely useless in agent's point of view (due to it's knowledge based on observations of the world). A deeper explanation on the contents of this paragraph (including interpretations and a complete introduction of definitions 12-13) can be found in [6].

## 5 Conclusions and Final Remarks

A base architecture for a BDI-based action-planning agent was given. An environment's description was formulated, an agent's way of perceiving the environment was and an action planning module and a preference choice module were described using objects from proposed *Profile Space*  $\mathcal{P}$ . A direct description of action preconditions and postconditions was presented and therefore a process of an agent's interaction with world was described.

Basics given in this paper will be used in future work. Among the problems for future research there are topics like: encapsulated knowledge extraction, time-scheme identification and recognition (corresponding to a way in which an agent extracts knowledge stored in a *ProfileBase*( $t$ )), effective action planning (basic concepts for an evaluation of a plan was given, still an action planning module needs an action planning algorithm or heuristic in order to reduce the number of plans for the evaluation process).

## References

1. BYLANDER T., *The computational complexity of propositional STRIPS planning*, Artificial Intelligence, 1994, vol. 69, 165-204.
2. COHEN P.R., LEVESQUE H.J., *Intention is choice with commitment*, Artificial Intelligence., 1990, vol.42, 213-269.
3. KATARZYNIAK R., *The language grounding problem and its relation to the internal structure of cognitive agents*, Journal of Universal Computer Science, 2005, vol. 11, no. 2, 357-374.
4. KATARZYNIAK R., NGUYEN T.N., *Reconciling inconsistent profiles of agent's knowledge states in distributed multiagent systems using consensus methods*. Systems Science, 2000, vol. 26, no 4, pp. 93-119.
5. LINDERN B. van, HOEK W. van der, MEYER J.-J. CH., *Formalizing abilities and opportunities of Agents*, Fundamenta Informatica, 1998, vol.34, 53-101.
6. POPEK G., KATARZYNIAK R., *An approach to reducing uncertainty of knowledge in planning actions of BDI agents*, Tch. Rep. Series SPR, Institute of Information Science and Engineering, Wrocław University of Technology, Wrocław 2006.
7. SHOHAM Y., *Agent-oriented programming*, Artificial Intelligence, Vol.60 (1993), pp.51-92.
8. SINGH M., *Multiagent systems: A theoretical framework for intentions, know-how, and communications*, Springer-Verlag, Heilderberg 1994.

# An Approach to Resolving Semantic Conflicts of Temporally-Vague Observations in Artificial Cognitive Agent

Wojciech Lorkiewicz

Institute of Information Science and Engineering  
Wrocław University of Technology  
50-370 Wrocław, Janiszewskiego 11/17, Poland  
wojciech.lorkiewicz@student.pwr.wroc.pl

**Abstract.** A population of artificial cognitive agents is studied, where agents gather their individual knowledge in strictly private knowledge bases. In this paper we present an approach which integrates the autonomously obtained temporally-vague observations forming the global system perspective of the observed reality. This global system view is then extended in order to reach highest possible consistency. An effective and simple procedure for extending consensus obtained observations is proposed and discussed.

## 1 Introduction

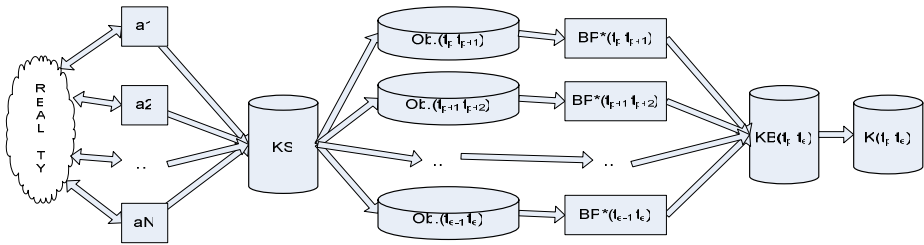
In this paper, we assume a population of artificial cognitive agents, cognitive in the sense of Denett (or Newell) where cognitive systems are understood as intentional systems [1,2,3,4]. Agent can be perceived as an encapsulated system, situated in some environment that is capable of flexible and autonomous actions in that environment in order to meet its design objectives. The agent is autonomous as regards to internal organization and behavioral abilities therefore its external and language behavior can be described with mental and anthropomorphic concepts. In consequence each formula formulated by such an entity can be perceived as assigned a certain meaning [5]. We assume that the agent is equipped with sensors, which allow him to examine the environment, and may build experience knowledge about the world it observes.

Further, we discuss a multiagent system similar to presented in [6], where it is assumed that each agent's perception of the world is related to a particular time point and is represented by a relational system. Agents collect their individual empirically verified knowledge, which is then referred to knowledge gathered by other autonomous agents. Using consensus methods, integrated view is determined, and this view represents global system knowledge. In this paper, we extend the above idea by introducing time-fuzzy base profiles referring not to a certain time point but to a certain time interval from system global time line. Such an extension seems to be a more universal model for a broader range of real time situations, because not all observations are discrete and related to a certain time point.

The integration of individual observations carried by autonomous agents allows agents to relate their experience knowledge to the global system perspective. Benefits of using such cognitive experience of cooperating agents are well known and rather unquestionable. It allows for example to share one agent experience with other agents.

We introduce the cycle (see Fig.1) of resolving semantic conflicts of temporally-vague observations in an artificial cognitive multiagent system:

1. All the time-fuzzy base profiles coming from autonomous agents in a certain global system time window must be collected and stored,
2. Next gathered time-fuzzy base profiles must be divided into smaller collection about which a consensus must be agreed upon,
3. For every such smaller collection a consensus time-fuzzy base profile must be determined and then collected and stored,
4. In the last step all the determined consensus time-fuzzy base profiles must be complemented in order to form global system knowledge about the observed reality.



**Fig. 1.** The cycle of the proposed module

As proposed in [7] and [8] we may perceive the introduced system as a corporate information system equipped with the knowledge integration and knowledge extraction module. Empirically gathered experience is integrated and then extracted what allows to obtain global corporate perspective and not just individual agents views. A specialized database (data warehouse) is introduced to integrate all experience of the individual agents (see Fig.1). Gathered experience may then be the global system information and can be used as a basis for system – end user communication.

Below, a class of multiagent systems is described, in which agents compute their empirically verified knowledge, resolve conflicts and reject mutual knowledge inconsistencies by determining consensus. It is assumed that agents represent their knowledge in a language of relational systems.

## 2 Model of External World and Observation

### 2.1 Observation, Internal Knowledge Base

We assume a simple dynamic system of objectively existing atom objects as an external world of artificial cognitive agents. The environment consists of finite, static and invariant in time set of objects. The dynamicity is introduced in the area of object properties. Every atom object in a certain time point or a certain time interval from the global environment time line can be attributed with a property  $P_i$ , may exhibit it or not, from the set of all properties  $\mathbf{P}=\{P_1, P_2, \dots, P_M\}$ . It should be underlined here that the states of these objects are a target of agent cognition.

It should be stated that world profiles do not have to be time disjoint (the time intervals to which they refer do not have to be disjoint). Nevertheless, when they are not disjoint they cannot be contradictory. Such a contradiction is understood as an existence of an atom object  $o$  such that in one world profile object  $o$  exhibits the property  $P_i$  and in the other it does not. We must assume that the states of external world cannot conflict with our common sense interpretation.

In this paper, the notion of fuzzy world profile, fuzzy base profile and fuzzy knowledge state will be introduced, based on traditional notions of world profile, base profile and knowledge state introduced in [9] and [5].

## 2.2 Observation, Internal Knowledge Base

The external states of the world are the subject of observations carried out by artificial cognitive agents. Therefore, we assume that every autonomous agent is able to perceive the current state of property  $P_i$  from the set of all properties  $\mathbf{P}$  in an atom object  $o$  from the set of all objects  $\mathbf{O}$ . Each observation must be related to a certain time interval  $[t_1, t_2]$  and is stored in an internal, strictly private, agent knowledge base. Such a set of all observations provides the structure and perceivable states of external world. We further assume that every observation concerning the objects from the set  $\mathbf{O}$  in time interval  $[t_1, t_2]$  related to properties from the set of all properties  $\mathbf{P}$  is a relation system.

The idea of a base profile concerning a certain time point has already been introduced elsewhere [9], but here we extend this notion to time interval perspective. In a given time interval  $[t_1, t_2]$  atom object may or may not be attributed with a certain property  $P_i$  from the set of all properties  $\mathbf{P}$  by agent  $a$  from the global set of all agents  $A$ . In consequence at each time interval  $[t_1, t_2]$  the state of external world perceived by agent  $a$  is given by the internal time-fuzzy base profile.

**Definition 1.** By the agents  $a$  time-fuzzy base profile we understand a relation system:

$$BP^a([t_1, t_2]) = \langle \mathbf{O}, P_1^+([t_1, t_2]), P_1^-([t_1, t_2]), P_1^\pm([t_1, t_2]), \dots, P_M^+([t_1, t_2]), P_M^-([t_1, t_2]), P_M^\pm([t_1, t_2]) \rangle \quad (1)$$

where  $\mathbf{O} = \{o_1, o_2, \dots, o_K\}$  is the set of all atom objects and  $\mathbf{P} = \{P_1, P_2, \dots, P_M\}$  is the set of all observable properties that can be attributed to objects from  $\mathbf{O}$ , and where below conditions hold:

- C1.  $\forall i \in 1, 2, \dots, M \quad P_i^+([t_1, t_2]) \cap P_i^-([t_1, t_2]) = \emptyset,$
- C2.  $\forall i \in 1, 2, \dots, M \quad P_i^\pm([t_1, t_2]) = \mathbf{O} / P_i^+([t_1, t_2]) \cup P_i^-([t_1, t_2]),$
- C3.  $\forall i \in 1, 2, \dots, M \forall o \in \mathbf{O} \quad o \in P_i^+([t_1, t_2])$  iff agent  $a$  in time interval  $[t_1, t_2]$  observed object  $o$  exhibits the property  $P_i,$
- C4.  $\forall i \in 1, 2, \dots, M \forall o \in \mathbf{O} \quad o \in P_i^-([t_1, t_2])$  iff agent  $a$  in time interval  $[t_1, t_2]$  observed object  $o$  not exhibits the property  $P_i.$

We will further, in the integration process, need the notion of consistent base profiles introduced in work [9]. Consistent base profiles can be treated as two non contradictory observations of reality concerning the set of objects **O** and set of properties **P**.

**Definition 2.** We treat base profiles  $BP([t_1, t_2])$ ,  $BP([t_3, t_4])$  as consistent if and only if:

$$\forall i \in 1, 2, \dots, M \quad P_i^+([t_1, t_2]) \cap P_i^-([t_3, t_4]) = \emptyset \wedge P_i^+([t_3, t_4]) \cap P_i^-([t_1, t_2]) = \emptyset. \quad (2)$$

Further the set of all observations concerning time interval  $[t_b, t_e]$  gathered by an agent  $a$  will be called the agents  $a$  knowledge time-fuzzy state of  $[t_b, t_e]$ . By analogy the set of different agents knowledge time-fuzzy states of  $[t_b, t_e]$  will be called system knowledge time-fuzzy state of  $[t_b, t_e]$ . On the figure (Fig. 1) the system knowledge state is represented by the *KS* database.

### 3 Integration of Internal Time-Fuzzy Base Profiles

It must be stressed that the internal time-fuzzy base profiles, individual observations, coming from multiple sources in order to form global system knowledge must be integrated. Because the incoming observations might be incomplete or even mutually contradictory therefore a consensus must be reached upon them. In order to perform such integration we introduce an internal base profiles integration procedure (see Fig. 1). In order to incorporate consensus method we will create, from the system knowledge state, a set of time disjoint system knowledge states. The disjoint time intervals will be calculated by a simple recursive procedure which we will call time decomposition.

**Definition 3.** By a time decomposition of time interval  $[t_b, t_e]$  with a given system knowledge state we will call such a division of  $[t_b, t_e]$  where following conditions hold:

1) The set **T** is defined as follows:

$$T = \{t_i : \exists a \in A, t_j \in [t_b, t_e] BP^a([t_j, t_i]) \in KS([t_b, t_e]) \vee BP^a([t_i, t_j]) \in KS([t_b, t_e])\}$$

$$2) [t_b, t_e] = \bigcup_{i=1}^{card(\mathbf{T})-1} [t_i, t_{i+1}], \quad 3) \forall i \neq j \quad [t_i, t_{i+1}] \cap [t_j, t_{j+1}] = \emptyset,$$

Time decomposition procedure can be described by the below three basic steps:

Step 1. For every base profile  $BP^a([t_k, t_l]) \in KS([t_b, t_e])$  we add  $t_k$  and  $t_l$  to the set **T**.

Step 2. For  $i=1, 2, 3, \dots, card(\mathbf{T})$  we define  $t^d_i = \min(T / \bigcup_{j=1}^{i-1} t^d_j)$ .

Step 3. For  $i=1, 2, \dots, card(\mathbf{T})-1$  we form the time interval  $[t^d_i, t^d_{i+1}]$ .

Further time intervals  $[t^d_i, t^d_{i+1}]$ ,  $i=1, 2, \dots, card(\mathbf{T})-1$  obtained by the time decomposition procedure will be called decomposed time intervals.

Based on decomposed time intervals the process of integration of individual knowledge states can be decomposed into several integration tasks, similarly to idea presented in [6]. We may notice that every obtained decomposed time interval defines

a set of internal time-fuzzy base profiles which contain information about a given time interval (see Fig.1). Therefore for every decomposed time interval we define the observation set consisting of individual agent observation about analyzed time interval.

**Definition 4.** As an observation set of time interval  $[t_i, t_{i+1}]$  we call:

$$Ob([t_i, t_{i+1}]) = \{BP^a([t_i, t_{i+1}]) : \exists a \in A \exists BP^a([t_l, t_k]) \quad [t_i, t_{i+1}] \subseteq [t_l, t_k]\}. \quad (3)$$

We may now define the subtask of integration of internal time-fuzzy base profiles coming from different autonomous agents as follows (see Fig.2 for example):

**Definition 5.** Having the observation set of decomposed time interval  $[t_i, t_{i+1}]$  by a consensus time-fuzzy base profile we will call a base profile:

$$BP^*([t_i, t_{i+1}]) = \langle O, P_1^{*+}([t_i, t_{i+1}]), P_1^{*-}([t_i, t_{i+1}]), P_1^{*\pm}([t_i, t_{i+1}]), \dots, P_M^{*+}([t_i, t_{i+1}]), P_M^{*-}([t_i, t_{i+1}]), P_M^{*\pm}([t_i, t_{i+1}]) \rangle. \quad (4)$$

Which satisfies all the below common sense postulates introduced and presented in [10]:

**Postulate PS1** (unanimity of knowledge). If object  $o$  exhibits (doesn't exhibit) the property  $P_i$  in each base profile  $BP([t_i, t_{i+1}]) \in Ob([t_i, t_{i+1}])$ , then in a consensus profile  $BP^*([t_i, t_{i+1}])$  exhibits (doesn't exhibit) the property  $P_i$ .

**Postulate PS2** (superior of knowledge). If object  $o$  exhibits (doesn't exhibit) the property  $P_i$  in at least one base profile  $BP([t_i, t_{i+1}]) \in Ob([t_i, t_{i+1}])$  and there doesn't exist base profile  $BP'([t_i, t_{i+1}]) \in Ob([t_i, t_{i+1}])$ , in which the object  $o$  doesn't exhibit (exhibits) the property  $P_i$ , then in a consensus profile  $BP^*([t_i, t_{i+1}])$  the object  $o$  exhibits (doesn't exhibit) the property  $P_i$ .

**Postulate PS3** (the rule of majority). If the number of base profiles  $BP([t_i, t_{i+1}]) \in Ob([t_i, t_{i+1}])$  in which the object  $o$  exhibits (doesn't exhibit) the property  $P_i$  is bigger then the number of base profiles  $BP'([t_i, t_{i+1}]) \in Ob([t_i, t_{i+1}])$ , in which the objects  $o$  doesn't exhibit (exhibits) the property  $P_i$ , then in a consensus profile  $BP^*([t_i, t_{i+1}])$  the object  $o$  exhibits (doesn't exhibit) the property  $P_i$ .

**Postulate PS4** (the rule of majority of knowledge). If the number of base profiles  $BP([t_i, t_{i+1}]) \in Ob([t_i, t_{i+1}])$  in which the object  $o$  exhibits (doesn't exhibit) the property  $P_i$  is bigger then the number of base profiles  $BP'([t_i, t_{i+1}]) \in Ob([t_i, t_{i+1}])$ , in which the objects  $o$  doesn't exhibit (exhibit) the property  $P_i$  and the state of an object  $o$  is unknown, then in a consensus profile  $BP^*([t_i, t_{i+1}])$  the object  $o$  exhibits (doesn't exhibit) the property  $P_i$ .

**Postulate PS5** (restrictive rule of majority of knowledge). If the number of base profiles  $BP([t_i, t_{i+1}]) \in Ob([t_i, t_{i+1}])$  in which the object  $o$  exhibits (doesn't exhibit) the property  $P_i$  is bigger then the sum of the number of base profiles  $BP([t_i, t_{i+1}]) \in Ob([t_i, t_{i+1}])$ , in which the objects  $o$  doesn't exhibit (exhibits) the property  $P_i$  and the base profiles  $BP([t_i, t_{i+1}]) \in Ob([t_i, t_{i+1}])$ , in which the state of an objects  $o$  is unknown and the number

of base profiles  $BP([t_i, t_{i+1}]) \in \text{Ob}([t_i, t_{i+1}])$ , in which the state of an object  $o$  is unknown, then in a consensus profile  $BP^*([t_i, t_{i+1}])$  the object  $o$  exhibits (doesn't exhibit) the property  $P_i$ .

Effective algorithm for reaching consensus for time point base profiles satisfying postulates PS1 to PS5 is presented and discussed in [10].

### 4 Extending Consensus Time-Fuzzy Base Profiles

The last step of presented procedure is intended to gather obtained consensus time-fuzzy base profiles (see Fig.1 KB) and extend them forming the global system knowledge base (see Fig.1 K). The intention of such an extension procedure is to join time-fuzzy neighbor profiles which are consistent (see Chapter 2). Since consistent base profiles represent coherent observations of reality we assume that the sum of such profiles represent the state of the reality that held for the time interval equal to the sum of time intervals of summed profiles.

**Definition 6.** By the sum of time-fuzzy base profiles  $BP([t_1, t_2])$  ,  $BP([t_3, t_4])$  we understand a relational system:

$$\begin{aligned}
 &BP([t_1, t_2]) \bar{\cup} BP([t_3, t_4]) = \\
 &< P_1^+([t_1, t_2]) \cup P_1^+([t_3, t_4]), P_1^-([t_1, t_2]) \cup P_1^-([t_3, t_4]), P_1^\pm, \dots, \\
 &P_M^+([t_1, t_2]) \cup P_M^+([t_3, t_4]), P_M^-([t_1, t_2]) \cup P_M^-([t_3, t_4]), P_M^\pm >
 \end{aligned}
 \tag{5}$$

Where  $P_i^\pm = \mathbf{O} / (P_i^+([t_1, t_2]) \cup P_i^+([t_3, t_4]) \cup P_i^-([t_1, t_2]) \cup P_i^-([t_3, t_4]))$  and  $t_2 \in [t_3, t_4]$ .

It should be stressed that the above definition of the sum of time-fuzzy base profiles does not guarantee that the resulting relation system is a well defined time-fuzzy base profile, it may not satisfy the condition C1 (see Definition 1). Theorem 1 describes this situation more precisely.

**Theorem 1.** The sum of time-fuzzy base profiles  $BP([t_1, t_2]) \bar{\cup} BP([t_3, t_4])$  is a well defined time-fuzzy, satisfies the conditions C1 to C4 (see Definition 1), iff the base profiles  $BP([t_1, t_2])$  and  $BP([t_3, t_4])$  are consistent.

The proof of introduced theorem is presented in details in [6], therefore we will omit it here. The above theorem leads to a simple conclusion.

**Conclusion 1.** If the sum of time-fuzzy base profiles  $\bar{\bigcup}_{i \in I} BP([t_i, t_{i+1}])$ ,  $I = \{1, 2, \dots, S-1\}$

is a well defined time-fuzzy base profile, satisfies the conditions C1 to C4 (see Definition 1) and we will add a base profile  $BP([t_0, t_1])$  then:

a) if  $\bar{\bigcup}_{i \in I} PB([t_i, t_{i+1}]) \bar{\cup} BP([t_0, t_1])$  satisfies conditions C1 to C4 it means that

every pair of time-fuzzy base profile  $BP([t_i, t_{i+1}])$ ,  $i \in I = \{0, 1, 2, \dots, S-1\}$  are consistent.

b) if  $\bigcup_{i \in I} PB([t_i, t_{i+1}]) \overline{\cap} BP([t_0, t_1])$  does not satisfy conditions C1 to C4 it means that

there exists a pair of base profiles  $BP([t_k, t_{k+1}])$  and  $BP([t_b, t_{b+1}])$  which are inconsistent.

The procedure of extending consensus time-fuzzy base profiles:

- Step 1. We generate the sequence  $A$  by sorting the set of observations  $OK$  in an ascending order due to bottom boundary of time interval about which the observations hold :  $A = \langle BP^*([t_b, t_{b+1}]), BP^*([t_{b+1}, t_{b+2}]), \dots, BP^*([t_{e-1}, t_e]) \rangle$ .
- Step 2. We take and erase the first element  $BP_R = BP^*([t_1, t_2])$  from the sequence  $A$  and add to the set  $T_R$  time interval  $[t_1, t_2]$ .
- Step 3. We check whether the sequence  $A$  is empty, stop condition. If the stop condition holds we add to the set of extended time-fuzzy base profiles  $K$  the  $BP_R$  base profile and we present the set  $K$  as a result. Otherwise we continue the procedure in step 4,
- Step 4. We take and erase the first element  $BP_R = BP^*([t_b, t_{b+1}])$  from the sequence  $A$ ,
- Step 5. We calculate  $W = BP_R \overline{\cap} PB([t_i, t_{i+1}])$ ,
- Step 6. We determine whether  $W$  satisfies conditions C1 to C4 (see Definition 1) If it holds we assign  $BP_R = W$ , add to the set  $T_R$  time interval  $[t_i, t_{i+1}]$  and continue to step 3. Otherwise we continue to step 7,
- Step 7. We add to the resulting set  $K$  base profile  $BP_R$ ,
- Step 8. We assign  $BP_R = BP^*([t_b, t_{b+1}])$  and  $T_R = [t_b, t_{b+1}]$ . We continue to step 3.

Analogous problem for time point based consensus base profiles is introduced in [6], where a simple procedure is presented and discussed.

As a result we obtain a set of extended consensus time-fuzzy base profiles which in a best way integrate the individual agents' observations. Such a system with extension module, as mentioned earlier, allows the end user to communicate with the system treated not as an individual agent but as a global multiagent structure. By the term communicate we understand the grounding mechanism described in [5] which can be easily integrated in presented environment.

## 5 Final Remarks

In this paper, the notions of time-fuzzy base profiles are introduced as an extension of classical base profile architecture. Such an extension allows to model wider range of real world situations. The integration and extension cycle of individual agent's internal knowledge in a cognitive multiagent system was proposed and discussed. Such a cycle may be seen as a part of a more complex global system. Knowledge integrated and extended in a proposed manner may also be the basis for grounding mechanisms. Further research in this area may involve the introduction of so called trust for every individual agent representing our degree of priorities of knowledge coming from this source. Such a value could be used as an additional factor for the determination of consensus time-fuzzy base profile.



## References

1. Dennet D.C., *Kinds of minds*, Basic Books, New York, 1996.
2. Cohen P.R., Levesque H.J., *Intention is Choice with Commitment*, *Artificial Intelligence*, Vol. 42, pp. 213-261, 1990
3. Wooldridge M., Müller J.P., Jennings N.R., (eds.), *Intelligent agents III, Agent, Theories, Architectures and Languages*, Proceedings of IJCAI'96, Budapest, Hungary, August 12-13, Springer, 1997
4. Newell A., *The unified theories of cognition*, Harvard University Press, Cambridge 1990.
5. Katarzyniak R., *The language grounding problem and its relation to the internal structure of cognitive agents*. *Journal of Universal Computer Science*, vol. 11, no. 2, pp. 357-374.
6. Lorkiewicz W., Katarzyniak R.P., *Resolving semantic conflicts of observations with crisp and vague timestamps*, Tech. Report SPR, Wrocław univ. of Technology, Inst. of Infor. Science & Engineering, Wrocław 2006 (in print).
7. Katarzyniak R.P., *Organizations as communicative cognitive agents*, Proceedings of the 7th International Workshop on Organizational Semiotics OS'04, Setubal, Portugal, July 2004, 309-324.
8. Katarzyniak R.P., Pieczyńska-Kuchtiak A., *Strategia wyboru komunikatu modalnego w dialogowym systemie wspomaganie decyzji (A Strategy for the choice of modal response In an interactive decision suport system)*, *Badania Operacyjne i Decyzje 1 (Operations Reasearch and Decisions)*, 2004, 21-35 (In polish).
9. Katarzyniak R.P., Nguyen N.T., *Reconciling inconsistent profiles of agent's knowledge states in distributed multiagent systems using consensus methods*, *Systems Science* 26(4), 2000, 93-119.
10. Pieczyńska-Kuchtiak A., *Experienced-based learning of semantic messages generation in resource-bounded environment*, *Systems Science* 30(3), 2004, 115-132

# Neural Network Approach for Learning of the World Structure by Cognitive Agents

Agnieszka Pieczyńska and Jarosław Drapała

Institute of Information Science and Engineering, Wrocław University of Technology,  
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland  
{agnieszka.pieczynska, jaroslaw.drapala}@pwr.wroc.pl

**Abstract.** In this work an original method for coping with agents' incomplete knowledge is introduced. This method called the algorithm for the messages generation is applied by the cognitive agents when the states of external objects can not be directly perceived. To approximate the current states of objects all agent's experience as temporal data base is taken into account. As a result of the algorithm the logic formulas with modal operators are generated. One of the steps of proposed algorithm is the classification of the observations. It is shown how neural network approach might be used in order to determined some tendencies to occurrence specific states of objects.

## 1 Introduction

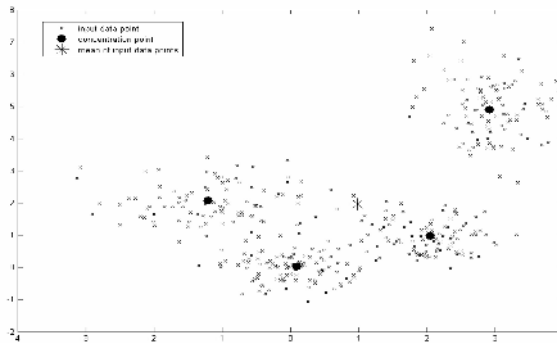
Distributed tasks solving in multiagent systems requires among other things communication between the members of population to exchange their opinions, integrate partial solutions or answer to the questions of resource bounded agents. It is assumed that cognitive agents during the process of communication use the language that let them express the opinions about the states of external objects in the form of logic formulas with modal operators. The content communicated by each message relates to the state of a particular object from external world in relation to two properties. Due to the acting in a dynamic environment and some difficulties with cognitive capacities it is possible that the current states of some objects from external world might be for an agent not known. In such situations there are two possible solutions considered: asking others from a population about their opinions or applying some internal mechanisms for coping with incomplete knowledge. In the first case received answers might be not identical and the agent must be equipped with some strategies for the integration of other agents attitudes in order to achieve opinions' agreement [12, 16]. However in the second case an agent approximates the current states of the objects taking as an input its overall experience stored in a private temporal database and applying the algorithm for the messages generation. An algorithm for the messages generation widely discussed in works [8, 15] relates formulas to the internal agents' knowledge states and reflects the process of symbol grounding [5]<sup>1</sup>. Each formula is treated by the cognitive agent as true, if and only if

---

<sup>1</sup> The symbol grounding problem deals with the question: How symbols can be used meaningfully [2, 10, 17, 18].

this formula is satisfied by the overall state of agent-encapsulated knowledge. Such an approach to understanding satisfaction relation is alternative to the commonly known extensional definition of the satisfaction formulas accepted within Tarskian theory of truth. In this sense the algorithm realizes an original way of semantic and interpreted language generation [6, 7, 11].

One of the steps of this algorithm is the classification of the observations. Then for each class of observation one representative called consensus profile is obtained. The current state of the object is assumed as in the consensus profile that is closest to the current base profile. In other words in the algorithm for the messages generation the agent tries to find some tendencies to occurrence some states of objects and if at the current time point the state of particular object is not known and must be approximated then the states of other objects from particular base profile are the context for generating external opinion. Median profile – a mean of input data points that is a concentration point with complex structure is determined for each class of perception. But it might be possible that inside of one class there are more than only one concentration point i.e. consensus profile that represents some subset of the base profiles (Fig.1).

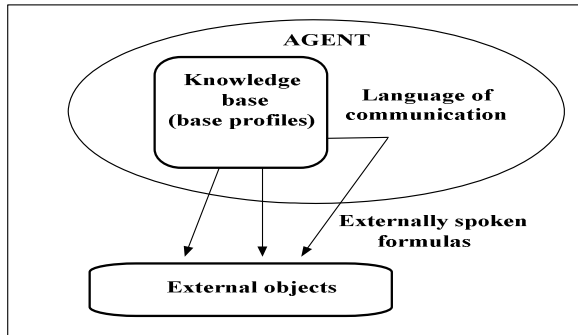


**Fig. 1.** An example of a class of perception consisted of the set of base profiles (small dots) and four concentration points

Let us note that even though a median profile (mean of input data) is almost equally distant from each subset of base profiles (input data points) it is not really the best representative because some information about some tendencies to occurrence well grounded states of objects are lost. In this connection to approximate the current states of objects for each class of perception all concentration points must be taken into account and as many distance values should be computed as many concentration points exist. Then the approximated state of object should be the same as in concentration point closest to the current base profile. Referring to these remarks a new proposal of the algorithm for the messages generation based on neural network approach for data classification and finding of the concentration points is introduced. A paper is organized as follows: in section 2 the general agent's structure is given. Section 3 presents the idea of the neural network based algorithm for the messages generation. In section 4 details of concentration points computing are presented. Finally in section 5 concluding remarks are given.

## 2 Internal Agent's Structure

It is assumed that the agent is independent entity equipped with sensors and situated in some world. This world is consisted of the atomic objects  $O = \{o_1, o_2, \dots, o_s\}$ . Each object  $o_s \in O$  at the particular time point  $t_k \in T$ ,  $T = \{t_0, t_1, t_2, \dots\}$  possesses or doesn't posses the property  $P_i \in P$ ,  $P = \{P_1, P_2, \dots, P_w\}$ . The states of the objects from  $O$  are a target of agent's cognitive processes [8]. The overall agent's knowledge is encapsulated in its body as a temporal database consisted of the set of observations. Each observation is related to the states of objects' at the particular time point. Each formula is consisted of the modal operator and two indices to the properties that are observed by the agent  $a \in A$  in the object  $o_s \in O$  at the time point  $t_k \in T$  or are estimated on the basis of agent's previous experiences applying the algorithm for the messages generation. The co-existence of these properties in objects from the set  $O$  is reflected by logic connectives such as: conjunction, exclusive alternative or alternative. For  $OP \in \{Bel, Pos, Know\}$  the agent's language is given as:  $\Omega = \{OP^a(P_i(o_s) \wedge P_j(o_s)), OP^a(P_i(o_s) \wedge \neg P_j(o_s)), OP^a(\neg P_i(o_s) \wedge P_j(o_s)), OP^a(\neg P_i(o_s) \wedge \neg P_j(o_s)), Bel^a(P_i(o_s) \vee P_j(o_s)), Bel^a(P_i(o_s) \vee P_j(o_s))\}$ . For example sending by the agent a message  $Bel^a(P_i(o_s) \wedge P_j(o_s))$  is equivalent spoken language: "I agent a believe that object  $o_s$  possesses both the property  $P_i$  and  $P_j$ ."



**Fig. 2.** Agent's knowledge base, externally spoken formulas and the external objects represent semiotic triangle that reflects the concept of formulas' epistemic satisfaction [17]

Modal operators express the level of agent's certainty with reference to the communicated content and logic connectives the level of unambiguosity of the states of objects. If an agent sends a message with conjunction as a logic connective then the state of an object  $o_s$  is unequivocally specified. In case of exclusive alternative as a connective the state of an object  $o_s$  is not unequivocally specified. Then two states of affairs are considered: 1) an object  $o_s$  possesses the property  $P_i$  and doesn't posses the property  $P_j$  either 2) object  $o_s$  doesn't posses the property  $P_i$  and possesses the property  $P_j$ . Finally if an agent sends a message with alternative as a logic connective then three equally possible states of affairs are considered: 1) an object  $o_s$  possesses the property  $P_i$  and possesses the property  $P_j$  either 2) an object  $o_s$  possesses the property

$P_i$  and doesn't possess the property  $P_j$  either 3) object  $o_s$  doesn't possess the property  $P_i$  and possesses the property  $P_j$ .

**Definition 1.** The overall agent's experience is represented by the temporal database consisted of the set of base profiles.

$$KS(t_c) = \{V(t_k): t_k, t_c \in T \text{ and } t_k \leq t_c\},$$

where  $V(t_k)$  is the vector base profile defined as:

$$V(t_k) = \langle v_{11}(t_k), v_{12}(t_k), \dots, v_{1W}(t_k), v_{21}(t_k), v_{22}(t_k), \dots, v_{2W}(t_k) \dots v_{SW}(t_k) \rangle$$

and  $v_{sj}(t_k)$  denotes the state of an object  $o_s$  in relation to the property  $P_j \in P$  at the time point  $t_k$  observable by the agent. Following interpretation is assumed:  $v_{sj}(t_k)=1$  if an object  $o_s$  possess the property  $P_j$ ;  $v_{sj}(t_k)=2$  if an object  $o_s$  doesn't possess the property  $P_j$  and  $v_{sj}(t_k)=3$  if an agent cannot observe the state of an objects  $o_s$  at the time point  $t_k$  in relation to the property  $P_j$ .

### 3 The Idea of the Algorithm for the Messages Generation

In works [8, 9, 15] the algorithm for the messages generation has been proposed. This algorithm represents a strategy for an approximation the states of objects in resource bounded environment. In this work the algorithm for the messages generation is developed and in order to discover some tendencies to occurrence a specific states of objects neural network approach is introduced. Generally the idea of this algorithm is as follows: at the current time point  $t_c$  the agent can not observe the state of a certain object  $o_s$  in relation to the properties  $P_i$  and  $P_j$ .

**Step 1.** The agent groups all the base profiles into four classes, the criterion of the objects' membership to the particular class  $G^m(t_c)$ ,  $m=1,2,3,4$  is the state of an object  $o_s$  in relation to the properties  $P_i$  and  $P_j$ :

- a)  $G^1(t_c) = \{V_g(t_k): t_k \leq t_c, V_g(t_k) \in KS(t_c), v_{si}(t_k), v_{sj}(t_k) \in V_g(t_k), v_{si}(t_k)=1 \text{ and } v_{sj}(t_k)=1\}$
- b)  $G^2(t_c) = \{V_g(t_k): t_k \leq t_c, V_g(t_k) \in KS(t_c), v_{si}(t_k), v_{sj}(t_k) \in V_g(t_k), v_{si}(t_k)=1 \text{ and } v_{sj}(t_k)=2\}$
- c)  $G^3(t_c) = \{V_g(t_k): t_k \leq t_c, V_g(t_k) \in KS(t_c), v_{si}(t_k), v_{sj}(t_k) \in V_g(t_k), v_{si}(t_k)=2 \text{ and } v_{sj}(t_k)=1\}$
- d)  $G^4(t_c) = \{V_g(t_k): t_k \leq t_c, V_g(t_k) \in KS(t_c), v_{si}(t_k), v_{sj}(t_k) \in V_g(t_k), v_{si}(t_k)=2 \text{ and } v_{sj}(t_k)=2\}$

where  $g \in \{1, 2, \dots, G\}$  with each class  $G^m(t_c)$ ,  $m=1,2,3,4$  modal language statements are correlated:  $OP^a(p_i(o_s) \wedge p_j(o_s))$  with  $G^1(t_c)$ ,  $OP^a(p_i(o_s) \wedge \neg p_j(o_s))$  with  $G^2(t_c)$ , with  $G^3(t_c)$ ,  $OP^a(\neg p_i(o_s) \wedge p_j(o_s))$  with  $G^4(t_c)$ ,  $OP^a(\neg p_i(o_s) \wedge \neg p_j(o_s))$ ,  $OP^a(p_i(o_s) \vee p_j(o_s))$  with  $G^2(t_c)$  and  $G^3(t_c)$ ,  $OP^a(p_i(o_s) \vee \neg p_j(o_s))$  with  $G^1(t_c)$ ,  $G^2(t_c)$  and  $G^3(t_c)$ , where  $OP \in \{Bel, Pos\}$ .

**Step 2.** For each class  $G^m(t_c)$ ,  $m=1,2,3,4$  the agent determines the set of concentration points:

$C^m(t_c) = \{C_n(t_c): \text{for each } G^m, n \in \{1, 2, \dots, N\}\}$ ,  $N$  denotes the number of concentration points in  $C^m(t_c)$ . As a result of the step 2 the set  $C = \{C^m(t_c): m=1,2,3,4\}$  of all concentration points is obtained.

**Step 3.** For each  $C_n(t_c) \in C^m(t_c)$  the agents computes a distance value  $d(C_n(t_c), V_g(t_c))$  that reflects the similarity between each concentration point  $C_n(t_c)$  and current base profile  $V_g(t_c)$ . The set  $D = \{d(C_n(t_c), V_g(t_c)) : C_n(t_c) \in C, C^m_n(t_c), V_g(t_c) \in KS(t_c)\}$  of all distance values is computed. The distance between a concentration point and a current base profiles can be understood by means of the costs of transformation one object into another [14]. In work [9] such method is presented. Also various coefficients might be applied such as: Dice's index, Jaccard's index, overlap or cosine measures [4].

**Step 4.** The agent chooses the concentration point  $C_{min}$  for which the distance value  $d(C_n(t_c), V_g(t_c))$ ,  $C_n(t_c) \in C^m(t_c)$ ,  $m \in \{1, 2, 3, 4\}$  is minimal. The approximated state of an object  $o_s$  in relation to the properties  $P_i$  and  $P_j$  is the same as in  $C_{min}$  and the external message is with modal operator of belief. Let us note that in this case all messages are with logic conjunction as a connective. Step 4 might be developed and another variant is to reject the concentration points for which a distance value  $d(C_n(t_c), V_g(t_c))$  is higher than threshold value  $\tau$ . As a result the agents determines the set  $C^D$  of the close sets  $C^m_d(t_c) \in C^m$ ,  $m \in \{1, 2, 3, 4\}$  of concentration points:  $C^D = \{C^m_d(t_c) : C^m_d(t_c) \in C^m(t_c), \text{ such that for each } C_n \in C^m_d(t_c) : d(C_n(t_c), V_g(t_c)) < \tau, m \in \{1, 2, 3, 4\}\}$ . All the concentration points in  $C^D$  are treated as equally close to the current base profile  $V_g(t_c)$ .

**Step 5.** The agent computes the set  $H = \{\alpha^m : \alpha^m, m \in \{1, 2, 3, 4\}$  is the number of close concentration points in each  $C^m_d(t_c) \in C^D$ .  $\alpha^m$  reflects the support strength for a state of object correlated with each class of observation  $G^m(t_c)$ ,  $m \in \{1, 2, 3, 4\}$ .

**Step 6.** The agent applies a decision procedure [15]. In this case it is necessary to verify if:

- The support coefficients (if belong to the set H):  $\alpha^2$  and  $\alpha^3$  or  $(\alpha^1, \alpha^2$  and  $\alpha^3)$  are close to each other then the message with exclusive alternative (or alternative) and Bel operator is generated.
- All the values from H (or some subset of H) are close to each other and it is difficult unambiguously to determine the state of an object  $o_s$ . Then all the messages correlated with support coefficients are generated with Pos operator and logic conjunction as a connective.
- There exists one coefficient value  $\alpha^m$  for which:  $\alpha^m - \alpha^l > \delta$ , for each  $\alpha^l \in H$ ,  $l, m \in \{1, 2, 3, 4\}$ ,  $m \neq l$  then the message with logic conjunction as a connective and belief operator Bel is generated.

## 4 Application of Self-organized Neural Networks

In step 2 of the algorithm for the messages generation for each class of observation  $G^m(t_c)$ ,  $m=1, 2, 3, 4$  the set  $C^m(t_c)$  of concentration points is determined. In order to solve this problem neural network is applied. Self-organized neural networks learning algorithms are a specific class of numeric algorithms, widely used in data sets clusterization [3, 13]. It is assumed that neural network and input data consist of points of the same type conventionally called neurons. In this work each point is represented by the base (vector) profile (see definition 1) and neural network is used to determine concentration points. Solution is obtained by performing learning algorithm starting from arbitrary chosen concentration points. Initially this set is

mostly far from a correct concentration points, but accuracy of the solution is improved after each iteration of the algorithm. By correct concentration point we understand such base (vector) profile that minimize the distance to all base profiles in particular region of base profiles. If the total number of iterations is sufficient i.e. next iterations don't improve the current result then the algorithm is ended.

General scheme of self-organized neural network learning algorithm proposed in this work is as follows:

**Input:** The class of base profiles  $G^m(t_c)$ . For more readability time references and indexes are omitted:  $G^m = \{V_1, V_2, \dots, V_G\}$  where for  $V_g \in G^m$ ,  $V_g = \langle v_g^1, v_g^2, \dots, v_g^R \rangle$ .

**Output:** For each class of observations the set of concentration points:  $C^m = \{C_1, C_2, \dots, C_N\}$ ,  $C^m \subset G^m$  where for  $C_n \in C^m$ ,  $C_n = \langle v_n^1, v_n^2, \dots, v_n^R \rangle$

**Step 1.** Set iterations number  $I$ , threshold  $\gamma \in [0, 1]$  and concentration points number  $N$  i.e. number of neurons.

**Step 2.** Chose arbitrarily initial solution  $C^m(0) = \{C_1(0), C_2(0), \dots, C_N(0)\}$ , set the values of parameter  $\eta$ .

**Step 3.** Transform initial solution by performing  $I$  iterations described in steps 4 to 7, for  $i=0, 1, \dots, I-1$ . Finally, perform step 8.

**Step 4.** Chose a random number  $g$  from the range  $[1, G]$ .

**Step 5.** Calculate distances  $d_n$  of vector profile  $V_g$  from each  $C_n(i)$ :

$$d_n = \sum_{r=1}^R (c_n^r(i) - v_g^r)^2, \quad n = 1, 2, \dots, N \tag{1}$$

**Step 6.** Chose the minimal value  $d_b(C_b(i), V_g)$  of all distance values  $d_n(C_n(i), V_g)$ ,  $C_n(i) \in C^m(i)$ ,  $b \neq n$ . For  $C_b(i)$  set  $h(n)=1$  else  $\forall_{n \in \{1, 2, \dots, N\}/b} h(n)=0$ .

**Step 7.** Replace  $C_n(i)$  by  $C_n(i+1)$  according to following formula:

$$C_n(i+1) = C_n(i) + \eta(i)h(n)[V_g - C_n(i)], \quad n = 1, 2, \dots, N \tag{2}$$

where  $\eta(i) \in [0, 1]$  is correction coefficient at the  $i$ -th iteration and  $h(n)$  is so called "neighborhood function".

**Step 8.** Replace values of all vectors  $C_n(I)$ ,  $n=1, 2, \dots, N$  by values calculated from equation  $|c_n^T + 1 - \gamma|$ .

Presented algorithm is called WTA (*Winner Takes it All*) [1]. In each  $i$ -th step only one vector  $w_b(i)$  for which distance  $d_b$  is smallest, increments values of its elements. Correction coefficient  $\eta(i)$  decreases exponentially as learning process progresses:

$$\eta(i) = \eta(0) \left( \frac{\eta(I)}{\eta(0)} \right)^{i/I} \tag{3}$$

Initial  $\eta(0)$  and final  $\eta(I)$  values of parameter  $\eta$  are chosen at step 2 of general self-organized neural network learning scheme.

## 5 Conclusion

The aim of this paper was to investigate the problem of grounding modal formulas generated individually by the agents after applying dedicated algorithm for the messages generation. This algorithm is used to approximate the current states of objects. In order to determine some tendencies to occurrence specific states of objects neural network is used and so called concentration points are determined. A point is understood as a agent's experienced knowledge about the states of objects related to the particular time point. For further research it is necessary to develop a decision procedure-last step of the algorithm taking into consideration the number of concentration points in each class of observation, their distances to the current base profiles and also their role as representative understood by means of the number of neighbouring base profiles.

## References

1. Barreto, G.,A., Identification and Control of Dynamical Systems using the Self-Organizing Map. IEEE Trans. Neural Networks, Vol.~15, September (2004) 1244-1259
2. Coradeschi S., Saffiotti A., An Introduction to the Anchoring Problem, Robotics and Autonomous Systems 43, (2003), 85-96.
3. Daszykowski, M., Walczak, W., Massart, D.L., On the Optimal Partitioning of Data with K-Means, Growing K-Means, Neural Gas, and Growing Neural Gas. J. Chem. Inf. Comput. Sci. (2002), 42, 1378-1389
4. Egghe, L., Michel, C., "Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques", *Information Processing and Management*, vol. 39, pp.771-807, 2003.
5. Harnad, S.:The Symbol Grounding Problem. *Physica*, 42, 335-236
6. Katarzyniak, R., Pieczynska-Kuchtiak, A.: Formal Modelling of the Semantics for Communication Languages in Systems of Believable Agents. In: Proc. of ISAT'2001, Szklarska Poreba, (2001), 174-181
7. Katarzyniak, R., Pieczynska-Kuchtiak, A.: Intentional Semantics for Logic Disjunctions, Alternatives and Cognitive agent's Belief. In: Proc. of the 14th International Conference on System Science, Wroclaw, Poland, (2001), 370-382
8. Katarzyniak, R., Pieczynska-Kuchtiak, A.: A Consensus Based Algorithm for Grounding Belief formulas in Internally Stored Perceptions. *Neural Network World*, 5, (2002) 671-682
9. Katarzyniak, R., Pieczynska-Kuchtiak, A.: Distance Measure Between Cognitive Agent's Stored Perceptions. In: Proc. of IASTED International Conference on Modelling, Identification and Control, MIC'2002, Innsbruck, Austria (2002) 517-522
10. Katarzyniak, R., Pieczynska-Kuchtiak, A.: Grounding Languages in Cognitive Agents and Robots. In: Proc. of Sixteenth International Conference on System Engineering, Coventry (2003) 332-337
11. Katarzyniak, R., Pieczynska-Kuchtiak, A.: Grounding and extracting modal responses in cognitive agents: AND query and states of incomplete knowledge. *International Journal of Applied Mathematics and Computer Science*, 14(2), (2004), 249-263.
12. Katarzyniak, R., Pieczynska-Kuchtiak, A.: An Approach to resolving semantic inconsistency of multiple prepositional attitudes. *Journal of Intelligent & Fuzzy Systems* 17(3), (2006), (to appear).



13. Ossowski, S., Neural Networks – algorithmic approach, (in polish), WNT Warsaw, (1996)
14. Nguyen, N.T.: Consensus System for Solving Conflicts in Distributed Systems. Information Sciences, 147, (2002) 91-122
15. Pieczyńska-Kuchtiak, A.: A Decision Function in the Algorithm for the Choice of Semantic Messages. In: Proc. of Information Systems Architecture and Technology, Poland (2002)
16. Pieczyńska-Kuchtiak, A., Towards measure of semantic correlation between messages in multiagent system, ICCS 2005, Lecture Notes on Computer Science, Kraków, (2004) pp. 567-574.
17. Vogt P., Anchoring of semiotics symbols, Robotics and Autonomous Systems, 43, (2003) 109-120

# Towards a Computational Framework for Modeling Semantic Interactions in Large Multiagent Communities

Krzysztof Juszczyszyn

Institute of Control and Systems Engineering  
Wroclaw University of Technology, Wroclaw, Poland  
krzysztof@pwr.wroc.pl

**Abstract.** The aim of this paper is to propose a framework for modeling semantic interactions in large multiagent communities. By semantic interaction we understand the act of modifying the internal knowledge representation (in the form of semantic network or ontology) as a result of communication between agents. The proposed framework takes into account the latest results investigating structure of large semantic networks and communication patters in diverse multiagent environments.

## 1 Introduction

In recent years we experience fast changes which push us to the limits of contemporary information and communication technologies. Rapid growth of infrastructure, knowledge sources and interoperability requirements creates a global net-centric architecture for information storing, delivery and processing. Web technologies provide a basis for the interchange of knowledge and services in such an environment and support the emergence of large-scale network structures: WWW itself, thematic information networks, virtual communities, market systems.

Development of dynamic intelligent services across the above systems is inevitably connected with so-called *semantic technologies* – functional capabilities that enable *both* humans and machines to create, discover, organize, share and process the meanings and knowledge [6]. This is achieved by the use of shared vocabularies or semantic nets. On the level of WWW this implies the adoption of the Semantic Web's XML-based standards for annotating and processing information, usually in the form of web ontologies [3]. Because ontologies are developed and managed independently the semantic mismatches between two or more ontologies are inevitable. Practical applications show that fully shared vocabularies are rather exceptional - a number of possible different semantic conflicts was identified by Shaw and Gaines [11], other classifications were addressed in [8]. The vision of Semantic Web allowing agents to publish and exchange ontologies requires strong mechanisms supporting ontology merging and alignment [7].

It should be stressed that any formal model aiming to investigate properties of the above systems has to take into account at least two factors:

- the architecture of the system (the pattern of connections and interactions between actors – humans and/or agents that constitute the system),
- the structure of the agent's vocabularies (ontologies) and their role in the abovementioned interactions.

As shown in the next sections there are many results investigating the structure of the contemporary networks and there are also some recent results which evaluate the large-scale structure of semantic nets and ontologies. But there are no works which deal with joining these issues under a common umbrella. The aim of this paper is to propose a framework for modeling *semantic interactions* in large multiagent communities. By semantic interaction we will understand the act of modifying the internal knowledge representation (in the form of semantic network or ontology) as a result of communication between agents. The rationale for such an environment is to investigate the conditions underlying the emergence of common vocabulary in the agents' community and the dynamics of the system.

In order to be compatible with the latest results the framework itself must integrate and formally represent the following components:

- Community structure (the architecture of links between actors).
- Community dynamics (the mechanism of formation of new links between actors).
- Communication model (the rules of choosing the communicating party).
- Semantic interaction model (the rules for establishing communication link and/or modification of internal knowledge representation as a result of the communication)

A framework for modelling semantic interactions in multiagent communities with respect to above issues will be presented in the following sections.

## 2 Structure and Formation of Small Worlds and Semantic Nets

Complex network structures emerge in many everyday situations among people (social networks), organizations, software agents, linked documents (WWW) and so on. Previous research has identified the most distinctive properties of such networks [14], [4]:

- Small diameter and average path length (of the order of  $\text{Log}(N)$  for  $N$  network nodes).
- High clustering (probability that the neighbors of any given node will be also each other's neighbors)
- A famous power-law (or scale-free) network node degree distribution.

These properties may be of use when simulating interaction between system components and building evolution models, and they form a basis of many robust and applicable theoretical results. It was shown that they influence the search strategies, communication and cooperation models, knowledge and innovation spreading etc. [12], [14].

Moreover, last results show that we may expect similar phenomena on the level of knowledge representation: in [12] a large-scale structure of the semantic networks of

three types was evaluated. It was shown that all the three (free word association network, Roget's thesaurus and WordNet lexical database) appear to be of small-world structure which (as a graph) may be characterized by sparse connectivity, short average paths and high node clustering – just like in the case of abovementioned networks. The Authors state [12]: *We argue that there are in fact compelling general principles governing the structure of network representations for natural language semantics, and that these structural principles have potentially significant implications for the processes of semantic growth and memory search.* In fact, from now on any computational model should take these results into account. A model of growing semantic network was also proposed but it was based on concept differentiation scheme and didn't assume multiple actors and interactions between them.

From the other hand the process of acquiring new concepts (concept learning) via communication was investigated in many other works, for example in [9] a mathematical model was used to simulate the emergence of coherent dictionary in a population of independent subjects (agents). The Authors proposed a well-known mechanism of language imitation as a self-organization factor. However, in these experiments a random communication and interaction strategy was assumed, which is not straightforward in real multiagent and social environments. As stated in the preceding section neither connection nor communication pattern between the agents are random. They show the properties of the scale-free net.

Now the challenge is to span a bridge between known properties of dynamic self-organizing agent societies and the semantic structures emerging within them. The growing and evolving semantic structures should be modified in result of interaction between agents constituting community of particular architecture.

### 3 The Semantic Interaction Framework

The architecture of the proposed framework consists of a set of agents (interpreted as software components as the framework is intended to investigate the Semantic Web environments), which of them is equipped with private vocabulary in the form of semantic net (or ontology). These structures may mutually overlap but there will also be inevitable difference. These vocabularies are suggested to be large-scale structures, which are then modified in the process of communication between the agents (Fig.1).

The activity of the system components is based on the following assumptions:

1. The architecture of connections (communication links) between agents conforms to the small world model (as shown in [14]). This assumption follows the results invoked in the preceding section – we may expect that spontaneously formed interaction networks will follow small world properties.
2. The individual semantic nets (ontologies) of the agents show scale-free properties as proved in [12].
3. The communication model is preferential (agents tend to connect to the hubs first, due to connectivity and information content).

- Semantic interactions between the agents follow the the *imitation model* [10]. This approach was investigated and documented by many researchers and serves as explanatory mechanism for the emergence of common vocabulary in communities of humans and social animals. It assumes that the agents change their private vocabularies on the basis of interactions with the others. In particular they may accommodate the meaning of the concepts used by their counterparts if they found it reasonable. In most simulations it is assumed that the imitation process is random (see [9]). In our approach an imitation based on the structure of semantic nets will be proposed.

Let's now list the parameters needed in our framework. Let  $A = \{A_1, A_2, \dots, A_n\}$  be a set of agents and  $O = \{O_1, O_2, \dots, O_n\}$  the set of their private ontologies. Each agent  $A_i$  uses ontology  $O_i$  as a formal conceptualization of particular domain of interest. We denote the set of concepts of ontology  $O_i$  as  $C_i = \{c_1^i, c_2^i, \dots, c_{m(i)}^i\}$  and the relations between them as  $R_i$ . Each agent  $A_i$  has also an associated utility function  $u_i : O \rightarrow [0,1]$  which is to express the potential attractiveness of the other agents as communication counterparts of  $A_i$ . In the simplest  $A_i$  may define a list of important concepts and  $u_i$  will return the result based only on the set comparison between set of concepts of given ontology and the list of topics  $A_i$  is interested in. This means that  $A_i$  is willing to communicate with the agents which have knowledge on the specific topics.

The network structures of the system are represented by communication graph represented by  $n \times n$  matrix  $G$  where an entry  $g_{ij}$  indicates the presence of directed link from the node (agent)  $A_i$  to  $A_j$ , and the graphs reflecting the structure of private agents' ontologies. Each of these ontologies  $O_i$  may be viewed as a graph  $SemNet_i$  with nodes corresponding to concepts from  $C_i$  and edges corresponding to relations from  $R_i$ . According to [12] we assume that these graphs show small world properties.

The following algorithm (consisting of the four steps executed repeatedly) comprises the idea of simulating semantic interactions in evolving multiagent small world environment:

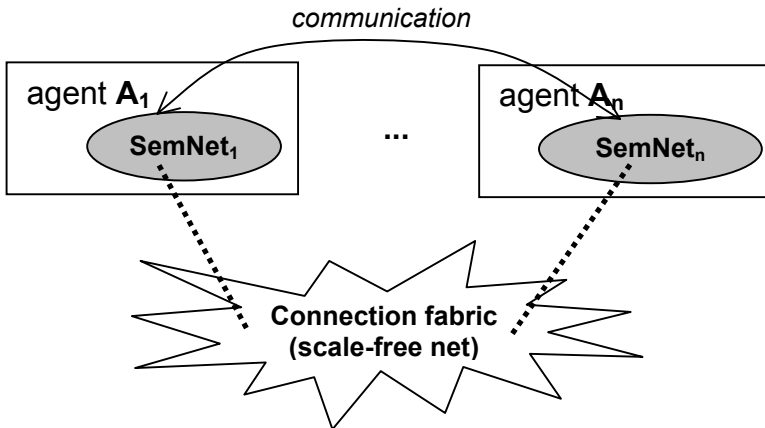


Fig. 1. The architecture of the multiagent system

Step 1:*Creating a new agent.*

When a new  $A_k$  agent is added to the system he must be connected to the others. According to results from [14] a newly-born agent is connected to its neighbors on the basis of preferential search strategy, which means that:

1. The agent chooses randomly a number of  $p$  nodes (called his *parents*, the value of  $p$  is a simulation parameter), then connects, statistically preferring those with the highest utility function  $u_k$  value. The  $G$  matrix is updated.
2. The agent then searches the parents' neighborhood in communication graph (of arbitrary diameter  $d$ ) finds more candidates to establish connection with, and connects to those with maximum utility value.

The above behavior results in formation of a classic scale-free pattern met in many artificial and social environments. We also assume that a new agent is equipped with his own semantic network (ontology)  $O_k$ . It is to be used in step 3.

Step 2:*The act of communication.*

In each discrete time moment  $t$  a random agent starts communication formulating a query composed of the concepts from its ontology. The query is passed to a counterpart chosen from between his neighbors in communication graph with probability proportional to the value of utility function. To infer about the equivalence of the any given pair of the concepts used both agents use the *semantic similarity function* (solutions discussed in [1] are good examples of such functions). At this point the two scenarios may occur:

1. If the value of similarity function exceed some threshold Step 3 of the algorithm (non random concept imitation) takes place.
2. If the value of similarity function is below threshold there are two possibilities: to start ontology negotiation process (like that proposed in [2]) and clarify the meaning of the concepts or announce communication failure. Note that it doesn't mean that communication between these two agents is not possible in the future – the evolution of the system may offer such possibility after a series of changes in agents' ontologies.

Step 3:*Performing a non-random imitation.*

In contrary to [9] the imitation is not random but takes place according to the following:

1. Imitation is possible if and only if the semantic similarity function used by the agents returns (for given pair of concepts from their ontologies  $O_i$  and  $O_j$ ) a threshold-exceeding value (the threshold itself is an arbitrary-valued parameter, depending of the similarity function used). This is interpreted as using two concepts for representing the same meaning and makes a condition of starting imitation process.

Imitation assumes that the better-connected (having greater node degree in its *SemNet* or: being part of greater number of relations) concept is more attractive for the use by the agents. Thus, a better-connected (as a node in its ontology graph, in terms of node degree and clustering coefficient) concept is being imitated by the agent who is the owner of less-connected concept.

**Step 4:***Evaluating the system properties.*

There is a large number of global properties which may be tracked when running simulation of such a system. They may be then used to investigate the evolution of multiagent system and to infer about possibility of reaching the semantic continuum [5],[13] (the state in which any of the two agents may seamlessly communicate).

The most important will be:

- *Semantic consistency for given concept*: the percentage of the agents' pairs which share that concept compared to all connected agent pairs.
- Standard agent network properties: *node degree distribution, clustering, average path length.*
- The overall *number of concepts* (it is probable that some will be eliminated during system evolution)
- The number of *competing concepts* which are concurrently used while having the same meaning.

In order to perform simulations of the above framework we must define the set of agents and their corresponding utility functions, the net of concepts (ontology) for any particular agent (this may be done on abstract level for obtaining quantitative results of simulation), the behavior of the concept similarity function and the static parameters (like thresholds). This will be done as the next stage of research.

## 4 Conclusions and Future Work

Mechanisms that govern evolution of emergent semantic structures in modern web-based multiagent environments are relatively new and not widely addressed research task. Its successful completion has potential to influence novel interconnection architectures (like Semantic Web and Semantic Grids) in many ways. The most interesting are:

- Creating knowledge and innovation spreading models.
- Developing intelligent search algorithms.
- Formulating the conditions for semantic integrity of distributed systems.
- Support for knowledge-based virtual organizations.

The further development of the proposed framework includes conducting experiments and investigating discovering the rules that govern evolution and behavior of the emerging Semantic Web environment and its underlying semantic network structures.

## References

1. Andrea, M., Egenhofer, M.: Determining Semantic Similarity Among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering* **15** (2003) 442–456.
2. Bailin, S.C. Truszkowski, W. *Ontology Negotiation between Agents Supporting Intelligent Information Management*, 5th International Conference on Autonomous Agents, Montreal, Canada, 2001, 13-20.

3. Berners Lee T., Hendler J., Lasilla O., The Semantic Web. Scientific American, May 2001
4. Carrington P., Scott J., Wasserman S. (eds.) Models and Methods in Social Networks Analysis, Cambridge University Press, 2005.
5. Cybenko G., Jiang G., Hendler J., Semantic Depth and Markup Complexity, IEEE Conference on System, Man and Cybernetics, 2003, 2138-2143.
6. Davis M., Semantic Wave 2006 - Part 1: Executive Guide to Billion Dollar Markets, Project10X Special Report, Washington, 2006.
7. Hendler, J.: Agents and the Semantic Web. IEEE Intelligent Systems **16(2)** (2001) 30-37
8. Hameed, A. et al.: Detecting Mismatches among Experts' Ontologies Acquired through Knowledge Elicitation. In: Proceedings of 21th International Conference on Knowledge Based Systems and Applied Artificial Intelligence ES2001, Cambridge, UK (2001) 9-24
9. Ke, J., Minett, J. W., Au, C-P., and Wang, W. S-Y. (2002) Self-organization and Selection in the Emergence of Vocabulary. Complexity, **7(3)**, 41-54.
10. Meltzoff, A. N., Prinz, W. The imitative mind: Development, evolution, and brain bases. Cambridge, England: Cambridge University Press, 2002.
11. Shaw, M.L.G., Gaines, B.R.: Comparing Conceptual Structures: Consensus, Conflict, Correspondence and Contrast. Knowledge Acquisition, **1(4)** (1989) 341-363
12. Steyvers M., Tannenbaum J., The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth, Cognitive Science, **29(1)**, (2005).
13. Uschold M., Gruninger M., Creating Semantically Integrated Communities on the World Wide Web, Invited Talk, Semantic Web Workshop, WWW Conference, Honolulu, USA, 2002.
14. Watts D, Strogatz S., Collective Dynamics of Small World Networks, Nature, **393**, 440-442.



# Grounding Crisp and Fuzzy Ontological Concepts in Artificial Cognitive Agents

Radosław Piotr Katarzyniak

Institute of Information Science and Engineering, Wrocław University of Technology,  
Wybrzeże Wyspiańskiego 27, 50-350 Wrocław  
radoslaw.katarzyniak@pwr.wroc.pl

**Abstract.** The grounding of crisp and fuzzy concepts is defined for the class of artificial cognitive agents equipped with means for recognizing states of external objects and language signs. Each concept is created in the internal cognitive space of the agent in the context of social communication. This communication makes it possible to correlate cognitive agent's perceptions with language signs generated by other members of the same population of agents. Related processes yield the so called semiotic relation consisting of recognized instances of semiotic triangles. A simple measure between concepts is suggested and used to define higher level relations of semantic synonymy, similarity, generalization and contradiction. In consequence an approach to modeling ontology creation for a particular class of agents and a particular language is defined.

## 1 Introduction

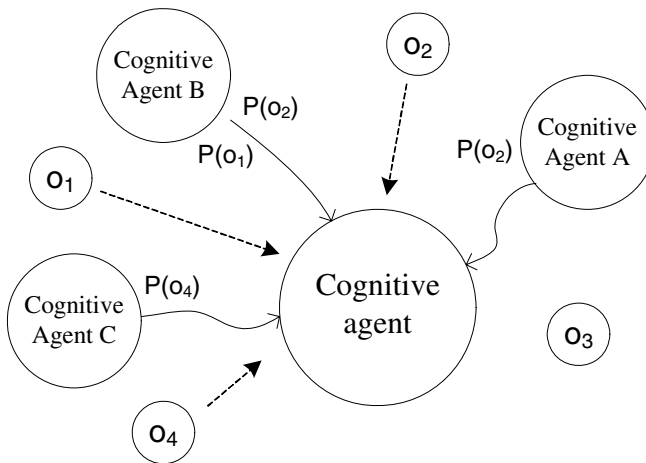
In the field of information systems and artificial intelligence ontologies are understood as formal specifications of domain conceptualizations. One of the main features of the majority of these definitions is that they do not refer ontologies to their underlying idea of knowledge subjects [2] and grounding [4]. The relation between ontology, which is a specification of knowledge, and any agent of this knowledge is fundamental due to the fact that conceptualizations are always created by particular cognitive minds [2].

In this paper an approach to modelling crisp and fuzzy conceptualizations of external world is discussed for a case of artificial cognitive agents introduced in other works [6]. This way ontologies that capture these concepts are grounded in actual empirical experiences of agents. All considered concepts are labelled by assigned language signs which are called language symbols due to their reference to particular meaning [1], [7], [8]. All theoretical notions introduced in this paper are given for an example case of a world, an artificial cognitive agent, and a simple language for communication. The higher ontological relations are defined on the basis of lower level notions of fuzzy and crisp concepts, and the additional idea of conceptual similarity.

The approach to grounding lower level concepts and higher ontological relations is similar to the process of social semiosis suggested in [7], [8] and applied to a class

of simple “agents” equipped with dedicated data structures for storing external worlds’ reflections [7], [8]. In that approach the authors assumed that language grounding would be a complex process carried out within an existing artificial population. Main steps proposed and assumptions made for the process of semiosis given in [7], [8] are:

- Artificial cognitive agents are assumed to be able to conceptualize externally originated data streams into internal mental representations. These data are collected through dedicated sensors which analyze incoming material and distribute it over dedicated knowledge representation structures. In natural cognitive agents this process is carried out on the basis of biological neural networks [3].
- All internal representations that are stored by cognitive agents induce higher level models of external world, provided that their strength is grounded by multiple instantiation of the same recognized shapes and order of observed objects.
- The above mentioned process of collecting reflections of external world is correlated with additional processing of language recognition. Additionally, artificial cognitive agents are assumed to be able to recognize and store correlations between particular instantiations of external worlds’ perceptions and accompanying language signs produced by other agents of the same population.
- Artificial cognitive agents learn the underlying correlation between all stored perceptions and language labels produced by other parties of the discourse and recognized simultaneously to certain observations (classes of observations). The output of such process of semiosis is the creation of internally grounded fundamental meaning of language signs that can be further process in order to create higher level representations.



**Fig. 1.** The general organization of environment for the process of semiosis

Figure 1 shows the general framework of environment within which the above mentioned process of semiosis will be considered below. In this environment an agent is given which collects both the perceptions of four objects in the world and the language signs incoming from other agents of population. This way other agents name the external world with language signs and teach the agent their socially accepted meaning. As the result signs are bounded step by step to particular content and become symbols for the cognitive agent.

## 2 Model of Semiotic Triangle

Let us introduce formal representations for particular elements of the environment described above. Let the cognitive agent be able to recognize five dimensions (properties)  $W_1, W_2, W_3, W_4$  and  $W_5$  of four objects  $o_1, o_2, o_3$  and  $o_4$  constituting the external world.

The language of signs used and recognized by all members of the population is given as the set  $L = L_1 \cup L_2 \cup L_3$ , where  $L_k = \{P_k(o_1), P_k(o_2), P_k(o_3), P_k(o_4)\}$ ,  $k=1,2,3$ . As it can be seen formulas of this language consists of subparts that can be treated as references to particular objects of the world and some aspects of these objects. For instance, in the language sign ' $P_2(o_1)$ ' the sub-string ' $o_1$ ' is assumed to be interpreted by all agents of the population as the reference to actual object  $o_1$  in the external world and the sign ' $P_2$ ' can be assigned to some of the dimensions  $W_1, W_2, W_3, W_4$ , and  $W_5$ . The first reference between ' $o_1$ ' and actual object  $o_1$  is further assumed as already developed in the cognitive agent. An approach to developing the other reference between ' $P_2$ ' and perception dimensions is the main target of suggested semiosis. The latter means that in the meaning of the complex language signs  $P_m(o_n)$  is assumed to be unknown for the considered cognitive agent and the creation of its meaning is the main target of learning processes.

Each individual observation realized by the artificial cognitive agent at the time point  $t$  is represented by the so called base profile [5], [6]:

$$BP(t) = \langle O, W_1^+(t), W_1^-(t), \dots, W_5^+(t), W_5^-(t) \rangle,$$

where:

- $O = \{o_1, o_2, o_3, o_4\}$  consists of conceptual representations of all objects of the world,
- For each  $i=1,2,\dots,5$ , both  $W_i^+(t) \subseteq O$  and  $W_i^-(t) \subseteq O$  hold.
- For each  $i=1,2,\dots,5$  and  $o \in O$  the relation  $o \in W_i^+(t)$  holds if and only if the agent perceived this object as assigned to the conceptual dimension  $W_i$  and the relation  $o \in W_i^-(t)$  holds if and only if the agent perceived this object as unassigned to the conceptual dimension  $W_i$ .
- For each  $i=1,2,\dots,5$  the condition  $W_i^+(t) \cap W_i^-(t) = \emptyset$  holds.

Each base profile is a formal representation of actual conceptualisation of particular perception developed by cognitive processes of external world recognition.

Let us assume that in case of our artificial cognitive agent particular base profiles can be sometimes developed simultaneously to the process of signs recognition. Obviously, this process is natural when natural agents (e.g. humans) are involved to

the semiotic process of learning the language. Namely, an agent perceives the world, creates related base profiles and at the same time recognizes particular signs of L produced by other agents in order to describe the current situation. This correlation between base profiles and recognized language signs can provide the cognitive agent with the meaning assigned to particular signs by the population involved in the discourse. Therefore each act of communication can be treated and stored as a pair of a base profile for particular perception and some language signs spoken by other members of the same population and recognized by the learning agent. The set of all these pairs is called semiotic relation and is formalized in the next section.

### 3 Semiotic Relation and Symbol Grounding

#### 3.1 Semiotic Relation

Let us assume that correlations between base profiles and their language descriptions given by external members of population is represented in the artificial cognitive agent as the so called semiotic relation:

**Definition 1.** Let the set  $T=\{t_1, t_2, \dots\}$  be the set of time point,  $L=\{P_k(o_j): k=1,2,3 \text{ and } j=1,2,3,4\}$  be the set of language signs used in communication and  $O=\{o_1,o_2,o_3,o_4\}$  be the set of objects' mental representations. Each subset SemiosisResult of the Cartesian product  $T \times 2^L \times 2^O \times 2^O \times 2^O \times 2^O \times 2^O$  stored in the cognitive agent is called the semiotic relation, where  $2^L$  and  $2^O$  are sets of all subsets of the language L and the set O, respectively.

In each tuple  $\langle t, AS(t), W_1(t), W_2(t), W_3(t), W_4(t), W_5(t) \rangle \in \text{SemiosisResult}$ :

- $t$  denotes a particular time point
- $AS(t)$  denotes a set of language signs recognized by the cognitive agents as produced by other members of population in order to point at the state of the world observed at the time point  $t$ ,
- for  $k=1,2,3,4,5$   $W_k(t)=W_k^+(t)$  where  $W_k^+(t)$  is a part of base profile  $BP(t)$ .

**Table 1.** An example of semiotic relation for an individual agent with  $W_k(t)=W_k^+(t), k=1, \dots, 5$

$t$	$AS(t)$	$W_1(t)$	$W_2(t)$	$W_3(t)$	$W_4(t)$	$W_5(t)$
$t_1$	$P_1(o_1), P_1(o_2), P_3(o_3)$	$o_1, o_2, o_4$	$o_1, o_2$	$o_3$	$o_4$	$o_3$
$t_2$	$P_1(o_1), P_1(o_2), P_3(o_1)$	$o_1, o_2$		$o_1, o_3$	$o_4$	$o_1$
$t_3$	$P_3(o_3), P_3(o_4)$	$o_2$	$o_1, o_4$	$o_1, o_2, o_3, o_4$		$o_3, o_4$
$t_4$	$P_3(o_4), P_2(o_2), P_2(o_3)$	$o_2$	$o_1$	$o_1, o_2, o_3$	$o_3$	$o_4$
$t_5$	$P_1(o_1), P_1(o_2)$	$o_1, o_2$	$o_1$	$o_1, o_3$		
$t_6$	$P_1(o_1), P_3(o_3), P_3(o_4)$	$o_1, o_2$	$o_1$	$o_1, o_3$		$o_3, o_4$

An example of the semiotic relation (particular result of semiosis) is given in Table 2. The tuple  $\langle t_2, \{P_1(o_1), P_1(o_2), P_3(o_1)\}, \{o_1, o_2\}, \emptyset, \{o_1, o_3\}, \{o_4\}, \{o_1\} \rangle$  represents the following content:

- At the time point  $t_2$  objects  $o_1$  and  $o_2$  were recognized as exhibiting the property  $W_1$ , no object was recognized as exhibiting the property  $W_2$ , objects  $o_1$  and  $o_3$  were perceived as exhibiting the property  $W_3$ , object  $o_4$  was perceived as exhibiting the property  $W_4$  and object  $o_1$  was perceived as exhibiting the property  $W_5$ .
- At the same time point  $t_2$  this state of the world was described by other members of the population with means of three language signs  $P_1(o_1)$ ,  $P_1(o_2)$  and  $P_3(o_1)$ .

Obviously, it is assumed that in case of the considered cognitive agent the above tuple contributes to its process of learning the meaning of language signs  $P_1(o_1)$ ,  $P_1(o_2)$  and  $P_3(o_1)$ .

### 3.2 Concept Creation and Symbol Grounding

The above notion of semiotic relation can be used to yield certain distribution of particular language signs over the set of assumed conceptual dimensions (properties)  $W = \{W_1, W_2, W_3, W_4, W_5\}$ . Such distribution reflects the content of subjective empirical experiences of the cognitive agent given by all instances of recognized communication of certain concepts for analyzing the external world.

To make it clear let us consider the label  $P_1$ . The related distribution of  $P_1$  over the set  $W$  results from Table 1 and is assumed to be given as a fuzzy set over the domain  $W$  being the output of the following strategy:

**Step 1.** ( $P_1$ -selection)

- For all**  $t = t_1, \dots, t_6$  **do if** the condition  $\forall o \in \{o_1, o_2, o_3, o_4\}. P_1(o) \notin AS(t)$  is fulfilled, **then** reject the tuple  $\langle AS(t), W_1(t), W_2(t), W_3(t), W_4(t), W_5(t) \rangle$  from Table 1.  
 Remove all occurrences of  $P_i(o), i \neq 1$  from  $AS(t)$  in the remaining tuples.

The result of  $P_1$ -selection applied to Table 1 is given in Table 2.

**Table 2.** The result of applying  $P_1$ -selection

$t$	$AS(t)$	$W_1(t)$	$W_2(t)$	$W_3(t)$	$W_4(t)$	$W_5(t)$
$t_1$	$P_1(o_1), P_1(o_2)$	$o_1, o_2, o_4$	$o_1, o_2$	$o_3$	$o_4$	$o_3$
$t_2$	$P_1(o_1), P_1(o_2)$	$o_1, o_2$		$o_1, o_3$	$o_4$	$o_1$
$t_5$	$P_1(o_1), P_1(o_2)$	$o_1, o_2$	$o_1$	$o_1, o_3$		
$t_6$	$P_1(o_1)$	$o_1, o_2$	$o_1$	$o_1, o_3$		$o_3, o_4$

**Step 2.** ( $P_1$ -unlabeled objects rejection)

- For all**  $t = t_1, t_2, t_5, t_6$  and  $o \in \{o_1, o_2, o_3, o_4\}$  **do**  
**if** the condition  $P_1(o) \notin AS(t)$  is fulfilled  
**then** reject all instances of  $o$  from  $W_i(t), i = 1, \dots, 5$ .

The result of  $P_1$ -unlabeled objects rejection for Table 2 is given in Table 3.

**Table 3.** The result of P<sub>1</sub>-unlabeled objects rejection

<i>t</i>	AS( <i>t</i> )	W <sub>1</sub> ( <i>t</i> )	W <sub>2</sub> ( <i>t</i> )	W <sub>3</sub> ( <i>t</i> )	W <sub>4</sub> ( <i>t</i> )	W <sub>5</sub> ( <i>t</i> )
t <sub>1</sub>	P <sub>1</sub> (o <sub>1</sub> ), P <sub>1</sub> (o <sub>2</sub> )	o <sub>1</sub> ,o <sub>2</sub>	o <sub>1</sub> ,o <sub>2</sub>			
t <sub>2</sub>	P <sub>1</sub> (o <sub>1</sub> ), P <sub>1</sub> (o <sub>2</sub> )	o <sub>1</sub> ,o <sub>2</sub>		o <sub>1</sub>		o <sub>1</sub>
t <sub>5</sub>	P <sub>1</sub> (o <sub>1</sub> ), P <sub>1</sub> (o <sub>2</sub> )	o <sub>1</sub> ,o <sub>2</sub>	o <sub>1</sub>	o <sub>1</sub>		
t <sub>6</sub>	P <sub>1</sub> (o <sub>1</sub> )	o <sub>1</sub> ,o <sub>2</sub>	o <sub>1</sub>	o <sub>1</sub>		

**Step 3.** (FP<sub>1</sub>-creation). The fuzzy set FP<sub>1</sub>={ (W<sub>1</sub>,w<sub>1</sub>), (W<sub>2</sub>,w<sub>2</sub>), (W<sub>3</sub>,w<sub>3</sub>), (W<sub>4</sub>,w<sub>4</sub>), (W<sub>5</sub>,w<sub>5</sub>) } is determined provided that:

$$w_k=N_k/N_{max}, k=1,2,\dots,5.$$

$$N_m=card(W_m(t_1))\cup card(W_m(t_2))\cup card(W_m(t_5))\cup card(W_m(t_6)), m=1,2,\dots,5,$$

$$N_{max}=\max\{N_1, N_2, N_3, N_4, N_5\}$$

It follows from the content of Table 3 that N<sub>1</sub>=8, N<sub>2</sub>=4, N<sub>3</sub>=3, N<sub>4</sub>=0, N<sub>5</sub>=1, N<sub>max</sub>=8, and the resulting fuzzy set is FP<sub>1</sub>={ (W<sub>1</sub>,1.000), (W<sub>2</sub>,0.500), (W<sub>3</sub>,0.375), (W<sub>4</sub>,0.000), (W<sub>5</sub>,0.125) }.

This set is treated as representation of empirically verified meaning of the sign P<sub>1</sub>. This meaning is socially accepted by other agents of population and has been communicated to the cognitive agent by these members of population in particular circumstances. In this sense the process of social communication induces a particular mental structure in the mind of the cognitive agent for an ontological concept. This concept is described by the fuzzy set FP<sub>1</sub> which represents a particular result of grounding the language sign P<sub>1</sub>. Obviously this concept is induced by the content of semiotic relation given in Table 1.

### 3.3 Classes of Concepts

The following notions of λ-fuzzy and λ-crisp concepts can be defined depending on the nature of fuzzy sets introduced in section 3.2.

**Definition 2.** For each P<sub>i</sub>, i=1,2,3 and λ∈ [0,1] the ontological concept labelled by P<sub>i</sub> is called λ-crisp if and only if card(FP<sub>i</sub><sup>λ</sup>)=1, where FP<sub>i</sub><sup>λ</sup>={x:(x,α)∈ FP<sub>i</sub>(W) and α>λ}. An ontological concept is called crisp if and only if it is 0-crisp.

**Definition 3.** For each P<sub>i</sub>, i=1,2,3 and λ∈ [0,1] the ontological concept labelled by P<sub>i</sub> is called λ-fuzzy if and only if card(FP<sub>i</sub><sup>λ</sup>)>1, where FP<sub>i</sub><sup>λ</sup>={x:(x,α)∈ FP<sub>i</sub>(W) and α>λ}. Each ontological concept is called fuzzy if and only if it is 0-fuzzy.

In our example the concept denoted by P<sub>1</sub> and represented by the fuzzy set FP<sub>1</sub>={ (W<sub>1</sub>,1.000), (W<sub>2</sub>,0.500), (W<sub>3</sub>,0.375), (W<sub>4</sub>,0.000), (W<sub>5</sub>,0.125) } is 0.000-fuzzy, 0.100-fuzzy, ..., 0.499-fuzzy, and 0.500-crisp. This concept is fuzzy and not crisp.

## 4 Higher Level Ontological Relations

### 4.1 Conceptual Similarity

Let the following approach to measuring the conceptual similarity between ontological concepts be given. Let two language signs  $P_i \neq P_j$  be assigned the fuzzy sets  $FP_i = \{(W_1, x_1), (W_2, x_2), (W_3, x_3), (W_4, x_4), (W_5, x_5)\}$  and  $FP_j = \{(W_1, y_1), (W_2, y_2), (W_3, y_3), (W_4, y_4), (W_5, y_5)\}$ , respectively. The conceptual similarity of these concepts can be described by the following function:

$$\text{sim}(FP_i, FP_j) = 1 - \frac{1}{5} \cdot \sqrt{\sum_{i=1}^5 (x_i - y_i)^2}$$

The notion of conceptual closeness is quite obvious and can be further used to define higher level ontological relations of semantic synonymy, similarity, generalization and contradiction.

### 4.2 Semantic Synonymy

Two different language symbols  $P_i$  and  $P_j$  are synonyms if and only if the conceptual similarity of ontological concepts assigned to these symbols during the process of semiosis is equal to 1. Formally, the requirement

$$\text{sim}(FP_i, FP_j) = 1$$

needs to be fulfilled.

### 4.3 Semantic Similarity

Two different language symbols  $P_i$  and  $P_j$  are semantically similar if and only if they are not synonyms and the conceptual similarity of ontological concepts assigned to these symbols is close to 1. Formally, the requirement

$$\text{sim}(FP_i, FP_j) \approx 1$$

needs to be fulfilled.

### 4.4 Semantic Generalization

The language symbols  $P_i$  is the generalisation of another language symbol  $P_j$  if and only if the meaning assigned  $P_j$  is contained by the meaning of the symbol  $P_i$ . Formally, the requirement

$$FP_j \subset^F FP_i$$

needs to be fulfilled, where the symbol  $\subset^F$  denotes the notion of fuzzy sets inclusion [9].

### 4.5 Semantic Contradiction

Two different language symbols  $P_i$  and  $P_j$  are antonyms if and only if the conceptual similarity of ontological concepts assigned to these symbols during the process of semiosis is equal to 0. Formally, the requirement

$$\text{sim}(FP_i, FP_j) = 0$$

needs to be fulfilled.

## 5 Final Remarks

In this paper the grounding of crisp and fuzzy concepts has been defined for the particular class of artificial cognitive agents. Each concept is assumed to be created in the internal cognitive space of the agent in the context of social communication. This work was partly suggested by the approach to modeling social process of artificial semiosis given in [7], [8] and is complementary to the research on symbol grounding and artificial language generation carried out in other works [6]. The suggested model can be practically applied to considered class of agents in order to automate at least some tasks related to learning the meaning of language and autonomous ontology generation. The notions of semantic synonymy, similarity, generalization and contradiction can be used to develop semantic network of language signs used by considered population of agents located in a particular world.

## References

1. Eco, U.: *La struttura assente*, Tascabili Bompiani, Milano (1991)
2. Fichte, J.G.: *Johann Gottlieb Fichtes sämtliche Werke.*, Veit & Comp., Berlin (1845-1846).
3. Freeman, W.J.: A neurobiological interpretation of semiotics: meaning, representation, and information., *Information Sciences*, Vol.124 (2000), 93-102.
4. Harnad, S.: The symbol grounding problem., *Physica D* 42, 335-346.
5. Katarzyniak, R., Nguyen, T.N.: Reconciling inconsistent profiles of agent's knowledge states in distributed multiagent systems using consensus methods., *Systems Science*, Vol. 26, No. 4 (2000), 93-119.
6. Katarzyniak, R.: The language grounding problem and its relation to the internal structure of cognitive agents., *Journal of Universal Computer Science*, Vol. 11, no. 2 (2005), 357-374.
7. Vogt, P.: Anchoring of semiotic symbols., *Robotics and Autonomous Systems*, Vol. 43, (2003), 109-120.
8. Vogt, P.: The Physical Grounding Symbol Problem., *Cognitive Systems Research*, Vol. 3, (2002), 429-457.
9. Zadeh, L.A. : Fuzzy sets., *Information and Control*, Vol. 8 (1965), 338-353.



# A Design of Hybrid Mobile Multimedia Game Content

Il Seok Ko<sup>1</sup> and Yun Ji Na<sup>2</sup>

<sup>1</sup> School of Computer and Multimedia, Dongguk University, 707 Seokjang-dong, Kyungju, Kyungsangbukdo, South Korea  
isko@dongguk.edu

<sup>2</sup> School of Internet Software, Honam University, 59-1 Seobong-dong, Gwangsan-gu, Gwangju 506-714, South Korea

**Abstract.** This paper proposes a hybrid mobile multimedia game content using a hybrid method. The proposed method can share linked materials in real-time between these methods and presents an effective solution for various limitations caused by the memory capacity of mobile terminals in the planning and production processes of Multi media game content. In addition, this method presents a real performance by applying it to the production of an actual mobile multimedia game.

**Keywords:** WAP method, download scheme, VM method, and actual mobile game.

## 1 Introduction

The MMC serviced in the present time can be classified as a WAP method [1, 2,3] and VM method [4,5,6]. In the case of the WAP method, it presents a limitation in the production of dynamic content even though it presents low criteria in memory capacities. However, a VM method presents the limitation of MMC capacities. Thus, an effective link between these two methods can improve the production quality of MMC. However, the MMC that uses a linked method with WAP  $\rightarrow$  VM and VM  $\rightarrow$  WAP has not been developed yet using the existing method [7,8,9].

It is true that the WAP  $\rightarrow$  VM method has only been used as a type for mobile games that introduce certain rankings due to the problem existing between platforms even though these two methods should be applied as WAP  $\rightarrow$  VM and VM  $\rightarrow$  WAP. In order to link these two methods, a VM  $\rightarrow$  WAP method is required in which data packets can be transmitted after cutting VM platforms in a mobile terminal in order to transmit data packets to VM  $\rightarrow$  WAP. As a result, a VM processor is not to be in an active state. Thus, a VM processor should be assigned as an active state in order to send data, which is generated by a VM  $\rightarrow$  WAP method, to WAP  $\rightarrow$  VM continuously. However, it is impossible to send data packets due to the fact that the VM processor is not assigned as an active state. Thus, the existing MMC transmits data packets that verify rankings by applying a VM  $\rightarrow$  WAP method, which is easy to transmit data, in order to verify the transmission.

This paper presents a scheme that can produce a large capacity of MMC by applying the offset between the limitation of the storage capacity of VM content in a

mobile phone and the weakness of the static section in WAP content using a link between the existing simple method of WAP and VM services and the MMC that applies a partial link between wired/wireless methods.

## 2 Linkage of VM and WAP

### 2.1 System Architecture

The proposed method uses a link method by using application modules and databases for the WAP and VM methods and consists of several modules that take charge of Automatic Application Download, Phone number identification, Data transmission DB server and Application, and DB and Application Linkage.

Each module can be used to a link process presented in Fig. 1. First, it can be applied by downloading the user's applications and MMC after verifying a user's phone number. It can then be transmitted using the WAP method. After transmitting various information used in the present operation to a DB server if changes in the method are required through a link in the data. After transmitting it to a VM method, the required MMC can be used by downloading it in which the existing data can be automatically linked to the stored database server. Because the DB server can be linked to the server existed in a mobile communication company at a Content Provider (CP), which uses the platform of a mobile communication company, it can be implemented at a CP server or configured as an independent CP.

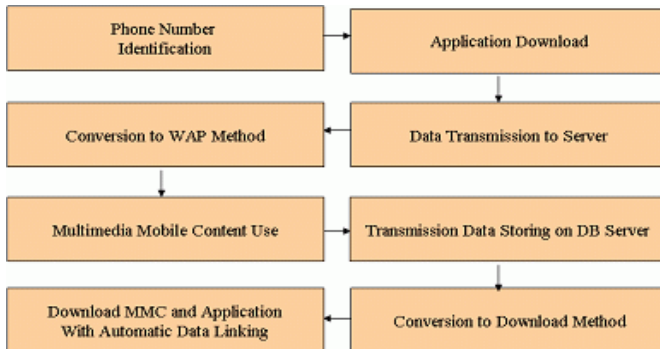


Fig. 1. Linking process

### 2.2 Functions of Each Module and Processors

#### 2.2.1 DB and Application Linking

User's information and ranking can be checked in DB by linking to the database for each platform using SQL sentences. Linked DB safely stores the value of ranking that is transmitted by the user to a CP server, and the DB is presented to display a WAP page according to the request of users.

Users first download applications and MMC to their mobile terminals using a VM method in which a transformation is required from a VM method to a WAP method in

order to link as a WAP method. A user terminal can be transmitted as a WAP method after transmitting various data to a CP server and storing it to the database of a CP server. The CP server stores user data to the DB. Then, users download data stored in a mobile terminal in the CP server. User's mobile terminals can use MMC using a link method without any losses to their data for both transitions of VM and WAP. Fig. 2 presents the procedure and data flow in a link method from VM to WAP.

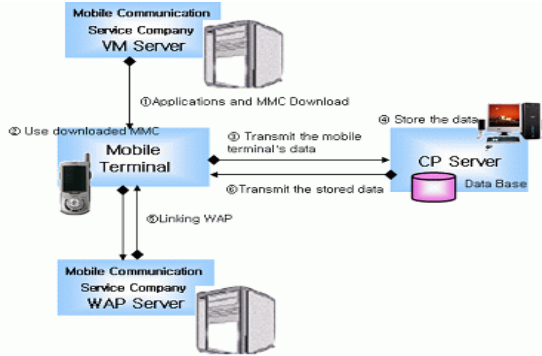


Fig. 2. DB and Application Linking (VM → WAP)

In addition, it is possible to link from a WAP method to a VM method using the database of a CP server. Fig. 4 presents the link from the WAP to the VM. Users who used multimedia mobile content through a WAP method can download the MMC stored in a VM server to their mobile terminals after transmitting their data stored in mobile terminals to the database of a CP server in order to store the data in the CP server. Then, mobile users can use the linked data by downloading the existing data stored in the DB of a CP server to their mobile terminals.

This method can achieve a continuous link for various materials stored in users' mobile terminals to VM→WAP method presented in Fig. 3 and WAP→VM method presented in Fig. 3.

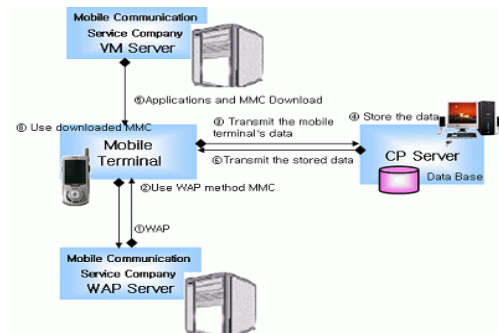


Fig. 3. DB and Application Linking (WAP → VM)

### 2.2.2 Phone Number Identification Module: Identification of Phone Numbers in a Server

This module updates the existing phone number in the present DB by transmitting a query to the DB by extracting a user's mobile phone number from the HTTP\_COOKIE header in a phone or adds a new number. Table 1 presents a part of sources to identify a phone number.

**Table 1.** Phone number identification

```
$Suid1=getenv(HTTP_COOKIE);
$Suid1=explode(";", $Suid1);
$Suid2=$Suid1[0];
$Suid2=explode("=", $Suid2);
$Suid2=$Suid2[1];
$Suid3=$Suid1[2];
$Suid3=explode("=", $Suid3);
$Suid3=$Suid3[1];
$Suid = $Suid3;
$Suid .= $Suid2;
if (!$Suid) $Suid = 'testuid011';
```

### 2.2.3 Automatic Application Download Module: Application Downloads Using a WAP Server

This module downloads certain applications automatically when a WAP server requires a download. The function used in this process is down.wmls#appl(). Table 2 presents a part of application download sources for a WAP server.

**Table 2.** Application download

```
header("Content-type: text/vnd.wap.wml");
header("Cache-Control:no-cache,must-revalidate");
header("Pragma: no-cache");
header("Expires: content=0");
echo ("<?xml version=1.0 encoding=KS_C_5601-1987 ?>");
echo ("<!DOCTYPE wml PUBLIC -//WAPFORUM//DTD WML 1.1//EN
http://www.wapforum.org/DTD/wml_1.1.xml>");
echo("<wml>");
echo ("<card newcontext=true title=application program name>");
echo "<p><do type=vnd.skmn1 label=appl_start>";
echo "<p><do type=vnd.skmn2 label=DOWNLOAD>";
echo "<p><do type=vnd.skmn3 label=appl_explanation>";
echo "<go href=down.wmls#appl2() />";
echo "</do></p>";
echo "</card></wml>";
```

### 2.2.4 Data Transmission DB Server and Application: Transmission of Various Data Used in Applications to a Server

This will transmit data to a DB server using a phone number function (GetSysMin ()), data extraction function (MakeStrStr ()), and server transmission function (BackToBrowser ()). Table 3 presents parts of sources to transmit data to a server in applications.

**Table 3.** Data transmission

```
GetSysMin(TextOut2);
// a function that receives phone number in application
MakeStrStr(TextOut, /* server information that is transmitted */
"http://m.gameneo.com/gvm/dino.php?ver=1&phonenumber=%s",TextOut2);
// A function to extract data to send to server
BackToBrowser(TextOut);
// A function to transmit extracted data to server
```

## 3 Applications: Mobile Multimedia Games

This paper develops game content that will be serviced based on a GVM platform used in SKT Telecom.( Korea company) by applying the method proposed in this paper. In the method used in this development, a method, which simply transmits ranking information, implements a local stage, and executes a game by downloading files for the stage to WAP→VM during the execution of the game using WAP. Then, the stored data is be transmitted to VM-WAP when the stage is cleared, and the data is stored to the database of a WAP server. The next stage is then processed.

Table 4 presents functions for each item, production tools, production platform, and used components for a main client. The tool used in this development was Visual C++ 6.0 by Microsoft. Table 5 presents the database table structure of a CP server. The database structure applied in a link is variable according to the characteristics of serviced MMC and objective of the link.

**Table 4.** Configuration of a main client

Item	Function	Work tool	Work platform	Component
Network processing	Connection to WAP Page	Visual C++	GVM	BackToBrowser
Graphic processing	Processing about images on each necessary logic	Visual C++	GVM	DrawMap, PlayerDraw, DinoDraw, DrawDownInterface
Game event	Event processing that was created on game	Visual C++	GVM	EVENT_KEYPRESS
User event	Event processing to have been happened by keys input of a user or other operation	Visual C++	GVM	GameMain
Sound processing	Sound output on each logic	Visual C++	GVM	SndReSet
Stored data processing	Data processing that must keep it after a game end	Visual C++	GVM	SaveData, LoadData

**Table 5.** DB table configuration of CP server

Field	Description
PhoneNum	Terminal number
StDate	First date that registered a score
UpDate	Date of enrollment recently
Version	Game version
Experience	User level & player experience price value
Money	User possess money

The component used in this game development consists of 5 groups as follows:

- Network processing: BackToBrowser.
- Events processing in a game: EVENT\_KEYPRESS and EVENT\_TIMEOUT.
- Recursive call in a game: SetTimer.
- Saving game data and load processing: ArrayToVar, GetUserNV, and PutUserNV.
- Graphics processing: CopyImageDir, CopyImagePal, CopyImageDirPal, DrawText, SaveLCD, and RestoreLCD

Because the transmission of VM→WAP is easy, the existing MMC uses this method in order to verify ranking. In addition, the ranking can be verified by transmitting data packets to VM→WAP method.

#### 4 Comparison and Analysis

Table 6 presents the comparison between the proposed and the existing method. In the case of the WAP method, it is difficult to present multimedia effects in MMC services even though it has few limitations in capacities and extensibility. In the case of the VM method, it presents some problems in capacities which are to the demerit of download based services, and that presents a lack of extensibility even though it can service various MMC services compared to the WAP method. However, the proposed method has the merit that the advantage of multimedia effects in a VM method can be maintained and presents extensibility due to no limitations in capacities. In addition,

**Table 6.** Comparison between the proposed method and the existing method

Item	WAP method	VM method	Proposed method
Multimedia effects	Text and simple multimedia effects	Various multimedia effects	High quality MMC that is composed high quality stage production is possible because it is possible continuously download with a WAP→ VM linking
Restriction of capacity	Capacity restriction is small because of continuously connected with server	Capacity restriction is large because of total MMC(GVM:128Kbyte)	Small restriction
Expansibility	Possible	Impossible	It can continuously add to stage like an online RPG game
Use purpose of linking	-	Ranking	Ranking, additional download stage

based on the results of the real application in mobile multimedia game, it is evident that the proposed method can be applied to certain rankings and stages when compared to the fact that the link of a VM method can be limitedly applied to certain rankings. It means that the proposed method can be applied to various applications that use MMC.

In addition, the existing method presents some difficulties in the development of games, which use various styles, due to the fact that this method cannot be linked to others except for the verification of rankings. Thus, the VM method presented in Table 8 was applied to produce a limited game due to the capacity limitation of a mobile phone. Although the WAP method presented no capacity limitations in a game, it has the limitation in the presentation of dynamic displays. However, the proposed method can implement a mobile game in a dynamic execution environment and extend the game due to the application of the merit of both WAP method and VM method.

## 5 Conclusions

This paper improved the limitation of the capacity movement that is a limitation in mobile multimedia services due to the maintaining of data in different methods by sharing data in real-time between the WAP and the VM. In addition, this study presented a new mobile multimedia game service model by applying the proposed method to a real mobile multimedia game. The proposed method was able to improve the limitation of the text and still image based WAP method. In addition, this paper achieved an improvement in the quality based on the performance due to the fact that certain interactive factors can be applied in a mobile multimedia game by solving the problem of the link in a VM method.

Moreover, the proposed method was able to extend other multimedia services, such as stock trading, news, and other various services while also apply it to other various multimedia services because it can supply actual information by linking it to huge materials and MMC.

## References

- [1] Gwo-Jen Hwang, Judy C.R. Tseng, Yu-San Huang, "I-WAP: An Intelligent WAP Site Management System," *IEEE Transactions on Mobile Computing*, Vol. 1, No. 2, pp. 82-95, April 2002
- [2] Wireless Application Protocol WAP 2.0 Technical White Paper, Wireless Application Protocol Forum, Jan. 2002
- [3] Huw Evans, Paul Ashworth, Getting Started with WAP and WML, SYBEX, Inc., 2001
- [4] SK-VM 2.x, <http://developer.xce.co.kr>
- [5] SinjiSoft GENEX, [http://www.sinjisoft.co.kr/html/gnex\\_gnex.htm](http://www.sinjisoft.co.kr/html/gnex_gnex.htm)
- [6] Qualcomm BREW, <http://www.qualcomm.com/brew>
- [7] J. Arreymbi, M. Dastbaz, "Issues in Delivering Multimedia Content to Mobile Devices," *Proc. of Sixth International Conference on Information Visualization (IV'02)*, pp. 622-626, July 2002
- [8] S.R. Subramanya, Byung K. Yi, "Mobile Content Customization," *IEEE Multimedia*, pp.103-104, October 2005
- [9] Nina D. Ziv, "Toward a New Paradigm of Innovation on the Mobile Platform: Redefining the Roles of Content Providers, Technology Companies, and Users," *Proc. of International Conference on Mobile Business (ICMB'05)*, pp.152-158, 2005

# Development of Oval Based Vulnerability Management Tool (OVMT) on a Distributed Network Environment\*

Geuk Lee<sup>1</sup>, Youngsup Kim<sup>1</sup>, and Sang Jo Youk<sup>2</sup>

<sup>1</sup> Dept. of Computer, Hannam University, DaeJeon, South Korea  
leegeuk@hannam.ac.kr, c4i2@is.hannam.ac.kr

<sup>2</sup> School of Multimedia, Hannam University, DaeJeon, South Korea  
youksj@paran.com

**Abstract.** This paper designs and implements an unified vulnerability assessment tool which can assess vulnerability of both a network and host systems together. The vulnerability database is constructed based on CVE to report vulnerabilities in standard. The network scanner is implemented using Nessus and the system scanner is implemented using OVAL. A tool managing both network scanner and system scanner is implemented and is controlled by Web manager with which user can assess systems at anywhere in a distributed environment through web if a web manager agent is installed. As a conclusion, the unified vulnerability assessment tool implemented in this paper can provide fast and more accurate vulnerability assessment and proper guidelines to corresponding vulnerabilities

**Keywords:** vulnerability, assessment, scanner, Nessus, OVAL.

## 1 Introduction

At present, we are living in highly digitized computer and information environment. Everything is automatically worked with computer chips and everyone communicates each other using computerized equipments and networks. And most information is transferred through the internet. But, unfortunately, malicious cyber attacks and critical damages are also increased. Therefore computer security managers have to continuously monitor the status of computer and network systems, periodically make a backup, and analyze every log information to keep the system secure from any kind of cyber attacks. Also they have to detect vulnerabilities of the system in advance to remove possible sources for the attacks.

Even though there are several vulnerability assessment tools, those have limitations. It is hard to share vulnerability information because each vendor developed vulnerability assessment tool using their own policies and no standard way to assess vulnerability. Assessment script is too difficult and complex to be read and updated. And existing vulnerability scanners are classified into one of network scanner and

---

\* This work was supported by a grand No.R12-2003-004-02003-0 from Korea Ministry of Commerce Industry and Energy.



system scanner. Therefore, standardized and unified vulnerability assessment tool is necessary. In this paper, an easier, more accurate and standard vulnerability assessment tool is designed and implemented. It also unifies network scanner and system scanner together.

## 2 Related Works

### 2.1 Vulnerability Assessment Tool

The vulnerability is weaknesses of computer system security. It is not a critical problem by itself to the system but it may provide environment for dangerous security attacks.

The vulnerability assessment tool is a tool which investigates security weaknesses broadly distributed in networks and host systems and analyzes and reports an assessment result using information of the vulnerability database. The vulnerability assessment tool is classified into a network scanner and a system(host) scanner[1].

The network scanner assesses vulnerabilities of network environment such as DoS, RPC, HTTP, SMTP, FTP, FINGER, and etc. Typical network scanners are Nessus, SAINT, ISS network scanner.

The system scanner assesses vulnerabilities within host systems. Because the system scanner works at a host level, it is mostly platform dependent and should be installed at all target hosts[2]. Most current system scanners, such as COPS, Tiger, STAT, have limitations of checking vulnerability in a distributed environment.

### 2.2 Vulnerability Database

The vulnerability database, which is one of the most important parts of the vulnerability assessment tool, is a database classifying, storing, and managing vulnerability information of computer systems. The vulnerability database is used to query about collected vulnerabilities and to report the assessment results[3]. The research of the vulnerability database is world-widely being performed, and some of leading research centers are Security Research Team COAST at University of Purdue and ICAT Meta Vulnerability Database of NIST.

### 2.3 OVAL(Open Vulnerability Assessment Language)

OVAL is a standard language to detect vulnerabilities of characteristics and environment information of local host systems. It provides a method to define and to detect vulnerability based on CVE which is a naming schema giving standard name to each vulnerability. OVAL definition can be described in XML script or SQL script. The vulnerability assessment tool use information in OVAL definition to find out existing vulnerabilities instead of simulation attacks to the system.

The system scanner collects vulnerability information of hosts using the system vulnerability assessment client installed on each host, analyzes those information using the OVAL interpreter, queries to the vulnerability database, and then reports the assessment results [4]. The system scanner can detect all the vulnerabilities registered

on CVE and describe vulnerabilities clearly and logically because the vulnerability assessment script is written in XML and SQL [3, 5].

OVAL definition includes three types of OVAL schemas. Definition schema keeps vulnerability information, System Characteristic schema represents information of a target system, and Result schema provides information about assessment results. There are different OVAL definitions since different operating platforms have different conditions of the presences of vulnerabilities. OVAL was suggested by MITRE and its standardization is currently being under going by security expert groups[3, 5].

### 3 Design and Implementation of OVMT

#### 3.1 System Structure of OVMT

OVMT consists of the vulnerability, a network scanner, a system scanner, UVAT(Unified Vulnerability Assessment Tool) which manages scanners to detect necessary vulnerability information and report assessment result, and the web manager which can remotely communicate with UVAT (Fig. 1).

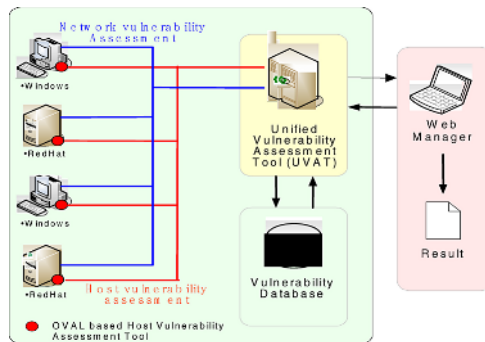


Fig. 1. System structure of OVMT

Only one network scanner agent is installed for a target network but the system scanner agent should be installed on each target host since the system scanner cannot be assessed in remote process. The web manager can access UVAT through internet from any computer on which web-manager agent is installed

Because of using NTP protocol and encrypting data packets in SSL, vulnerability information transferred within OVMT is surely secure against packet sniffing[4].

#### 3.2 Vulnerability Database

The vulnerability database consists of schemas described in Table 1. It uses CVE ID as an identifier of the vulnerability, and provides 100% CVE compatible with other vulnerability databases because of using CVE. It can be optimized by parsing I-CAT DB, and includes web client which contains a module performing real-time update of the database.

**Table 1.** Entities of the vulnerability database

Entity	Contents
Vulnerability	General contents of the vulnerability. CVE ID, Type, Risk Level, Damage Type, Domain Type, requirement for attack, and etc
Related Reference	Corresponding reference information to the vulnerability
Vulnerable Software	Information of vulnerable Software. Same as a reference entity, one vulnerability may include more than one vulnerable software.
Vulnerability Name	Name of the vulnerability. Classified into English and Korean
Vulnerability Summary	Summary of the vulnerability
Vulnerability Description	Description of the vulnerability
Vulnerability Solution	Solution of the vulnerability
Vulnerability Author	User who updated information of the vulnerability
User	User information

### 3.3 UVAT (Unified Vulnerability Assessment Tool)

UVAT is designed to control the network scanner agent and the system scanner agents together simultaneously. When UVAT receives a message from the web manager, UVAT logs in to corresponding scanner agent, gives an order of detecting vulnerability information, queries to the vulnerability database, and reports the result to the web manager. Because the network scanner using Nessus takes client-server structure, UVAT acts as a client of Nessus. Also UVAT can a client of system scanner using OVAL.

### 3.4 Network Vulnerability Assessment Tool

The network scanner is implemented using Nessus[7]. Nessus describes vulnerability information in plug-in scripts. Because the plug-in scripts can periodically and rapidly be updated by the user group, Nessus can diagnoses large scale networks. The network scanner communicates with UVAT using NTP(Nessus Transfer Protocol) which is ASCII based communication protocol for Nessus agent and clients. UVAT uses NTP v.1.2 and overall procedure is below.

UVAT asks access to the Nessus agent using NTP v.1.2. Once successfully connected, Nessus sends the plug-in lists to UVAT and UVAT selects and returns a proper plug-in to Nessus. Then, Nessus performs vulnerability assessment and replies the vulnerability information to UVAT.

### 3.5 System Vulnerability Assessment Tool

The system scanner is implemented based on OVAL on RedHat 9 Linux platform. The system scanner includes Data File of OVAL definitions, Data File Verification module, OVAL Interpreter, System Information Collecting module, Reporting module, Log Management module, and OS Platform(Fig. 2).

Data File is a set of OVAL definition in XML script, and OVAL definition is used as an input data to OVAL scanner. OVAL definition has a name in a form of "OVAL + ID number" which is assigned to each OVAL definition in sequential order of announcement[4].

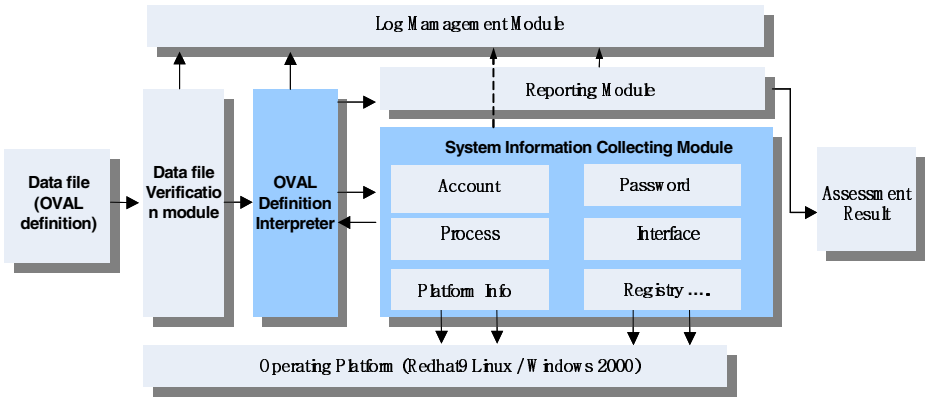


Fig. 2. Overall Structure of System Vulnerability Assessment Tool

Data File Verification module checks effectiveness of OVAL definition using MD5 hash which is supported by OVAL homepage(<http://www.oval.mitre.org>).

The OVAL Definition Interpreter interprets the OVAL definition, decides what kinds of information to be collected, makes System Information Collecting module collect the information, compares collected information with conditions of vulnerability described in the OVAL definition, and then forwards results to the Reporting module. The System Information Collecting module collects host information such as system, file, network demon, process, RPM package, shadow password, operating platform, and etc. Log Management module handles all logs created during execution time of OVAL. It is implemented using “log4J library” which is developed in process in Java.

### 3.6 Web-Manager

The web manager communicates with both scanners using UVAT. At any computers in which the web manager is installed, the vulnerability assessment is available in real-time just after logging in the web manager. The web manager can adjust the environment setup and plug-in setup for assessing the vulnerability.

The management console of the web manager is divided into five parts. User can control UVAT using Menu and Toolbar. The Session part displays session selected, and the View part shows session information and vulnerability information, The Console View part represents information such as log messages.

## 4 Results of Vulnerability Assessment

### 4.1 Network Vulnerability Assessment

Fig. 3 is a real time dialogue box displaying current progress of network vulnerability assessment. The left section displays a list of target hosts, and the right section shows

details of current situation of assessment such as a host being assessed, a current stage being performed, vulnerability information being checked, number and security level of the vulnerabilities, and etc.

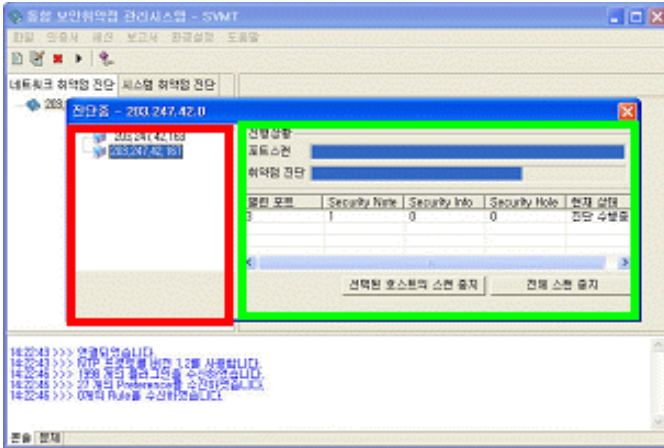


Fig. 3. Real time dialogue box

Fig. 4 is a result of the network vulnerability assessment by performing a port scanning. It shows a possibility of remote log-in using a NULL session and a guideline to prevent the NULL session

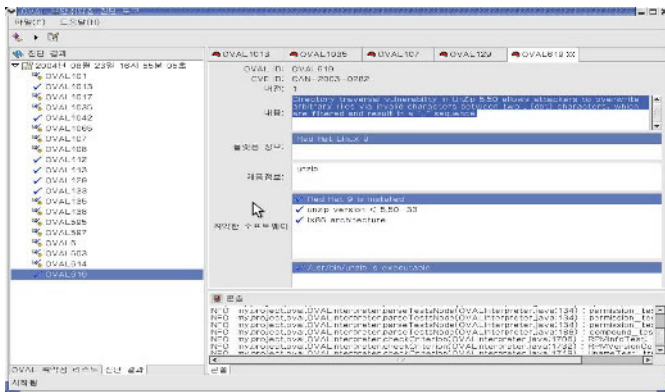


Fig. 4. Snapshot of Network vulnerability Assessment

### 4.2 System Vulnerability Assessment

Fig.5 is a photocopy of result of the system vulnerability assessment. The left section displays a list of OVAL definition checked by the system scanner. Each OVAL definition on the list is colored one of Red and Green, and the red colored OVAL

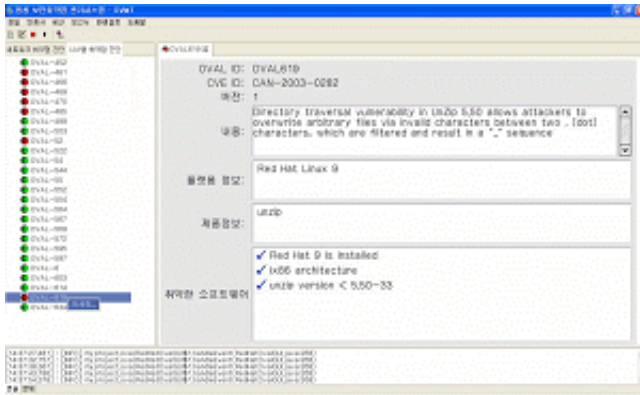


Fig. 5. Snapshot of System vulnerability Assessment

definition means the vulnerability is detected. The right section shows corresponding vulnerability information. When clicking the right mouse button of any OVAL definition on the list, an assessment result of the selected OVAL definition is listed on the right section.

Fig. 5 shows vulnerability of CAN-2003-0282 as a result of OVAL definition 619. It is a vulnerability using /usr/bin/unzip file which version is 2.50-30 or below on RedHat Linux version 9.

## 5 Conclusion

In this paper, OVAL based vulnerability management tool is designed and implemented to remotely detect the vulnerabilities of several systems(Windows 2000, Redhat 9) and to provide proper guidelines through WEB system on the distributed network environments.

The standard vulnerability database is constructed based on CVE. Network vulnerability assessment is implemented using Nessus. UVAT can remotely control Nessus scanner, set a range of assessment, and save a result in a file in order to manage assessment result continuously.

System vulnerability assessment is implemented using OVAL. It can diagnosis Linux and Window 2000 Server and provides standardized vulnerability information and corresponding guidelines.

OVMT has several strong points. First of all, Speed of Vulnerability assessment is faster and easier than other existing tools do since OVMT can assess networks and systems at a same time and provide vulnerabilities and guidelines together. OVMT can detect newest vulnerability because vulnerability assessment scripts are periodically updated; the network vulnerability script is by Nessus user group, and the system vulnerability script is by MITRE. OVMT increases an interoperability with other vulnerability database by applying CVE. At last, it can provides easier and more visible GUI.

## References

1. Jung Hee Kim, "The Trends of International Standardization on Vulnerability Assessment of Computer", [http://www.secureosforum.org/download/data\\_read.htm?id=56&start=0](http://www.secureosforum.org/download/data_read.htm?id=56&start=0), The Trends in International Information Security, Korea Information Security Agency, June 2003
2. UNIX Security Checklist v2.0, [http://www.cert.org/tech\\_tips/usc20\\_full.html](http://www.cert.org/tech_tips/usc20_full.html)
3. Ragi Guirguis, Network and Host-based Vulnerability Assessments,
4. <http://www.securitytechnet.com/resource/security/consulting/1200.pdf>, February 2004.
5. Young Mi Gwon, Hui Jae Lee, Geuk Lee, " A Vulnerability Assessment Tool Based on OVAL in Linux System", IFIP International Conference, LNCS3222, pp653-660, October 2004
6. M. Wojcik, T. Bergeron, T. Wittbold and R.Roberge, Introduction to OVAL,
7. <http://oval.mitre.org/documents/>, MITRE Corporation, November 2003.
8. Introduction to CVE, The Key to Information Sharing,
9. [http://cve.mitre.org/docs/docs2000/key\\_to\\_info\\_shar.pdf](http://cve.mitre.org/docs/docs2000/key_to_info_shar.pdf)
10. Introduction to Nessus, <http://www.securityfocus.com/infocus/1741>

# A Study on a Design of Efficient Electronic Commerce System

Yun Ji Na<sup>1</sup>, Il Seok Ko<sup>2</sup>, and Jong Min Kwak<sup>3</sup>

<sup>1</sup> Department of Internet Software, Honam University, 59-1 Seobong-Dong, Gwangsan-gu, Gwangju 506-741, South Korea  
yjna@honam.ac.kr

<sup>2</sup> School of Computer and Multimedia, Dongguk University, 707 Seokjang-dong, Kyungju, Kyongsangbukdo, South Korea  
isko@dongguk.edu

<sup>3</sup> ChungCheong dot Com, 40-10 Bokdae-Dong, Heungdok-gu, Chongju, South Korea  
webmaster@ccilbo.com

**Abstract.** In this study, an e-commerce system with the capability of web cache based on the split area in a hierarchical structure is designed. The performance is then assessed through experiment. The local server for the proposed system is a duplicated server for the e-commerce system. The advantage of location improves the latency time. The system's response speed is also improved due to dispersion of the entire systems load. The performances of the system and algorithm are analyzed with an experiment. With these experiment results, the algorithm is compared with previous replacement algorithms.

**Keywords:** E-commerce, Network traffic, Web server.

## 1 Introduction

The developed IT technology and the rapid spread of Internet are not only expanding the e-commerce system suddenly but also occasioning a sharp increase in the number of users. These current situations bring about the heavier network traffic in addition to a sudden increase of load to the e-commerce system, and also incur the unnecessary squandering of the system resources as the repeated requests for the same object occupy a large portion of the network bandwidth. Therefore, it's necessary to study the e-commerce system in consideration of the operational efficiency and response speed. There are several factors that have impact upon the user's response speed on the Internet as follows: size of the object, advantageous position, status of the network traffic, and performance of the server. To improve the physical factors such as traffic condition, performance of the server, the required cost will be relatively considerable. For that reason, we shall require improvement of the response speed through the dispersion of the load and traffic, rather than betterment of each individual system's own performance. Our efforts to fulfill the needs will just have to go through an innovative improvement in network traffic condition and server performance.

In this study, we designed an e-commerce system with the capability of web cache based on the split area in a hierarchical structure and then assessed the performance through an experimental trial. The local server for the proposed system is, as a



replicated server for the e-commerce system, capable of shortening latency time by the system's advantageous location as well as improving response speed through the dispersion of the entire system's load. A Design was devised to decrease impact that the repetitive request for the same object have upon the network bandwidth through the Web Cache function of the local server.

The web cache that is proposed in this study provides the benefit of improving the hitting ratio against a certain object requested by the client compared with the existing LRU technique. We found, in consequence, that it can contribute more improved response speed and quick adaptability toward the customer satisfaction of the e-commerce system requirement.

## 2 A Design of E-Commerce System

### 2.1 Configuration of System

The proposed system consists of multiple servers variably depending on the system load. The e-commerce system enables the distribution of traffic among the multiple servers - the component elements of a web server through load balancing (L/B). Figure 4 shows the experimental model of e-commerce system.

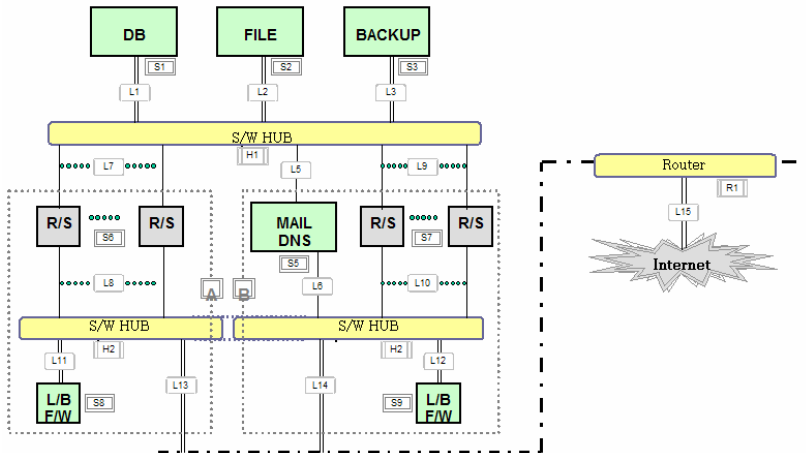


Fig. 1. Experimental Model of e-Commerce System

Listed below are features and functions of the component servers listed in Figure 1

- DB Server: It manages a variety of Databases such as merchandise information, customer information, registration information server over e-commerce sites operating on the Real Server and provides services for a request of a Mail Server.

- Case Base Server: It is used to provide services for the customer who uses the e-commerce system through the case base. Its setup may be independent of or include a File Server.

- Backup Server: It is responsible for the second function to perform backup for a variety of DB servers and used to make clear the stability of data in the middle- or large-scale e-commerce system though being omitted sometimes in setting up a small-scale e-commerce system.
- Real Server: It plays a part in providing services to the customer directly through diverse capabilities over the web site such as membership, purchase, approval, etc.. Also it is set up with plural servers in consideration of distributing loads.
- Load Balancer: It distributes loads among the plural component servers. Most web servers are composed of multiple sub-servers, and as such can be fulfilled through the efficient distribution of traffic between the lower servers that construct a web server by way of load balancing.
- Local Server: It provides the capability of caching for a request of a user which connects with each local server in addition to the function of main system replication. This will provide a good remedy of latency time against the user’s requirement.

**Table 1.** Component Elements and System Specifications of the Experimental Model

Component	Number	System Setup
DB Server	S1	CPU: Alpha 21264 667Mhz*2 Memory: 1GB HDD1: 9GB*1 HDD2: RAID Controller, 18.2GB*4 4 Ethernet (100Mbps)
File Server	S2	CPU Alpha 21264 667Mhz*2 Memory: 1GB HDD1: 9GB*1 HDD2: RAID Controller, 18.2GB*4 4 Ethernet (100Mbps)
Backup Server	S3	CPU: Alpha 21264 667Mhz Memory: 500MB HDD1: 9GB*1 HDD2: RAID Controller, 18.2GB*4 4 Ethernet (100Mbps)
Real Server	S6, S7	CPU: Alpha 21264 600Mhz Memory: 256MB HDD1: 9.1GB*1 2 Ethernet (100Mbps)
Load Balancer	S8, S9	CPU: Alpha 21264 600Mhz Memory: 500MB HDD: 9.1GB*1 2 Ethernet (100Mbps)
Mail/DNS server	S5	CPU: Alpha 21264 667Mhz Memory: 1GB HDD1: 9GB*1 HDD2: RAID Controller, 18.2GB*4 4 Ethernet (100Mbps)
Switching Hub	H1, H2	24 Port Fast Ethernet switching HUB
Network	L15	100Mbps

Table 1 provides the list of system specifications and component s as applied for the experimental model. Experimental equipment was based on an alpha processor

mounted system. As seen through Table 1, the experimental system was not prepared in the latest environment of equipment and network but optimal configuration of the system could make it clear to predict a potential improvement of its performance. As an experimental model of system implementation for e-commerce, available specifications are equivalent to Apache Web Server for the environment of Red Hat Linux7; hp4 for the server script language in Java Script and HTML; and Mysql for the database management system. Also, a speedy response is an absolute requirement to the DB server and file server for the customer and merchandise information service. Therefore, to improve the response speed, we have completed modeling a system with two mountable CPUs. Sever that Alpha Processor was loaded is a Processor product of Samsung Electronics Company. In the prototype experiment, the server with two processors specified to Alpha 21264 667Mhz CPU is used as DB server and the file server Is equivalent to a model which is specifically mounted with Alpha 21264 600Mhz CPU, 256MB Memory, and 9.1GB\*1 Hard Disk. For the real server and load balancing, we applied the same system model as this one. But we use neither mail/DNS server nor backup server that would have no effect on the assessment of performance.

## 2.2 Web Caching

The object replacement algorithm that considers an object size is required in order to increase the efficiency of web caching through the previous log analysis. The idea of a proposal algorithm is as follows. First, it is possible to classify objects based on size characteristics, and to manage the divided storage of a cache efficiently. Second, reference characteristics of an object are variable. Therefore, efficiency based on the divided size of cache storage is varied, too.

The number of division scope, the volume of scope to be allocated to divided scope, and the determination of size to classify an object have an important influence on the performance of web caching in this algorithm. Storage scope of the object that has an influence on web caching must be assigned in order to increase the hit ratio of cache.

The object storage scope of 10k or above is divided into scope LARGE, and the object storage scope of 10k or less is divided into scope SMALL. When the object requests of a client arrive, the cache administrator confirms the existence of an object in the corresponding division scope according to the size of an object. The service of the object that a client required would be provided if a cache-hit occurs. Then, the cache administrator updates the time record on when it was used so that high ranking is assigned to this object in an LRU-MIN replacement.

If a cache miss occurs, the cache server requests the corresponding URL for the service of the object, and the object is transmitted to the cache server. Then, a transmitted object is classified into a corresponding grade according to size, and a cache administrator confirms whether there is a space for this object to be saved in a cache scope of a corresponding grade.

If there is a space to save in the cache scope, this object is saved, and the object is saved by LRU-MIN replacement algorithm if it is not there. Then, the web object saved in each scope is substituted among the objects of the same grade. Also, a time

record of the newly arrived object is saved, and a high ranking is assigned to this object in an LRU-MIN replacement process.

As was mentioned in the previous reference characteristics analysis of an object, the reference characteristics and the heterogeneity of web objects would be affected by the characteristics of web service and the user's aging characteristics and user's academic background, as well as a timing factor. The web service that includes many different kinds of multimedia data and the object reference of a comparatively young age user increase the object reference to large size. According to this, the object reference characteristic has an extreme variation. Therefore, the size of cache scope division must be varied.

The proposed algorithm is similar to LRU and LRU-MIN, SIZE in the reference tendency but there are differences in the following points. First, LRU is referring to the time immediately before an object was referred to and to the size of an object among the past object reference information, and it is not reflecting a reference frequency and the heterogeneity of an object. Second, basically, LRU-MIN operates with LRU equally. But the point that substitutes a small size object for it at first in order to substitute the minimum object for it is different in reflection on a size of an object. In the worst case, a lot of small-sized objects are removed by one large-sized object. The algorithm of SIZE improved this issue by replacing the greatest object among objects of cache storage scope for new object. But LRU-MIN and SIZE are not reflecting heterogeneity of object, either. Third, the proposed algorithm can reflect the size and heterogeneity of an object sufficiently.

This study experimented on the performance of a cache scope divided by 6:4 and 7:3. But the efficiency of web caching may be increased more if the division scope of variable size is used according to reference characteristics of an object than by division scope of size that has been fixed.

### 3 Analysis

Generally, for the performance evaluation measure of the caching algorithm, hit ratio is used. And for the performance evaluation of the e-commerce system, response speed is mostly used.

In this experiment, the performance of the proposed algorithm is evaluated by comparing an average gain ration of object-hit. And the performance of the proposed system is evaluated by comparing a response speed.

Response speed  $RT$  have the several response speed notations,  $RT_{ch}$ (Response speed of cache-hit) on the cache-hit and response speed  $RT_{cm}$ (Response speed of cache miss) on the cache miss. Response speed for an object request of a client has the following delay factors.

- ①  $TDT_{client\_to\_cache}$ : Transmission delay time that occurs when a client requests an object to cache server
- ②  $DDT_{cache}$ : Delay time required for a determination of cache-hit or cache miss of cache server
- ③  $SDT_{cache}$ : The delay time required for a search of an object saved in Large or Small scope of cache

④  $TDT_{cache\_to\_client}$ : Delay time required when an object is transmitted from cache server to a client

⑤  $TDT_{cache\_to\_URL}$ : Transmission delay time required when cache server requests an object to URL

⑥  $TDT_{URL\_to\_cache}$ : Delay time needed when an object is transmitted from URL to cache server

⑦  $RDT_{cache}$ : Delay time that occurs in cache server when an object is replaced

1) A case of cache-hit

$$RT_{ch} = TDT_{client\_to\_cache} + DDT_{cache} + SDT_{cache} + TDT_{cache\_to\_client} - \text{Formula (1)}$$

2) A case of cache miss

$$RT_{cm} = TDT_{client\_to\_cache} + DDT_{cache} + TDT_{cache\_to\_URL} + TDT_{URL\_to\_cache} + RDT_{cache} + TDT_{cache\_to\_client} - \text{Formula (2)}$$

The response speed has a close relationship with the object-hit ratio. Four delay factors occur with response speed of a hit object in cache as in Formula(1), but six delay factors occur following the pattern of Formula(2) in the response speed of a miss object. Among these delay factors,  $TDT_{client\_to\_cache}$ ,  $TDT_{cache\_to\_client}$ ,  $TDT_{cache\_to\_URL}$ , and  $TDT_{URL\_to\_cache}$  are affected by the physical environment of networks. Therefore, delay time gets longer than cache-hit for the cache miss because of the many influences of the physical environment. Also, the  $RDT_{cache}$  is the delay time that occurs in cache server when objects replaced have a lot of influences on the performance of web caching in Formula(2). Therefore, if we increase the object-hit ratio, we can improve the response speed .

First, we measured object-hit ratio. The experiment method is as follows. 70% on the cache scope was assigned first to a LARGE grade, and 30% was assigned to a SMALL grade. And the experiments were conducted on object-hit performance of this algorithm and LRU, LRU-MIN, SIZE. Second, 60% on the cache scope was assigned to a LARGE grade, and 40% was assigned to a SMALL grade. Also, we experimented on the performance of these algorithms.

Response speed is the time required to provide the requested web object to a client(customer). Figure 2- Figure 4 show the results of the experiment on response speed. And Figure 10- Figure 11 show the gain ratio on response speed.

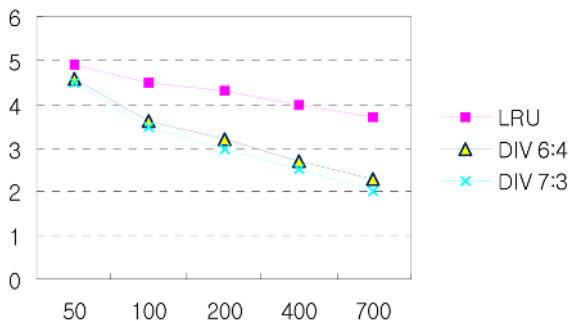


Fig. 2. Response speed (sec.): compare with LRU

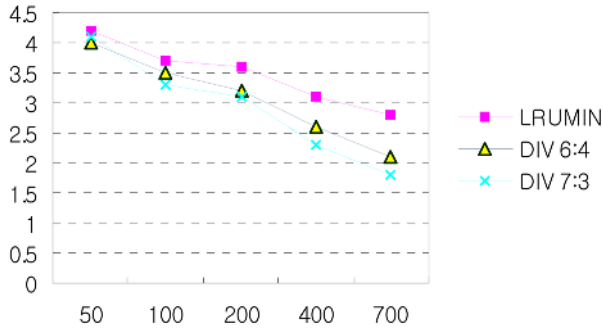


Fig. 3. Response speed (sec.): compare with LRU-MIN

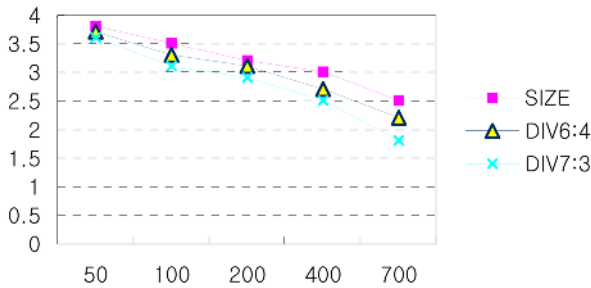


Fig. 4. Response speed (sec.): compare with SIZE

Experiments were conducted on response speed performance of the supposed system with LRU, LRU-MIN, SIZE. We reached the following conclusion by the experiments results of a response speed.

1) As the capacity of cache grew larger, the response speed performance of the proposed system is more efficient than traditional replacement algorithms.

2) As for the gain ratio of response speed, we can get 30% or above performance enhancement than LRU. Also, we can get 15% or above performance enhancement than LRUMIN and SIZE. The reason that performance enhancement of gain ratio is higher than performance enhancement of object hit ratio originated in size of the object which user refer to. There was comparatively a lot of reference on large-sized object in the experiment. According to this, response speed affected the delay time of the network greatly.

## 4 Conclusions

This study includes a design of hierarchical e-commerce system using the local server in order to decrease the decline of the system’s response speed due to various geographical factors; proposes a partition-based web cache algorithm for the exclusive cache server in order to remedy latency due to object size; and verifies the usefulness of the system and algorithm.

Based on the experiment results, the proposed web cache algorithm indicates a higher hit ratio than the existing cache algorithm and therefore guarantees improved response speed and more flexibility for the client's various requests. It also prevents a decline of response speed caused by the multiplex requests from multiple users concurrently using the e-commerce system.

The e-commerce system developed for this study produces some additional expense required for the configuration of the local server. But it is a possible trade-off to balance that cost against replacement costs for improving the performance in the existing system and relevant costs for enhancing the performance of network. In the future, the adaptability of the proposed system shall be verified through the application in the actual field of e-commerce. Moreover, a user-centered performance of the system by the wide application of the intelligent agent architecture for e-commerce will need to be developed. The issue of the intensive request for an object on the e-commerce system will be revisited in a further analysis of the customer's preference. In conclusion, reflecting the customer's preference on the cache algorithm will result in even a higher efficiency of the cache.

## References

- [1] Abrams M., Standridge C. R., Abdulla G., Williams S. and Fox E. A., 1995. Caching Proxies: Limitations and Potentials, Proc. 4th Int'l World Wide Conf.
- [2] Almcida V., Bestavros A., Crovella M. and Oliveira A., 1996. Characterizing Reference Locality in the WWW, In Proc. of the 4th Int'l Conf. on Parallel and Distributed Information Systems.
- [3] Bolot J. C. and Hoschka P., 1996. Performance engineering of the World-Wide Web: Application to dimensioning and cache design," Proc. of the 5th Int'l WWW Conf.
- [4] Cardellini V., Colajanni M. and Yu P. S., 1999. Dynamic Load Balancing on Web-Server Systems, IEEE Internet Computing, pp.28-39.
- [5] Kangasharju J., Ross K. W., 1999. A Clustering Structure for Reliable Multi-casting, Computer Communications and Networks, 1999. Proceedings, Eight International Conference, pp.378-383.
- [6] Leem C. S. and e-Biz Lab., 2000. e-Business File. YoungJin Biz.com, Seoul, Korea
- [7] Rayport J. F. and Jaworski B. J., 2000. e-Commerce. McGraw-Hill international edition.
- [8] Yua S. T. and Liu A., 2000. Next-generation Agent-enabled Comparison Shopping, Expert Systems with Applications 18, pp.283-297.
- [9] Zwass V., 1996. Electronic commerce: structures and issues. International journal of electronic commerce.

# A Scheduling Middleware for Data Intensive Applications on a Grid\*

Moo-hun Lee<sup>1</sup>, Jang-uk In<sup>2</sup>, and Eui-in Choi<sup>1</sup>

<sup>1</sup>Dept. of Computer Engineering, Hannam University,  
133 Ojeong-Dong, Daedeok-Gu, Daejeon, 306-791, Korea  
{mhlee, eichoi}@dblab.hannam.ac.kr

<sup>2</sup>Dept. of Computer and Information Science and Engineering,  
University of Florida, Gainesville, FL 32611, USA  
juin@cise.ufl.edu

**Abstract.** A grid consists of high-end computational, storage, and network resources that, while known a priori, are dynamic with respect to activity and availability. Efficient scheduling of requests to use grid resources must adapt to this dynamic environment while meeting administrative policies. This paper discusses the necessary requirements of such a scheduler and proposes a framework that can administrate grid policies and schedule complex and data intensive scientific applications. We present early experimental results for proposed a framework that effectively utilizes other grid infrastructure such as workflow management systems and execution systems. These results demonstrate that proposed a framework can effectively schedule work across a large number of distributed clusters that are owned by multiple units in a virtual organization.

## 1 Introduction

Grid computing is becoming a popular way of providing high performance computing for many data intensive, scientific applications. The realm of grid computing is beyond parallel or distributed computing, requiring the management of a large number of heterogeneous resources with varying, distinct policies and controlled by multiple organizations. Grid computing allows a number of competitive and/or collaborative organizations to share mutual resources, including documents, software, computers, data and sensors, to seamlessly process data and computationally intensive applications [1, 4].

Realizing the potential of grid computing requires the utilization of these dynamic resources. The execution of user applications must simultaneously satisfy both job execution constraints and system usage policies. Although many scheduling techniques for various computing systems exist [6, 7, 8, 9, 10, 12, 15] traditional scheduling systems are inappropriate for scheduling tasks onto grid resources. First, grid resources are geographically distributed and heterogeneous in nature. One of the central concepts of a grid is that of a virtual organization (VO) [4], which is a group of consumers and producers united in their secure use of distributed high-end computational resources

---

\* This work was supported by a grand from Ministry of Commerce, Industry and Energy.



towards a common goal. Second, these grid resources have decentralized ownership and different local scheduling policies dependent on their VO. Third, the dynamic load and availability of the resources require mechanisms for discovering and characterizing their status continually [4].

In the following sections, we describe a prototype Scheduling framework that incorporates these unique grid characteristics, and present methods for integrating this framework with related infrastructure for workflow management and execution. Section 2 discusses the unique requirements of a scheduling infrastructure for grids. Section 3 outlines a new scheduling system that provides all the scheduling requirements described in the previous section. Section 4 describes other grid services related to the scheduling service. Section 5 reviews currently available grid scheduling systems, and compare them with proposed a framework. Future works and conclusion are discussed on Section 6.

## 2 Requirements of a Grid-Scheduling Infrastructure

A grid is a unique computing environment. To efficiently schedule jobs, a grid scheduling system must have access to important information about the grid environment and its dynamic usage. Additionally, the scheduling system must meet certain fault tolerance and customizability requirements. This section outlines the different types of information the scheduling framework must utilize and the requirements a scheduler must satisfy.

### 2.1 Information Requirements

A core requirement for scheduling in the dynamic grid environment is to successfully map tasks onto dynamically changing resource environment while maximizing the overall efficiency of the system. The scheduling algorithm that performs this mapping must consider several factors when making its decision. Seven factors significantly affect this scheduling decision:

**Execution time estimation.** Because of the heterogeneous nature of grid resources, their real execution performance differs from the optimal performance characterized by analytic benchmarking [7]. However, the real execution time can be effectively estimated, even on heterogeneous grid resources, by statistically analyzing the performance during the past executions [11].

**Usage policies.** Policies, including authentication, authorization, and application constraints are important factors for maintaining resource ownership and security. The set of possible constraints on job execution can be various and can change significantly over time. These constraints can include different values for each job.

**Grid weather.** The scheduling system must keep track of dynamically changed load and availability of grid resources. In addition, faults and failures of grid resources are certain to occur. The state of all critical grid components must be monitored, and the information should be available to the scheduling system.

**Resource descriptions.** Due to the heterogeneous nature of the grid, descriptions of grid resource properties are vital. Such descriptions include configuration information such as pre-installed application software, execution environment information such as paths to local scratch spaces, as well as hardware information.

**Replica management.** The scheduling system must arrange for the necessary input data of any task to be present at its execution site. Individual data locations can have different performance characteristics and access control policies. A grid replica management service must discover these characteristics and provide a list of the available replicas for the scheduler to make a replica selection.

**Past and future dependencies of the application.** Grid task submission is often expressed as a set of dependent subtasks and modeled as a Directed Acyclic Graph (DAG). In this case, the subtasks are represented by nodes in a graph, and the dependencies by branches. When allocating resources to the subtasks, inter-task dependencies affect the required data movement among resources.

## 2.2 System Requirements

While the kinds of information above should be available to the system for efficient grid scheduling, the following requirements must be satisfied in order to provide efficient scheduling services to a grid Virtual Organization (VO) community.

**Distributed, fault-tolerant scheduling.** Clearly, scheduling is a critical function of the grid middleware. Without a working scheduling system (human or otherwise), all processing on the grid would quickly cease. Thus, any scheduling infrastructure must be strongly fault-tolerant and recoverable in the inevitable case of failures. This need for fault tolerance consequently gives rise to a need for a distributed scheduling system.

**Customizability.** Within the grid, many different VOs will interact within the grid environment and each of these VOs will have different application requirements. The scheduling system must be customizable enough to allow each organization with the flexibility to optimize the system for their particular needs.

**Interoperability with other scheduling systems.** Any single scheduling system is unlikely to provide a unique solution for all VOs. In order to allow cooperation at the level of VOs, for example in a hierarchy of VOs or among VO peers, the scheduling system within any single VO should be able to route jobs to or accept jobs from external VOs subject to policy and grid information constraints.

**Quality of service(QoS).** Multiple qualities of service may be desirable as there are potentially different types of users. There are users who are running small jobs that care about quick turnaround time or interactive behavior from the underlying system. On the other hand, large production runs may be acceptably executed as batch jobs.

## 3 Proposed a Framework

**Fig. 1** shows a general architecture of client and server.

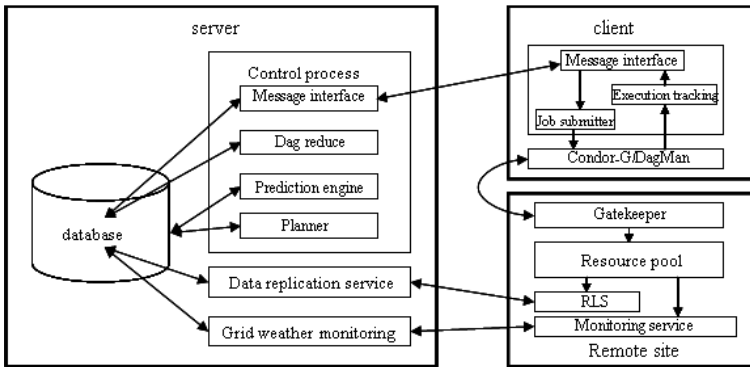


Fig. 1. general architecture

### 3.1 Highlights of Architecture

**Easily accessible system.** The scheduling system consists of two parties, the client and the server. The separation supports system accessibility and portability. The client is a lightweight portable scheduling agent that represents the server for processing scheduling requests. The client also interacts with a user for scheduling procedure. It provides an abstract layer to the scheduling service, and supports a customized interface to accommodate user specific functionalities.

**Automated procedure and modulated architecture.** System consists of multiple modules that perform a series of refinement on a scheduling request. The procedure begins from a ‘start’, and the final state should be annotated with ‘finished’, which indicates the resource allocation has been made to a request. Each module takes a request, and changes its state according to the functionality of itself.

**Robust and recoverable system.** The server adopts database infrastructure to manage scheduling procedure. Database tables support inter-process communication among scheduling modules in the system. A module reads scheduling state of a request from the tables, edits the state, and writes the modification to the tables. It also supports fault tolerance by making the system easily recoverable from internal component failure.

**User interactive system.** System supports user interaction to its resource allocation procedure. A user submits a request with QoS requirement. The requirement may specify resource usage amount and period. It is challenging to be able to satisfy the specification in a dynamically changing grid environment.

**Platform independent interoperable system.** A scheduling system should expect the interaction with systems on various kinds of platforms in a heterogeneous environment. System adapts communication protocols based on XML such as SOAP and XML-RPC to satisfy the requirement.

### 3.2 Client

Client interacts with both the scheduling server that allocates resources for task execution, and a grid resource management system such as DAGMan/Condor-G [14]. To

begin the scheduling procedure, a user passes execution request to client. The request is in the form of an abstract DAG that is produced by a workflow planner such as the Chimera Virtual Data System [5]. The abstract plan describes the logical I/O dependencies within a group of jobs. The client sends scheduling request to the server with a message containing the DAG and client information. After receiving resource allocation decision from the server, the client creates an appropriate request submission file according to the decision. The client submits the file to the grid resource management system.

In order to achieve a user’s QoS requirement proposed a framework implements interactive resource allocation. The client as a scheduling agent negotiates QoS satisfaction level with the user. Proposed a framework presents the user resource allocation decision such as estimated execution time, resource reservation period and amount according to the current grid resource status. Then the user should decide acceptance of the suggestion.

The tracking module in the client keeps track of execution status of submitted jobs. If the execution is held or killed on remote sites, then the client reports the status change to the server, and requests re-planning of the killed or held jobs. The client also sends the job cancellation message to the remote sites on which the held jobs are located. **Table 1** shows the functionalities and proposed a framework API’s.

**Table 1.** Client functionalities for interactive job scheduling and execution tracking

Functions	proposed a framework API’s	Parameters
Execution request	job_submit (String file_loc, String sender) //Client submits this request to client	file_loc: abstract dag file location sender: request sender information
Scheduling request	send_request (String dagXML, String msgType, String sender) // Client sends this request to server	dagXML: dag in XML format msgType: request type sender: request sender information
Admission control	send_msg (String msgXML, String msgType) send_msg (String msgXML, String msgType, String sender) //User interact for resource allocation	msgXML: message in XML format msgType: message type sender: message sender info.
Submission request	createSubmission (String jobInfo) //Client create a submission file. submit_job (String rescAlloc, String submitter) //Client send the file to DAGMan/Condor-G	JobInfo: scheduling decision information rescAlloc: job submission file Submitter: job submitter information
Execution tracking	updateStatus (int jobId) String status = getJobStatus (int jobId) //Update the status of the job with the information //from a grid resource management service	jobId: ID of a job that is currently running on a grid resource. The ID is assigned by the grid resource management system. status: the status of job, which the grid resource management system provides.

### 3.3 Server

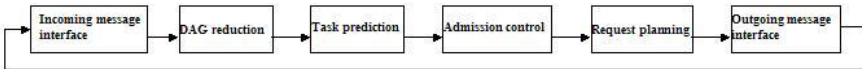
Server as a major component of the system performs several functions. The current version of the server supports the following functions. First, it decides how best to allocate those resources to complete the requests. Second, it maintains catalogs of data, executables and their replicas. Third, it provides estimates for the completion time of the requests on these resources. Fourth, the server monitors the status of its resources.

System adapts finite automation for scheduling status management. The scheduler moves a DAG through predefined states to complete resource allocation to the jobs in the DAG. The order of states in the table indicates the scheduling procedure. The server has a control process, which completes the scheduling by managing several inner service modules such as resource monitoring interface, replicate management interface, prediction, message handling, DAG reducing and planning. Each module performs its corresponding function to a DAG or a job, and changes the state to the next according to the predefined order of states. **Table 2** shows server functions described in this section.

**Table 2.** Server functions for resource allocation

Modules	proposed a framework API's	Parameters
Message Handling Module	String incMsg = inc_msg_wrapper () out_msg_wrapper (String msg) msg_parsing (String msg) msg_send (String msg, String msgType String dest, String sender) //These functions are to send, receive and parse messages. //The message handling module is a gateway to server	incMsg: incoming message msg: message in XML format msgType: message type dest: receiver information sender: sender information
DAG Reducer	dag_reducing (int dagId) //For each of all the jobs in the dag, call replica management //service to check if all the outputs of the job exist. If exist, //then reduce the job and all the precedence of the job.	dagId: ID of a dag in dag table. The state of the dag is unreduced
Prediction Engine	String est_info = exec_prediction(int jobId) //This function provides estimated information such as //execution time, resource usage (CPU, storage etc.)	est_info: estimated data in XML string format jobId: ID of a job in job table
Planner	Planning (int jobId, String strategy) //It is to allocate resources to jobs according to scheduling //strategy.	JobId: ID of job to be planned strategy:scheduling algorithm

**Control Process.** The main function of the control process is to launch the necessary server-side service modules to process resource allocation to a request. The control process functions as finite state automation performs. Stateful entities such as DAG's and Jobs are operated and modified by scheduling modules. This architecture for Server, in which the control process awakens modules for processing stateful entities, provides an extensible and easily configurable system for future work. **Fig. 2** shows overall structure of control process in proposed a framework. The controller checks the state of jobs that are currently in the procedure of scheduling.



**Fig. 2.** Overall structure of control process

**Message Handling Module.** The message handling function is to provide a layer of abstraction between the internal representation of the scheduling procedure and the

external processes. Additionally, the function is responsible for maintaining a list of all the currently connected clients, for ensuring that the list is kept accurate and for directing I/O from the various internal components to these various clients. The server maintains database tables for storing incoming and outgoing messages. Control process invokes incoming or outgoing message interfaces to the tables for retrieving, parsing and sending the messages.

**DAG Reducer.** The DAG reducer reads an incoming DAG, and eliminates previously completed jobs in the DAG. Such jobs can be identified with the use of a replica catalog. The DAG reducer simply checks for the existence of the output files of each job, and if they all exist, the job and all precedence of the job can be deleted. The reducer consults replica location service for the existence and location of the data.

**Prediction Engine.** The prediction engine will provide estimates of resource use. It estimates the resource requirements of the job based upon historical information, size of input if any, and/or user provided requirements. In the first implementation, this is constrained to overall execution time by application and site. A simple average and variance calculation is used to provide an initial estimation scheme; this method could be made more intelligent and robust in future implementations. When the prediction engine is called, it selects a DAG for prediction, estimates the completion time of each job, and finally from this data, estimates the total completion time of the DAG.

**Planner.** The planner module creates an execution plan of the job whose input data are available. According to the data dependency a job should wait until all the precedent finish to generate output files. The execution plan includes several steps:

1. *Choose a set of jobs that are ready for execution according to the input data availability.*
2. *Decide the optimal resources for the job execution. The planner makes resource allocation decision for each of the ready jobs. The scheduling is based on resource status and usage policy information, job execution prediction, and I/O dependency information.*
3. *Decide whether it is necessary to transfer input files to the execution site. If necessary, choose the optimal transfer source for the input files.*
4. *Decide whether the output files must be copied to persistent storage. If necessary, arrange for those transfers.*

After the execution plan for a job has been created, the planner creates an outgoing message with the planning information, and passes the message to a message-handling module.

## 4 Relationship with Other Grid Services

In this section we discuss the interaction between proposed a framework scheduling service and other grid computing services.

## 4.1 Data Replication Service

The data replication service is designed to provide efficient replica management. The Globus Replica Location Service (RLS) [2] provides both replica information and index servers in a hierarchal fashion. In addition, GridFTP [13] is Grid Security Infrastructure (GSI) enabled FTP protocol that provides necessary security and file transfer functions. Initial implementations of proposed a framework will both make use of RLS and GridFTP for replica and data management. **Table 3** shows proposed a framework API's for accessing replica information through RLS.

**Table 3.** Proposed a framework API's for accessing data replicas through RLS service. In the table PFN or pfn represents physical file name, and lfn means logical file name.

API's	Parameters
Vector pfn = getPFN (String lfn) //It returns PFN mappings for the given lfn.	pfn: a list of physical file names lfn: logical file name
createMapping (String lfn, Sstring pfn) //It creates mapping for the given lfn and //pfn in the RSL service database.	lfn: logical file name pfn: physical file name

## 4.2 Grid Monitoring Interface

The resource allocation decision made by the planner and the replication site selection by the data replication service depends on the information provided through the monitoring interface of proposed a framework. As such, the interface provides a buffer between external monitoring services and the proposed a framework scheduling system. In order to accommodate the wide-variety of grid monitoring services, the interface is developed as an SDK so that specific implementations are easily constructed.

Grid monitoring service will be used to track resource-use parameters including CPU load, disk usage, and bandwidth; however, in addition, a grid monitoring service could possibly also collect policy information provided from each site, including resource cost functions and VO resource use limits. Additional interface modules will be developed to gather VO-centric policy information, which may be published and maintained by a VO in a centralized repository.

## 5 Related Works

In this section we review currently available grid scheduling systems, and compare them with proposed a framework. Proposed a framework supports additionally necessary functionalities in the scheduling middleware, while most of them provide proficient features.

### 5.1 Pegasus

Pegasus [3] is a configurable system that can map and execute DAGs on a grid. Currently, Pegasus has two configurations. The first is integrated with the Chimera Virtual Data System. The Pegasus system receives an abstract DAG file from Chimera.

In its second configuration, the Pegasus system performs both the abstract and concrete planning simultaneously and independently of Chimera. It then uses AI planning techniques to choose a series of data movement and job execution stages that aims to optimally produce the desired output. The result of the AI planning process is a concrete plan (similar to the concrete plan in the first configuration) that is submitted to DAGMan for execution.

## 5.2 Condor

The Condor Team continues to develop Condor-G and DAGMan. Recently, to improve its just-in-time planning ability, DAGMan has been extended to provide a call-out to a customizable, external procedure just before job execution. This call-out functionality allows a remote procedure to modify the job description file and alter where and how the job will be executed. However, as DAGMan increase in functionality, DAGMan itself could become a scheduling client and communicate through this and other callouts to the scheduling server directly.

## 6 Conclusions and Future Work

A novel grid scheduling framework for computing has been proposed in this paper and an initial implementation presented. Resource scheduling is a critical issue in executing large-scale data intensive applications in a grid. Due to the characteristics of grid resources, we believe that traditional scheduling algorithms are not suitable for grid computing. This document outlines several important characteristics of a grid scheduling framework including execution time estimation, dynamic workflow planning, enforcement of policy and QoS requirements, VO-wide optimization of throughput, and a fully distributed, fault tolerant system.

Our proposed system currently implements many of the characteristics outlined above and provides distinct functionalities, such as dynamic workflow planning and just-in-time scheduling in a grid environment. It can leverage existing monitoring and execution management systems. In addition, the highly customizable client-server framework can easily accommodate user specific functionality or integrate other scheduling algorithms, enhancing the resulting system. This is due to a flexible architecture that allows for the concurrent development of modules that can effectively manipulate a common representation for the application workflows. The workflows are stored persistently in database using this representation allowing for development of a variety of reporting abilities for effective grid administration.

## References

1. Avery, P., Foster, I. The GriPhyN Project: Towards Petascale Virtual-Data Grids. The 2000 NSF Information and Technology Research Program 2000.
2. Chervenak, A., et al, Giggle: A Framework for Constructing Scalable Replica Location Services. To appear in Proceedings of SC2002 Conference, November 2002.
3. Deelman, E., Blythe, J., Gil, Y., Kesselman, C. Pegasus: Planning for Execution in Grids. Technical Report GriPhyN-2002-20, Nov. 2002.



4. Foster, I., Kesselman, C., Tuecke, S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International J. Supercomputer Applications*, 15(3), 2001.
5. Foster, I., Voeckler, J., Wilde, M., Zhao, Y. Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation. The 14th International Conference on Scientific and Statistical Database Management (SSDBM 2002), 2002.
6. Gerasoulis, A., Yang, T. On the Granularity and Clustering of Directed Acyclic Task Graphs. *IEEE Trans. Parallel and Distributed Systems* 5(9), 951-967, 1994.
7. Ghafoor, A., Yang, J. A Distributed Heterogeneous Supercomputing Management System. *Computer*, 26(6), 78-86, June 1993.
8. M. Kaddoura and S. Ranka. Runtime Support for Parallelization of Data-Parallel Applications on Adaptive and Nonuniform Environments, *Journal of Parallel and Distributed Computing (Special Issue on Workstation Clusters and Network-based Computing)*, June 1997, pp. 163-168.
9. Karypis, G., Kumar, V. A fast and high quality multilevel scheme for partitioning irregular graphs. Technical Report, Department of Computer Science, University of Minnesota, 1995.
10. Kwok, Y., Ahmad, I. Static Scheduling Algorithms for Allocating Directed Task Graphs to Multiprocessors. *ACM Computing Surveys*, 31(4), December 1999
11. Li, Y. A., Antonio, J. K., Siegel, H. J., Tan, M., Watson, D. W. Determining the Execution Time Distribution for a Data Parallel Program in a Heterogeneous Computing Environment. *Journal of Parallel and Distributed Computing* 44 (1), 35-52, 1997.
12. S. Ranka, M. Kaddoura, A. Wang, and G. C. Fox. Heterogeneous Computing on Scalable Heterogeneous Systems, *Proceedings of Supercomputing 93*, pp. 763--764.
13. Thomas Sandholm, Jarek Gawor, Globus Toolkit 3 Core – Agrid Service Container Framework, <http://www-unix.globus.org/toolkit/documentation.html>
14. Douglas Thain, Todd Tannenbaum, and Miron Livny, “Condor and the Grid”, in Fran Berman, Anthony J.G. Hey, Geoffrey Fox, editors, *Grid Computing: Making The Global Infrastructure a Reality*, John Wiley, 2003. ISBN: 0-470-85319-0
15. Yang, T., Gerasoulis, A. DSC: Scheduling parallel tasks on an unbounded number of processors. *IEEE Trans. Parallel and Distributed Systems*, 5(9), 951-967, 1994.

# Development of a Monitoring Module for ITS (Intrusion Tolerant System)

Wankyung Kim<sup>1</sup>, Wooyoung Soh<sup>1</sup>, Hwangrae Kim<sup>2</sup>, and Jinsub Park<sup>3</sup>

<sup>1</sup> Department of Computer Engineering, Hannam University,  
{wankk12, wsoh}@hannam.ac.kr

<sup>2</sup> Department of Computer Engineering, Kongju National University  
plusone@kongju.ac.kr

<sup>3</sup> Department of Computer Engineering, Daejeon University  
Daejeon, S. Korea  
jspark@dju.ac.kr

**Abstract.** Security systems need be tested on the network, when they are Recently the cyber terror has become a crucial issue of the national security over the world. Cyber terror happens to attack the critical infrastructures that might result in a malfunction of a nationwide information network. Several techniques against such cyber terror have been proposed including firewall system, intrusion detection system, virtual private network, and intrusion tolerant system (ITS). Among those ITS is the one that tries to deliver useful services even if an attacker with malicious purpose were able to intrude the system. This paper implements a monitoring module of the system security information for ITS to provide continuous services. This module monitors the information of the hardware, memory, CPU processes, network traffic and others. This information being monitored will be reported to the administrator to maintain the ITS's services.

## 1 Introduction

Many of the information security techniques such as intrusion detection and intrusion interruption, which are against the security violation accidents, have been developed. These techniques can be used for the known vulnerabilities, but they have limitations that cannot effectively detect unknown vulnerabilities and attacks. Aside from these, it could also cause some serious problems by stopping services when violation accident happened. There is a requirement on how to confront the unknown vulnerabilities and attacks and ITS is presented as one of the confrontation method [1].

ITS provides application that tries to deliver useful service, even if an attacker with malicious purpose were able to intrude the system [2]. ITS can be constructed by applying various kinds of technique to satisfy the relative characteristic of the system. Several techniques have been used to detect and prevent intrusions, and ITS is used to keep the essential services running even if an attack succeeds. Intrusion tolerant is required in systems that must complete important service necessary in any situation. ITS must be able to prepare more detections than other security systems.

This paper is organized as follows. Section 2 presents the location and role of ITS in information security. Section 3 describes the chosen important services, relocation

of resource and structure of a model proposed by KISA. Section 4 also presents the design and implementation of monitoring technique of the resource information at real time to keep essential service and to stop others. Section 5 describes conclusion

## 2 ITS (Intrusion Tolerant System)

ITS is an information security technique like vulnerability analysis, IDS and Firewall that attempts to countermeasure attacks and intrusions to information system.

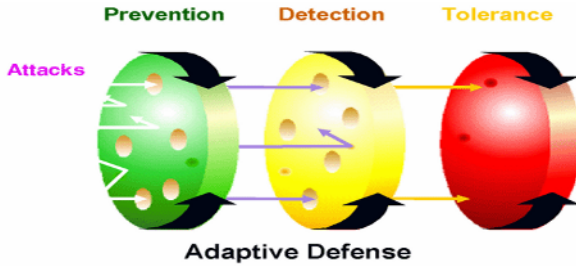


Fig. 1. Layers level of ITS

Fig. 1. shows how information security techniques are applied to secure information infrastructure. The first layer is Prevention layer. This layer presents the prevention of attack like the vulnerability analysis of system, firewall technique and others. The second layer is the detection layer. This layer presents countermeasure technique and detection from intrusion attacks through the prevention layer. Nowadays, most information security techniques are implementing these two layers. However, these two layers can be compromised by unknown vulnerability attacks. The last layer is the Tolerance layer, the Intrusion Tolerant System using the reproduction system, and presents countermeasure against unknown vulnerability attacks. This layer countermeasures attacker who has some high skills in intrusion technique [3]).

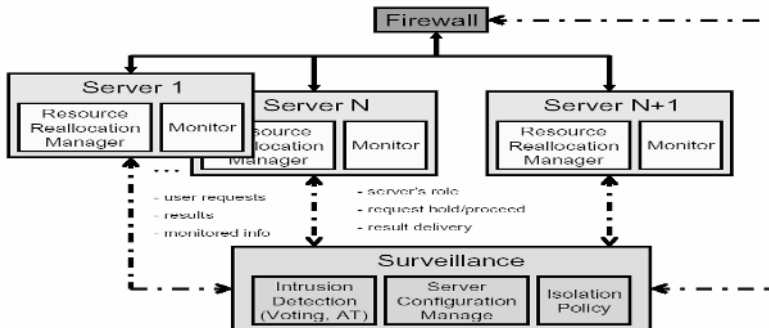


Fig. 2. The Structure of Overlapping System

### 3 A Model of ITS

A model of ITS developed by KISA (Korea Information Security Agency) provides a way of keeping a chosen essential service to proceed by relocating of resources even if one of the computing nodes is intruded [4]. If a node is intruded, it will lose network resources that communicate with other computer resources such as processes, memory and others. To adapt on this situation, the model by KISA advises server administrators to choose important service before intrusion occurs. That means, if secured minimum resources can execute the selected service, the administrator can keep the service even after an intrusion succeeded. Certainly other services could give up.

Relocation of resources is working on the same node. If an essential service shorts resources, another service will give its spare resource to the essential service and then it stops running. The resources are relocated dynamically and the essential service works on. But, it is hard to estimate how much CPU processes, memory and network resources are needed. So, after define the Baseline and in normal case, the amount of resources being by essential service that exceeds the baseline. The accumulation time of baseline is a critical value to estimate the exceed baseline, and to correspond in change of momentary resources amount being used.

If it could not secure necessary resources to keep the essential services on the same node, present N+1 redundancy server to keep the essential service working [5]. N+1 redundancy structure Fig. 2. is a way that uses receptivity and choice method together; this is consisted of N active nodes and a standby back-up node.

A back-up node prepares after the first intrusion, and at the same time surveillance and control the extra network which approach the limited normal users while solving the problem, so reconstruction is possible in a very short time and could provide against another intrusion.

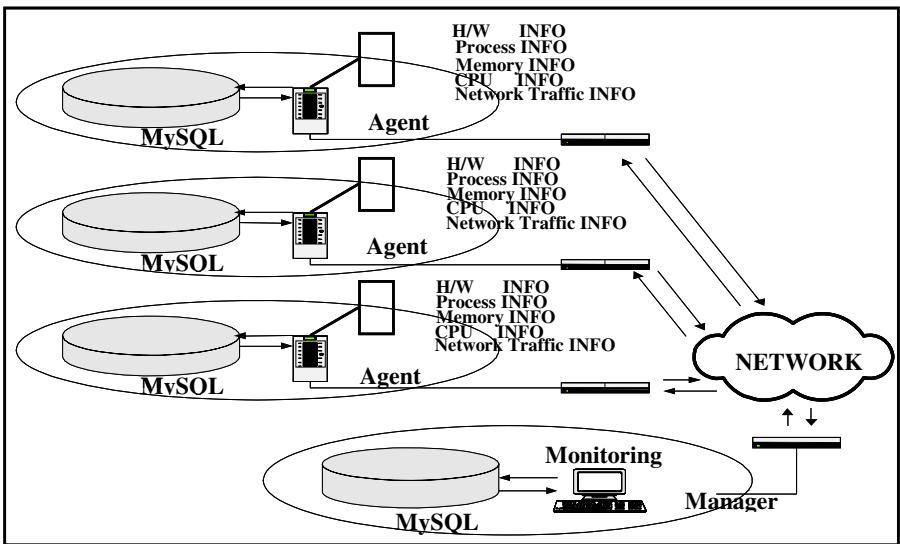


Fig. 3. Structure of Network for Monitoring Module

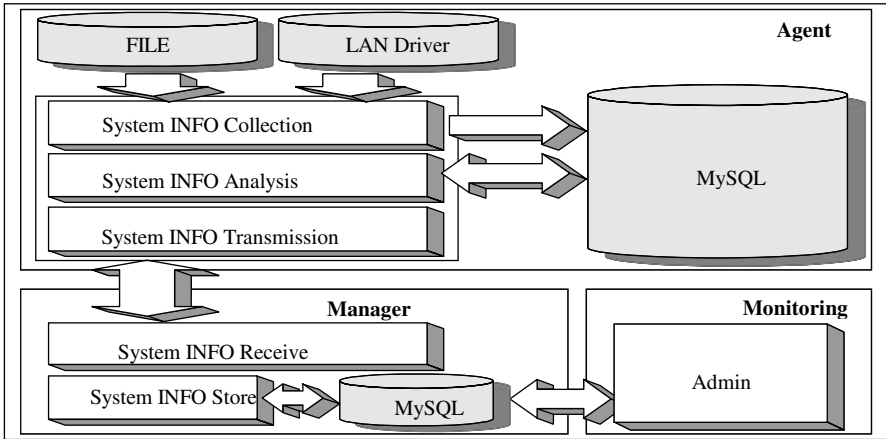


Fig. 4. Implementation of Security Information Monitoring

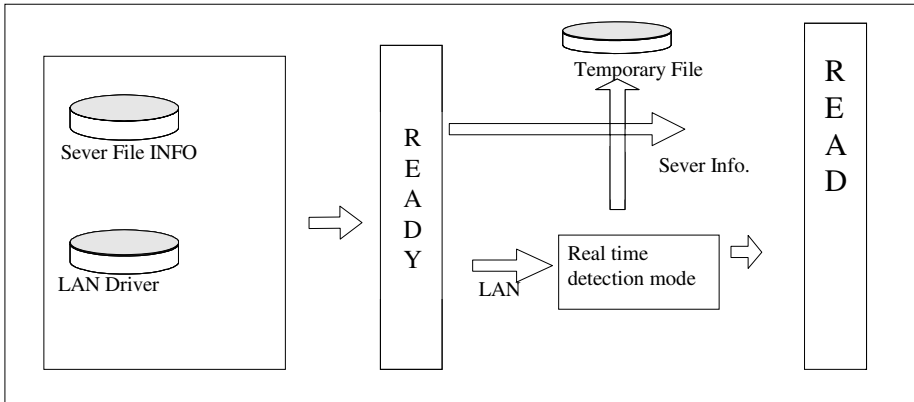


Fig. 5. Agent System of Information Collection

This paper implements a monitoring module for the system security information to have a continuous services in ITS.

#### 4 Design and Implementation of the Monitoring Module

This paper implemented an Agent and Manager that could be applied in network environment. Agent can be found to be many subnets. It saves principal security information of the system through socket communication from constitutive Agent on real time. And a monitoring manager module is in GUI. Fig. 4. presents the whole system structure that is implemented in this paper. It also presents the system that gets resources information. This system structure is divided into three parts, the Agent part, Manager part, and the Monitoring part. Agent module analyzes and collects

principal information of the system from LAN driver or system file, and saves the system's security information to Agent module DB.

Agent module takes charge all the resource collection function. Each Agent server collects various data of CPU possession, the amount of memory being used, hardware information and network traffic information, and then submits to Manager Server in real time. In Fig. 5., it collects data from a file in the main server system and reads the amount of traffic information and analysis network traffic. Agent module is implemented in LINUX system.

In LINUX, system information directory is in /proc, the CPU information is in /proc/stat file and the memory information is in /proc/meminfo. The amount of transmission and reception of packet for networking is in /proc/net/dev. There are a lot of hardware information, for example, the CPU information is in /proc/cpuinfo, PCI information is in /proc/pci and the hard disk information is in /proc/ide

The Manager module waits for Agent to connect on standby, and gets system information data from Agent after connecting. These data are analyzed and saved in memory on each field. Fig. 6. presents the data reception function of system between the Agent and Manager modules.

Monitoring Module requires and receives various system information data from DB that is processed by Manager. Monitoring Module create various statistical information and the system information data supplies administrator a monitoring function about each Agent system. Fig. 7. shows system information monitoring function between Agent and Manager Modules. The Manager UI is implemented by Delphi, Fig. 8. shows Agent IP address on left screen after connected. Agent IP address part shows sources and devices being used by Agent server. Also it shows information of hard disk partition, amount of network traffic. Fig. 9. shows device condition of Agent server such as hardware, CPU, and hard disk. In this UI, we analysis amount of network traffic and notice a reason when network overload is happened. This UI is going to use ITS for necessary services and amount of resources basically.

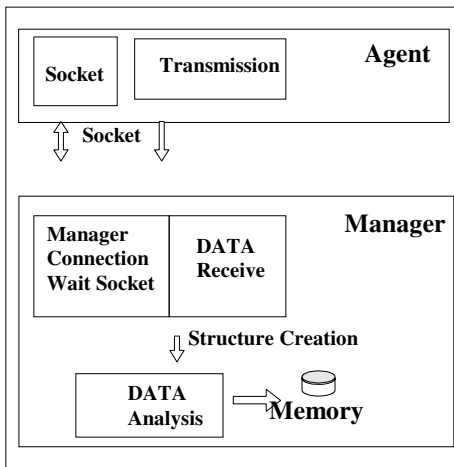


Fig. 6. Manager System of Information Reception

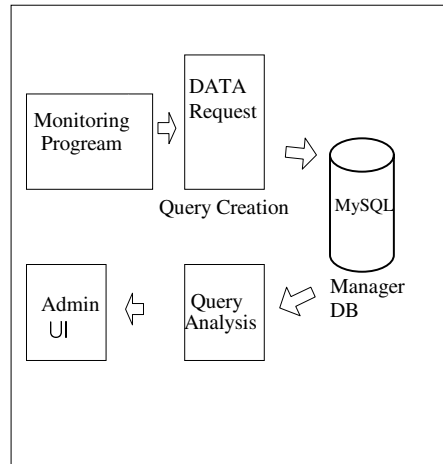


Fig. 7. Request of Monitoring Module Information

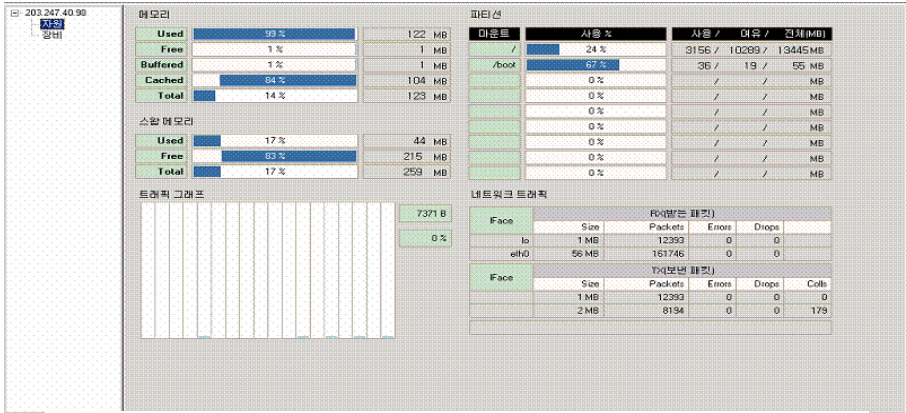


Fig. 8. The UI that showing Usage of Agent Server Resource

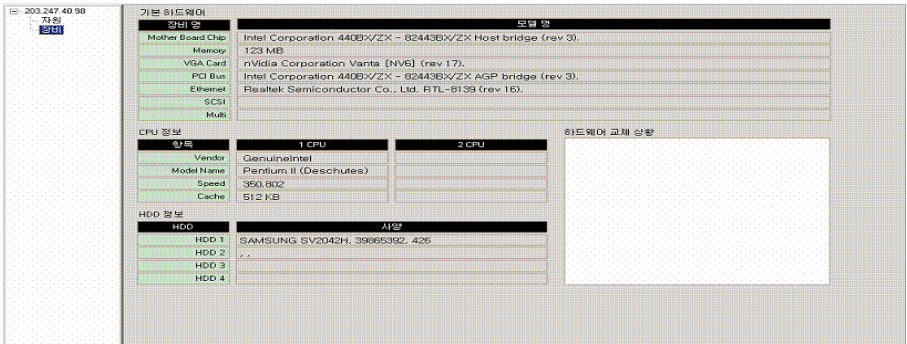


Fig. 9. The UI that Showing Usage of Agent Server Device

## 5 Conclusion

Many kinds of information secure techniques such as intrusion detection and intrusion interruption techniques, which prevent violation accidents, have been developed. These techniques show good results in detecting vulnerabilities which already known, but there are limitations of these techniques which cannot effectively detect unknown vulnerabilities and attacks. There is a requirement on how to confront the unknown vulnerabilities and attacks and Intrusion Tolerant System (ITS) is presented as one of the confrontation method.

This paper implements a monitoring module for the system security information to have a continuous services in ITS. This module monitors the information of the hardware, memory, CPU processes, network traffic and others. This information being monitored will be reported to the administrator to maintain the system's services. The Agent of this monitoring module is working only on Linux system, so, it is necessary another Agent module for Unix and Windows system.

## References

- [1] Chung Sub Choi, Kyoung Gu Lee and Hong Kun Kim, "Study of Intrusion Tolerant Technology", Korea Institute of Security & Cryptology, Vol. 13, No 1, February 2003.
- [2] Matti A. Hiltunen, et. al., "Survivability through Customization and Adaptability: The Cactus Approach", DARPA Information Survivability Conference & Exposition, 2000.
- [3] <http://www.kisa.or.kr>
- [4] Byung Jun Min, "Development of Computer Security Technique", kisa, Symposium on Information Security, July 2002.
- [5] Byung Jun Kim and Sung Gee Kim, "Proposal of Scheme Relocation Resource and Server Duplicate against DoS", Korea Information Processing Society, Vol. 10-A, No. 1, March 2003.



# Design of DRM-LMS Model in M-Learning Environment\*

Mingyun Kang<sup>1</sup>, Seoksoo Kim<sup>1,\*</sup>, Gil-Cheol Park<sup>1</sup>, Geuk Lee<sup>2</sup>, and Minwook Kil<sup>3</sup>

<sup>1</sup>Hannam University, Department of Multimedia Engineering, Postfach , 306 791  
133 Ojeong-Dong, Daedeok-Gu, Daejeon, South Korea  
{Kang}card7s@paran.com, {Kim}sskim@hannam.ac.kr

<sup>2</sup>Hannam University, Department of Computer Engineering, Postfach , 306 791  
133 Ojeong-Dong, Daedeok-Gu, Daejeon, South Korea  
leegeuk@hannam.ac.kr

<sup>3</sup>Mun Kyung College, Dept. of Computer & Image Information, MunKyung, South Korea  
mwkil@mkc.ac.kr

**Abstract.** The development of the wireless Internet and digital contents is activating Internet business and, in this situation, most online contents providers are supplying diverse contents attracting users' attention. These contents include not only texts but also multimedia such as music, images and videos. Various multimedia data are being applied to numerous areas like Internet broadcasting, education, news, sports, tourist information and experts' consulting and creating new services in the virtual space. The present study examined M-learning, which is e-learning system in wireless Internet environment, and analyzed DRM technology for contents protection in M-learning. We planned a DRM-LMS system focused on contents protection, which has been insufficient until now, by designing basic components of e-learning user registration, course registration, report evaluation, personal learning schedule and classroom platform as well as course contents service using DRM to support mobile systems in wireless environment.

## 1 Introduction

Information technology has brought revolutionary changes in business and is causing digital reforms in education as well. Changes in education represented by e-Learning are transforming not only the technological aspect of educational environment but also its paradigm. Matti Hamalainen said, "Education in the past drew people into education but now it brings education to people." The rise of the Internet crated many new businesses and expanded existing ones. In addition, under the concept of mobile computing, Internet environment is switching from fixed one to into wireless and mobile one, which enables e-learning in the field. Education system using mobile devices is called M-learning. M-learning will be a part of next-generation e-learning. With evolution to M-learning, various digital education contents are being developed, and the convenience of digital information is expected to increase demands for digital education contents and to facilitate the production and distribution of digital education contents.

---

\* "This work was supported by a grant No. (R12-2003-004-03003-0) from Ministry of Commerce, Industry and Energy".

\* Corresponding author.

Turning the age of the wireless Internet, wireless Internet technology will be improved in addition to M-learning and this will accelerate the growth of the digital contents industry. The growth of digital contents distribution will activate various types of businesses including e-commerce, content production and electronic payment. Digital information, however, is easily reproduced, modified and distributed illegally and these illegal actions bring loss to copyright holders and distributors. Difficulty in revenue creation is a serious problem in the contents industry. Moreover, as contents are distributed and its prices are paid through the Internet, critical problems are emerging with regard to security. Thus, we need to develop a system to prevent illegal distribution and protect the revenues of copyright holders and distributors [1], [2].

## **2 Related Works**

### **2.1 Necessity of Contents Protection**

With the development of digital contents accelerated by the progress of the wireless Internet, Internet business is booming and online contents providers are supplying a variety of contents stimulating users' interest. These contents are texts as well as multimedia such as music, image and video. Multimedia data are applied to countless areas including Internet broadcasting, education, news, sports, tourist information and experts' consulting, creating new services in the virtual space. Moreover, companies are struggling to move toward knowledge-based management strategies and, to support the transition, are building various systems like KMS, EDMS and PDM and accumulating intellectual assets in the systems. Thank to such efforts, employees become able to share knowledge and find information promptly and conveniently. The government is also accelerating the digitization of information (library informatization by the Ministry of Culture and Tourism, knowledge resource networking project by the Ministry of Information and Communication, etc.). Furthermore, the super-speed information communication network is spreading digital paradigm as a new lifestyle throughout the whole society, and the emergence of digital broadcasting, digital signature, electronic libraries, electronic books, electronic settlement, electronic payment, etc. is an aspect representing the new paradigm. Compared to traditional analogue ones, digital contents have many advantages in terms of production, processing, publication and distribution but copyright protection for digital products is a critical issue because digital contents can be easily reproduced. Copyright protection in digital paradigm cannot be provided with legal/institutional systems applied to analogue contents. Thus, we need a new legal system for copyright protection as well as technological devices for practical protection of digital contents. In response to the urgent demand, methods for protecting digital contents copyrights and preventing illegal distribution are under development. MIT selected this type of technologies as one of 10 promising technologies and many researches began to be made on this area as an independent academic discipline. As digital contents business rises as a core next-generation industry, DRM-related R&D and commercialization are essential tasks to be carried out [3],[4].

## 2.2 DRM Technologies

DRM (Digital Rights Management) is to provide services related to the production, distribution and utilization of digital contents including the prevention of illegal use of documents, automatic billing, payment agent and additional services through encrypting documents and controlling access rights so that only authorized users can open the documents according to their right. DRM is mostly focused on preventing at the root the illegal reproduction of contents by limiting access or illegal distribution by tracing contents through various types of copyright information and user ID embedded in the contents. Simply speaking, DRM technology protects contents by distributing them in an encrypted form. Contents exist in an encrypted form and are decrypted temporarily by authorized users and, even if they are reproduced, they cannot be opened by unauthorized users. In this way, the technology suppresses illegal reproduction [5].

Encryption for contents protection is called packaging. Commercial digital contents products are protected through packaging. Protected contents are delivered to consumers safely through contents suppliers and distributors and they can be used only when the consumers satisfy specific conditions (payment of price, etc.). Of course, protected contents cannot be accessed by those who fail to meet the conditions. All transactions and uses of contents are reported to the DRM server and this data shows how and by whom the contents have been traded and used. Thus DRM technology, applied to contents e-commerce models, prevents illegal reproduction of digital contents, supports the value chain that guarantees proper profits promised among those involved in the transaction of digital contents and enables suppliers to adopt various business models as well as flexible service models in response to consumers' demands. Various concepts of technologies are required depending on application area and security level. The following five categories of technologies are most essential in companies' efforts to prevent information leakage.

- (1) Encryption technologies: Various encryption technologies are used including encryption, electronic signature and authentication and key distribution to authenticate contents and contents users, enforce transaction and usage rules and confirm transactions and uses (non-repudiation). Before publication, contents are protected safely through packaging. Packed data contain contents, metadata and decryption information. A key used in contents encryption is processed for safe protection so that only the authorized user (or user's system) can access, and generates decryption information. Metadata defines business rules on the distribution and use of contents and the rules are also protected cryptographically to prevent alteration or modification [6],[7].
- (2) Key distribution and management: Safe key management and distribution mechanism is required to guarantee the reliability of encryption technology used to protect contents. The biggest characteristic that distinguishes the key management in DRM from other encryption systems is that the key should be kept from all users including the supplier, the distributor and the consumer. DRM key distribution methods can be divided into the symmetric key method and the public key method. In the symmetric key method, load is concentrated on a key distribution server and the server is involved in all contents transactions. On the other hand, the public key method is advantageous in terms of distribution, scalability and in-

teroperability but it requires public key infrastructure (PKI). Therefore, a proper key management mechanism should be selected according to the characteristics of contents and application environment. For example, key management mechanism where load is concentrated on a key distribution server is not desirable if contents are distributed extensively and many role objects are involved in the flow of contents distribution as in the distribution of electronic books and music.

- (3) TRM (Tamper Resistant Module): A factor that hinders contents protection is that the contents must be decrypted at a moment for use. If the decryption key or decrypted contents are exposed to users in processing or using the contents, contents may leak out without breaking into the encryption technology. TRM is a software or hardware module like a black box that hides detailed operations and stops its operation if it is modified. In DRM, software reengineering using a debugging tool is prevented by applying TRM technology to modules dealing with information on access rights, keys and decrypted contents. TRM is expected to be used as a key technology determining the safety of DRM system.
- (4) Digital watermarking: Watermarking technology, which has been spotlighted as one of copyright protection technologies in the last three years, inserts an invisible mark into digital contents as an evidence of ownership. Watermarking has difficulties in finding a profit model because it is applicable only after illegal reproduction has been made and the safety of its algorithm has not been proved. Recently a new soft watermarking technology was commercialized that is used to prevent the fabrication of certificates. Because detection is possible only after an incident has happened, however, the technology is not used widely. Currently watermarks are commonly used in printed documents.
- (5) Hooking: When an application (e.g. PowerPoint) is executed, it occupies a memory space, uploads necessary functions and controls actions made by the user. If the user makes a specific action (copy & paste, print, save, capture, etc.) the application replaces the address of DLL uploaded to the memory with that provided by DRM solution and controls the functions of the application and, by doing so, controls the use of documents according to the user's access right. This is memory hooking widely used in DRM solutions [1],[5],[8].

### 3 Design of LMS Model

MDRM-learning is a system designed to protect contents using DRM technology based on e-learning and M-learning. We designed Basic structure and 12 components.

- (1) Account Manager: Register a user or a group of users and change information on registered users.
- (2) Notice Manager: Have multiple bulletin boards and manage notices, lectures, questions and other tasks related to bulletin board management.
- (3) Report up: Upload reports, distinguishing managers and learners.
- (4) Study Evaluation: Perform course application, report evaluation, online assessment, questionnaire survey, etc
- (5) Course Manager: Register courses through self-paced method, instructor-led method, etc.
- (6) Student Info: Provide students with services such as personal information and mail.

- (7) Student Study: Show timetables and progress of learning for students' learning management.
- (8) Student Util: Form study groups and connects to the education broadcasting station.
- (9) Curriculum Manager: Set curriculums.
- (10) Course: Register and change students and course managers for each course.
- (11) Contents Provide: Contents service component based on DRM.
- (12) Course Access: Register students, lecturers and course managers.

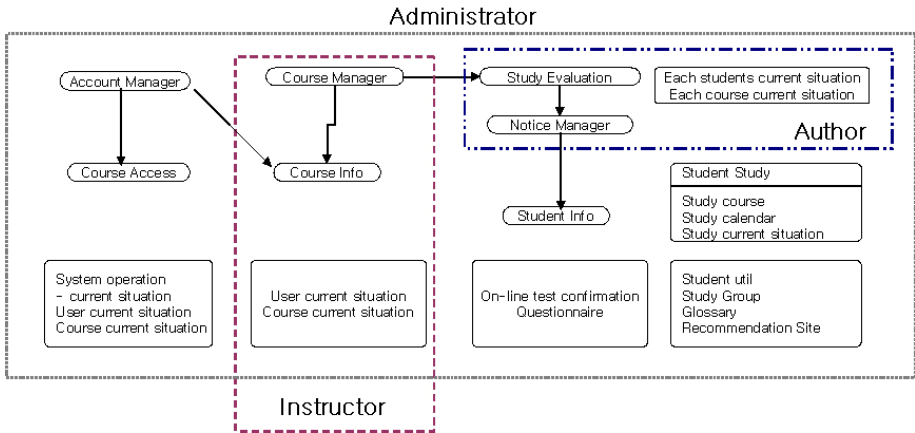


Fig. 1. Structure of DRM-LMS system

### 3.1 User Registration

MDRM-learning includes additional items to user information so that it can be used by all institutions (companies, schools, etc.) that may execute education. The user data is again divided into data of students, course managers and lecturers and, for the

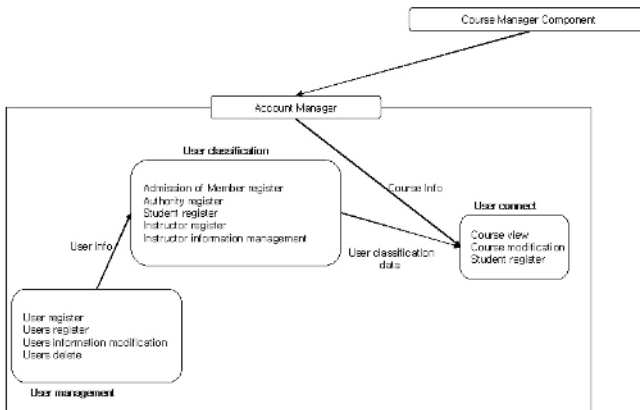


Fig. 2. User registration procedure

division, we also need information on courses. Course Manager Component transfers Course info to Account Manager Component. Based on the information, the User Connect module connects courses to users. This procedure can be diagrammed as in Figure 2.

### 3.2 Course Registration

In MDRM-learning, a course can be executed in the self-paced method or in the instructor-led method. In the self-paced method, the learner checks his/her progress and completes learning by having a test for each learning material. In the instructor-led method, learning is made together with the lecturer and activities such as reports and online assessment occur in virtual classroom environment. In addition, teaching plans can be announced by open flow or step by step. Open flow is a tree structure, showing the entire course at once, and the step-by-step method announces teaching plans one by one over a period of time. In order to set lectures, curriculums are defined first and different types of learning are executed. This procedure is as in Figure 3.

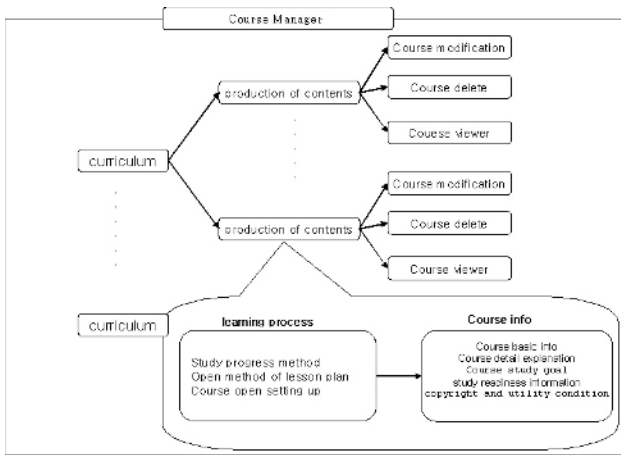


Fig. 3. Course registration procedure

### 3.3 Report Evaluation

Report up uses bulletin boards and distinguishes between managers and learners. Learners can upload but can read, modify and delete only what they have uploaded. Managers can access all posted in the bulletin boards. This can be represented as in Figure 4.

### 3.4 Personal Learning Schedule

This component shows information on currently attending courses and allows learners to make monthly and weekly plans to manage their personal schedule. In addition, this component shows the progress of each course. These functions are represented in Figure 5.

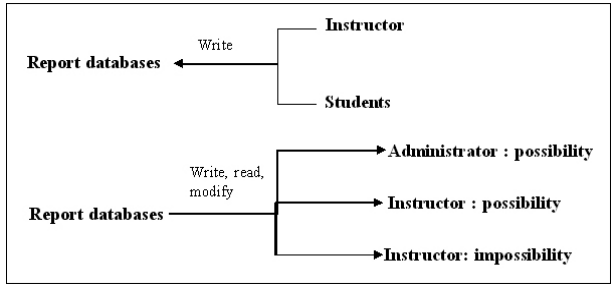


Fig. 4. Report evaluation

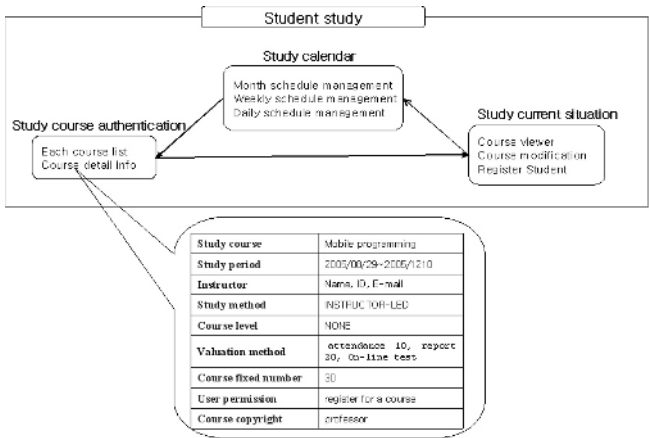


Fig. 5. Personal learning schedule

### 3.5 Course Contents Service Using DRM

This shows the inner structure of DRM contents distribution component. DRM server is largely composed of encryption module, license management module, key management module (SSL module), transaction data management module, etc. and DRM client is composed of license key management module, decryption module, trace response protocol, etc. DRM server includes SSL server, encrypts files from the software encryption code developer, makes license keys and distribute them online. By the request of contents from a client, the server checks if payment has been made, has the encryption module on the Web encrypt the contents, allows download and records the information into the database. Using the information, DRM client manages user license, communicates with SSL client and checks the license. DRM client contains SSL client, which decrypts the encrypted file downloaded from DRM server and certified by the license and sends information on the user's mobile device to the server. Using the mobile device information, the server manages the user. Using the components and their functions listed in Table 1 and 2, we designed MDRM-learning for mobile environment.

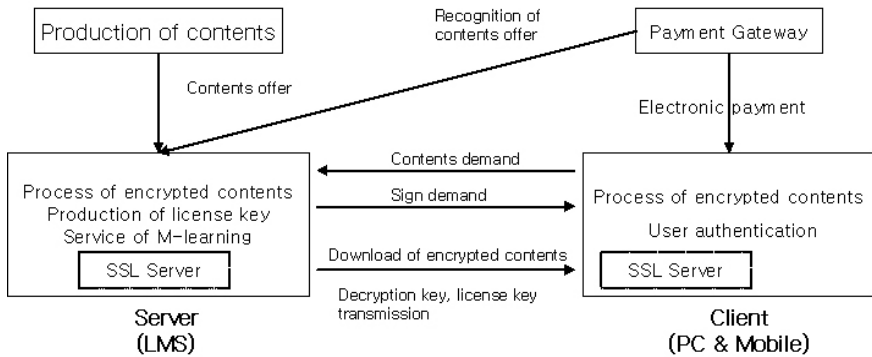


Fig. 6. DRM model for LMS

## 4 Conclusions

The development of the wireless Internet and digital contents is activating Internet business and, in this situation, most online contents providers are supplying diverse contents attracting user's attention. These contents include not only texts but also multimedia such as music, images and videos. Various multimedia data are being applied to numerous areas like Internet broadcasting, education, news, sports, tourist information and expert's consulting and creating new services in the virtual space. The present study examined M-learning, which is e-learning system in wireless Internet environment, and analyzed DRM technology for contents protection in M-learning. We planned a MDRM-learning system focused on contents protection, which has been insufficient until now, by designing basic components of e-learning user registration, course registration, report evaluation, personal learning schedule and classroom platform as well as course contents service using DRM to support mobile systems in wireless environment. Further research is necessary on DRM and relevant technologies focused on contents security in education and wireless Internet environment as well as on education systems using these technologies.

## References

1. Ho-young Kim, Jin-woo Kim: An Empirical Research on Important Factors of Mobile Internet Usage. The Korea Society of Management Information Systems, Vol.12 No.3,(2002)
2. Yong-gyu Lee: A study on mobile DRM system technology in wireless internet service platform. Master's thesis (2005)
3. Eun-mi Im.: Design Strategies for the Tutorial Module to construct Intelligent Tutoring System. Master's thesis (1999)
4. Jun Lee. : Learning Content Management System based e-Learning Development and Application. Korean Association for Educational Information & Media, Vol.8 No.2, (2002)
5. Digital Rights Management Candidate Version 2.0., Open Mobile Alliance DRM W/G, 2004
6. Platform technique of wireless Internet. KTF technique education data, (2003).5
7. Korea Institute of Information Security & Cryptology. : Next a generation network security technique. Korea Information Security Agency. (2002)
8. DRM Content Format Candidate Version 2.0., Open Mobile Alliance DRM W/G. (2004)



# SBEAVER: A Tool for Modeling Business Vocabularies and Business Rules

Maurizio De Tommasi and Angelo Corallo

e-Business Management School - ISUFI  
University of Lecce

via per Monteroni sn, 73100 Lecce (Italy)

{maurizio.detommasi, angelo.corallo}@ebms.unile.it

**Abstract.** Methodologies in software development are typically applied when a problem is already formulated and described. Software developers transform requirements into code with a relatively repetitive process. The actual difficulty lies in describing business needs and expected functionalities. Stakeholders involved in software development can express their ideas using a language close to them, but they usually are not able to formalize these concepts in a clear and unambiguous way. In this paper, we introduce a new tool intended primarily for business analysts and modelers who want to formalize their business knowledge using a business oriented notation based on natural language and fact-oriented approach. Moreover, the capability to map models to formal logic allows automatically generation of IT system design artifacts bridging the existing language gap between business and IT.

## 1 Introduction

E-business adoption represents a unique opportunity for enterprises to gain competitiveness by improving their business performance at a local level and helping them take advantage of global market opportunities. A major stumbling block for enterprises (especially Small and Medium Enterprises) in adopting new ICT is the lack of a common operational notion describing business needs and ICT solutions. The need for natural language as a means for business modeling is particularly important in order to allow enterprises to author, validate, and dynamically define and redefine in the underlying IT-systems their products, services, prices, policies and terms [6]. As such, following the MDA roadmap [10], the direct involvement of business actors as well as the adoption of a business oriented formal language in business modeling would represent a concrete attempt to align the business strategy with IT infrastructure [5]. The main purpose of natural language modeling approach is hence to make natural language suitable for conceptual modeling. In other words, it aims at designing analytic processes able to produce a simple syntax and to reduce ambiguity and vagueness, preserving language completeness and essential meaning [3].

In addition, formalizing business policies and rules allows for automated transformations enabling creation of corresponding Platform Independent or Platform

Specific Model elements (PIMs or PSMs) in different software and systems architectures to support the design and construction of the system (business applications, workflow, etc.). Furthermore, formal business models support evolution of the system design as business facts and rules change, thus allowing alignment between business strategy and the underlying IT infrastructure [9].

In this paper we present a tool called SBEAVER which allows business analysts or experts to create formal and interchangeable business models in accordance with the OMG's *Semantics of Business Vocabulary and Business Rules* (SBVR) standard. It is a rich text editor which allows the business modeler to type structured sentences and business rules through an easy-to-use graphical interface. The editor guides the modeller in the process of creating a business model in a computation-independent fashion, avoiding technical modeling formalisms typically based on the object oriented modeling paradigm as used by IT system designers and technical people. The resulting models serves two different purposes. On the one hand, they are semantically rich shared representation of business knowledge useful as guidance for business and input to IT system specifications. On the other hand, the models are suitable for easy interchange among people and organizations or software tools and repositories.

## 2 Business Modeling Approach

*The Semantics of Business Vocabulary and Business Rules* [12] defines a meta-model conceptualized for business people and designed to be used for business purposes, independently of information systems designs. Its first aim is to allow business vocabularies construction and business rules definitions by business people in business language (i.e. natural language, common graphics, and tables), enabling their interchange among organizations.

SBVR is self-describing: the foundation that makes up SBVR itself is represented through SBVR Vocabularies and their related rules. The SBVR Vocabulary is extensible: since it is a vocabulary, it can be included in other vocabularies in order to create an extended SBVR Vocabulary. The latter can, for example, add new symbol for existing concepts or add new concepts along with symbols that represent them. In this way, even if the SBVR Vocabulary is based on English language, it is possible to create an alternative SBVR Vocabulary based on a different language; it should provide symbols from the different language for the concepts represented in the SBVR Vocabulary.

### 2.1 Business Vocabularies and Rules

A business vocabulary contains all the specialized terms and definitions of concepts that a given organization or community uses in their talking and writing in the course of doing business. A SBVR business vocabulary provides capabilities to define taxonomies, categorization schemes, Thesauri including synonyms, abbreviations. It provides also the ability to specify definitions formally and

unambiguously in terms of other definitions in the business vocabulary as well as the ability to define connections between concepts (fact types). The SBVR follows a common-sense definition of business rule: *rule that is under business jurisdiction* [12]. ‘Under business jurisdiction’ is taken to mean that the business can enact, revise and discontinue business rules as it sees fit. Rules serve as criteria for making decisions. They are statements that define or constrain some aspect of the business [14]. In SBVR, rules are always constructed by applying necessity or obligation to fact types.

In order to enable the use of SBVR artifacts in information systems design, SBVR is underpinned by First Order Predicate Logic, but it also provides an extended formalization for higher-order types, that uses a restricted version of higher-order logic. The formal semantics of SBVR is based on various formal approaches: typed predicate logic, arithmetic, set and bag theory, with some results from modal logic.

## 2.2 Semantic Interchange

As stated above, a business vocabulary provides a means of recording and communicating facts. Following OMG’s Model Driven Architecture [11], a business vocabulary developed as an information system independent model (Computation Independent Model or CIM) is used to drive the creation of a platform independent MOF model representing these facts. The MOF model is, in turn, used to drive generation of Java interfaces and an XML schema [7]. The SBVR Metamodel is intended to provide for standardized data interfaces and data interchange among tools that collect, organize, analyze and use Business vocabularies and rules, as well as tools that bind business vocabularies and rules to other models and implementations.

In order to address semantic interchange of business vocabularies and rules SBVR specification defines a Vocabulary-to-MOF/XMI Rule Set that governs how a business vocabulary is mapped to a MOF 2 model. An XML Schema could be then generated based on XMI 2.1. The resulting SBVR model is intended, not for business people, but for software engineers that build tools for business people. The SBVR Metamodel is includable and extendable in models that address various business domains. That the SBVR Metamodel is generated without manual intervention guarantees that it accurately represents the concepts of the SBVR Vocabularies.

SBVR explicitly uses the fact-oriented approach in order to standardize business-level data interchange. This approach implies the creation of a separate class for each type of fact that can be expressed; in other words, each fact is represented by its own object, and not by an attribute of some other object.

## 2.3 Structured English Notation

The most common means to express definitions and business rules is through statements. Even if there are numberless ways to use a language in order to

express them, SBVR specification introduces a small number of English structures and common words in order to provide a simple and automatic mapping to SBVR concepts. In fact, since SBVR aims to be a powerful means to express and interchange business semantics, its primary focus is to provide the means to express each possible statement in an unambiguous form. For this reason, SBVR specification defines a SBVR Structured English. This means that each statement represented in terms of SBVR Vocabulary has a structured form and, in particular, can be represented through a logical formulation (i.e. the SBVR representation of formal logic). Consequently, all formal definitions and rules stated using the SBVR Structured English can be automatically interpreted in order to create MOF and/or XMI representations. However, it is important remembering that SBVR Structured English is just one of many possible notations that can map to the SBVR Metamodel.

The SBVR notation is characterized by some specific font styles and some keywords, each with its formal meaning. Moreover, it describes some structures to be used in order to define each vocabulary entry and rule. Fig. 1 shows some example of Structured English business rules.

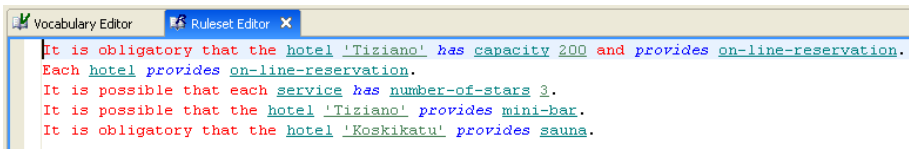


Fig. 1. SBEAVER: Structured English business rules

### 3 SBEAVER: A Business Modeling Tool

In order to enable creation of SBVR business models (in terms of business vocabulary and business rules, as explained in the previous section) we developed an open source tool called SBEAVER. It allows to define concepts, fact types and business rules following the SBVR semantics in accordance with the Structured English notation and its mapping to formal logic. Sentences representing fact types and business rules are parsed and validated using a *LL* parser with multi-token lookahead<sup>1</sup> in order to have an in-memory representation of the model conforming to the SBVR Metamodel. The resulting in-memory model representation is then suitable for further transformations.

SBEAVER is designed primarily for business analysts and modelers working in enterprises, who want to specify business policies and rules precisely and using a non technical notation. Their business view is the enterprise business view, or perhaps a view of part of the business. Nevertheless, mapping to formal logic

<sup>1</sup> An *LL* parser is a table-based top-down parser for a subset of the context-free grammars. It parses the input from Left to right, and constructs a Leftmost derivation of the sentence. An *LL* parser is called an *LL(k)* parser if it uses *k* tokens of look-ahead when parsing a sentence.

allows to translate natural language based artifacts into different representation formats (e.g. MOF/XML, XML or UML).

Currently we support automatic generation of XMI XML Schema and Prolog. This feature makes the tool suitable for users such as system engineers, integrators and developers responsible for designing, integrating and implementing IT systems following the business rules and contents provided by business people. Automating transformation of software artifacts from business rules represents a basis for validating the system design by business people at the business level of understanding.

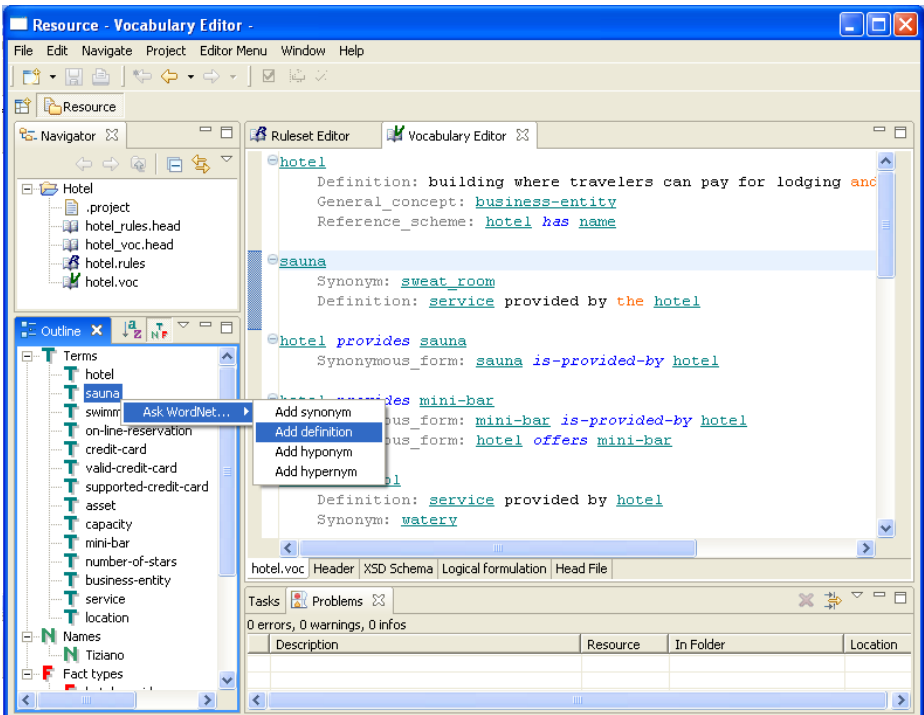


Fig. 2. The SBEAVER main window

### 3.1 Main Features

SBEAVER key features are:

- Business knowledge representation: SBEAVER provides a tool for formalizing the semantics of business knowledge using the Structured English notation.
- Portability: SBEAVER is written in Java for portability and could be installed on many different operating systems without code changes. SBEAVER comes with all source code with EPL licence which can be modified or tailored to meet a user's specific needs.

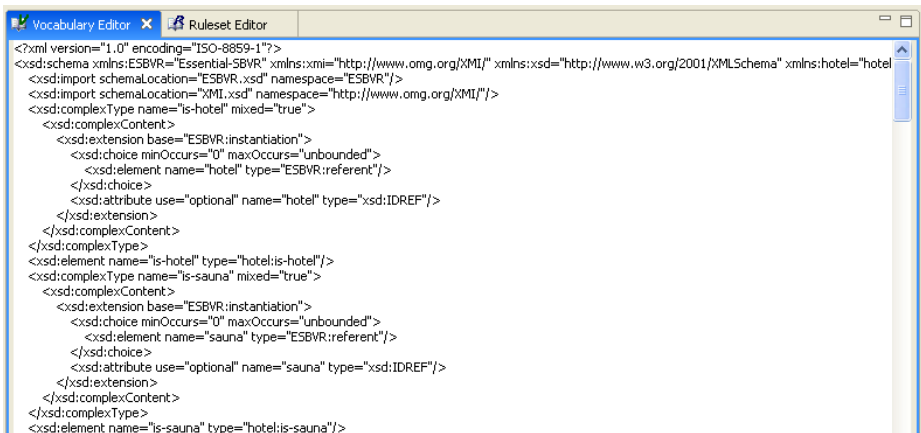
- Integration and Extensibility: SBEAVER is based on Eclipse platform which allows to easily extend or integrate the tool by a user through the use of a well-defined framework.
- Model validation: SBEAVER includes a number of features to support the verification and validation of business models including semantic analysis of rule to determine possible inconsistencies.

### 3.2 Functional Architecture

Fig. 2 shows the SBEAVER main window. In particular, from a functional architecture point of view SBEAVER provides:

- Presentation and user modification of SBVR Structured English business vocabularies and business rules.
- Standard text editing operations (cut, copy, paste, find, replace).
- Automatic syntax highlighting following the SBVR Structured English font styles.
- Content assist to allow easy and fast creation of contents.
- Hierarchical navigation of vocabulary, as shown in the left bottom column in Fig. 2.
- Dictionary (at present we embed WordNet dictionary) support for synonyms, hypernyms, hyponyms, meronyms, definitions.
- Mapping from in-memory Java representation of vocabulary and rules to logical representation.
- XMI XML Schema serialization allowing interchangeability of models between software tools.

Fig. 3 shows the XSD schema generated automatically starting from the model shown in the main window in Fig. 2.



```

<?xml version="1.0" encoding="ISO-8859-1"?>
<xsd:schema xmlns:ESBVR="Essential-SBVR" xmlns:xmi="http://www.omg.org/XMI" xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:hotel="hotel"
<xsd:import schemaLocation="ESBVR.xsd" namespace="ESBVR"/>
<xsd:import schemaLocation="XMI.xsd" namespace="http://www.omg.org/XMI"/>
<xsd:complexType name="is-hotel" mixed="true">
  <xsd:complexContent>
    <xsd:extension base="ESBVR:instantiation">
      <xsd:choice minOccurs="0" maxOccurs="unbounded">
        <xsd:element name="hotel" type="ESBVR:referent"/>
      </xsd:choice>
      <xsd:attribute use="optional" name="hotel" type="xsd:IDREF"/>
    </xsd:extension>
  </xsd:complexContent>
</xsd:complexType>
<xsd:element name="is-hotel" type="hotel:is-hotel"/>
<xsd:complexType name="is-sauna" mixed="true">
  <xsd:complexContent>
    <xsd:extension base="ESBVR:instantiation">
      <xsd:choice minOccurs="0" maxOccurs="unbounded">
        <xsd:element name="sauna" type="ESBVR:referent"/>
      </xsd:choice>
      <xsd:attribute use="optional" name="sauna" type="xsd:IDREF"/>
    </xsd:extension>
  </xsd:complexContent>
</xsd:complexType>
<xsd:element name="is-sauna" type="hotel:is-sauna"/>

```

**Fig. 3.** SBEAVER generates automatically XSD Schemas

## 4 Related Work

As stated in [3], in the past, the use of natural language to influence software design has been mainly tacit and informal.

In the late 1970's, Halsteads [8] categorization of operands and operators from natural language elements raised the possibility that software theory and quantitative analysis might extend to natural language, suggesting the existence of a mapping from natural language to computational primitives.

In the early 1980's, Abbott [1] proposed an approach to Ada program design based on linguistic analysis of informal strategies written in English.

In 1989, Saeki, Horai, and Enomoto [13] proposed a software design process based on natural language, complementing the work by Abbott. They focused on the identification of dynamic system behavior as expressed by the verbs in a natural language description. Their work offers many useful ideas regarding the information needed to represent the relationships between natural language elements and some rules for selecting message senders and receivers.

In 1990, Carasik, Johnson, Patterson, and Von Glahn [4] pointed out the limitations of entity-relationship models for defining semantic intensions arguing for the usage of conceptual modeling languages and knowledge representation techniques to formally represent meaning. However, conceptual models need not diverge so far from natural language as to be unintelligible to stakeholders.

In 1992, Cordes and Carver made one of the first attempts to apply automated tools to requirements analysis and the automatic generation of object models from requirements documents. While the translation of the initial requirements into a suitable knowledge base requires human interaction to resolve ambiguities, the translation of the domain knowledge into object models is automated. However, they recognized that the translation of formalized knowledge into object models is sensitive to the quality of the initial requirements specification. Still, their process and tools can help a requirements analyst begin to bridge the gap between informal requirements and formal software models.

The RECORD (REquirements COLLECTION, Reuse and Documentation) [2] project has recently integrated several tools and techniques, including natural language processing, with the aim of providing a complete solution for requirements collection, analysis, management and object-oriented modeling.

## 5 Conclusions

Natural language is generally used by business organizations in order to describe themselves and their rules. Nevertheless, even if complex constructs and ambiguous forms of expression provide a great communicative power, they usually make this description unclear and informal. Conversely system requirements gathering and creation of machine-readable documents need a higher degree of precision and formality, with a consequent loss in richness of meaning and expressions.

In this paper, we have presented SBEAVER, an easy-to-use and extensible business modelling tool supporting the Semantics of Business Vocabulary and

Business Rules standard and allowing business modelers to capture and formalize business knowledge in a fact-oriented and natural language approach. We have also shown how semantically rich computation independent business models can be automatically translated into other system design models thanks to the SBEAVER's capability to map Structured English to formal logic. This feature makes SBEAVER suitable for different users which, at different level of understanding, share meanings relevant to the business that drive the creation of IT systems. However, there are still many possibilities for improvement.

First, although we already use one of the most important English dictionary, we need to widen the selection of dictionaries as well as the selection of language notations (currently we support only the Structured English), thus enabling multilinguality.

Second, we recognize that alternative graphical notations are likely important to represent business knowledge in a more compact manner in order to facilitate readability of textual models.

Third, since SBEAVER is in its preliminary phase, implemented functionalities need to be further tested and stabilized as well as new features need to be developed in order to improve usability and content development.

Finally, we note that an obvious barrier remains: because of the new approach to business modeling, sufficient knowledge of the main characteristics of SBVR is a necessary prerequisite for effective use of our tool.

## References

1. R. Abbott. Program design by informal english descriptions. *Communications of the ACM*, 26(11):882–894, Nov 1983.
2. J. Börstler. User-centered requirements engineering in record - an overview. In *Proceedings NWPEN'96, the Nordic Workshop on Programming Environment Research.*, pages 149–156, Aalborg, Denmark, May 1996.
3. N. Boyd. Using natural language in software development. *Journal Of Object Oriented Programming - JOOP*, 11(9):45–55, Feb 1999.
4. R. Carasik, S. Johnson, D. Patterson, and G. Von Glahn. Towards a domain description grammar: An application of linguistic semantics. *ACM SIGSOFT Software Engineering Notes.*, 15(5):28–43, Oct 1990.
5. B. Connell. Web Services Management in Action: Aligning IT with Business Objectives. Westglobal <http://www.westglobal.com>, 2003.
6. M. De Tommasi, V. Cisternino, and A. Corallo. A rule-based and computation-independent business modelling language for digital business ecosystems. In *KES (1)*, pages 134–141, 2005.
7. D. S. Frankel, editor. *Model Driven Architecture Applying MDA to Enterprise Computing*. Wiley Publishing inc., 2003.
8. M. Halstead, editor. *Elements of Software Science*. Elsevier North-Holland, Inc., New York, NY, 1977.
9. Hendryxs & Associates. Integrating Computation Independent Business Modeling Languages into the MDA with UML 2. document ad/03-01-32 <http://www.omg.org/cgi-bin/doc?ad/03-01-32>, Jan 2003.
10. A. Kleppe, J. Warmer, and W. Bast, editors. *The Model Driven Architecture; Practice and Promise*. Addison Wesley, 2003.



11. J. Miller and J. Mukerji. MDA Guide Version 1.0.1. OMG <http://www.omg.org>, Jun 2003.
12. OMG. Semantics of Business Vocabulary and Business Rules (SBVR), ver1.0, draft adopted specification. document dtc/05-11-01 <http://www.omg.org/cgi-bin/doc?dtc/05-11-01>, Nov 2005.
13. M. Saeki, H. Horai, and H. Enomoto. Software development process from natural language specification. In *Proceedings of the 11th International Conference on Software Engineering (ICSE-11)*: IEEE Computer Society Press, 1989.
14. The Business Rules Group. Defining Business Rules - What Are They Really? Final Report, revision 1.3. BRG <http://www.businessrulesgroup.org>, 2000.

# A Semantic Recommender Engine Enabling an eTourism Scenario

Angelo Corallo, Gianluca Lorenzo, and Gianluca Solazzo

e-Business Management School - ISUFI  
University of Lecce

via per Monteroni sn, 73100 Lecce (Italy)

{angelo.corallo, gianluca.lorenzo, gianluca.solazzo}@ebms.unile.it

**Abstract.** This work approaches the problem of delivering services in a personalized way in an eTourism scenario. Our research, on one side, exploits semantic annotation of either services and user profiles to add a layer of business description that allows the system to supply the most suitable service to the user who requested it. On the other side, this work aims to extend Service Oriented Architecture with the use of semantics and ontologies to enable e-business relations in the tourism applicative domain. This approach is adopted in the MAIS project<sup>1</sup>, in which a Service Oriented Architecture has been developed. Our model relies on semantic description of services, rule-based user profile and the use of semantic matching algorithms.

## 1 Introduction

Nowadays the ICTs enable to access information systems in an ubiquitous way, through various kind of interaction devices (PCs, laptops, palmtops, cellular phones, etc.), increasing their economic and social impact [8]. The MAIS project main research goal is to provide a flexible environment to adapt the interaction and provide information and services according to every changing requirements, execution contexts, and user needs. Our contribution within the project draw on Recommender Systems, which are e-Commerce applications, nowadays widely used. Recommendations can be used to support the dynamic composition of e-services using the User Knowledge collected by the system and processed in an appropriate way.

Recommender Systems became very popular in the 90's, offering a solution to the problem of information overload in the World Wide Web. In a few years, many approaches have been developed and used, and each of them presents some benefits and some disadvantages. Recommender Systems are able to learn over time user preferences and, through their analysis, are automatically able to identify and propose to the user relevant products or services. Recommender Systems are also able to track dynamically how single user interests change by

---

<sup>1</sup> This work was partially supported by the Italian Ministry of Research - Basic Research Fund(FIRB), within MAIS project, Multi-channel Adaptive Information System. Website at <http://www.mais-project.it>

building a user profile from his preferences. They can "observe" the behaviour of the user during his interaction with the information system, building and updating a user's profile preferences. The acquisition of user knowledge is very important because it is necessary to collect quite a wide amount of information in order to grant the correctness of the profiles.

Actual Recommender Systems can be divided in three categories [2]:

- Content-based Recommender Systems: in the content-based approach, the system tries to recommend items similar to those in which a given user has indicated interest in the past. [1].
- Collaborative Recommender Systems: in the collaborative approach, the system identifies users whose preferences are similar to those of the given user and recommends items they have liked. Recommendations for a user are made on the basis of similarities to other users [4,11].
- Hybrid Recommender Systems: this approach tries to leverage the positive aspects of both content-based and collaborative-filtering systems avoiding their drawbacks. Generally, in order to determine recommendations a hybrid Recommender System implements algorithms that use both the content and an item's attributes as well as user's opinions [3].

Within Recommender Systems user knowledge can be obtained in explicit and implicit way too. Implicit knowledge acquisition is the preferred way to collect information due to its low impact on user interaction with the system. Transparent monitoring of user activity is useful to discover behavioural data. Explicit knowledge acquisition requires that the user periodically interacts with the system in order to provide feedback. This kind of knowledge has a high degree of confidence because it is directly provided from the user and it is not obtained after an interpretative process [6].

One of the applicative domains in MAIS is the tourism that suits very well to the experimentation of Service Oriented Architecture. We see these issues in details in the section 2. In the section 3 we describe the functionalities of our Semantic Recommender Engine and its role in the MAIS architecture, while in section 4 and 5 we describe in depth the role of semantics either in the recommendation generation process and in the description of service and user profiles.

## 2 The e-Business and Its Evolution in the eTourism Sector

The fast growth and diffusion of the web, its effect on trade, on economic transaction and on communication system for people and organisation, allowed and boosted the creation and reinforcement of the Information Technology Infrastructure. These phenomena allowed organizations to grow globally, increasing efficiency and reducing coordination costs. The organizations geographically distributed start to interact and to trade globally, enhancing the growth of network of distributed companies.

The Internet traffic doubles each year and more than 40% of that traffic is related to business activity [12]. Despite this growth, e-business still accounts for a relatively small share of the total transactions. Using the widest range of B2B transactions, including established EDI (Electronic Data Interchange) as well as Internet transactions, it emerged that the total on-line transactions in the year 2000 were generally 8% of the total business transactions. The situation is worse for SMEs, since they lag behind larger firms in Internet transactions. Small organisations are not ready to use Internet extensively as a business tool, because they are not able to overcome the barriers that prevent them to adopt e-business technologies [7]. E-Business adoption represents a unique opportunity for SME, as through e-business it is possible to reduce transaction costs and increase the speed and reliability of transactions, reduce inefficiencies resulting from lack of co-ordination between firms in the value chain, reduce information asymmetries between buyers and suppliers [8].

## 2.1 Why e-Business is Successful in the eTourism Sector

It could be very useful to adopt an e-business strategy to enable the integration and reconfiguration of the process of cultural, environmental, touristic and local resources in order to fulfil requests for touristic flows. Moreover, new ICTs enable dynamic and personalized fruition of the cultural heritage and the integration of a wide experience required by the tourist. It can be realized by integrating the disaggregate offers, supporting a strong service personalization, reaching the user/tourist in a pervasive way.

E-business can support the integration of local resources in a distinctive, adaptive and proactive offering system. Indeed, e-business has already reached a wide diffusion in the tourism domain. In fact, according to the "European e-business Report 2004" [8], the tourism is the first in the online trading, with the 30% of the european companies who already sell through Internet producing from 5% to 25% of their total returns. The tourism is also growing from the e-business transactions perspectives but it is still growing slowly. The touristic-cultural system presents some peculiar characteristics:

- touristic-cultural products are perceived like integrated and inclusive experiences, that are estimated in their wholeness by a tourist.
- touristic-cultural demand is a compound experience made by several elements, supplied by many complementary and concurrent organisations (belonging to different sectors), and that makes the offer diversified and highly fragmented;
- the production and the fruition of touristic-cultural products are usually contemporary and strongly interdependent.

In this context, tourist-cultural operators have to collaborate dynamically in a creative way, being also strongly proactive in their offering strategies. The challenge is to create a business scenario that can include many self-organizing agents, assisted by systems that enable in a loosely invasive way the relations

between the actors in order to make the tourist-cultural experience satisfying, holistic and rewarding.

In this scenario, the technological Service Oriented Architecture (SOA) have undertaken a key role. The application of SOA architecture in Tourism domain needs to extend the specific characteristics of SOA. Using SOA is not possible to cover company or product aspects necessary in real business scenario. For example using a SOA in an e-business scenario, user that owns a service, can register his business profile through a complex process that needs the help of computing experts; his business profile is mainly overlapping on the description of his services and does not contains a business-like representation of the product or service offered. On the opposite side, users who want to trade, have to select e-business partners basing their choice on a description that mainly covers technical aspects. Using Semantic Web Service approach is possible to extend service description in order to cover also business aspects. In order to leverage the powerful potentiality of the SOA, it is necessary to imagine a framework that focuses on domain knowledge and allows users to interact as they really do in actual life. Semantic Web and Recommendation System allows to describe services, adding business and domain information and to retrieval specific service basing on expressed or implicit users request.

### 3 Semantic Recommender Engine

The scenario described above suits perfectly the MAIS project, and the MAIS Architecture that is Service Oriented. MAIS services can be abstract or concrete. An abstract service is a service with a description completely independent from any kind of implementation. Its function is to represent a service archetype and not a invocable service. A concrete service can be defined as the implementation of an abstract service with the addition of all the concrete aspects that makes it invocable. Web Service orchestration and composition is made by a flexible invocation of available services selected on the basis of their functional properties and quality of service. The MAIS platform uses semantics just for choosing a set of concrete services that referred to a single abstract service which match mainly with functional parameters and quality of service. We considered the opportunity to use semantic structures such as domain ontologies in order to enrich web service description by introducing business characteristics of a service related to a specific domain. We also defined a semantic layer on the service descriptions that can be used to express and explicit user needs in terms of which service is more suitable to his profile. Having these enriched information on services and user profile, we can improve system capabilities to retrieve, select and compose services that are really similar to user needs and preferences contained in his profile. In addition to that, it must be considered that user needs and preferences on services are more often implicit, and they should be extracted from the Knowledge Base of the system. The way we added semantics is through semantic annotation, that provides a richer and more formal service description in order to enable a more accurate discovery mechanism.

For the reasons explained above, an "environment", able to collect and manage the user knowledge and the service descriptions, has been developed within the MAIS architecture. The Environment, called "MAIS Recommendation Environment", applies user profiling and information filtering techniques to services and adaptive information systems. It can be considered an Hybrid Recommender System as it mixes a content based approaches on e-service modelling, with the collaborative approach taking into account some parameters that specify the overall behaviour of all the MAIS users. The Recommendation Environment performs also an implicit knowledge acquisition, by monitoring significant business events, raised by the MAIS platform, and by deriving from them the behaviour of each user. The recommendation technique developed is based on a semantic matching algorithm, that uses ontologies in order to calculate semantic distance between concept to identify which e-service is the most suitable to a user. This approach is based on Paolucci's work [9].

Within the MAIS Recommendation Environment, a "Semantic Recommender Engine" component has to choose the most suitable service, while a "data mining" component has to create and to manage user profiles by collecting all the business events generated by the MAIS platform and that will feed the user profile updating process.

## 4 Services Profiles and User profiles

In our approach, an e-service is an entity described from two different but complementary perspectives: a computational perspective and a business one. From the computational perspective we see the e-service as the set of descriptions necessary to its right functioning. So we consider languages and descriptions for its technological definition (WSDL, SOAP etc.), but also for workflow orchestration, services composition, negotiation, substitution, invocation. From the business perspective, we see the e-service as a description of the provider and of the feature of the service itself, in a way easily understandable for a business people, in order to provide an added value for services developing, selection, composition and aggregation.

Thus, a complete description of a web service should provide all the necessary information about "what" the service does, "how" it works and how to invoke its functionalities. It also has to be consistent with real characteristics implemented by a service and contain enough information to allow a correct and more efficient execution of the discovery process. From this perspective, a service description can be composed mainly by three sections:

- A human-readable description of generic characteristics of the service (name, short textual description of the functionalities, provider's name, etc);
- A description of its interface by means of a list of functional attributes: input, output, pre and post-conditions;
- A set of extra functional attributes (i.e. quality of service and other additional attributes related to a particular instance of a service).

We are interested in the Semantics of extra functional parameters, that consists in a formal description of a set of QoS parameters which are domain independent and a set of additional parameters which are specific to a domain and to a service. It is used to select the more appropriate service, given a user profile.

Here it is an example of semantic description of extra functional parameters made with OWL-S that is a OWL-based Web service ontology [5].

```
<profileHierarchy:HotelReservation rdf:ID="HotelReservation" >
<profile:serviceParameter>
  <profile:ServiceParameter rdf:ID="TvInRoom" >
    <profile:serviceParameterName>TvInRoom</profile:serviceParameterName>
    <profile:sParameter rdf:resource="http://localhost.localdomain/ontHotel.rdfs#tv" />
  </profile:ServiceParameter>
</profile:serviceParameter>
<profile:serviceParameter>
  <profile:ServiceParameter rdf:ID="TypeOfRestaurant" >
    <profile:serviceParameterName>TypeOfRestaurant</profile:serviceParameterName>
    <profile:sParameter rdf:resource="http://localhost.localdomain/ontHotel.rdfs#
internationalMenuRestaurant" />
  </profile:ServiceParameter>
</profile:serviceParameter>
</profileHierarchy:HotelReservation>
```

All the *sParameter* contained in the Service Profile are concepts belonging to an ontology. For our work, the ontology we are using is the MAIS Tourism Ontology, written in OWL and that is the domain ontology which models the tourist resources.

The other input for the Recommender Engine is the user profile that is the part disposed to contain the information enabling ranking process. This part will contain the information about services extra-functional parameters and can be obtained by the analysis of the events caught by the system during the interaction session with the user. Having a correlation between behavioural data and static attributes of a user, it is possible to extract information on the typologies of user behaviour in the form of rules. Rules are a raw form of knowledge and they will describe in the user profile two or more actions or preferences of the user. An example of rule is:

```
IF
  the user is a student, requesting from Rome, ask for a service belonging to ServiceCategory HotelReservation
THEN
  the user prefers 2stars category, room with tv, conditioning air, EthnicMenuRestaurant
```

Recommender Engine will verify the premise of the rule (that is the user, who requested for a service belonging to that ServiceCategory, is a student and his request comes from Rome) and will use the preferences contained in the rule consequence to estimate the similarity degree between these preferences and the concrete services. It is important to underline that the Recommender Engine is not able to manage rule with two different ServiceCategory. This is also out of our scope because what we want to obtain is a ranking of concrete services that must belong to the same service category. Finally, to each rule can be associated some parameters that will describe how the rule is relevant for an user taking into account either his overall behaviour and overall behaviour of all the MAIS users.

## 5 Ontology-Based Matching Algorithm

The way the recommender engine calculates a similarity, between a concrete service business description and the description of an user who requested for the corresponding abstract service, takes into account the semantics associated to them. In our discussion, we will present the algorithm based on the evaluation of the *DegreeOfMatch()* function, that is an example of feature-based similarity approach developed by Paolucci et al. [5]. This approach is extended in order to evaluate the semantic similarity on  $n$  different ontologies. The application to the MAIS platform expects the use of one domain ontology. More in particular, we assume that for each e-service, identified by the Concrete Service Invocator through the MAIS e-service Ontology and through the MAIS Registry, there is a Semantic Description File associated.

Thus to  $Y$ , that is a set of concrete e-services functionally equivalent, it exists  $Y'$  that is the corresponding set of concrete e-services description  $SP$ . Given the Ontology  $\Omega$ , given  $SP$  that is the generic Service Profile Description File, subset of  $\Omega$  and given  $UP$  that is the generic User Profile File, subset of  $\Omega$  and hold by the MAIS platform, we can define the function *SemSimilarity()*:

$$SemSimilarity_{\Omega}(SP \in Y', UP) \rightarrow r \in [0, 1]$$

Thus, *SemSimilarity()* is a function that returns real number value representing the similarity degree between its arguments by measuring the semantic similarity on all the extra-functional parameters contained in the user profile and in the service profile. This function will calculate the degree of semantic similarity among each concrete e-services business description and the user profile, allowing a ranking among concrete e-services with the same functionalities.

We should observe that the Service Profile Description File  $SP$ , can be defined as:  $SP = \{SP_1, SP_2, SP_3, \dots, SP_n\}$  where each  $SP_i$  is a concept of the  $\Omega$  ontology. In the same way the generic User Profile File  $UP$  can be defined as:  $UP = \{UP_1, UP_2, UP_3, \dots, UP_n\}$  where each  $UP_i$  is a concept of the  $\Omega$  ontology. In order to define the *SemSimilarity()* function it is necessary to identify a function whose role is to measure the degree of match among two different concepts in the same ontology. This function will be applied in order to measure the relation between concepts contained in the user profile and in the service description.

$$DegreeOfMatch_{\Omega}(SP_i \in SP, UP_i \in UP) \rightarrow r \in [0, 1]$$

In particular, the *DegreeOfMatch* is given by the minimum distance between concepts in the ontology view and it is possible to distinguish four different kinds of matches [5]:

1. an exact match can occur in two cases: in the simplest situation when two concepts coincide, and also when the concept specified in the user profile is a direct specialization (first level specialization) of the concept specified in the service description and contained in the ontology;



2. plugIn occurs when the concept specified in the service description is a direct specialization of the concept specified in the user profile. This kind of relation is weaker than the previous;
3. subsumes occurs when the concept specified in the user profile is a specialization of the concept specified in the service description;
4. fail occurs when no transitive relation exists between two specific concepts.

The four cases the degreeOfMatch can assume will be associated to discrete values, considering that the preferable degree is the "exact" and the less preferable is the "subsumes".

In order to define *SemSimilarity()* it might be necessary to introduce some rules in order to be able to choose the best match and to reduce the computational complexity of the algorithm. If one of the additional attributes contained in User Profile matches with more than one of the Service Profiles attributes, we must choose the best combination. In general, we can say that given a user profile with  $n$  attributes and a Service Profile with  $m$  attributes, there will be  $n \times m$  pairings: we must choose the  $n$  distinct pairings with the highest value. With this assumption, given  $UP = \{UP_1, UP_2, UP_3, \dots, UP_n\}$  and  $SP = \{SP_1, SP_2, SP_3, \dots, SP_m\}$ :

$$NSemSimilarity_{\Omega}(SP, UP) = \Sigma_{i=1}^n [max_{j=1}^m (DegreeOfMatch_{\Omega}(SP_j, UP_i))]$$

The value returned by the *NSemSimilarity* is normalized with the number of preferences contained in the user profile:

$$SemSimilarity_{\Omega}(SP, UP) = \frac{NSemSimilarity_{\Omega}(SP, UP)}{n}$$

We can generalize this result to a complex condition. We could have to deal with a MAIS based environment which need different reference ontologies in order to be effective. In this specific condition is quite reasonable that both the Service Profile Description File and the User Profile File will contain concept from the different Ontologies.

Let  $\Omega$  be the set of Ontologies used in the system and  $\Omega_{\alpha}$  the generic ontology of  $\Omega$  such that  $\Omega = \{\Omega_1, \dots, \Omega_{\eta}\}$ . A specific Service Profile Description File is defined as  $SP = \{SP_1, SP_2, SP_3, \dots, SP_m\}$  where each  $SP_i$  is a concept of a given ontology  $\Omega_j$ :

$$SP = \{SP_1, \dots, SP_g \in \Omega_1, SP_{g+1}, \dots, SP_l \in \Omega_2, \dots, SP_{q+1}, \dots, SP_m \in \Omega_{\eta}\}$$

In the same way a specific User Profile is defined as  $UP = \{UP_1, UP_2, UP_3, \dots, UP_t\}$  where each  $UP_i$  is a concept of a given ontology  $\Omega_j$ :

$$UP = \{UP_1, \dots, UP_p \in \Omega_1, UP_{p+1}, \dots, UP_r \in \Omega_2, \dots, UP_{s+1}, \dots, UP_t \in \Omega_{\eta}\}$$

we can define:

$$NSemSimilarity_{\Omega_{\alpha}}(SP, UP) = \Sigma_{i=1}^g [max_{j=1}^p (DegreeOfMatch_{\Omega_{\alpha}}(SP_j, UP_i))]$$

Taking in to account that, according to the definition of *degreeOfMatch*, the *degreeOfMatch* among concepts of different ontology is always zero, we can conclude that:

$$\begin{aligned} NSemSimilarity_{\Omega}(SP, UP) &= \sum_{\rho=1}^{\eta} NSemSimilarity_{\Omega_{\rho}}(SP, UP) \\ &= \sum_{i=1}^g [max_{j=1}^p (DegreeOfMatch_{\Omega_1}(SP_j, UP_i))] + \dots \\ &\dots + \sum_{i=q+1}^m [max_{j=s+1}^t (DegreeOfMatch_{\Omega_n}(SP_j, UP_i))] \end{aligned}$$

This result is important since allow us to improve the effectiveness of the selection adding a specific weight  $\omega_{\rho}$  to each ontology, according to its perceived value for the customer.

$$NSemSimilarity_{\Omega}(SP, UP) = \sum_{\rho=1}^{\eta} \omega_{\rho} NSemSimilarity_{\Omega_{\rho}}(SP, UP)$$

with  $\sum_{\rho=1}^{\eta} \omega_{\rho} = 1$ .

Finally, the value returned by the *NSemSimilarity* is normalized with the number of preferences contained in the user profile.

$$SemSimilarity_{\Omega}(SP, UP) = \frac{NSemSimilarity_{\Omega}(SP, UP)}{n}$$

As an example we can take in account a MAIS implementation which includes two different Ontologies; we can suppose that in this context the User Profile and the Service Profile are semantically annotated using additional attributes such as the e-services QoS parameters and additional parameters related to business. We can define the *NSemSimilarity* function related to the QoS ontology as:

$$NSemSimilarity_{QoS}(SP, UP) = QoSSS(SP, UP)$$

and in the same way we can define the *NSemSimilarity* function related to the Additional Parameter ontology as:

$$NSemSimilarity_{AddPar}(SP, UP) = AddSS(SP, UP)$$

Using the previous result the *SemSimilarity*() could be defined as:

$$NSemSimilarity(SP, UP) = [\omega_1 QoSSS(SP, X) + \omega_2 AddSS(SP, X)]$$

with  $\omega_1 + \omega_2 = 1$ , in which *QoSSS* is the semantic similarity function calculated on the quality of service [10] parameters, whereas *AddSS* is the semantic similarity function measured on the additional attributes of a service.

This final result shall be normalised in order to have a [0,1] value:

$$SemSimilarity(SP, UP) = \frac{(NSemSimilarity(SP, UP))}{n}$$

## 6 Conclusion

With our work we aim to continue to explore the new field for the application of user profiling techniques and information filtering systems to web services and adaptive information systems. Starting from the assumption that service capabilities, that enable delivery in a personalized way, are those information related to the business capabilities of a service, we pointed out the need to add to technological description of a service a semantic layer for the business description. This is very important and can be very effective in the eTourism application domain because of its peculiar characteristics: fragmentation ed heterogeneity of the sector and high distribution required to the offering systems.

Next steps will be to test the recommendation environment, by measuring its accuracy in the generation of recommendations.

## References

1. B. Arslan and F. Ricci, editors. *Workshop on Recommendation and Personalization in eCommerce: Case-Based Session Modeling and Personalization in a Travel Advisory System (RPEC'02)*, Malaga, Spain, May 28th 2002.
2. M. Balabanovi and Y. Shoham. Fab, content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.
3. R. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.
4. J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM*, 40(3):77–87, 1997.
5. D. Martin, M. Paolucci, S. Mclraith, M. Burstain, D. MacDermott, D. MacGuinness, B. Parsia, T. Payne, M. Sabou, M. Solanki, N. Srinivasan, and K. Sycara, editors. *Proceedings of the 1st International Workshop on Semantic Web Services and Web Process Composition:Bringing Semantics to Web Services: The OWL-S approach (SWSWPC 2004)*, San Diego, California, USA, July 06-09, 2004.
6. D. Nichols. Implicit rating and filtering. In *Proceedings of 5th DELOS Workshop on Filtering and Collaborative Filtering*, pages 31–36. ERCIM, 1998.
7. OECD. Science, Technology and Industry Outlook: Drivers of Growth: Information Technology, Innovation and Entrepreneurship. OECD <http://www.oecd.org>, 2001.
8. OECD. ICT e-business and SME. OECD <http://www.oecd.org>, Jan 2005.
9. M. Paolucci, T. Kawamura, T. R. Payne, and K. P. Sycara. Semantic matching of web services capabilities. In *ISWC '02: Proceedings of First International Semantic Web Conference on The Semantic Web*, pages 333–347, London, UK, 2002. Springer-Verlag.
10. S. Ran. A model for web services discovery with QoS. *SIGecom Exch.*, 4(1):1–10, 2003.
11. J. B. Schafer, J. Konstan, and J. Riedi. Recommender systems in e-commerce. In *EC '99: Proceedings of the 1st ACM conference on Electronic commerce*, pages 158–166, New York, NY, USA, 1999. ACM Press.
12. United Nations Conference on Trade and Development, editors. *E-Commerce and Development Report*. United Nations Publications, 2003.

# A Legal Architecture for Digital Ecosystems

Virginia Cisternino, Angelo Corallo, and Gianluca Solazzo

e-Business Management Section - ISUFI

University of Lecce

via per Monteroni sn, 73100 Lecce (Italy)

{virginia.cisternino, angelo.corallo, gianluca.solazzo}@ebms.unile.it

**Abstract.** This work approaches the problem of the lack of competence, certainty in the reciprocal trading and awareness of regulatory issues for Small and Medium Enterprises that cooperate to dynamically exchange resources, applications, services and knowledge in an e-business context. It is presented the approach adopted in the DBE project<sup>1</sup>, in which it is fundamental to promote and boost trust relations between the DBE-adopters and the potential ones, supporting the creation of services and transactions that will occur respecting local regulatory requirement. In this paper we will show how we addresses the previous requirements to design a legal architecture for digital ecosystems, model the knowledge related to contracts, agreement and regulatory issues, grant effectiveness to the general model by designing and implementing a process for the capture, modification and extension of this knowledge.

## 1 Introduction

The high degree of changeability characterizing new markets context forces firms to transform themselves, becoming extremely reactive, flexible and focused on core competences. To obtain this result, during last years several organizations tended to adopt the *networked enterprise model* [4], that exploits new digital technologies in order to overcome organizational boundaries and focus on value creating activities. Nevertheless, even if practice is demonstrating the actual effectiveness of this model, its adoption seems to be very slow and limited only to those organizations able to strongly invest in IT systems [7].

The DBE project aims at overcoming the aforementioned difficulties by creating a new way of conceiving co-evolution among organization and technology that shifts from a mechanistic way of organizing business (based on static view of the market) to a new organicistic approach (based on Mathematics, Physics and biological Science models) and an approach for technology development unrelated to inter organizational issues to new paradigms in which technology and organization are related variables enabling innovative way of collaborating and competing[6].

---

<sup>1</sup> This work was supported by the Digital Business Ecosystem project, that is an EU VI framework Integrated Project. Website at <http://www.digital-ecosystem.org>

Another of the main objective in the project is to support the transaction of the Small and Medium Enterprises (SMEs) toward e-business, fostering local economic growth through new forms of dynamic business interaction and global co-operation among organizations and business communities enabled by digital ecosystem technologies. The project aims also at creating an integrated and distributed network of local digital ecosystems for SMEs that cooperate to dynamically exchange resources, applications, services and knowledge.

In order to achieve that, it is necessary to promote and boost trust relations between the DBE adopters and the potential ones, supporting the creation of services and transactions that will occur respecting local regulatory requirements. However, though those requirements aim at granting certainty in the reciprocal trading, often the lack of competence and awareness about regulatory issues creates reluctance and uncertainty between SMEs respect to the e-business. As obvious consequence, there is the necessity of a regulatory environment, able to help SMEs to solve legal and regulatory issues, increasing benefits for SMEs that will join a digital ecosystem.

Though several legal supporting ICT tools and standards tried to cover such issues as digital signatures, contracts, dispute resolution as well as law practice support tools, there is a lack of an integrated architecture based on open source tools aiming at supporting SMEs legally in e-Commerce activities, particularly in managing contracts[5,1]. In this regard, this work wants to provide a useful contribution by depicting the initial analysis for the creation of a legal architecture supporting and integrating contract creation tools as well as a knowledge base of regulatory issues.

This paper is structured as follows: in section 2, scenarios for the DBE Legal Architecture are presented, while in section 3 the DBE Legal Architecture is presented and the Knowledge Base model is detailed. In section 4, we focus on a specific layer of the architecture that should enable mechanisms for knowledge creation and management.

## 2 A Scenario for the DBE Legal Architecture

The main reason behind the realization of the Legal Architecture is represented by the necessity to support DBE-adopters' users in creating and managing the contract life cycle. Within a simple starting scenario, DBE-adopter SMEs just have necessity to sign contracts for their business, and on the other hand, to guarantee legal compliance and effectiveness to contracts by using the knowledge base for the legal and regulatory issues.

The complexity of interactions between SMEs makes this scenario to need another actor, in charge to maintain the Knowledge Base (KB) of legal and regulatory issues, who could be a centralized institution, such as a Government Body. Besides centralized institutions, the Legal Architecture can also support the creation of a particular SME typology, which can be called *Lawyer SMEs* and which could perform the KB maintaining and have the responsibility and authority to ensure legal compliance to contracts signed by DBE-adopter SMEs.

A Lawyer SME should aim, first of all, at the growth of the trust in the DBE and not necessary should aim at the creation and signature of contracts. The Business Model that allows the creation of Lawyer SMEs is the one adopted by the consultings firms. Lawyer SMEs can help a DBE-adopter SME:

- in the creation of contract templates or contracts, receiving a fee from them;
- to authorize and to guarantee the contract (signed by different SMEs) by ensuring its compliance in time, receiving a fee from the DBE-adopter SMEs.
- to maintain for free the Knowledge Base of the DBE Legal Architecture, as upgrading service that they can offer to the DBE-adopter SMEs.

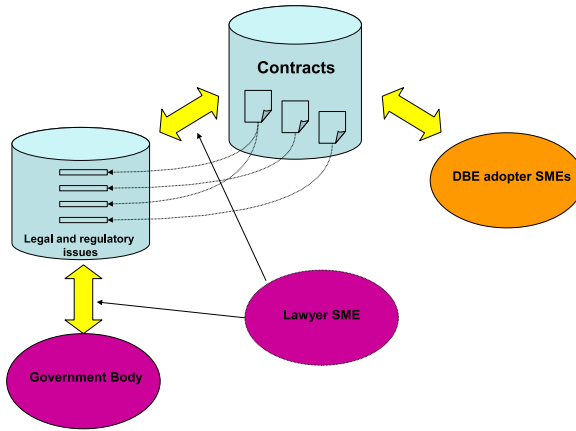


Fig. 1. Scenario

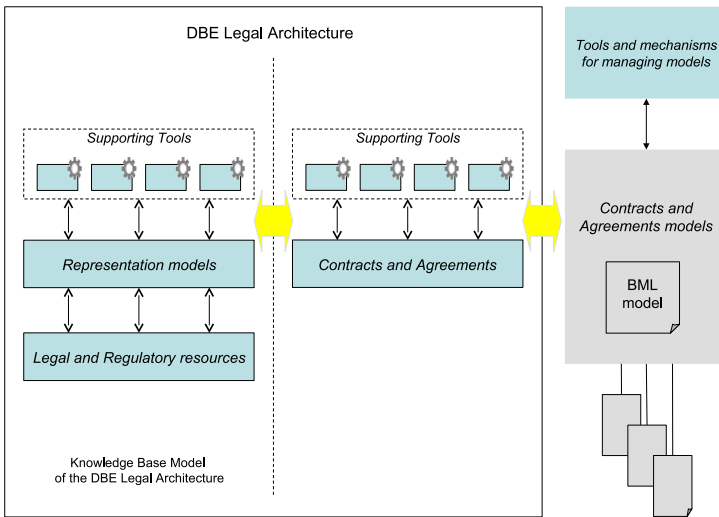
### 3 The DBE Legal Architecture

As previously said, through the DBE Legal Architecture, the DBE-adopter SMEs will be able to access the regulatory knowledge relevant to their needs, and make their contracts legally effective linking them with the proper legal resources. As a consequence, the DBE Legal Architecture can be thought as logically split into two main areas, respectively focused on legal and regulatory knowledge and contracts and agreements (CA), as shown in fig.2.

The first area represents the core element within the architecture. It aims at collecting the legal resources and providing the suitable mechanisms to represent, manage and exploit such resources in an effective way. This means that it is oriented toward actors interested in maintaining the knowledge base (i.e. government bodies and/or lawyer SMEs) as well as other actors that need to access the content of the knowledge base (i.e. DBE adopter SMEs and/or Lawyer SMEs). All the elements encompassed in such area contribute to define the Knowledge Base Model of the Regulatory Issue. Even though the needs addressed by the Knowledge Base Model, as well as its characteristics and components, will be analyzed in the following sections, it is possible to figure out the general structure

of the area, focusing on its main components. As shown in the layered model depicted in the following figure, the involved actors use some specific supporting tools in order to deal with legal resources to reach their specific goals. Anyway, the interaction between such tools and the content of the knowledge base is not direct. An intermediate layer is in fact needed to represent the legal resources and enable their effective management and exploitation.

The second area of the Legal Architecture is mostly focused on contracts and agreements (CA). It aims at enabling lawyer and DBE adopter SMEs to prepare, negotiate and finalise contracts for e-business transactions, providing them with specific supporting tools (e.g. contract creation tools). It is interesting to notice how such area outlines the edge between the Legal Architecture and the external DBE environment. In particular, a contract can have a model, represented through an instance of the BML metamodel[8]. It is planned that these contract templates will be stored in the DBE Knowledge Base[2] and can be accessed by the DBE execution environment when DBE services are executed. This means that a strong relationship exists between the Legal Architecture and BML, as well as with the DBE Knowledge Base.



**Fig. 2.** The DBE Legal Architecture

### 3.1 Knowledge Base Model for the DBE Legal Architecture

The way the Knowledge Base Model is organized, counts of three main layers:

- a distributed knowledge base;
- a set of models for the legal and regulatory issues;
- a set of tools and mechanisms for managing and extracting the knowledge.

A knowledge base, in relation to information technology (IT), is usually defined as a machine-readable resource for the dissemination of information, generally online or with the capacity to be put online. It is used to optimize information collection, organization, and retrieval for an organization, or for the general public. This layer represents the back-end persistent infrastructure that enduringly stores the DBE Legal Architecture's knowledge. As a consequence, it provides the basic services needed to manage such knowledge. The Knowledge Base of the DBE Legal Architecture will contain all the legal and regulatory resources enabling the creation of the legal and regulatory parts of a contract between DBE users. Concerning the scope of the knowledge base, it is possible to refer to:

- generic issues: fundamental regulatory issues for the DBE generic layer, supporting modelling activities for the basic e-services portfolio;
- sector specific issues: more complex range of issues drawn from sector-specific implementation of DBE;
- localised issues: more complex range of issues drawn from local implementation of DBE. Legal Resources should be made available to the system in a standard format enabling interoperability.

Besides the legal resources, the DBE Legal Architecture KB will also contain a raw form of knowledge, held by the system, about these resources. This form of knowledge shall be stored in a structured way and be accessible from each actor.

The intermediate layer is represented by a set of models that enable knowledge extraction and management mechanisms. They include all the structures (semantically enriched) useful to represent the resource in the KB. In particular, this layer is built upon the taxonomies developed for the knowledge base of the regulatory issues, which were intended to identify, classify and assess regulatory issues relevant to the DBE vision. The taxonomies will be used as a framework for investigating regulatory issues arising in sector-specific and local implementation cases and are intended to evolve with the research findings from these stages. The taxonomies are comprised of three basic dimensions used for categorising and organizing research inquiry into regulatory concerns relevant to the DBE vision. The first dimension is that of trust types (X, Y and Z, further subdivided by the DBE layer facets and types of commercial relationships).

The second dimension uses the building blocks of regulatory trust (privacy and consumer protection, e-signatures and security, jurisdiction and consumer protection). The final dimension of the taxonomy draws on the various operational perspectives from which the regulatory issues may be considered in the DBE context (these include the perspective of DBE relationships, DBE actors and software lifecycles) [3].

The third layer contains software tools, interfaces etc. that will implement specific mechanisms for extracting and managing the knowledge of the DBE Legal Architecture. More specifically, in order to provide a generic user with the capabilities specified through the KB requirements of the previous section, this layer will contain:



- a semantic indexer, for uploading legal resources and adding semantic content to them;
- a semantic and syntactic Navigator, for browsing the content of the KB with the help of its models and representations;
- a CA Validator, for providing contracts and contract models with a link to the knowledge base content.

The specific features of these tools, as well as the details related to their design, will be described in the next section.

## 4 Tools for the DBE Legal Architecture

The knowledge capture mechanism will be realized by the functionalities offered by several software tools. In this section we will describe these tools giving an insight on their internal processes.

### 4.1 Semantic Indexer

The Semantic Indexer enables the association of semantic content related to the legal and regulatory issue model to the Knowledge Base (taxonomies) of the Legal Architecture (resources). This task will be performed through the creation of assertions that contains concepts belonging to the taxonomies. The assertions will be linked to each Xpath of a resource in the Knowledge Base. We will make use of semantic assertion in the form of "Xpath is.about concept[instance]". In the following figure, the main activities of the semantic indexing process are represented.

To index a resource in the Knowledge Base, first of all we load the taxonomy (step 1) with which we want to index, then retrieve all the resources and select

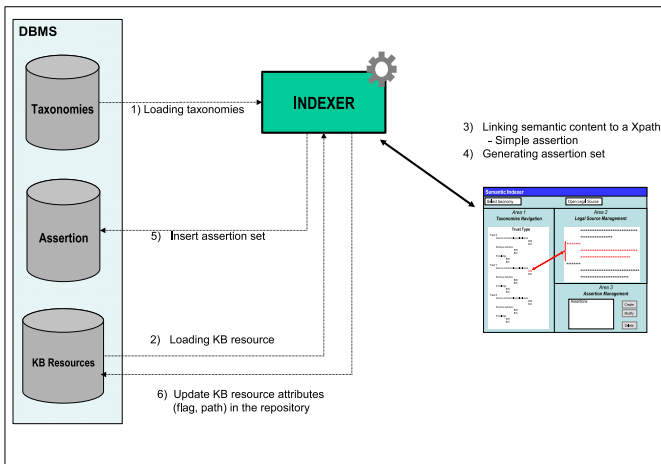


Fig. 3. Semantic indexing process

the one we want to index (step 2). It will be possible to index the entire resource or just a part of it. The following step will be the creation of one or more semantic assertion that create the link between the Xpath and a concept of the taxonomy (step 3-4). This process can be execute more times with the same resource in the Knowledge Base, creating several assertions using different taxonomies. The generated assertions will be stored in a database (step 5) and as consequence the attributes of the resource will be updated (step 6). Assertions can be modified, deleted and added to a resource of the Knowledge Base.

## 4.2 Navigator

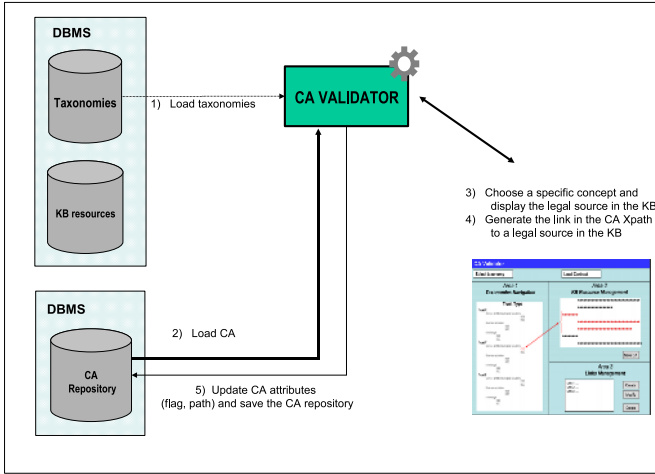
The Navigator is the tool that will search and browse the knowledge base of the Legal Architecture by a *syntactic search* and a *semantic search*. Thus, the Syntactic Navigator is thought to be a full-text research engine that will retrieve a query string, inserted by the user, searching in all the resources in the Knowledge Base (so in both indexed and not indexed resources).

The Semantic Navigation enables a user to browse directly the taxonomy and to retrieve all the resources in the Knowledge Base (for which assertions has been previously created) related to a specific concept. If the Syntactic Navigation is chosen, a user has to insert his query string (step A1) so that the navigator can search and retrieve (step A2) the resource in the Knowledge Base that contains the string wanted. Finally, the list of the resources will be displayed (step A3). To perform a Semantic Navigation, it is necessary to load the taxonomy through which we want to navigate (step B1), pointing out the concept we are interested in (step B2). Doing so, all the resources in the Knowledge Base with semantic assertions that contain that particular concept will be retrieved (step B3-B4). More in particular, for a specific resource the Xpaths contained will be displayed (step B5).

## 4.3 CA Validator

The Contract&Agreement Validator is a tool that enables the user to support its operations in ensuring legal validity to either an existing contract or an empty contract template. It means that the Validator has to be primarily the tool that should support the user who is in charged to give legal validity to a contract. On the other hand a contract validated by a Lawyer SME by using the Validator can be considered, from a DBE-adopter SME's point of view, as a legally valid contract, and it can be visualized using the validator to display the legal resources, which the contract has been linked to. More specifically, the CA Validator is a tool that supports a legal expert of a Lawyer SME, when he/she gives a legal validity to a contract, by linking to the entire contract (or just to part of it) one or more links to a resource in the DBE Legal Architecture Knowledge Base. This tool offers three functionalities:

- the Semantic Navigation of the KB through the taxonomy;
- the creation of link to the KB resources in the Xpath of the contract;
- Navigation of Contracts.



**Fig. 4.** The legal validation process of a contract

While the first and second functionalities are intended to Lawyer SME (in the real validation process), the last is designed for a legal expert of DBE adopter SME that wants know and read all contracts signed by SME. Once the contract is linked with one or more resource of the KB, it can be stored as a contract legally supported by the KB. As shown in figure 5, the legal validation process starts with the load of the taxonomy (1) and then of the contract we want to validate (step 2). By the legal validator user interface, it is possible to visualize the resources related to a specific concept in the taxonomy (step 3). Once a resource is chosen, it is possible to create links to that resource in the Xpath of the contract (step 4). It is also possible to visualize already existing links, modify or delete them. The process ends with the storage of the contract in the CA repository (step 5).

## 5 Conclusion

In this paper, we first focused on the importance to support DBE SMEs in their business activities by providing them with a legal framework capable to support them in their e-business activity. Then we presented the general legal architecture proposed within the DBE project. The Legal Architecture we designed is composed by three layers that model and implement any kind of interactions between all the actors involved. We designed the KB model that represents the basis of the technological infrastructure, that will support the Legal Architecture itself. Effectiveness to the architecture is granted by three processes we modelled, to extract and manage the knowledge of the KB repository, and more specifically, these processes can be implemented within a specific layer of the Legal Architecture, providing DBE users with a common access to the ecosystem legal knowledge.

Next steps will be to implement an effective infrastructure for the tools and the KB repository and test the entire architecture in a digital ecosystem.

## References

1. A. Boer, R. Hoekstra, and R. Winkels. *Metalex: Legislation in xml*, 2002.
2. Nektarios Gioldasis, Nikos Pappas, Fotis G. Kazasis, George Anestis, and Stavros Christodoulakis. A p2p and soa infrastructure for distributed ontology-based knowledge management. In *DELOS Workshop: Digital Library Architectures*, pages 93–104, 2004.
3. G. Gow, K. Glushkova, and S. Elaluf-Calderwood. Generic layer knowledge base. Technical report, London School of Economics, UK, april 2005.
4. Castells M. *The rise of the network society*. Blackwell Publishers ltd., 1996.
5. Z. Milosevic, A. Berry, A. Bond, and K. Raymond. Supporting business contracts in open distributed systems, 1995.
6. P. Dini et al. The Digital Ecosystems Research Vision: 2010 and Beyond. <http://www.digital-ecosystems.org>, 2005.
7. W. W. Powell. Neither market nor hierarchy: Network forms of organization. *Research in Organizational Behavior*, 12:295–336, 1990.
8. M. De Tommasi, Virginia Cisternino, and Angelo Corallo. A rule-based and computation-independent business modelling language for digital business ecosystems. In *KES (1)*, pages 134–141, 2005.

# Evolutionary ANNs for Improving Accuracy and Efficiency in Document Classification Methods

Antonia Azzini and Paolo Ceravolo

Università degli Studi di Milano - Dipartimento di Tecnologie dell'Informazione  
Via Bramante, 65 - 26013 Crema - Italy  
{azzini, ceravolo}@dti.unimi.it

**Abstract.** Approaches to document classification belong to two major families: similarity-based (crisp) classification methods and neural networks (gradual) ones. For gradual techniques, a major open issue is controlling search space dimension. While similarity-based methods identify clusters based on the same number of variables used for document encoding, neural networks automatically identify variables that cause distinctions among clusters. Therefore, the variables' number may vary depending on the documents structure and content, and is difficult to estimate it a priori. This paper proposes a hybrid classification method suitable for heterogeneous document bases like the ones commonly encountered in business and knowledge management applications. Our method is based on an evolutionary algorithm for tuning both neural network's structure and weights. While searching the optimal neural network's configuration it is possible to determine the minimal number of variables to be used in order to classify the given set of documents.

**Keywords:** Ontology Construction, Formal Concept Analysis, Fuzzy Bags, Neural Networks, Genetic Algorithms.

## 1 Introduction

Many methods have been proposed for classifying a set of documents under a categorization. The general idea is that documents having *similar features* belong to the same class.; but taking into account a high number of features would result in an unacceptably wide search space. In order to reduce the search space, the documents have to be transformed from the full text version to a document vector which describes the contents of the document. These vectors can be analyzed and classified with different techniques. Typical examples are neural networks [10], [7], bayesian classifiers [9] and clustering methods [11].

In general, we distinguish between *similarity-based* (crisp) and *neural networks-based* (NN) (gradual) methods. In similarity based methods the number of variables used in order to identify clusters is the same number of variables used for document encoding. These methods are applied to data sets where documents within the same cluster have a high degree of association, while clusters are relatively distinct from one another. On the other hand, fuzzy clustering and NN can work with data sets having less compact and cohesive clusters, because a document can

be associated to a cluster with different degree. The result is that these methods are less susceptible to noisy input data, making them more suitable to a number of knowledge extraction and management applications.

Cluster analysis typically consists of two distinct problems: (i) the determination of the number of clusters; and (ii) the definition of the variables used in order to describe document instances. This second point is very relevant, because variables used for encoding document instances provide the rationale behind the clustering and define the dimension of the search space. It is easy to see that similarity-based methods ensure close control on the variables used for classification, since they are the same selected for documents' encoding. For this reason, similarity-based methods generally have a high level of *accuracy*, i.e. after the set-up phase the frequency of misclassification is low. On the contrary, NN-based methods automatically define the boundaries among clusters; doing so, they provide a high level of *efficiency* because they classify document relying on the minimum number of variables.

In this paper we hybridize these approaches in order to obtain a system having both high *accuracy* and *efficiency*. In our approach, similarity based clustering (SBC) is first executed in order to obtain a classification, then the result are fed into a NN in order to identify the minimal number of variables to be used in the classification. During the training phase we feed the NN with a set of input documents associated to the right output to be returned. Intuition tells us that by identifying the optimal NN configuration we can determine the minimal number of variables to be used in order to classify the set of documents at hand. The main problem related to this idea is that the configuration of a neural network is usually a time-consuming task, requiring an appropriate knowledge.

We tackle this problem by adopting an evolutionary approach that employs evolutionary algorithms (EAs) to select the neural network's configuration. Our technique allows all aspects of NN design to be taken into account at once and does not require expert knowledge of the problem. Through the use of EAs, the problem of designing a NN is regarded as an optimization problem. Some EAs search over the topology space, or for the optimal learning parameters, while others focus on weight optimization. We rely on simultaneous evolution of different aspects of a NN, combining architecture and weight evolution.

This work is organized as follows: Section 2 provides an introduction to evolutionary NNs. Section 3 introduces the document set used in this works and describes the clustering method used for clustering documents. Section 4 explains how ANNs can obtain the same classification optimizing the research space. A sample application of our technique to a XML dataset is also described.

## 2 Evolutionary NNs

The aim of designing NNs based on EAs is both finding an optimal network architecture and training the network on a given data set. The evolutionary approach is designed to be able to take advantage of the backpropagation (BP) algorithm if that is possible and beneficial; however, it can also do without it.

The basic idea is to exploit the ability of the EA to find a solution close enough to the global optimum, together with the ability of the BP algorithm finely tune a nearly-optimal solution and reach the nearest local minimum. This research was prompted by two bio-informatics applications considering the design of neural engine controllers and brain wave signal processing. Recently it has been adapted to business applications like financial modeling. This research is described in detail in [2]; here, we only give an outline of the approach, in order to introduce our technique for document classification. The overall evolutionary process can be described by the following pseudo-code:

1. Initialize the population, either by generating new random individuals or by loading a previously saved population.
2. Create for each genotype the corresponding MLP, and calculate its mean square error (mse), its cost and its fitness values.
3. Save the best individual as the best-so-far individual.
4. While not termination condition do
  - (a) Apply genetic operators to each network.
  - (b) Decode each new genotype into the corresponding network.
  - (c) Compute the fitness value for each network.
  - (d) Save statistics.

The initial population is seeded with random networks initialized with different number of hidden layers and neurons for each individual, that is encoded in a structure in which are maintained basic information about string codification of topology and weights. The dimension of the entire population created is a parameter of the algorithm, but usually, a population of more than 50 individual has been evolved in the neuro-genetic approach. The fitness of an individual depends both on its accuracy and on its cost, therefore, it is proportional to the value of the mean square error and to the cost of the considered network. Firstly, each network is trained with backpropagation algorithm and successively genetic algorithms are applied in order to evolve networks weights and topologies together. The genetic algorithms implemented in the neuro-genetic approach specify a selection strategy, in order to choose and evolve better individuals, reproduction operators, like crossover and two kinds of mutation, different for weights and topologies, and finally, replacement operator, that deletes worst cases obtained during the entire evolutionary process.

For each generation the fitness of all individuals is computed in order to determine the better solution.

### 3 A Similarity Based Classifier

The document corpus used for our experimentation is composed of a set of bibliography entries in an XML format. In particular we consider entries from the **DBLP** and from the **Sigmoid** databases. This kind of corpus requires a clustering method especially tailored for semi-structured data format. In this work we adopted the XML-oriented method exposed in [6]. The main idea of this method

is to use *fuzzy bags* in order to encode structural information of semi-structured documents. Fuzzy bags are a straightforward extension of fuzzy sets, where each element can have multiple instances [14]. In a fuzzy set each element is associated to a membership value. This way, individual terms can be represented as a single entry, while still taking into account cardinality differences; also, irrelevant terms can be easily discarded. This encoding can be very useful in order to extend the traditional representation of documents represented by term vectors. In a fuzzy bags a term with different structural positions in the document can be associated with different membership values. This encoding is used in order to group similar documents by means of distance measures, such as the well-known cosine distance. These groups of documents, loosely called clusters, represent typical classes of the domain. A further step is to order domain classes into a hierarchy. Some approaches rely on preliminary linguistic processing, where terms are analyzed, aggregated into concepts, and progressively organized according to taxonomical relations and rules ([1], [8], [12]). Other extraction methodologies apply *Formal Concept Analysis* (FCA), a time-honored technique used to build hierarchies of common subset of attributes from a set of data items [13].

### 3.1 Representing Term Vectors as Fuzzy Bags

In this section we provide a method for encoding XML documents in terms of fuzzy bags. This construction is based on the notion of *fuzzy integer*, equivalent to the cardinality of a fuzzy bag. In [3] a complete discussion on fuzzy integers and the corresponding inclusion, union and intersection operators is provided, and a set of operations on fuzzy bag is proposed. Our encoding of an XML tree as a fuzzy bag is computed as follows: supposing that every tag  $x_i$  has associated a membership degree  $\mu(x_i)$ , initially equal to 1 ( $\mu(x_i) = 1$ ), we divide every membership degree by its nesting level  $L$  (where  $L_{\text{ROOT}} = 1$ ), obtaining a “lossy” representation of the XML tree that only takes into consideration the original nesting level of tags<sup>1</sup>. Our flat encoding is also sensitive to structural differences between XML documents containing the same group of tags in different position. Fig. 1 shows a cluster of XML documents encoded as follows:

$$\begin{aligned} A &= \{1/R, 0.5/a, 0.5/b\}, \\ B &= \{1/R, 0.5/a, 0.5/b, 0.5/c\}, \\ C &= \{1/R, 0.5/a, 0.5/b, 0.3/c, 0.3/c\}, \\ D &= \{1/R, 0.5/a, 0.5/b, 0.3/c, 0.3/c\}, \\ E &= \{1/R, 0.5/a, 0.5/b, 0.3/c\}. \end{aligned}$$

The XML documents can be encoded in terms of fuzzy integer as follows:

$$\begin{aligned} A &= \{\{0.5/1\} * a, \{0.5/1\} * b\}, \\ B &= \{\{0.5/1\} * a, \{0.5/1\} * b, \{0.5/1\} * c\}, \\ C &= \{\{0.5/1\} * a, \{0.5/1\} * b, \{0.3/2\} * c\}, \end{aligned}$$

<sup>1</sup> The content of leaf tags is contextually saved independently from the fuzzy bag representation.



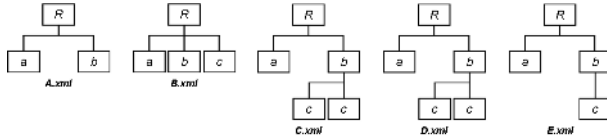


Fig. 1. Representation of XML documents in a cluster

$$D = \{\{0.5/1\} * a, \{0.5/1\} * b, \{0.3/2\} * c\},$$

$$E = \{\{0.5/1\} * a, \{0.5/1\} * b, \{0.3/1\} * c\}.$$

### 3.2 Clustering Fuzzy Bags

After this encoding phase a clustering of the document set is executed. As far as structural similarity is concerned, we need one that is monotonic with respect to element addition and removal[4], such as the *Jaccard* norm:

$$S(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{1}$$

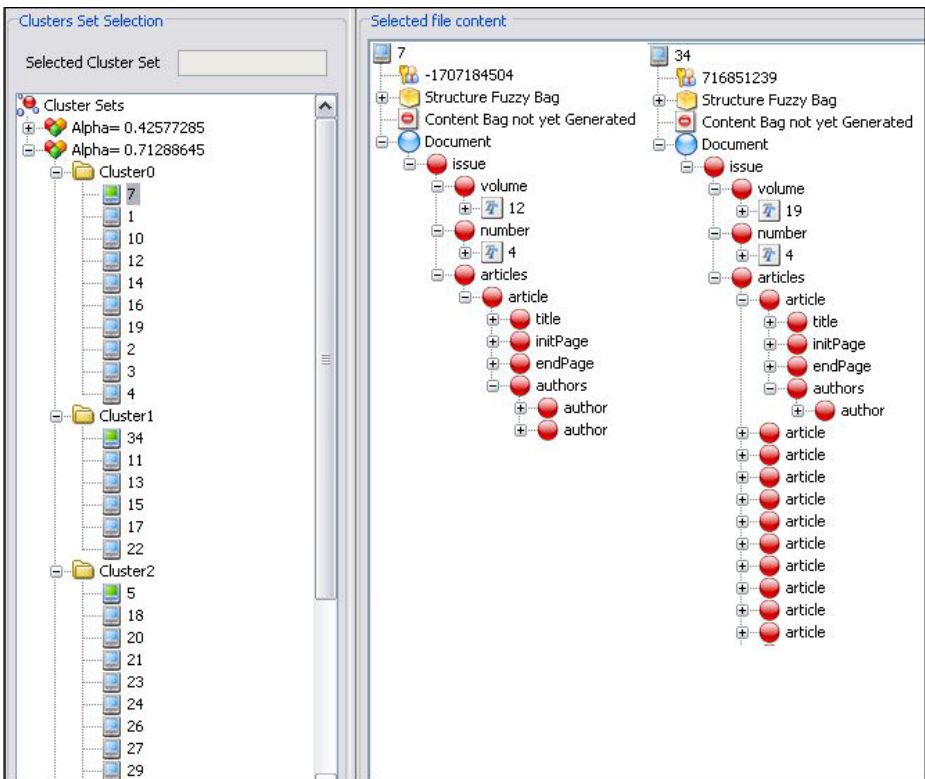
Bags are clustered using  $S$  as a similarity, by means of a hybrid  $k$ -Means and NN clustering algorithm. Specifically, our method uses a tunable  $\alpha$  threshold for the clustering process, in order to avoid suggesting an initial number of clusters ( $k$ ). This way, some well-known problems of the  $k$ -Means algorithm are entirely avoided. Our clustering algorithm compares all items with the centroid of each cluster, considering only the top similarity value. If this value is bigger than  $\alpha$ , the data item is inserted in the cluster; otherwise a new empty cluster is generated and the item is inserted into it. Each cluster is identified by a representative, called *cluster-head*, usually equivalent to the intersection of all documents belonging to the cluster. In order to organize *cluster-heads* into a hierarchy we implement a FCA technique. Following the FCA method we take *cluster-heads* as the context to be analyzed. The notion of a formal context  $\mathcal{K}$ , is defined as a triple  $(\mathcal{D}, \mathcal{A}, \mathcal{R})$ , where  $\mathcal{D}$  represents the data items (called *documents* in the FCA terminology),  $\mathcal{A}$  the attributes, and  $\mathcal{R}$  is a binary relation  $\mathcal{R} \subseteq \mathcal{D} \times \mathcal{A}$ . The final goal we are pursuing is building a hierarchy based on intersections among cluster-heads. In FCA, documents intersections are interpreted as formal concepts. A concept lattice can be obtained evaluating two functions:  $f$  and  $g$ . Function  $f$  maps a set of documents into set of common attributes, whereas  $g$  does the same for attribute sets. In order to compute these formal concepts we have to evaluate the function  $g$ , checking if the codomain of  $f(X)$  is equivalent to the codomain of  $g(Y)$ . In our setting  $g$  and  $f$  can be interpreted as  $t$ -norm, as explained in [5].

## 4 Combining SBC with ENNs

The process we propose is composed of the following phases. Firstly, we apply a similarity based clustering on a set of documents; the result of these classification

are used for the training of the NN. Then, different configurations of the NN are randomly generated. The neuro-genetic approach progressively evolves those configurations obtaining the NN configuration that represents the best solution. The result is a NN classifying our corpus of documents with the minimal number of variables but following the classification driven by the SBC.

Running an algorithm following the lines described in section 3 on our sample data, we obtain the classification shown in Fig. 2. The example provided here is based on the result obtained running OntoExtractor, a tool developed by the Knowledge Management Group of the university of Milan. This tool prompts on the left frame clusters resulting from the execution. The documents belonging to a cluster are listed below it. The first document in the list is the *cluster-head*, and its inner structure can be shown in the right frame.



**Fig. 2.** The clustering result produced by OntoExtractor

As shown in Figure2, clusters are mainly distinguished by the number of tags *article*, that identify issues describing journals or proceedings. Since the best solution found by the evolutionary process defines structure and weights of the neural network that works as the best classifier, it will rely mainly on these tags

to differentiate clusters. In other words, the tag *article* will become the main feature used by the neural network for cluster identification.

Initially, all networks of the population are trained with backpropagation algorithm, basing on training set of classification previously defined. During the next generations, networks' evolution is driven by a genetic algorithm applying selection, crossover, mutation and replacement operations. New individuals are created during the evolutionary steps and are trained with the same datasets. For each generation, the best individual (i.e. neural network) found is saved; the result of all evolutionary process is a NN classifying our corpus of documents with the minimal number of variables.

## 5 Conclusions

In this paper we proposed the combination of SBC and ENNs for improving accuracy and efficiency of document classification. A prove of concept of the rationale behind our methodology was provided testing a simple document set. We claim the encouraging results obtained justify further testing to be executed on a more sophisticated document corpus of heterogeneous business documents.

## Acknowledgment

This work was partly funded by the Italian Ministry of Research Fund for Basic Research (FIRE) under projects RBAU01CLNB\_001 "Knowledge Management for the Web Infrastructure" (KIWI).

## References

1. K. Ahmad and A. E. Davies, Weirddness in Special-language Text: Welsh Radioactive Chemicals Texts as an Exemplar. Internationales Institut fr Terminologieforschung Journal: 22–52, 1994.
2. A. Azzini and A.G.B. Tettamanzi, 'A neural evolutionary approach to financial modeling', in *GECCO 2006 - Genetic and Evolutionary Computation Conference*, ed., to appear, (July 8-12 2006).
3. P. Bosch and D. Rocacher, The set of fuzzy rational numbers and flexible querying. *Fuzzy Sets and Systems*, 2005.
4. B. Bouchon-Meunier, M. Rifqi, and S. Bothorel, "Towards general measures of comparison of objects," *Fuzzy Sets Syst.*, vol. 84, no. 2, pp. 143–153, 1996.
5. P. Ceravolo, E. Damiani, M. Viviani, Extending formal concept analysis by fuzzy bags. Proceedings of IPMU, Paris, 2006.
6. P. Ceravolo, A. Corallo, E. Damiani, G. Elia, M. Viviani, A. Zilli. Bottom-up extraction and maintenance of ontology-based metadata. in: *Fuzzy Logic and the Semantic Web*, Elie Sanchez ed., Elsevier, 2006.
7. Farkas, Jennifer. Generating Document Clusters Using Thesauri and Neural Networks. Canadian Conference on Electrical and Computer Engineering, Vol.2, p. 710-713, 1994.

8. D. Faure and T. Poibeau. First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In Proceedings of the ECAI Workshop on Ontology Learning, 2000.
9. M. Iwayama and T. Tokunaga. Hierarchical bayesian clustering for automatic text classification. IJCAI-95. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol.2, p. 1322-7, 1995.
10. W. Li, B. Lee, F. Krausz and K. Sahin. Text classification by a neural network. Proceedings of the 1991 Summer Computer Simulation Conference. Twenty-Third Annual Summer Computer Simulation Conference, p. 313-318, 1991.
11. Y. Li and A. K. Jain. Classification of text documents. Proceedings of IJCAI-95. International conference on pattern recognition, IEEE computer society press, Vol 14, p. 1295-1297, 1998.
12. U. Hahn and K. G. Marko, Ontology and lexicon evolution by text understanding. In Proceedings of the ECAI Workshop on Machine Learning and Natural Language Processing for Ontology Engineering, 2002.
13. R. Wille, Restructuring lattice theory: An approach based on hierarchies of concepts, *Ordered Sets*: 445-470, 1982.
14. R. Yager, Cardinality of fuzzy sets via bags, *Math. Modelling*, 9(6): 441-446, 1987.

# A Methodology for Determining the Creditability of Recommending Agents

Omar Khadeer Hussain<sup>1</sup>, Elizabeth Chang<sup>1</sup>,  
Farookh Khadeer Hussain<sup>1</sup>, and Tharam S. Dillon<sup>2</sup>

<sup>1</sup> School of Information Systems, Curtin University of Technology, Perth, Australia  
{omar.hussain, elizabeth.chang,  
farookh.hussain}@cbs.curtin.edu.au

<sup>2</sup> Faculty of Information Technology, University of Technology, Sydney, Australia  
tharam@it.uts.edu.au

**Abstract.** The trusting agent in order to analyze the Risk that could be present in its future interaction with a trusted agent might solicit for its recommendations from other agents. Based on the recommendations achieved the trusting agent can decide whether to interact or not with the trusted agent. If the trusting agent decides to proceed, then after its interaction it should adjust the creditability of the recommending agents, whose recommendation it considered. Doing this would help the future trusting agents to classify the recommending agents according to their trustworthiness and ignore those which are un-trustworthy. In this paper we propose such an approach by which the trusting agent adjusts the credibility of the recommending agent after its interaction depending on the recommendation that it gave.

## 1 Introduction

The Australian and New Zealand Standard on Risk Management, AS/NZS 4360:2004, states that Risk Identification is at the heart of risk management [1]. Hence, risk should be identified according to the context of the transaction in order to analyze and manage it better. Risk analysis is the science of evaluating risks resulting from past, current, anticipated or future activities. The amount of risk inherent in a transaction must be understood or analyzed before a transaction begins. This also applies to the transactions in the field of e-commerce and peer-to-peer business. Risk is a combination of: a) the uncertainty of an outcome; and b) the cost of the outcomes when it occurs, usually the loss incurred.

Analyzing risk is important in e-commerce transactions and there is a whole body of literature based in rational economics that argues that the decision to buy is based on the risk-adjusted cost-benefit analysis [2]. Thus, it commands a central role in any discussion of e-commerce that is related to a transaction. Risk plays a central role in deciding whether to proceed with a transaction or not. It can broadly be classified as an attribute of decision making that reflects the variance of its possible outcomes.

We have developed an approach by which the trusting agent can determine the risk beforehand that may be present in its interaction with a trusted agent by assimilating the recommendations communicated by the recommending agents. In this paper, we propose a novel approach where the trusting agent, after its interaction with the

trusted agent, adjusts the creditability of the recommending agents depending on the recommendations communicated by them. This paper is organized into five sections. In section 2 we discuss previous related work and formalize the problem; in section 3 we propose a methodology for the creditability adjustment of the recommending agents; in section 4 we explain the methodology through example and in section 5 we conclude the paper.

## 2 Related Work

From the above discussion it can be concluded that risk analysis is necessary before carrying out a business interaction. In order to analyze the risk in an interaction, we defined the term *Riskiness* in Hussain, Chang, Hussain & Dillon [3]. Riskiness is defined as *the numerical value that is assigned by the trusting agent to the trusted agent after the interaction, which shows its level of Risk on the Riskiness scale.* The Riskiness scale, as shown in Figure 1, has seven levels ranging from -1 to 5. The Riskiness scale has six levels to represent each type of risk and one level to represent *Unknown Risk*.










Riskiness Levels	Magnitude of Risk	Riskiness Value	Star Rating
Unknown Risk	-	- 1	Not Displayed
Totally Risky	91 - 100 % of Risk	0	Not Displayed
Extremely Risky	71 - 90 % of Risk	1	From  to 
Largely Risky	51 - 70 % of Risk	2	From  to 
Risky	26 - 50 % of Risk	3	From  to 
Largely UnRisky	11 - 25 % of Risk	4	From  to 
UnRisky	0 - 10 % of Risk	5	From  to 

Fig. 1. The Riskiness scale

The Riskiness value assigned by the trusting agent to the trusted agent depends on the level of un-commitment versus promised commitment. The promised commitment is the expected behaviour that is defined by the trusting agent according to its criteria before initiating its interaction with the trusted agent. This is the behavior by which the trusted agent is expected to behave in the interaction. Criteria are defined as the set of factors or bases that the trusting agent wants in its interaction with the trusted agent and later against which it will determine the un-commitment behavior of the trusted agent. The greater the degree of un-commitment behavior with the promised commitment the higher the Risk present in it and vice versa. The process by which the trusting agent assigns a Riskiness value to the trusted agent depending on the level of its un-committed behaviour is defined in Hussain et al. [3] and we will not be discussing it here due to space limitation.

But the Riskiness value assigned to the trusted agent by the trusting agent is after its interaction with it. As mentioned in section 1, the decision to proceed in an interaction is based on the risk adjusted cost benefit analysis. It would be much easier for the trusting agent to decide whether to proceed or not in an interaction with a trusted agent, or to choose a trusted agent to interact with among a set of possible

trusted agents, if it knows beforehand the possible level of risk that could be present in its future interaction. In order for the trusting agent to analyze the risk beforehand that could be present, two scenarios arise. They are:

**Scenario 1:** If the trusting agent has interacted with the trusted agent previously in the same context using the same criteria, then it can determine the risk that could be present in their interaction by analyzing the Riskiness value that it assigned to the trusted agent in its previous interactions.

**Scenario 2:** If the trusting agent has not interacted with the trusted agent previously OR in the same context with the same criteria as that of its future interaction then it can determine the risk that could be present in their future interaction by analyzing the reputation of the trusted agent in that particular context and criteria. Reputation of the trusted agent can be used as an alternative when its Riskiness value is not known. Reputation of a possible trusted agent can be determined by the trusting agent through soliciting for its recommendations from other agents in the context of its future interaction with it. The agents who have interacted with the particular trusted agent in question and in context similar to that, for which the recommendations are being solicited for, reply back with their recommendation. The agents replying back with the recommendations are called *Recommending Agents*. Once the trusting agent receives the recommendations, it can assimilate them according to its criteria and determine the reputation or the Riskiness value of the trusted agent. Based on the reputation determined, the trusting agent can analyze the possible risk that could be present in interacting with the trusted agent and can decide whether to interact with it or not.

We have developed a methodology in Hussain, Chang, Hussain, Dillon and Soh [4] where a trusting agent can determine the Riskiness value or the reputation of a trusted agent in each criterion of its future interaction by assimilating the recommendations. We proposed that while assimilating the recommendations the trusting agent considers only Trustworthy and Unknown recommendations and omits the untrustworthy recommendations in order to reduce the Risk. We will not be discussing the methodology in this paper due to space limitation.

After determining the riskiness or reputation of the trusted agent if the trusting agent decides to proceed in an interaction then, based on the outcome of its interaction, it should adjust the creditability of all the recommending agents whose recommendation assisted in the decision. Adjusting the creditability of the recommending agents would help any future trusting agent soliciting recommendations, to determine whether the recommending agent is giving a viable recommendation or not. As mentioned in Hussain et al. [4], the creditability of the recommending agents is the RRP value that is associated with each recommending agent. This value determines whether the recommendation communicated by that agent is trustworthy or not. We will discuss the process of adjusting the creditability or the RRP of the recommending agents in the next section.

### 3 Adjusting Creditability of the Recommending Agents

In order to get a thorough understanding of the problem, let us consider this example:

A trusting agent 'A' wants to interact with a trusted agent 'B' over context 'C' on 20/04/2006. The criteria in its interaction are C1, C2 and C3. The trusting agent 'A'

has not had any previous interactions with the trusted agent 'B' in the context of its future interaction. To analyze the risk before proceeding in a business interaction with it, agent A solicits recommendations from other agents in the range of the past month. Hence, the *time space* [5] is of one month. 'A' divides the time space into 2 *time slots* [5] each of 15 days i.e. one slot from 21/03/2006 to 04/04/2006 and the second time slot from 05/04/2006 to 19/04/2006. Of the recommendations received in the time space of 21/03/2006 to 19/04/2006 the trusting agent should apply greater importance to recommendations in the time slot of 05/04/2006 to 19/04/2006 as it is near the *time spot* [5] of its future interaction with the trusted agent.

Let us suppose that replies are received from agents 'D', 'E', 'F' and 'G' in the form of Risk set [6]. All these agents have interacted with the trusted agent 'B' previously over context 'C'. Let us suppose:

**Recommendation from agent 'D':**

{Agent 'D', Agent 'B', C, 4, 4, ((C1, 1) (C5, 0)), 3, \$1000, 13/04/2006, 15/04/2006, -1}

**Recommendation from agent 'E':**

{Agent 'E', Agent 'B', C, 3, 4, ((C5, 1) (C6, 1)), 4, \$500, 10/04/2006, 11/04/2006, 1}

**Recommendation from agent 'F':**

{Agent 'F', Agent 'B', C, 3, 2, ((C1, 1)(C6, 1)), 2, \$200, 22/03/2006, 25/03/2006, UNKNOWN}

**Recommendation from agent 'G':**

{Agent 'G', Agent 'B', C, 4, 5, ((C1, 1) (C3, 1)), 5, \$1200, 15/4/2006, 16/4/2006, 2}

From the recommendations it can be seen that:

1. Recommendation from agent 'D' is a trustworthy recommendation as its RRP is in the range of (-1, 1) and the criteria in which it interacted with the 'B' is C1 and C5.
2. The recommendation from agent 'E' is trustworthy but the criteria in its interaction are C5 and C6. It is baseless for 'A' to consider this recommendation as the criteria do not match, even though the context is same
3. The RRP of 'F' is unknown and the criteria of its recommendation are C1 and C6.
4. The criteria of the recommendation from 'G' are C1 and C3 but the recommendation is un-trustworthy as its RRP is not in the range of (-1, 1). The trusting agent 'A' will not consider this recommendation.

Once the trusting agent receives the recommendations then it can utilize the methodology mentioned in Hussain et al. [4] and determine the Riskiness value of the trusted agent in each criteria of its interaction i.e. C1, C2 and C3.

For explanation sake, let us assume that the trusting agent 'A' after analyzing the risk proceeds in the interaction with the trusted agent 'B'. Then after completing its interaction, 'A' should adjust the credibility or the RRP of the agents from which it took recommendations namely, agents 'D' and 'F' depending on its outcome and the what those agents recommended.

We think that the creditability of a recommending agent should be adjusted by considering only the criterion in which it offered recommendations and not by considering the criterions of the whole interaction. For example, agent 'F' is communicating its recommendation in criteria C1 and C6. The trusting agent 'A' wants recommendations in criterion C1, C2 and C3. Hence, it will take recommendation from 'F' only in criterion C1. After its interaction with the trusted



agent ‘B’, the trusting agent ‘A’ should adjust the creditability of agent ‘F’ only according to the outcome of criterion C1 and not according to the whole criterions i.e. C1, C2 and C3.

Adjusting the creditability or the RRP of a recommending agent ‘R’ after an interaction, can be done by determining the deviation in the criterion commitment that the trusting agent found out after interacting with the trusted agent, and what ‘R’ recommended to the trusting agent for that particular criterion before its interaction. The deviation in the recommendation, when weighed with the trustworthiness of ‘R’ and the significance of the criterion gives the accurate measure of adjustment that is to be done to the RRP of the recommending agent.

The adjustment to be made to ‘R’s’ credibility according to the recommendation that it gave to the trusting agent ‘TA’ in criterion ‘C’ can be determined by:

$$A_C^R = (\alpha * ((\gamma * (\text{Com}_{TA}^C - \text{Com}_R^C)) + (\delta * (\text{Com}_{TA}^C - \text{Com}_R^C)))) + (\beta * ((\gamma * (\text{Com}_{TA}^C - \text{Com}_R^C)) + (\delta * (\text{Com}_{TA}^C - \text{Com}_R^C)))) \quad (1)$$

Where:

$\text{Com}_{TA}^C$  represents the commitment level of the trusted agent determined by the trusting agent ‘TA’ in criterion ‘C’ after the interaction,

$\text{Com}_R^C$  represents the commitment level recommended by the recommending agent ‘R’ to the trusting agent ‘TA’ for the trusted agent in criterion ‘C’ before the interaction,

$\gamma$  and  $\delta$  are the variables attached to the parts of the equation which gives more weight to the creditability adjustment if the recommendation was in the recent time slot of the trusting agent’s future interaction with the trusted agent as compared to the far recent ones respectively. In general  $\gamma > \delta$  and  $\gamma + \delta = 1$ ,

$\alpha$  and  $\beta$  are the variables attached to the parts of the equation which will gives more weight to the credibility adjustment if the recommendation was from a trustworthy recommending agent as compared to from an unknown recommending agent respectively. In general  $\alpha > \beta$ , and  $\alpha + \beta = 1$ .

The first part of the above equation calculates the adjustment to be made in the recommending agent ‘R’ credibility if it was a *Trustworthy recommending agent* while giving this recommendation. Similarly the second part of the same equation is used to calculate the adjustment to be made in agent ‘R’ credibility if it was an *Unknown recommending agent* while giving this recommendation. The trusting agent in Hussain et al. [4] considers only these two types of recommendations when it assimilates them before starting an interaction to determine the reputation of the trusted agent and, hence, adjusts the creditability of only those after its interaction.

If there is a discrepancy between the commitment level that the trusting agent found out after its interaction and the commitment level communicated by the recommending agent prior to interaction, then we believe that the degree of adjustment in the creditability of the recommending agent should be more if it was a trustworthy recommending agent as compared to if it was an unknown recommending agent. This is done by the weights  $\alpha$  and  $\beta$  attached to equation 1. It is because in the methodology discussed in Hussain et al. [4] that the trusting agent, while determining the reputation of the trusted agent by assimilating the recommendations of the recommending agents, gives more weight to recommendations from agents who are

*trustworthy* in giving them as compared to ones who are *unknown*. Hence, if there is any deviation in the recommendation then after its interaction the credibility adjustment too should be done accordingly.

Once the credibility adjustment to be done for a recommending agent ‘R’ in criterion ‘C’ has been determined, then it should be weighed with the significance of that criterion according to the trusting agent. All the criteria of an interaction will not be of equal importance or significance. The significance of each criterion in an interaction might depend on the degree to which it influences the successful outcome of the interaction according to the trusting agent. So the credibility adjustment of the recommending agent in a criterion too should be done according to its significance. The possible levels of significance for a criterion are shown in Table 1.

**Table 1.** Showing the significance level of each criterion

Significance level of the Criterion (Sc)	Significance Rating and Semantics of the level
1	Minor Significance
2	Moderate Significance
3	Large Significance
4	Major Significance
5	High or Extreme Significance

Determining the adjustment  $A^R$  to be made to the recommending agent’s ‘R’ credibility in a criterion according to its significance:

$$A^R = \frac{1}{TS} \left( \sum_{i=1}^n S_{Ci} * A^R_{Ci} \right) \tag{2}$$

Where  $A^R$  denotes the final adjustment to be made to the credibility of the recommending agent ‘R’,

TS represent the total significance of the criterions in the interaction according to the trusting agent,

$S_{Ci}$  represents the significance of the criterion ‘Ci’ in which the recommending agent gave its recommendation,

$A^R_{Ci}$  represents the adjustment to be made to the recommending agent’s ‘R’ credibility according to the recommendation that it gave to the trusting agent ‘TA’ in criterion ‘Ci’.

Finally, adjusting the credibility of the recommending agent:

$$RRP_{NEW\ R} = RRP_{OLD\ R} \oplus A^R \tag{3}$$

Where  $\oplus$  is the adjustment operator.

$RRP_{NEW\ R}$  will become the Riskiness value of ‘R’ when it is communicating recommendation any time in the future. The proposed concept will become clear when we explain it by using an example in the next section.

### 4 Illustrating with a Real World Example

Continuing our discussion from the previous section, the trusting agent ‘A’ in order to determine the reputation of the trusted agent ‘B’ solicits for recommendations

from other agents in the time space of 1 month. The trusting agent ‘A’ divided the time space into 2 time slots each of 15 days. It gets recommendations from agents ‘D’, ‘E’, ‘F’ and ‘G’. The trusting agent will consider recommendations from Agent ‘D’ and Agent ‘F’ as they both are communicating recommendations in the criterions of its interest. The rest of them are either un-trustworthy recommendations or deal with other criterions. Let us assume that the trusting agent ‘A’ after determining the reputation of the trusted agent ‘B’ decides to proceed ahead in the interaction with it.

Further let us assume that after the interaction, the commitment level found out for the trusted agent ‘B’ by the trusting agent ‘A’ in the criterions C1, C2 and C3 of its interaction are 0, 1 and 1 respectively. Further, the significance of the criterions C1, C2 and C3 according to the trusting agent ‘A’ is 4, 3 and 5 respectively.

In order to adjust the creditability of the recommending agents ‘D’ and ‘F’ according to the recommendation that they gave, the trusting peer has to first determine the deviation in the criterion commitment level that it found out and what the recommending agents recommended for that criterion.

#### 4.1 Adjusting Creditability of Agent ‘D’

Determining and representing in table 2 the deviation in the commitment level for criterion C1 between what the trusting agent ‘A’ found after its interaction and what the recommending agent ‘D’ recommended:

**Table 2.** Determining deviation in recommendation of Agent ‘D’

Criterion C1	Commitment Level
Commitment level determined by the trusting agent	0
Commitment level recommended by the recommending agent	1

As can be seen from the risk set, agent ‘D’ is a trustworthy recommending agent and its recommendation is in the recent time slot of the trusting agent’s future interaction with the trusted agent. Let us assume that the trusting agent ‘A’ gives a weight of 0.8 and 0.2 to recommendations from trustworthy and unknown recommending agents and a weight of 0.6 and 0.4 to the recommendations in the recent time slot and in the far recent ones respectively.

Utilizing equation 1 to determine the adjustment to be made to the creditability of the recommending agent ‘D’ according to the recommendation it gave in criterion C1:

$$A_{C1}^D = (0.9 * (0.6 * (0-1))) + 0 + 0 + 0 = - 0.54$$

Using equation 2 to determine the adjustment to be made to the recommending agent’s ‘D’ creditability according to the significance of the criterion:

$$A^D = \frac{1}{12} (4 * -0.54) = -0.18$$

Finally, adjusting the creditability of the recommending agent by utilizing equation 3:

$$RRP_{NEW D} = -1 \oplus (-0.18) = -1.18$$

### 4.2 Adjusting Creditability of Agent ‘F’

Determining and representing in table 3 the deviation in the commitment level for criterion C1 between what the trusting agent ‘A’ found after its interaction and what the recommending agent ‘F’ recommended:

**Table 3.** Determining deviation in recommendation of Agent ‘F’

Criterion C1	Commitment Level
Commitment level determined by the trusting agent	0
Commitment level recommended by the recommending agent	1

From the risk set it can be seen that agent F’s RRP is unknown and its recommendation is in the far recent time slot of the trusting agent’s future interaction with the trusted agent.

Utilizing equation 1 to determine the adjustment to be made to the creditability of the recommending agent ‘F’ according to the recommendation it gave in criterion C1:

$$A_{C1}^F = 0 + 0 + 0 + (0.1 * (0.4 * (0-1))) = -0.04$$

Using equation 2 to weigh  $A_{C1}^F$  with the significance of the criterion:

$$A^F = \frac{1}{12} (4 * -0.04) = -0.01$$

Finally, adjusting the creditability of the recommending agent by utilizing equation 3:

$$RRP_{NEW F} = -0.01$$

From the above examples it can be seen that:

1. The RRP of agent ‘D’ has become -1.18 due to the incorrect recommendation that it gave to the trusting agent in criterion C1.
2. The RRP of agent ‘F’ has changed from Unknown to -0.01 after the interaction.

## 5 Conclusion

In this paper, we proposed a novel approach by which the trusting agent can adjust the creditability of the recommending agents after its interaction. This would considerably help the future trusting agents, soliciting for recommendations to classify them according to its trustworthiness and discard those which are untrustworthy. In our approach the creditability of the recommending agents is adjusted by considering only those criterions in which they communicated their recommendation and not by considering the total criteria of the interaction. By doing so, the recommending agent is adjusted with the accurate adjustment to its credibility that it deserves according to the recommendation that it gave.

## References

1. Cooper, D.F., 'The Australian and New Zealand Standard on Risk Management, AS/NZS 4360:2004', Tutorial Notes: Broadleaf Capital International Pty Ltd. (2004) Available: [http://www.broadleaf.com.au/tutorials/Tut\\_Standard.pdf](http://www.broadleaf.com.au/tutorials/Tut_Standard.pdf)
2. Greenland, S., 'Bounding analysis as an inadequately specified methodology', *Risk Analysis* vol. 24, no. 5, (2004), 1085-1092.
3. Hussain, O.K., Chang E., Hussain, F.K & Dillon, T.S., 'A methodology for risk measurements in e-transactions', *Special Issue of International Journal of Computer System, Science and Engineering*, CRL Publishing Ltd, UK, (2006), 17-31.
4. Hussain, O.K., Chang E., Hussain, F.K, Dillon, T.S. & Soh, B., 'Context and time based riskiness assessment for decision making', *Proceedings of the International Conference on Internet and Web Applications and Services (ICIW)*, Guadeloupe, French Caribbean, (2006), pp. 104-109.
5. Hussain, O.K., Chang E., Hussain, F.K, Dillon, T.S. & Soh, B., 'Predicting the dynamic nature of risk', *Proceedings of the 4th ACS/IEEE International Conference on Computer Systems and Applications*, Dubai/Sharjah, UAE, (2006), pp. 500-507.
6. Hussain, O.K., Chang E., Hussain, F.K, Dillon, T.S., and Soh, B., 'Modelling the Risk Relationships and Defining the Risk Set' *Proceedings COLLECTeR Latam Chile* (2005), 1-9.

# How to Solve a Multicriterion Problem for Which Pareto Dominance Relationship Cannot Be Applied? A Case Study from Medicine

Crina Grosan<sup>1</sup>, Ajith Abraham<sup>2</sup>, Stefan Tigan<sup>3</sup>, and Tae-Gyu Chang<sup>1</sup>

<sup>1</sup> Department of Computer Science

Babeş-Bolyai University, Cluj-Napoca, 3400, Romania

<sup>2</sup> IITA Professorship Program, School of Computer Science and Engineering

Chung-Ang University, Seoul 156-756, Korea

<sup>3</sup> Department of Medicine

University Iuliu Hatieganu, Cluj-Napoca, 3400, Romania

`cgrosan@cs.ubbcluj.ro`

**Abstract.** The most common way to deal with a multiobjective optimization problem is to apply Pareto dominance relationship between solutions. The question is: how can we make a decision for a multiobjective problem if we cannot use the conventional Pareto dominance for ranking solutions? We will exemplify this by considering a multicriterion problem for a medical domain problem. Trigeminal Neuralgia (TN) is a pain that is described as among the most acute known to mankind. TN produces excruciating, lightning strikes of facial pain, typically near the nose, lips, eyes or ears. Essential trigeminal neuralgia has questioned treatment methods. We consider five different treatment methods of the essential trigeminal neuralgia for evaluation under several criteria. We give a multiple criteria procedure using evolutionary algorithms for ranking the treatment methods of the essential trigeminal neuralgia for the set of all evaluation criteria. Results obtained by our approach using a very simple method are the same as the results obtained by applying weighted sum method (which requires lots of domain expert input). The advantage of the new proposed technique is that it does not require any additional information about the problem (like weights for each criteria in the case of weighted sum approach).

## 1 Introduction

It is generally accepted that multicriterion optimization in its present sense originated towards the end of the last century when Pareto (1848-1923) presented a qualitative definition for the optimality concept in economic problems with several competing criteria.

Instead of one scalar objective function, usually several conflicting and often non-commensurable (i.e. such quantities which have different units) criteria appear in an optimization problem. This situation forces the designer to look for a good compromise solution by considering tradeoffs between the competing criteria. Consequently, he must take a decision-maker's role in an interactive design

process where typically several optimization problems must be solved. Multicriterion (multiobjective, Pareto, vector) optimization offers a flexible approach for the designer to treat such an overall decision-making problem in a systematic way [9], [10].

There are some particular situations for which Pareto dominance cannot be applied. This paper analyzes a multiobjective optimization problem for a medical domain problem. As evident from the considered test data, Pareto dominance cannot be applied in its initial form for classifying these treatments. The result will be that all solutions are non-dominated (which, in fact, means that all treatments are equal). This paper proposes an evolutionary algorithm to rank these treatments. A new dominance concept between two solutions is used. Results obtained are similar to the ones obtained by applying the weighted sum method.

The paper is structured as follows. Section 2 provides a short description of the Trigeminal Neuralgia. Sections 3 and 4 explain the scope of the present research and the motivation of the work done. Section 5 introduces and explains our approach. Section 6 presents the weighted sum method used for comparing the results of the proposed approach. Section 7 is dedicated to the experiments. Section 8 contains discussions and conclusions of the paper.

## 2 What Is Trigeminal Neuralgia?

Trigeminal neuralgia, also called tic douloureux, is a condition that affects the trigeminal nerve (the 5th cranial nerve), one of the largest nerves in the head. The trigeminal nerve is responsible for sending impulses of touch, pain, pressure, and temperature to the brain from the face, jaw, gums, forehead, and around the eyes. Trigeminal neuralgia is characterized by a sudden, severe, electric shock-like or stabbing pain typically felt on one side of the jaw or cheek. The disorder is more common in women than in men and rarely affects anyone younger than 50. The pain produced by trigeminal neuralgia is excruciating, perhaps the worst pain known to human beings. The attacks of pain, which generally last several seconds and may be repeated one after the other, may be triggered by talking, brushing teeth, touching the face, chewing, or swallowing. The attacks may come and go throughout the day and last for days, weeks, or months at a time, and then disappear for months or years [1], [5], [7]. Treatment for trigeminal neuralgia typically includes anticonvulsant medications such as carbamazepine or phenytoin. Baclofen, clonazepam, gabapentin, and valproic acid may also be effective and may be used in combination to achieve pain relief.

## 3 Scope of Our Research

The problem studied is the treatment of essential trigeminal neuralgia. For the treatment of essential trigeminal neuralgia many methods can be applied. The chronic evolution of the disease, its idiopathic character and the variable response to different treatment methods creates many disputes in the scientific world. The

evaluation of the treatment methods from multiple points of view is difficult and has a high degree of subjectivity. The complex and original study with many patients, over the usual number from related studies, can contribute greatly to the evolution of this domain. The problem is to rank these treatments subject to multiple criteria.

## 4 Motivation of the Work

The problem of effectively ranking several treatments for Trigeminal Neuralgia could be formulated as a multiobjective optimization problem due to the number of different criteria which have to be satisfied simultaneously. The case analyzed in this research is a real problem [2]. The most common approaches of a multi-objective optimization problem use the concept of Pareto dominance as defined below:

**Definition** (*Pareto dominance*).

Consider a maximization problem. Let  $x, y$  be two decision vectors (solutions) from the definition domain. Solution  $x$  *dominate*  $y$  (also written as  $x \succ y$ ) if and only if the following conditions are fulfilled:

- (i)  $f_i(x) \geq f_i(y), \forall i = 1, 2, \dots, n,$
- (ii)  $\exists j \in \{1, 2, \dots, n\}: f_j(x) > f_j(y).$

$n$  denotes the number of objectives. That is, a feasible vector  $x$  is Pareto optimal if no feasible vector  $y$  can increase some criterion without causing a simultaneous decrease in at least one other criterion.

As evident from the experiment section (Table 1), if we are applying the classical Pareto definition in order to obtain a hierarchy of treatments, all solutions will appear as non-dominated. This way, it is difficult to say one solution is better than the other. Consequently, any of the existing algorithms dealing with multiobjective problems from a Pareto dominance perspective cannot be applied. One solution is to use some of the traditional mathematical approaches which combine objectives and reduce the problem to a single objective optimization problem. But in this situation, additional information about the problem is required. For instance, every common approach will need details about the importance of each of the criteria. In this situation, a weight (i.e. a real number between 0 and 1) will be assigned to each criterion. This weight represent the importance (or the percentage) of that criteria between all criteria considered (sum of all these weights is equal to 1).

But in the case analyzed in this paper, all criteria are important since all of them are direct consequences of a treatment applied. So, finding a weight for each criterion is an extra task and can be sometimes difficult to assign.

## 5 Proposed Approach

An Evolutionary Algorithm (EA) [3], [4] approach is proposed for solving this problem. The population is initially randomly generated over the search space



which is the definition domain. By applying genetic operators (like selection, mutation, crossover, etc.) these solutions (called also chromosome, individuals) are improved. Each individual from population is evaluated by using a quality (fitness) function. Using this quality the best individuals are selected at each generation.

Since, in our case, the final solution has to be a hierarchy of the treatments used, a chromosome (or a solution) will consist of a string representing a permutation of these treatments. For each pair of consecutive genes we will compute the number of objectives for which one is better than the other. We finally want to obtain a decreasing order of the treatments efficiency. The fitness (quality) of a chromosome will be equal to the number of treatments which are not arranged in a decreasing order (as compared with the successor). The quality zero certifies that the treatments are decreasingly arranged while taking into account the efficiency.

The algorithm works as follows: The initial population is generated. The only genetic operator used is mutation which consists of exchanging values of two genes randomly generated. Each individual is affected by mutation with a given probability. Parent and offspring are compared using the dominance concept presented above. The best between parent and offspring will be kept in the population of the next generation. This process is repeated for a given number of generations.

## 6 Weighted Sum Approach

The weighted-sum method is a traditional, popular method that parametrically changes the weights among objective functions to obtain the Pareto front [6]. Let us consider we have the objective functions  $f_1, f_2, \dots, f_n$ . This method takes each objective function and multiplies it by a fraction of one, the "weighting coefficient" which is represented by  $w_i$ . The modified functions are then added together to obtain a single cost function, which can easily be solved using any method which can be applied for single objective optimization.

Mathematically, the new function is written as:

$$\sum_{i=1}^n w_i f_i,$$

$$\text{where } 0 \leq w_i \leq 1$$

$$\text{and } \sum_{i=1}^n w_i = 1.$$

The initial work using the weighted sum method was done by Zadeh [11]. The method is simple to understand and easy to implement. The weight itself reflects the relative importance (preference) among the objective functions under consideration. However, there are several disadvantages of this technique:

- The user always has to specify the weights values for functions and sometimes this will not have any relationship with the importance of the objectives;

- Non-convex parts of the Pareto set cannot be obtained by minimizing convex combinations of the objectives;
- A single solution is obtained at one time. If we are interested in obtaining a set of feasible solutions, the algorithm has to be run several times. This also, is not a warranty that the solutions obtained in different runs are different.

## 7 Case Study

We made a clinical study of the following treatment methods of essential trigeminal neuralgia: *infiltrations with streptomycin, low level laser therapy, treatment by skin graft, treatment by sciatic nerve graft, treatment by neurectomy*[2]. Among these treatments, neurectomy was considered a mutilating treatment and the other methods were considered conservative. The research was done on a number of 251 patients suffering from essential trigeminal neuralgia and took over 8 years. The data used in experiment represents real data and are adapted from [2]. In order to mark out the effects and results of these treatments seven evaluation criteria were considered: *hospitalization period, remission period, pain relief, decrease in the number of crises, decrease in pain level, decrease of the pain area, decrease in medication*. The application of the treatments was based on regular techniques and personal contributions. The evaluation matrix case is presented in Table 1.

**Table 1.** Data considered

CRITERIA	Criterion type	TREATMENT				
		Infiltrations with streptomycin	Low level laser therapy	Treatment by skin graft	Treatment by sciatic nerve graft	Treatment by neurectomy
		$T_1$	$T_2$	$T_3$	$T_4$	$T_5$
Hospitalization period	$C_1$ <i>min</i>	12.143	13.625	15.093	15.417	16.778
Remission period	$C_2$ <i>min</i>	9.964	10.453	12.07	11.889	12.022
Pain relief	$C_3$ <i>max</i>	21	18.34	25.2	34.742	18.457
Number of crises	$C_4$ <i>max</i>	6.667	6.688	11.209	9.75	7.244
Pain level	$C_5$ <i>max</i>	3.423	3.281	3.558	3.833	3.156
Pain area	$C_6$ <i>max</i>	0.904	0.937	0.937	0.978	0.848
Medication	$C_7$ <i>max</i>	364.286	442.188	655.814	586.111	255.556

### 7.1 Weighted Sum Approach Evaluation

As evident from Table 1, first two objectives have to be minimized and last 5 objectives have to be maximized. Weighted sum approach considers all objectives as having the same optimization criterion (minimization or maximization). For this purpose, we consider  $-f_1$  and  $-f_2$  instead of  $f_1$  and  $f_2$ . This way, all objectives have to be maximized. In order to apply the weighted sum method, we need to specify a weight for each criterion. For the seven studied criteria we established specific values in the interval 0.02 and 0.54. For the hospitalization period and for the remission period the values were more relevant as they decreased. For the other evaluated parameters, higher values expressed a good efficiency of the evaluated treatment method. Table 2 contains the values of the weights for each objective.

**Table 2.** Values of weights

	Criterion						
	$C_1$	$C_1$	$C_1$	$C_1$	$C_1$	$C_1$	$C_1$
<b>Weight</b>	<b>0.08</b>	<b>0.06</b>	<b>0.54</b>	<b>0.08</b>	<b>0.17</b>	<b>0.05</b>	<b>0.02</b>

Results obtained by applying weighted sum approach are presented in Table 3.

**Table 3.** Results obtained by applying weighted sum approach

Treatment	Criterion							Weighted sum
	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$	
$T_1$	12.143	9.964	21.0	6.667	3.423	0.904	364.286	<b>18.21691</b>
$T_2$	13.635	10.453	18.34	6.688	3.281	0.922	442.188	<b>18.16829</b>
$T_3$	15.093	12.07	25.2	11.209	3.558	0.937	655.814	<b>26.34107</b>
$T_4$	15.417	11.889	34.742	9.75	3.833	0.978	586.111	<b>30.01671</b>
$T_5$	16.778	12.022	18.457	7.244	3.156	0.848	255.556	<b>14.17278</b>
<b>Weights</b>	0.08	0.06	0.54	0.08	0.17	0.05	0.02	

As evident from Table 3, the ranking of the above treatments in decreasing order of its efficiency obtained by applying weighted sum approach is:  $T_4$  (Sciatic nerve graft),  $T_3$  (Skin graft),  $T_1$  (Streptomycin),  $T_2$  (Laser) and  $T_5$  (Neurectomy).

### 7.2 Proposed Approach Evaluation

A sample solution obtained by the evolutionary algorithm used for the treatments ranking is:

$$(T_5, T_4, T_2, T_1, T_3).$$

Base on the relationships between solutions as presented in Table 4, the fitness of this sample solution which is given by the number of treatments (genes) which are not in a decreasing order is equal to 3.

**Table 4.** Number of objectives (criteria) in which treatment  $T_i$  dominates treatment  $T_j$  (out of seven)

<b>Treatment</b>	$T_1$	$T_2$	$T_3$	$T_4$	$T_5$
$T_1$	–	4	2	2	6
$T_2$	3	–	2	2	5
$T_3$	5	5	–	3	7
$T_4$	5	5	4	–	7
$T_5$	1	2	0	0	–

The parameters used by the evolutionary algorithm are:

- Population size: 20
- Number of generations: 20
- Mutation probability: 0.5.
- Crossover probability: 0.7.

The hierarchy of treatments efficiency obtained is:  $T_4, T_3, T_1, T_2, T_5$ .

As evident from these experiments, both algorithms obtained same hierarchy for the treatments. The evolutionary algorithm is very simple and works very well for the considered data. Evolutionary algorithms can detect several solutions in a single run and the user could select the desired solution based on the problem constraints, feasibility and other criteria. But weighted sum approach (and other approaches for multiobjective optimization which are not using evolutionary computation) has to be applied several times in order to obtain multiple solutions. The common procedure involves exchanging objective's weights. But this can be problematic for the case studied due to the importance of objectives. For instance, we considered decreasing of pain criteria as being a very important one (weight is 0.54) and medication period (or hospitalization period) as having less importance (weights are 0.08 and 0.02 respectively). If we will exchange pain weight and medication weight we cannot have any warranty of the result quality.

## 8 Discussions and Conclusions

The patients had different responses to the same treatment method during the treatment period. This observation can be also found in other studies such as [8]. The inadequate dosing of medications or treatments can also lead to failure [12]. By applying evolutionary algorithms, the ranking of treatments efficiency obtained is similar to the one obtained by applying a standard mathematical approach for multiobjective optimization, namely the weighted sum approach. But the advantage is that we do not require any additional information about the problem while weighted sum approach involves a weight for each objective. This task can be sometimes difficult to achieve due to the objectives importance.

By combining all objectives in a single objective function (and transforming the multiobjective optimization problem into a single objective one) at most one solution could be obtained by the execution of the algorithm. In order to obtain

multiple solutions, we have to apply the algorithm several times. Even then, we cannot be sure that all solutions are different. Running time required is another disadvantage of the weighted sum approach.

## Acknowledgements

This research was supported by the International Joint Research Grant of the Institute of Information Technology Assessment foreign professor invitation program of the Ministry of Information and Communication, Korea. Authors would also like to thank Radu Campian, Grigore Baciut and Mihaela Baciut of the Department of Maxillofacial Surgery, University of Medicine and Pharmacy, Iuliu Hatieganu Cluj-Napoca, for the initial contributions of this research [2].

## References

1. Apfelbaum, R.I. Trigeminal Neuralgia : Vascular Decompression. Carter and Spetzler - Neurovascular Surgery. Mc Graw Hill. International edition, pp. 1107-18, 1995.
2. Campian, R., Baciut, G., Baciut, M., Tigan, S. Pain evaluation in essential trigeminal neuralgia of essential trigeminal neuralgia treatments, Applied Medical Informatics, 15(3-4) pp. 21-25, 2004.
3. Goldberg, D.E. Genetic algorithms in search, optimization and machine learning. Addison Wesley, Reading, MA, 1989.
4. Holland, J. Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, 1975.
5. Jannetta, P.J. Microvascular Decompression of The Trigeminal Nerve for Tic Douloureux. Youmans - Neurological Surgery. Saunders Company, Fourth edition, 5, pp 3404 - 3415, 1996.
6. Kim, I.Y., de Weck, O.L. Adaptive weighted-sum method for bi-objective optimization: Pareto front generation. Structural and Multidisciplinary Optimization, 29, pp. 149-158, 2005.
7. Kondo, A. Follow-up Results in Microvascular Decompression in Trigeminal Neuralgia and Hemifacial Spasm. Neurosurgery, vol. 40; pp. 46 - 52, 1997.
8. Lee K.H. Facial pain: trigeminal neuralgia, Annals of the Academy of Medicine, Singapore, 22(2): 193-196, 1993.
9. Stadler, W. A Survey of Multicriteria Optimization, or the Vector Maximum Problem. JOTA 29, 1-52, 1979.
10. Steuer, R.E. Multiple Criteria Optimization: Theory, Computation and Application. New York: Wiley, 1986.
11. Zadeh, L. Optimality and Non-Scalar-Valued Performance Criteria. IEEE Transaction on Automation Control 8, 59-60, 1963.
12. Zakrzewska J.M., Patsalos P.N., Drugs used in the management of the trigeminal neuralgia, Oral Surgery, Oral Medicine, Oral Pathology, 74(4): 439-450, 1992.

# Interpretation of Group Measurements of Validation Data Using Fuzzy Techniques in an Object-Oriented Approach

Eduardo Mosqueira-Rey and Vicente Moret-Bonillo

Department of Computer Science, University of A Coruña, A Coruña, 15071, Spain  
{eduardo, civmoret}@udc.es

**Abstract.** Fuzzy logic is particularly indicated for representing the uncertainty associated with the processes that take place in non-probabilistic systems. It is also useful for the linguistic quantification of sets in which the classification of concepts and events is affected by semantic imprecision. We describe a fuzzy method designed to assist with interpretations of group measurements obtained from validation data for intelligent systems operating in complex domains. The method described was implemented applying an object-oriented paradigm. The suitability of using fuzzy methods and an object-oriented approach are both discussed in the article.

## 1 Introduction

Intelligent systems that perform in complex domains like medicine do not usually have a standard reference to serve as a validation criterion. The solution to this problem is to perform a validation against the expert, with the opinions of human experts acting as reference. It is generally preferred to implement a validation against a group of experts, not only because a group ensures that the results are more objective, but also because the consistency between the interpretations of the different experts can be analysed. The problem with this approach is that the amount of data to be analysed is usually excessive, and so statistical measures to facilitate the analysis are required.

The statistical measures used for quantitative validations can be classified in one of three groups: pair measurements, group measurements, and agreement ratios. Agreement ratios are used when a standard reference exists, and so interpretation implies no particular difficulty. Pair and group measurements, on the other hand, are used when the validation criterion is provided by a group of experts. Examples of pair measurements are the agreement index and kappa measurements. Examples of group measurements - which are constructed from pair measurements - include cluster analysis, multidimensional scaling (MDS), and Williams measurements. A detailed description of these measurements can be found in [1].

The aim of this work is to describe a fuzzy method for assisting with interpretations of group measurements obtained from validation data. The method described was implemented applying an object-oriented paradigm.

## 2 Fuzzy Interpretation of Validation Measures

Our proposed interpretation system includes two different modules (Fig. 1): an *algorithmic module*, which starts off from the unprocessed statistical measurement data and produce as output high-level information with a bearing on the working environment; and a *heuristic module* that processes this high-level information so as to obtain the final results of the interpretation. The fact that the algorithmic module filters and processes the basic data in order to convert this into high-level data permits the rules for the heuristic module to be defined with economy of expression.

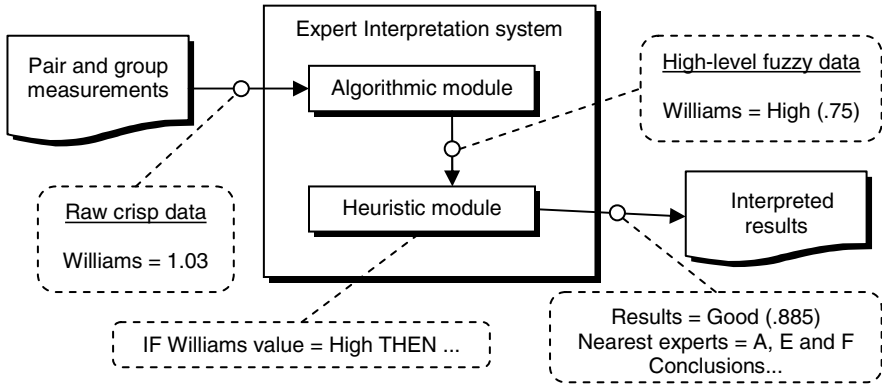


Fig. 1. Expert interpretation system modules, inputs and outputs

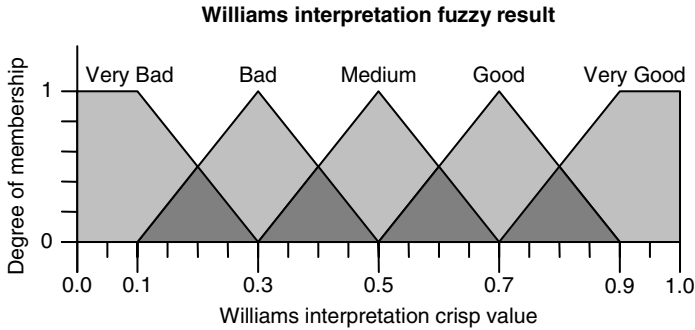
Previous work in this field by the authors [2] demonstrated the suitability of a fuzzy approach to dealing with the vagueness of notions expressed using linguistic labels, such as, for example, “the Williams index value is HIGH”. Below we describe the algorithm and heuristic analysis process for the different group measurements considered.

### 2.1 Williams Measurements

The Williams measurement [3] endeavours to determine whether an isolated expert agrees with the experts in a concrete reference group to the extent that the experts in the said group agree amongst themselves (represented by a value above or equal to 1).

An algorithmic analysis of the Williams measurement results enables three parameters to be identified: (1) the result of the measurement  $I_n$  for the intelligent system, (2) the dispersion of the data when the measurement  $I_n$  is calculated for all the experts, and (3) the relative position of the intelligent system in relation to the results of the experts. The data dispersion and the relative position of the intelligent system will serve to correct any possible distortion in the results of the Williams index if there are experts in disagreement in the reference group.

These three parameters are fuzzified and become the input to a Fuzzy Associative Memory (FAM) [4]. The result obtained is also a fuzzy variable depicted in Fig. 2.



**Fig. 2.** Classification of the results of a group measurement (in this case, Williams)

## 2.2 Multi-dimensional Scaling

Multidimensional scaling or MDS [5] is a data analysis technique that permits the representation of the experts as points in a geometric space, in which the distances that separate the different experts are indirectly proportional to their similarities as expressed by a specific pairs' measurement.

An algorithmic analysis of MDS results enables two parameters to be identified: (1) distance to the origin of coordinates and (2) distance to the nearest experts.

Distance to the origin is important because in MDS, the origin acts as the 'mass centre', which can be interpreted as consensus between the different experts. Distance to the nearest experts is used to check whether or not the intelligent system is contained within a cluster of experts. It is measured as the mean of the distance to the  $n/2$  nearest experts of the  $n$  experts available. These parameters are the input into a two-dimensional FAM cube and the outcome is classified in accordance with the criteria depicted in Fig. 2.

## 2.3 Cluster Analysis

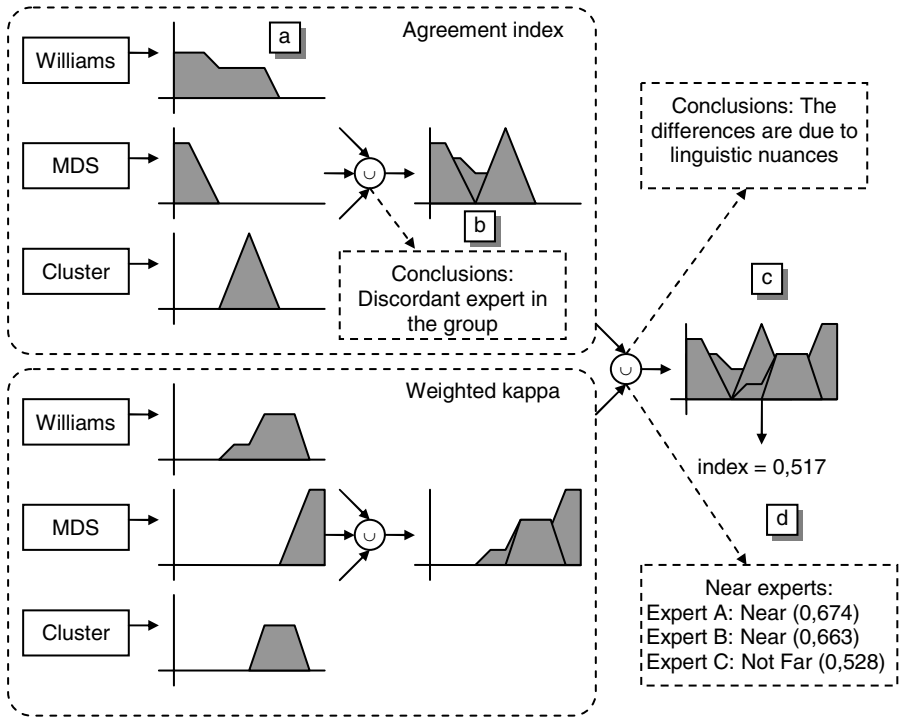
Cluster analysis [6], has as its objective the grouping of experts in a hierarchical tree in accordance with the similarity of their interpretations. The cluster analysis algorithm employed is that known by the acronym SAHN (Sequential, Agglomerative, Hierarchical and Non-overlapping).

An algorithmic analysis of cluster results enables two parameters to be identified: (1) the number of experts in the cluster of the intelligent system in relation to the number of experts in the other clusters, and (2) the level of aggregation of the intelligent system in its cluster in relation to the level of aggregation for the experts in the same cluster.

In this case, we endeavour to check whether the intelligent system is within the main group of experts and at what level it joins the experts in this group. To analyse these issues, it is essential to know the cluster to which the isolated expert belongs, and to do this, we have to determine which partitioning of the dendrogram for the hierarchical cluster analysis to use. A simple but highly effective method is to cut the dendrogram at the point that produces the greatest difference between two consecutive points of union [7].



The overall fuzzy application process is illustrated in Fig. 3.



**Fig. 3.** An example of the fuzzy application process: (a) interpretation of a group measurement, (b) union of interpretations of several group measurements and drawing of conclusions, (c) union of the final results of the group measurements for different pair measurements and drawing of conclusions, and (d) creation of the final index and classification of the experts according to distance

### 3 The Object-Oriented Approach

The validation system was developed applying an object-oriented philosophy, given the advantages implied, namely, extensibility, reusability, easily maintained, scalability, etc. Design was based on the use of patterns, as effective and reusable solutions to the programming problems that tend to occur [8].

Fig. 4 shows a diagram of the classes corresponding to a calculation of the group measurements (pair measurement functioning is analogous).

In this case we have a common interface named GroupMeasure from which the concrete classes that implement the different group measures (Williams, MDS, etc.) are derived. The result of a group measure calculation is a GroupResult derived class. The pattern applied for a pair table to calculate a group measure is called 'strategy', and its process is as follows: PairTable makes a request to GroupMeasure to make a

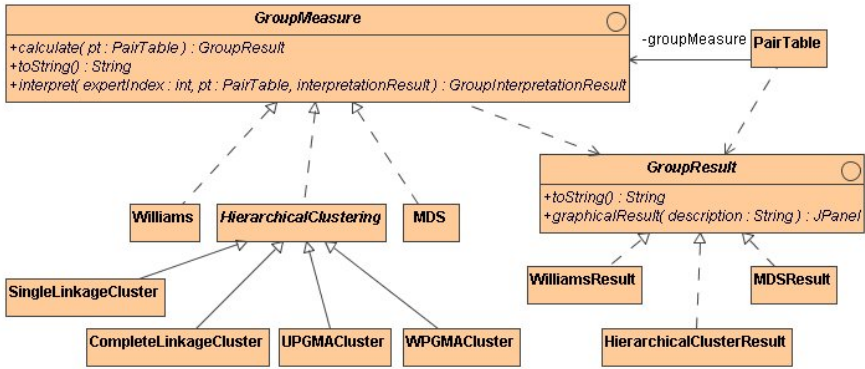


Fig. 4. Diagram of classes for the calculation of the group measurements

calculation on the basis of the data stored in this table. PairTable does not know with which measurement it is communicating but knows that it will implement a calculation method that will produce a result (GroupResult) that can be represented both textually and graphically. In reality, the classes that perform the calculations are sub-classes of GroupMeasure (e.g. Williams, MDS, etc.), but PairTable is unaware of the existence of these classes, and this facilitates the inclusion of new group measurement, by simply adding suitable GroupMeasure and GroupResult sub-classes.

In the interpretation of the group measurements a similar strategy to that used for calculation was pursued, illustrated in Fig. 5. In this case, there is also a common interface named GroupInterpretation from which the concrete classes that implement the interpretation of the different group measures are derived. The process followed is again a ‘strategy’ pattern whose dynamic interaction is depicted in Fig. 6: the interpretation system (InterpretationExpert) has a list of pair measures from which a PairTable can be calculated, as stated before. The interpretationExpert asks to the different GroupMeasure objects available to interpret the data that are stored in the PairTable. The nature of these GroupMeasure objects is unknown, although it is

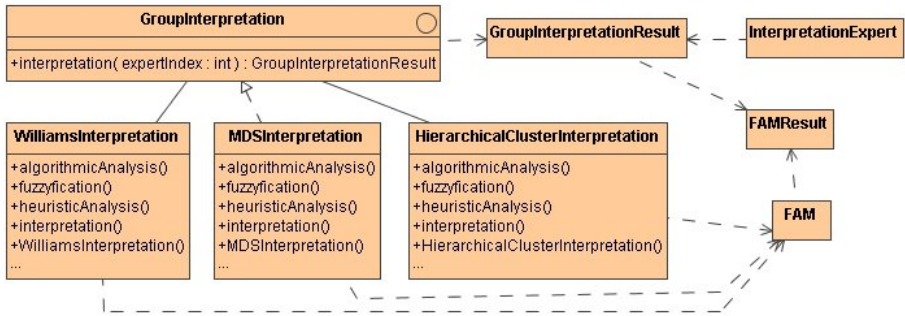
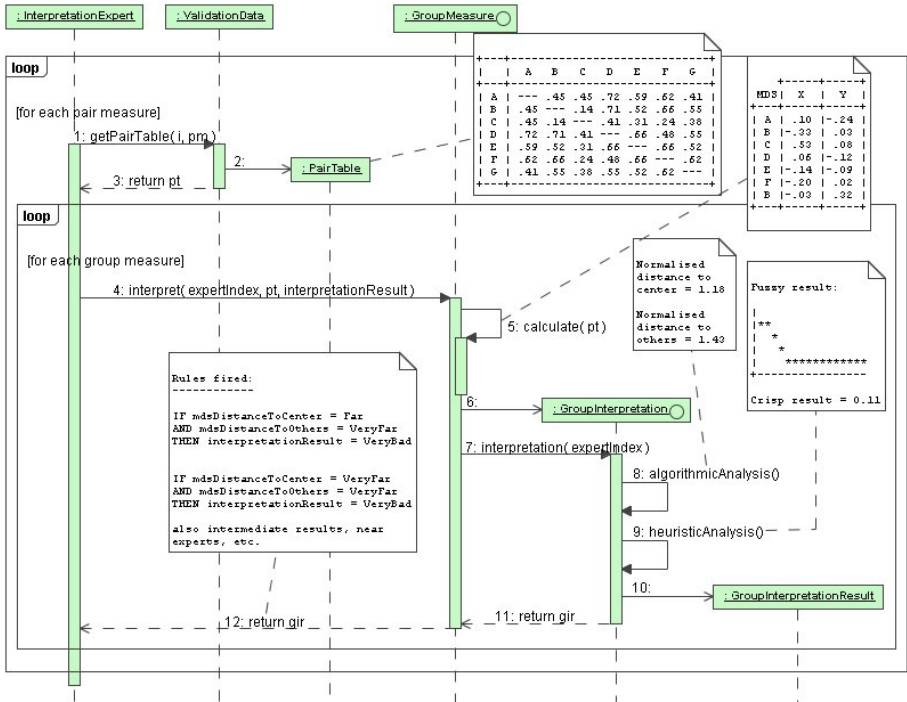


Fig. 5. Diagram of classes for the interpretation of the group measurements



**Fig. 6.** Sequence diagram for the interpretation of the group measurements. The notes show the values obtained if the GroupMeasure were MDS. The process is as follows: Initially for each pair measure a PairTable is constructed (messages 1-3). Using this PairTable an interpretation is required for each GroupMeasure (4). The GroupMeasure calculate itself with the data of the PairTable (5) and asks to its corresponding GroupInterpretation object to perform its interpretation (6-10). The final result is a GroupInterpretationResult object that is returned to the InterpretationExpert (11-12).

known that an interpretation method can be invoked. Each group measurement will delegate its interpretation to an object which will perform the tasks of the algorithmic module and of the heuristic model described above (sub-classes of Group Interpretation as WilliamsInterpretation, MDSInterpretation, etc.). The outcome of the interpretation (GroupInterpretationResult) is returned to the interpretation system that launched the process. As before, the fact that this object is unaware of the calculations being performed ensures that a new group measurement is interpreted simply by adding a suitable GroupInterpretation sub-class.

### 4 Discussion and Conclusions

In this work we have described a fuzzy approach to interpreting group measurements obtained from validation data. The philosophy of the system is based on being able to easily integrate new measurements, in both the calculation and interpretation

processes. This makes for a flexible system in which new measurements can be included, although if its interpretation is not available, this measurement will not be taken into account in interpretation. Moreover, the system takes advantage of the knowledge it holds in regard to some of the measurements, and so it can check a series of data for known measurements in order to extract conclusions (e.g., “If we are *close to the centre* and the Williams value is *high*, then we have a *consensus*”). An approach based on object-orientation and the use of design patterns is not only highly appropriate, it also facilitates the design of more reliable and reusable software. It also facilitates the design of software that is capable of adaptation, which is a crucial issue when working with artificial intelligence in complex domains.

**Acknowledgments.** This project has been partially funded by the Spanish Inter-Ministerial Commission for Science and Technology (CICYT) via research project TIN2005-04653 (co-funded by the European Regional Development Fund, ERDF).

## References

1. Mosqueira-Rey, E., Moret-Bonillo, V.: Validation of Intelligent Systems: A Critical Study and a Tool. *Expert Systems with Applications* **18** (2000) 1–16
2. Mosqueira-Rey, E., Moret-Bonillo, V.: Intelligent interpretation of validation data. *Expert Systems with Applications* **23** (2002) 189–205
3. Williams, G.W.: Comparing the joint agreement of several raters with another rater. *Biometrics* **32** (1976) 619–627
4. Negnevitsky, M.: *Artificial Intelligence: A Guide to Intelligent Systems*. Addison-Wesley, Harlow, England (2002).
5. Borg, I., Groenen, P.: *Modern Multidimensional Scaling*. Springer-Verlag, New York (1997)
6. Dubes, R.C.: Cluster analysis and related issues., in *Handbook of Pattern Recognition and Computer Vision*, C.H. Chen, L.F. Pau and P.S.P. Wang (eds.), World Scientific Publishing Company, River Edge, NJ (1993) 3–32
7. Everitt, B.S.: *Cluster Analysis* (3<sup>rd</sup> ed.). Arnold, London (1993).
8. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, Boston (1995)

# Accuracy of Neural Network Classifiers as a Property of the Size of the Data Set

Patricia S. Crowther and Robert J. Cox

School of Information Sciences and Engineering,  
University of Canberra,  
ACT 2601, Australia  
trishc@webone.com.au, Robert.Cox@canberra.edu.au

**Abstract.** It is well-known that the accuracy of a neural network classifier increases as the number of data points in the training set increases. A previous researcher has proposed a general mathematical model that explains the relationship between training sample size and predictive power. We examine this model using artificially generated data sets containing varying numbers of data points and some real world data sets. We find the model works well when large numbers of data points are available for training, but presents practical difficulties when the amount of available data is small and the data set is difficult to classify.

## 1 Introduction

Artificial neural networks (ANNs) are used to find a generalised solution with a subset of data from a problem domain. It is well-known that the accuracy of a neural network classifier increases as the number of data points in the sample set increases. When training a new neural network from a sample set, the data is commonly subdivided into three sets; train – the data used to train the neural network, test – used to stop the training thus preventing overfitting, and validation, used to measure the accuracy of the resultant neural network. The validation set must be independent of the train and test sets.

Whilst it may be desirable to obtain a very large sample set, this is not always possible; eg the amount of available data for kidney transplants or lung disease is limited. It would be useful to be able to gauge how well a given neural network was able to classify data on the basis of the size of the sample set, particularly where the available data was limited. It is our aim to develop an objective measure of the amount of data required to provide meaningful classification for a given data set.

Previous researchers developed a mathematical model for predicting the relationship between size of training sample and predictive power [1]. In developing their model, they used three real world data sets and then validated the performance of the model on the same data sets, using different sample sizes in the validation phase.

We examine that model using artificially generated data and some real world data, to determine whether the model may be applicable to neural network classification problems in general, or only to the three data sets used by the original researchers.

We briefly describe the model postulated in [1], then describe the methodology used in our own experiments, present our results, then evaluate them against the model and present our conclusions.

## 2 Description of Model

The general model from [1] is given by

$$\frac{dP(p)}{dp} = k(T - P(p)) \tag{1}$$

where  $P(p)$  is the predictive power for a training sample of size  $p$

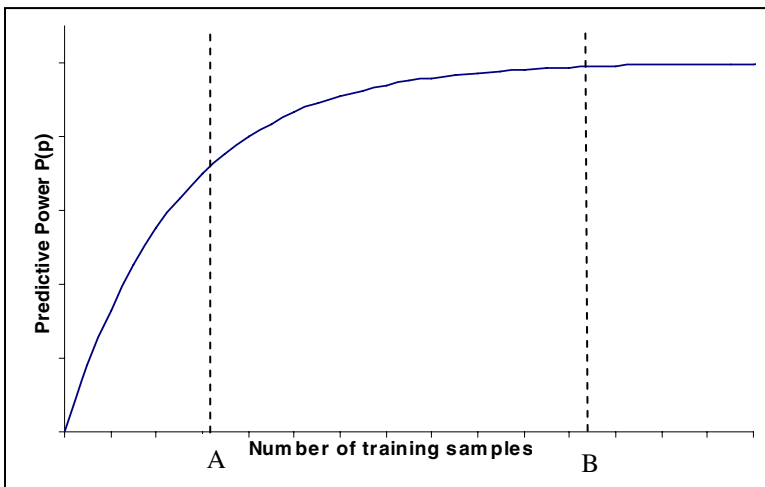
$T$  is the efficiency threshold – the point at which no significant improvement in predictive power can be seen

$k$  is the efficiency rate – the rate of improvement in predictive power per unit increase in efficiency.

The solution of (1)

$$P(p) = T(1 - e^{-kp}) + P(0)e^{-kp} \tag{2}$$

is graphed in figure 1, (for  $T = 1$ ,  $k = 0.004$  and  $P(0) = 0$ , for training samples from 0 to 2000). Predictive power is in the range 0 to 1. The rate of improvement from 0 to A is rapid, then declines from point A to B. Past point B, no significant improvement in predictive power is observed – the efficiency threshold is approached.



**Fig. 1.** General relationship between the size of training sample and the predictive power of a classifier

The value  $P(0)$  is the predictive power when no training data is used and is related to the learnability of a classification technique.

The application of the general relationship given in (1) and (2) has some difficulties, however. Although we have used specific values of  $k$ ,  $T$  and  $P(0)$  to produce the graph, one does not know these values for a given classifier, so available data must be fitted to the curve to evaluate the effectiveness of the model. In order to test the applicability of the model given in (2), the researchers in [1] used three different training set sample sizes (100, 500 and 1000) with a non-linear optimisation routine. The particular routine used was not specified in their paper.

We have assumed, as did the original researchers, that the rate of change of predictive power is proportional to  $(T-P(p))$ , rather than some other measure such as the square of  $(T-P(p))$  or the square root of  $(T-P(p))$ . Although it is possible that the experimental data could be fitted to other mathematical curves, examination of other possible relationships, though of interest, is beyond the scope of this paper.

### 3 Methodology

Initial experiments were conducted on artificially generated data sets, an approach used by other researchers [2]. We use two measures of accuracy, observed accuracy (OAcc) and true accuracy (Tacc). Oacc is the percentage of correctly classified data achieved by a trained neural network measured against the validation set. TAcc is the accuracy as reported by the trained ANN when the results are compared against the function used to generate the data [3].

We randomly generated data using the discriminator function

$$y = \frac{1}{2} \sin \frac{3\pi x}{2} + \frac{1}{2} \quad (3)$$

Data dimension was 2 ( $x$  and  $y$  values), with two classification outputs (above or below the discriminator). The data domain was in the range  $0 \leq x, y \leq 1$ . Data was uniformly distributed across the problem domain. Errors were introduced in a known percentage for each experiment.

As with [3,4,5], a data point was defined to be in error if it was above the discriminator when it should have been below, and vice versa. Errors were uniformly distributed across the data. The number of points in the training set was varied from 50 to 3000 points in steps of 50. The testing set, generated separately from the training set with a different random seed, was the same size as the corresponding training set. A validation set of 1000 data points, also independently randomly generated, was used for each experiment.

Each point on the accompanying graphs or tables represents the mean of 20 runs, each with different initial random weights. We used an automated training algorithm for MLP ANNs [3,4,5] so each run had the potential to have different training parameters (nodes, epochs and training constant). Experiments were conducted with error rates of 0%, 10%, 25% and 35%.

We then fitted the model with three points obtained using training sample sizes of 100, 500 and 1000 (to be consistent with the researchers in [1]), using the non-linear optimization method from the 'Solver' add-in provided in Microsoft Excel 2003, (conjugate gradient method), solving for the minimum of the sum of the squares of the differences between the actual and predicted values of each of the three points.

To evaluate the model on real world data, we used four data sets: Abalone, Cancer, Dermatology and Weed Seed. All input data was normalised into the range 0 to 1.

The Abalone data set [6,7] consists of 4177 samples in which abalone shellfish are classified into one of three age related groups based on eight measurements.

Wisconsin Breast Cancer [6,8] consists of 699 instances, 16 of which have incomplete data and were removed, leaving 683 instances. Ten attributes are listed, but one, an ID was discarded, leaving nine to classify a tumor as malignant or benign.

The Dermatology data set [6, 9] consists of 366 instances, eight of which have unknown values for age and so have been disregarded for this study, leaving 358 instances with 34 attributes. Instances are classified into six cases, psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis and pityriasis rubra pilaris.

The Weed Seed data set [10] consists of 398 instances; it classifies weed seeds into one of ten types, based on seven measurements of dimensions of the seeds.

For the real world data sets, one third of the instances were randomly extracted for use as the validation data set. The train and test sets were extracted randomly from the remaining data. For Abalone, the number of points in each of the training and test sets was varied from 50 to 1350 in steps of 50; for the remaining data sets the number of training points was varied from 10 to one third of the total points in the data set in steps of 10. Each point on the graph (or in the tables) represents the mean of 20 runs, each with a different random selection of training, testing and validation sets.

To fit the model to the data, three points were chosen for each data set. As Abalone had sufficient available points, we used the values obtained with 100, 500 and 1000 training points. For Cancer and Dermatology, data points were chosen at 10, 50 and 100 points in the training set. For Weed Seed we selected three different sets of three points, to show some of the difficulties in using this model on small data sets.

## 4 Results and Model Validation

Results from artificial data sets are shown in Table 1 and Figure 2. Values for  $T$ ,  $k$  and  $P(0)$  are given in table 1, along with predicted values of  $p$  (number of training samples) for the fitted  $T$ , maximum actual accuracy and number of training samples at which it is achieved. The mean absolute error is the mean of the differences between predicted  $P(p)$  and actual  $P(p)$  for each training sample used in an experiment.

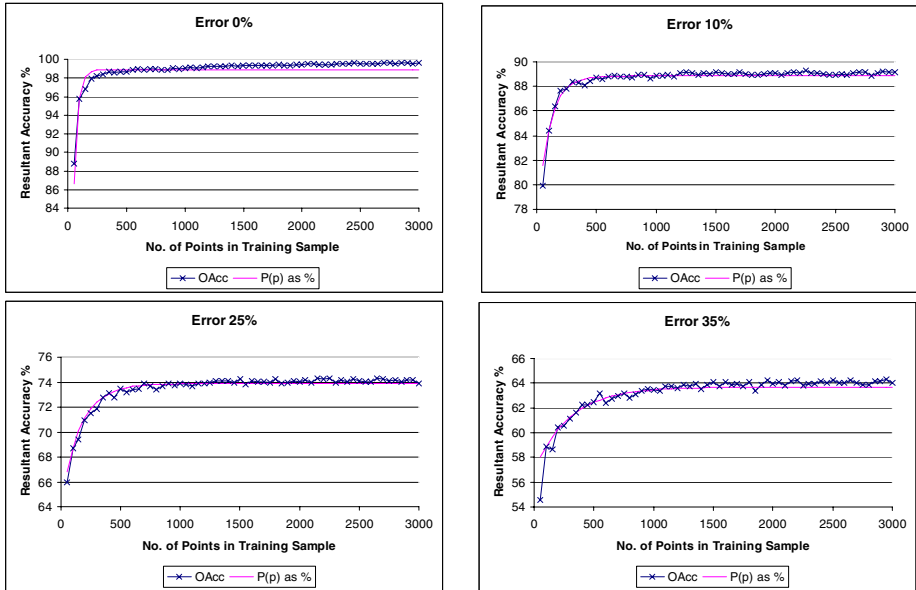
**Table 1.** Results for artificial data

<b>% error</b>	<b>k</b>	<b>P(0) as %</b>	<b>T as %</b>	<b><math>p</math> for given T</b>	<b>Max Actual Acc %</b>	<b># Points at Max %</b>	<b>Mean Abs Error</b>
0%	0.02759	50.00	98.91	516	99.65	3000	0.27
10%	0.01000	76.85	88.83	728	89.26	2250	0.17
25%	0.00624	64.18	73.88	1121	74.31	2650	0.18
35%	0.00337	56.96	63.69	2108	64.32	2950	0.24

In all four cases, the actual accuracy conforms well to the predictive model; though the actual accuracy exceeds the predicted accuracy slightly (by no more than 0.5)



once the efficiency threshold is reached. The low values for mean absolute error lend weight to the hypothesis that predictive power follows the mathematical formula (2) for data sets other than those used in [1].



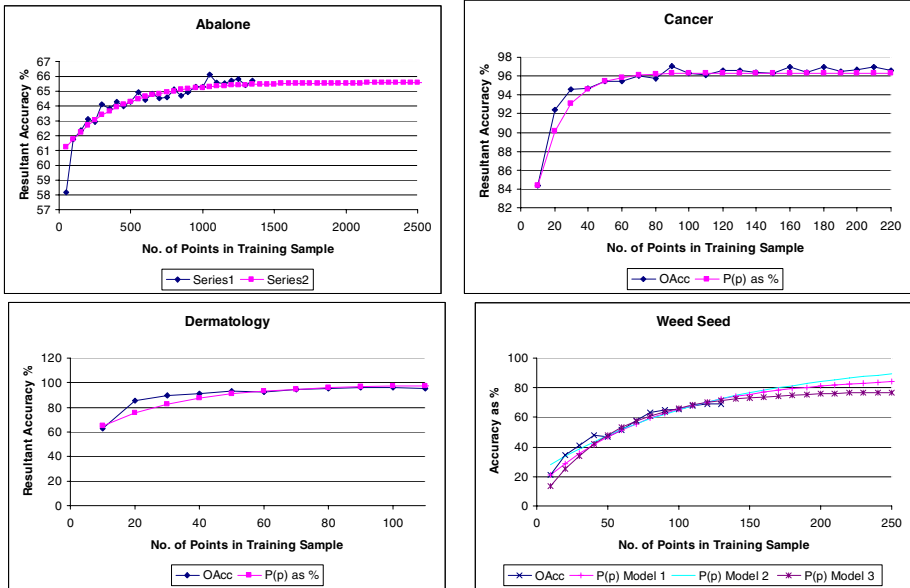
**Fig. 2.** Accuracy of classifier vs number of points in the training set for 0%, 10%, 25% and 35% errors, shown with the predictive model achieved by fitting points at 100, 500 and 1000

The results for real world data, shown in Table 2 and Figure 3, demonstrate some of the difficulties in applying the predictive model to data sets with little data. We have extended the predictive range for Abalone and Weed Seed beyond the actual number of points for training to illustrate these difficulties more readily.

**Table 2.** Results for real world data

Dataset	k	P(0) as %	T as %	p for given T	Max Actual Acc %	p at Max %	Mean Abs error
Abalone	0.00276	60.57	65.57	2323	66.15	1050	0.35
Cancer	0.06517	73.46	96.32	121	97.01	90	0.42
Dermatology	0.03726	50.00	98.42	240	96.0	100	2.77
Weed Seed1	0.01227	12.29	87.65	732	69.02	130	2.37
Weed Seed2	0.00799	22.26	100.0	1209	69.02	130	2.33
Weed Seed3	0.01944	00.00	77.40	471	69.02	130	3.10

A visual inspection of the Abalone graph up to 1350 training points does not yield a clear picture as to whether the efficiency threshold has been reached. Superimposing the predictive model, as in figure 3, shows that this threshold is very likely to have been reached, which is confirmed by the low mean absolute error. The results for Cancer, which also has a low mean absolute error, confirm the model’s applicability.



**Fig. 3.** Accuracy of classifier vs number of points in the training set for real world data sets, shown with the predictive model(s)

Dermatology provided only 110 points for training; Weed Seed 130 points. For Dermatology, this yielded a model with a relatively high mean absolute error, calling into question the model’s applicability. A visual inspection of the graph however, indicates the efficiency threshold is very likely to have been reached. For Weed Seed we examined three possible models. Model 1 used points at 10, 50 and 100 training samples; model 2 used 20, 50 and 100; model 3 used 50, 80 and 120 samples. None of the graphs for Weed Seed appear to have reached the threshold for any of the three models. The amount of data needed to reach the threshold (column 5 in Table 2) could be twice, three or even four times that available. All Weed Seed data appears to fit in the range 0 to A in Figure 1. It is our belief that when the three data points used to fit the model fall within this zone, the model cannot be reliably fitted.

## 5 Conclusion

We have examined the predictive model postulated in [1] to determine if it could be generally applicable. We find that for data sets where large amounts of data are

available it provides a good indication of how much data is needed to perform meaningful classification and provides a measure (the efficiency threshold,  $T$ ) of the predictive power of classification. We find that when the amount of data available is limited, but the problem is readily classifiable (eg Cancer) the method is applicable. We find that attempting to apply the model to a difficult classification problem where the amount of data is very limited has practical difficulties, but that these difficulties clearly point to the need to obtain more data in these cases (eg Weed Seed).

## Acknowledgement

We are grateful for assistance provided by Dr Ian Lisle.

## References

1. Natthaphan Boonyanunta & Panlop Zeepongsekul. Predicting the Relationship Between the Size of Training Sample and the Predictive Power of Classifiers. Knowledge-Based Intelligent Information and Engineering Systems 8<sup>th</sup> International Conference, KES 2004 Wellington, New Zealand, Sept 2004 Proc., Part III p 529-535, Springer-Verlag 2004.
2. Kuncheva L.I., S.T. Hadjitodorov, Using Diversity in Cluster Ensembles, Proc. IEEE International Conference SMC (2) 2004, The Hague, The Netherlands, 2004, 1214-1219
3. Cox, R.J. & Crowther, P.S. An Empirical Investigation into the Error Characteristics of Neural Networks, Proceedings AISAT 2004 The 2<sup>nd</sup> International Conference on Artificial Intelligence in Science and Technology, Hobart, Australia, 21-25 November 2004 p 92-97.
4. Crowther, P., Cox, R. & Sharma, D. A Study of the Radial Basis Function Neural Network Classifiers using Known Data of Varying Accuracy and Complexity. Knowledge-Based Intelligent Information and Engineering Systems 8<sup>th</sup> International Conference, KES 2004 Wellington, New Zealand, Sept 2004 Proc., Part III p 210-216, Springer-Verlag 2004.
5. Crowther, P.S. & Cox, R.J. A Method for Optimal Division of Data Sets for use in Neural Networks. Knowledge-Based Intelligent Information and Engineering Systems 9<sup>th</sup> International Conference, KES 2005 Melbourne, Australia, Sept 2005 Proc., Part IV p 1-7, Springer-Verlag 2005.
6. Machine learning repository: <http://www.ics.uci.edu/~mllearn/MLSummary.html>
7. Nash, Warwick J., Sellers, Tracy L., Talbot, Simon R., Cawthorn, Andrew J. & Ford, Wes B. The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait. Sea Fisheries Division Technical Report #48 (1994).
8. Mangasarian, O.L. & Wolberg, W.H. Cancer Diagnosis via Linear Programming. SIAM News, vol 23 no. 5, September 1990 p1 & p18.
9. Guvenir, H.A., Demiroz, G. & Ilter N. Learning Differential Diagnosis of Erythematous Squamous Diseases using Voting Feature Intervals. Artificial Intelligence in Medicine vol. 13, no. 3 (1998) pp 147-165.
10. Cox, R, Clark, D & Richardson, A. An Investigation into the Effect of Ensemble Size and Voting Threshold on the Accuracy of Neural Network Ensembles. The 12<sup>th</sup> Australian Joint Conference on Artificial Intelligence (AI'99), Sydney, Dec 1999, pp 268-277.

# Investigating Security in Multi-tree Based Technique in RFID Systems

Xu Huang<sup>1</sup> and Dharmendra Sharma<sup>2</sup>

<sup>1</sup> School of Information Sciences and Engineering, University of Canberra, ACT, 2617  
Australia

<sup>1</sup> Xu.Huang@canberra.edu.au, <sup>2</sup> Dharmendra.Sharma@canberra.edu.au

**Abstract.** Radio frequency identification (RFID) systems based on low-cost computing devices are small, inexpensive microchips capable of transmitting unique identifiers wirelessly, which make it dramatically increases the ability of the organization to acquire a vast array of data about the location and properties of any entity that can be physically tagged and wirelessly scanned within certain technical limitations. However, its security has been a major project to be focused on for RFID systems. In this paper the “tree-based” technique will be carefully investigated and a multi-tree group method is first discussed and the outcomes of investigations suggest that the overall probability that the whole attack succeeds is sharply dropped.

## 1 Introduction

Radio frequency identification (RFID) technology is not fundamentally new and concerns a whole range of applications. The first RFID application may have been the Royal British Air Force’s “Identify Friend or Foe” system, which was used during the Second World War to identify friendly aircrafts. RFID can be applied to a variety of tasks, structures, work systems and contexts along the value chain, including business-to-business (B-2-B) logistics, internal operations, business-to-consumer (B-2-C) marketing, and after-sales service applications [1-6]. However, the boom that RFID technology enjoys today is basically due to the standardization [7] and development of low cost devices.

As every wireless device, RFID systems bring with them security and privacy issues to all those people who have been working on this area. Security issues rely on classic attacks, namely denial of service, impersonation of tags or channel eavesdropping. These attacks are rendered more practicable because of the tags’ lack of computational and storage capacity. There are many papers investigated those issues in various ways [7, 8, 10, 11, 12]. Today’s challenge is to find protocols which allow authorized parties to identify the tags without an adversary being able to track them, thus getting to the root of the privacy problem [8, 13]. It is well known that the reason that it is not a correct way to use well known authentication protocols due to the fact that such protocols do not preserve the privacy of the provider. Asymmetric cryptography could easily solve this problem, but it is too heavy to be implemented within a tag.

As [8] discussed that if  $n$  is the number of tags managed by the system,  $O(n)$  cryptographic operations are required in order to identify one tag for a traditional method,

while [9] makes a reduction of the system’s workload from  $O(n)$  to  $O(\log n)$ . Also [12] presented a protocol, relying on hash chains, can make the comparable working load for the system without degrading privacy in comparison with the method provided by [9].

Considering the method presented by [12] needs constructing hash chains from each  $n$  initial values [12], the lifetime of the tag is a priori limited to  $m$  identifications, when a tag is scanned by a reader, its field in the database can be refreshed. Therefore, threshold  $m$  is the number of read operations on a single tag between two updates of the database. So if an attacker can tamper with a tag, she/he is not able to tack its past events. However, the whole cost increasing must be the tradeoff for non-degrading privacy.

In this paper, in order to keep tag system lower cost, we are focusing on the popular “tree based” technique to establish multi-tree technique and the analyses show that under some conditions the whole attack succeeds will be very low (much less than 0.1%) and would be a closer real system.

## 2 Tree Based Technique

In the following analyses, in order to compare the results with [8, 9], we are going to take the same system, namely the system relies on a single computer that takes  $\theta = 2^{-23}$  second to carry out a cryptographic operation, either hashing or encrypting a 128-bit blocks. The tag number is  $2^{20}$ . Identification of several tags is therefore sequential. Current implementations allow a single reader to read several hundreds of tags per second, which means that the system should spend at the most a few milliseconds to identify on tag.

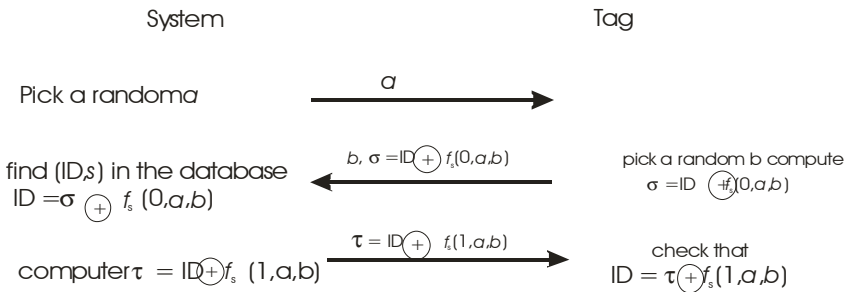


Fig. 1. The protocol of Molnar and Wagner

Let  $\text{ID}$  be the tag’s identifier that is stored in both the database of the system and the tag. They also share a secret key  $s$ . To initiate the authentication, the reader sends a nonce  $a$  to the tag. Next, the tag picks a random  $b$  and answers  $\sigma = \text{ID} \oplus f_s(0, a, b)$ , where  $f_s$  is a pseudo-random function. We may show the protocol in Figure 1.

The technique suggested by [9] relies on a tree structure in order to reduce the identification complexity. As shown in Figure 2 that let  $n$  be the number of tags managed by the system and  $l := \lceil \log_{\delta} n \rceil$  be the depth of the tree with a branching factor  $\delta$ .

Each edge in the tree is valued with a randomly chosen secret  $r_{i,j}$  where  $i$  is the level in the tree and  $j$  is the number of the branch. For example, we have  $T_5$  on Figure 2 is  $[r_{1,1}, r_{2,5}]$ . The average number of cryptographic operations required to identify one tag for traditional  $n$  tags is  $n/2$  and we need  $t_{tr} = (n\theta)/2$ . But in Figure 2, at each level  $i$ , the system have to search in a set of  $d$  secrets for the one matching the tag's secret. Given that  $[s_1, \dots, s_l]$  denotes a secret, the system has thus to compute  $\delta/2$  times  $f_s(0, a, b)$  on average at level  $i$ , meaning that  $\delta/2$  operations are required in order to identify one tag. We have  $t_{MW} = [(\delta\theta)/2]\log_\delta n$ .

When an attacker is able to tamper one tag with some tag, there are three situations: (1) the attacker has one tag  $T_0$  she/he can tamper with and thus obtain its complete secret (e.g. as the owner). For the sake of calculation simplicity, we assume that  $T_0$  is put back does not significantly affect the results (2) the attacker chooses a target tag  $T$  and can query it as much as she/he can but not tamper with it, (3) given two tags  $T_1$  and  $T_2$  and  $T \in \{T_1, T_2\}$ , if the attacker can definitely know which one is  $T$ , he/she is succeeds via querying not tampering them. It is the case that we assume that the attacker cannot carry out an exhaustive search over the secret space. He only guess a block of secret of a given tag is to query it with the blocks of secret he obtained by tampering with some tag. When she/he tampers with only one tag and obtains only one block of secret per level in the tree. If either  $T_1$  or  $T_2$  (but not both) have the same block as  $T_0$ , she/he is able to determine which of them is  $T$ . If neither  $T_1$  nor  $T_2$  has the same block as  $T_0$ , he/she cannot answer, even she/he can move on the next level of the tree because the authentication of the reader succeeds. We denote the secrets of  $T, T_0, T_1$  and  $T_2$  by  $[s_1, \dots, s_l], [s_1^0, \dots, s_l^0], [s_1^1, \dots, s_l^1],$  and  $[s_1^2, \dots, s_l^2]$  respectively. We consider a given level  $i$  where  $s_i^1$  and  $s_i^2$  are in the same sub-tree. We have four issues:

$C_i^1 = ((s_i^0 = s_i^1) \wedge (s_i^0 \neq s_i^2))$  then the attack succeeds,

$C_i^2 = ((s_i^0 \neq s_i^1) \wedge (s_i^0 = s_i^2))$  then the attack succeeds,

$C_i^3 = ((s_i^0 \neq s_i^1) \wedge (s_i^0 \neq s_i^2))$  then the attack definitivdy fails,

$C_i^4 = (s_i^0 = s_i^1 = s_i^2)$  then the attack fails at the level  $i$  but can move onto level  $i + 1$ ,

When the number of tags in the system is large, we assume that

$$P(C_i^1) = P(s_i^0 = s_i^1) \times P(s_i^0 \neq s_i^2)$$

The same assumption also applies to  $C_i^2, C_i^3,$  and  $C_i^4$ , thus we have:

$$P(C_i^1 \vee C_i^2) = \frac{2(\delta - 1)}{\delta^2} \quad (1 \leq i \leq l) \quad \text{and} \quad P(C_i^4) = \frac{1}{\delta^2}$$

The overall probability  $P$  that the whole attack succeeds is as follows

$$P(C_1^1 \vee C_1^2) + \sum_{i=2}^l [(P(C_i^1 \vee C_i^2) \times \prod_{j=1}^{i-1} P(C_j^4))] = 2(\delta - 1) \frac{1 - (\frac{1}{\delta^2})^l}{1 - \frac{1}{\delta^2}} \frac{1}{\delta^2}$$

Because of  $\delta^l = n$  hence, we have: 
$$P = \frac{2}{\delta + 1} \left(1 - \frac{1}{n^2}\right) \tag{1}$$

When an attacker is able to tamper more than one tag with some tag, we have five cases as below:

- $C_i^1 = ((s_i^1 \in K_i) \wedge (s_i^2 \in \Lambda_i))$  then the attack succeeds
- $C_i^2 = ((s_i^1 \in \Lambda_i) \wedge (s_i^2 \in K_i))$  then the attack succeeds
- $C_i^3 = ((s_i^1 \in K_i) \wedge (s_i^2 \in K_i) \wedge (s_i^1 \neq s_i^2))$  then the attack succeeds
- $C_i^4 = ((s_i^1 \in \Lambda_i) \wedge (s_i^2 \in \Lambda_i))$  then the attack definitively fails
- $C_i^5 = ((s_i^1 \in K_i) \wedge (s_i^2 \in K_i) \wedge (s_i^1 = s_i^2))$  then the attack at level  $i$  fails but can go level  $i + 1$

where  $K_i$  denotes the set of blocks of this (one-level) subtree that are known by the attacker and  $\Lambda_i$  denotes the set of those which are unknown by the attacker.  $k_i$  denotes the number of blocks in  $K_i$ . Thus, we have all  $i$  such that  $1 \leq i \leq l$ :

$$P(C_i^1 \vee C_i^2 \vee C_i^3) = \frac{2k_i}{\delta} \left(1 - \frac{k_i}{\delta}\right) + \left(\frac{k_i}{\delta}\right)^2 \left(1 - \frac{1}{k_i}\right) \text{ and } P(C_i^5) = \frac{k_i}{\delta^2}$$

So the overall probability  $P$  that the attack succeeds is obtained as below:

$$\begin{aligned} P &= P(C_1^1 \vee C_1^2 \vee C_1^3) + \sum_{i=2}^l [(P(C_i^1 \vee C_i^2 \vee C_i^3) \times \prod_{j=1}^{i-1} P(C_j^5)] \\ &= \frac{k_1}{\delta^2} (2\delta - k_1 - 1) + \sum_{i=2}^l \left[ \frac{k_i}{\delta^2} (2\delta - k_i - 1) \prod_{j=1}^{i-1} \frac{k_j}{\delta^2} \right] \end{aligned} \tag{2}$$

We have relation between  $k_1$  and  $k_0$ , the number of tags tampered with by the attacker at the level 0 as follows:

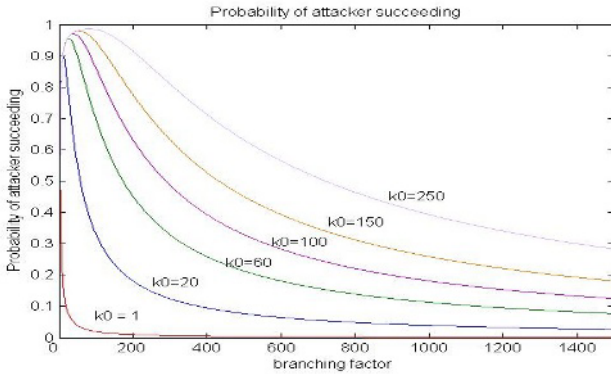
$$k_1 = \delta \left[ 1 - \left(1 - \frac{1}{\delta}\right)^{k_0} \right] \text{ and } k_i = \delta \left[ 1 - \left(1 - \frac{1}{\delta}\right)^{g(k_i)} \right] \quad (2 \leq i \leq l) \text{ with } g(k_i) = k_0 \prod_{j=1}^{i-1} \frac{1}{k_j} \tag{3}$$

### 3 Multi-tree Technique

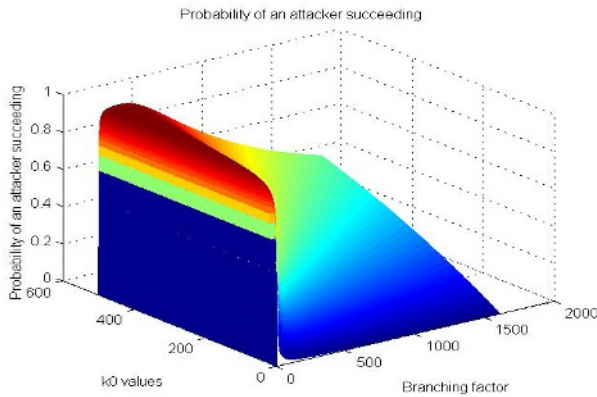
Before we discuss the multi-tree technique, we need to check equations (1) and (2). Figure 1 shows the probability of an attacker succeeds at different  $k_0$ s ( $k_0 = 1, 20, 40, 60, 100, 150$  and  $250$ ) with the branching factors change up to 1600. It is clear that when the branching factor is small, as shown in Figure 1 that it would extremely insecurity (almost regardless the “branching factors”). Therefore, we should always think about to increase the branch numbers to make sure the system at a reasonable security level. It is also obviously that if the attacker has many tags to be tampered, he/she will obtain enough information to get what he/she wanted, which is shown in the different  $k_0$  curves in Figure 1.

Figure 2 shows how the parameters “branching factor” and “ $k_0$ ” control the probability of an attacker succeeding. It is clear to show that if  $k_0$  is about 250, the system

will be in a vulnerability state (the probability of an attacker of succeeding is about > 60%) regardless branching factor, which is due to attacker has enough information to break the system. Figure 3 shows the probability of an attacker of succeeding controlled by  $k_0$  and branching factors. Also we can see if branching number is very small, the system would be extremely vulnerable (>70%) no matter  $k_0$  is, which is the attacker will use even very small sample to obtain the information to break the system.



**Fig. 2.** The probability of an attacker succeeds for the different branching factor under different  $k_0$  ( $k_0=1, 20, 60, 100, 150$  and  $250$ )



**Fig. 3.** The probability of an attacker of succeeding controlled by  $k_0$  and branching factors

Therefore we have an idea: to build multi-tree resulting in  $k_0$  as small as possible and keep the branching factor as big as we can. Let us to establish a multi-tree frame as a matrix, say size is  $w \times q$ , each element will have a sub-tree and total tree will be  $n = w \times q$ , each sub-tree has  $\delta$  branches with  $l$  levels, as shown in Figure 4.



$$\text{multi - tree} = \begin{bmatrix} T_{w_1,1} & T_{w_1,2} & \dots & T_{w_1,q_1} \\ T_{w_2,1} & T_{w_2,2} & \dots & T_{w_2,q_2} \\ \dots & \dots & \dots & \dots \\ T_{w_i,1} & \dots & \dots & T_{w_i,q_i} \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

Fig. 4. sub-tree frame

where each sub-tree, say  $T_{ij}$ , will be constructed shown in Figure 5.

Therefore we may have  $n = \delta_1^{11} . \delta_2^{12} \dots$ , which makes the  $k_0$  that were talking about above divide by factors, if we take each sub-tree (Figure 5) is the same for our simplest case and total  $y$  groups, the final sub-tree  $k_0$  would be  $k'_0 = k_0 / y$ . In this case we have very small probability of an attacker succeeding. For example if we take the results form [8],  $k_0 = 1$  with 1000 tags the probability of an attacker succeeding is 0.1%. Now if we divided 500 sub-tree, under the above condition, the final probability of an attacker succeeding is 0.05%, which is a reasonable case.

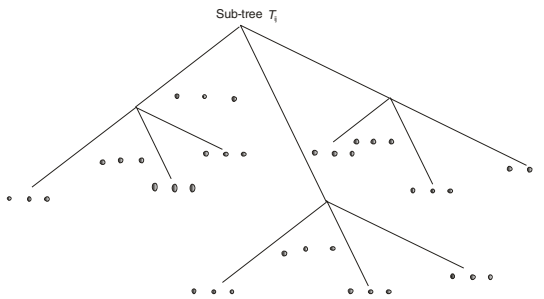


Fig. 5. Sub-tree construction

## 4 Conclusion

We have focus on the very popular tree based technique to keep the lower cost RFID systems staying at the lower price and first presented “sub-tree’ construction frame. The carefully analysis show that our RFID systems can be acceptable (the probability of an attacker succeeding is 0.1%. less than 0.02%) for a real system.

## References

1. G. P. Huber, “Organizational Learning: Contributing Processes and the Literatures,” Org. Sci., vol. 2, no. 1 pp88-115, 1991.
2. V. Standford, “Pervasive Computing Goes the Last Hundred Feet with RFID Systems,” IEEE Perv. Comp. April-May 2003.

3. R. Angeles, "RFID Technologies: Supply-Chain Applications and Implementation Issues," *Info. Sys. Mgmt.*, Winter pp51-65, 2005.
4. C. A. Thompson, "Radio Frequency Tags for Identifying Legitimate Drug Products Discussed by Tech Industry," *Amer. J. Health-Sys. Pharm.*, 61, 14. July 15, pp1430-1431, 2004.
5. G. Yang and S. I. Jarvenpaa, "Trust and Radio Frequency Identification (RFID) Adoption within an Alliance," In R. Sprague (Ed.), *Proc 38<sup>th</sup> Hawaii Intl. Conf. Sys. Sci.*, Big Island, HI, January, pp855-864, IEEE Comp. Soc. Press. Los Alamitos, CA, USA. 2005.
6. Steve Bono, Matthew Green, Adam Stubblefield, Ari Juels, Avi Rubin, and Michael Szydlo, "Security analysis of a cryptographically-enable RFID device," in 14<sup>th</sup> USENIX Security Symposium, pp.1-16, Baltimore, Maryland, USA, July-August 2005. USENIX.
7. Electronic Product Code Global Inc. <http://www.epcglobalinc.org>.
8. Gildas Avoine, "Security and privacy in RFID systems," Online bibliography available at <http://lasecwww.epfl.ch/~gavoine/rfid/>.
9. David Molnar and David Wagner, "Privacy and security in library RFID: Issues, practices, and architectures," In Birgit Pfizmann and Peng Liu, editors, *Conference on Computer and Communications Security-CCS'04*, pp210-219, Washington, DC, USA, October 2004. ACM Press.
10. Philippe Golle, Markus Jakobsson, Ari Juels and Paul Syverson, "Universal reencryption for mixnets," In Tatsuki Okamoto, editor, *The Cryptographers*, Track at the RSA Conference, CT-RSA, vol. 2964 of *Lecture Notes in Computer Science*, pp.163-178, San Francisco, California, USA, Feb. 2004. Springer-Verlag.
11. Dirk Henrico and Paul Muller, "Tacking security and privacy issues in radio frequency identification devices," In Alois Ferscha and Friedemann Mattern, editors, *Pervasive Computing*, Vol 3001 of *Lecture Notes in Computer Science*, pp219-224, Vienna Austria, April 2004. Springer-Verlag.
12. Miyako Ohkubo, Koutarou Suzuki, and Shingo Kinoshita, "Cryptographic approach to "privacy-friendly" tags," In *RFID Privacy Workshop*, MIT, Massachusetts, USA, Nov. 2003.
13. Gildas Avoine and Philippe Oechslin, "A scalable and provably secure hash based RFID protocol," In *International Workshop on Pervasive Computing and Communication Security*, PerSec 2005, pp110-114, Kauai Island, Hawaii, USA March 2005. IEEE Computer Society Press.

# An Improved ALOHA Algorithm for RFID Tag Identification

Xu Huang

School of Information Sciences and Engineering, University of Canberra, ACT, 2601  
Australia  
Xu.Huang@canberra.edu.au

**Abstract.** One of problems that we faced for Radio frequency identification (RFID) systems is that the collision between tags which lowers the efficiency of the RFID systems. One of the popular anti-collision algorithms is ALOHA-type algorithms, which are simple and have good performance when the number of tags to be read is reasonable small. In this paper, an enhanced dynamic framed slotted ALOHA algorithm for RFID is extended to a more general close form, hence reference [9] becomes only one special case described in our paper. The improved ALOHA algorithm for RFID tag identifications has much larger capacity to handle the cases when the tag numbers increasing largely while every frame keeping the optimum system efficiency 35.5%. Simulation results show that the proposed algorithm improves the efficiency nicely in comparison with the conventional algorithms keeping >80% as the previous paper shown.

## 1 Introduction

Organizations utilize modern information systems (IS) to acquire, interpret, retain, and distribute information [1]. Technological innovations in information technology (IT) continue to improve the cost-performance capabilities of organizations to perform these four basic IS tasks. Intelligent agents and knowledge management systems allow managers to interpret data and information to create useful managerial knowledge [2-3]. Technical improvements in storage media allow firms to amass vast data warehouses, while ever increasing processing power allows managers to mine their data for useful information about their operations, existing customers, and potential markets. Further, advances in technology-based real-time information gathering and decision support systems promote real-time decision making that allow organizations to refine their operational performance.

A new technology emerges that provides a major shift in the cost-performance capabilities of one of these four basic IS tasks. Radio frequency identification (RFID) is one such technology that dramatically changes the capabilities of the organization to acquire a vast array of data about the location and properties of any entity that can be physically tagged and wirelessly scanned within certain technical limitations. RFID technology has been noted as an example of the emergence of inexpensive and highly effective pervasive computers that will have dramatic impacts on individuals, organizations, and society [2]. RFID allows the tagged entity to become a mobile, intelligent, communicating component of the organization's overall information infrastructure. In

addition, the combination of the tagged mobile entity, the reader, the hardware infrastructure, and the software that processes the data makes RFID systems a new type of interorganizational system (IOS) that crosses company boundaries, resulting in new opportunities to transform the supply chain for real-time optimization.

One of reasons allowing RFID to become such important role in pervasive computers is because RFID system that has advantages of contact-less type and can hold more data than the others, such as bar code system. Nevertheless, RFID has its limitations such as the problem of identified data clearness, the slow progress of RFID standardization, etc. One of the largest limitations in RFID system is its low tag identification efficiency by tag collision [4-7].

Tag collision is the event that the RFID reader cannot identify the data of tag when more than one tag occupies the same RF communication channel simultaneously. Many researchers have been discussing this problem in various ways, some existing methods seem to have to increase data transmission speed by extending frequency bandwidth to increase tag identification efficiency via minimizing tag collision. Unfortunately it is not a good way to get this problem done due to frequency band being always limited. The most widely used ways are framed slotted ALOHA algorithm and binary search algorithm. Since it is simple implementation, framed slotted ALOHA algorithm is used frequently [7, 9]. For example, Type A of ISO/IEC 18000-6 and 13.56 MHz ISM band EPC Class 1 use Framed Slotted ALOHA algorithm and Type B of ISO/IEC 18000-6 and 900 MHz EPC Class 0 use binary search algorithm.

As most RFID systems use passive tags, frame sizes are limited in the framed slotted ALOHA algorithm [13]. In this algorithm, a tag randomly selects a slot number in the frame and responds to the reader using the slot number it selected. When the number of tags is small, in this method, the probability of tag collision is low, so the time used to identify the all tags is relatively short. But as the number of tags increases, the probability of tag collision becomes higher and the time used to identify the tags increases rapidly. Su-Ryun et al. [9] raised an enhanced dynamic framed slotted ALOHA algorithm for RFID tag identification to obtain slot efficiency by more than 85% when tags is about 1000 with the frame size up to 256 slots. In this paper a close mathematic form was obtained, hence the method presented in [9] becomes one of special cases of our description and the capability of our method to handle vary large tags is largely increased.

## 2 Basic Framed Slotted ALOHA (BFSA) Algorithms

Let us first to define some related terms for our following discussions. Slotted ALOHA algorithm is the tag identification method that each tag transmits its serial number to the reader in the slot of a frame and the reader identifies the tag when it receives the serial number of the tag without collision. A time slot is a time interval that tags transmit their serial number. The reader identifies a tag when a time slot is occupied by only one tag. The current RFID system uses a kind of slotted ALOHA known by framed slotted ALOHA algorithm. A frame is a time interval between requests of a reader and consists of a number of slots. A reader cycle is tag identifying process that consists of a frame. A read cycle is tag identifying process that consists of a frame.

Let us check the basic framed slotted ALOHA (BFSA) algorithm. In BFSA, the reader offers information to the tags about the frame size and the random number that is used to select a slot in the frame. Each tag selects a slot number for access using the random number and responds to the slot number in the frame [9]. As shown in Figure 1 for an example that the frame size is set to three slots and Tag 1 and Tag 3 simultaneously transmit their serial number in Slot 1. As tag collision, Tag 1, 2, 3 and 5 must respond next request of the reader. The reader can identify Tag 4 in the first reader cycle because there is only one tag response in the time Slot 3.

**Table 1.** The process of BFSA algorithm

Downlink	Request	①	②	③	Request	①	②	③
Uplink		Collision	Collision	11110101		Collision	10110010	10110011
Tag 1	→	10110010			→	→	10110010	
Tag 2	→		10100011		→	10100011		
Tag 3	→	10110011			→		→	10110011
Tag 4	→			11110101				
Tag 5	→		10111010		→	10111010		

Since the frame size of BFSA algorithm is fixed, its implementation is simple but less efficiency of tag identification. For example, there is no tag may be identified through the reading cycle due to many tags filling in all the slots with collision. It may also waste the time if a large size frame is used for very small number of tags.

### 3 Enhanced Dynamic Framed Slotted ALOHA (EDFSA) Algorithms

It is obviously to have a chance to improve BFSA by changing the frame size from fixed size to variable size depending on the information such as the number of slots and tags. DFSA algorithm has several versions as mentioned in [7, 9, 12].

There are some dynamic framed slotted ALOHA algorithms, such as to establish upper and lower thresholds for a frame [1], when the number of slots with collision is over the upper threshold, the reader increases the frame size, if collision probability is smaller than the lower threshold, the reader decreases the frame size. But the optimum upper and lower thresholds always are difficulty for this method. In [9], so-called system efficiency is defined and estimating of unread tags is made then a decision of the frame size is to be made after that. If the tag number is very large the restrict condition is made according to system efficiency to keep the optimal number of tags responds to the given frame size. However, it was only presenting a approximated result based on the definition of a system efficiency as “the number of slots filled with one tag divided by current frame size”. An example with tag number 354

and slots number of 256 per frame was presented even up to 1000 tags was mentioned in a curve. What the exactly relation between the estimated tag number and the slot number to keep the system efficiency being in optimum condition how to build a different optimum system efficiency depending on unread tags and slots will be discussed in the following section.

#### 4 Improved Enhanced Dynamic Framed Slotted ALOHA (IEDFSA) Algorithms

Generally in framed slotted ALOHA anti-collision method, the system efficiency being to decrease as the number of responding tags becomes larger. Assume that the reader uses a frame size of  $N$  of slots and the number of responding tags, denoted by  $n$ . The probability that  $k$  tags exist in one given slot is a binomial distribution as below:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{1}$$

where  $X$  is random variable,  $p$  is a success probability. Since we are looking at one slot of the frame for the responding tag and every slot in a frame has equal probability to get the responding tag, which means that we have  $p = 1/N$ . Thus we, we highlight the binomial with “ $B$ ” rather than “ $P$ ” (also making it easier to compare this with [9]) have

$$B_{n, \frac{1}{N}} = \binom{n}{k} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{n-k} \tag{2}$$

Therefore, the expected number of read tags during one read cycle is given by the follows with the fact that we are focusing on one tag is responded:

$$a_1^{N,n} = N \cdot B_{n, \frac{1}{N}}(1) = N \cdot n \left(\frac{1}{N}\right) \left(1 - \frac{1}{N}\right)^{n-1} \tag{3}$$

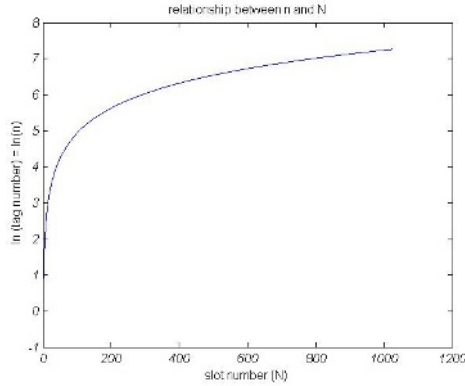
where,  $a_k^{N,n}$  denotes the number of slots with  $k$  tags with the frame size of  $N$  and  $n$  unread tags. Therefore, following the definition of the system efficiency given by [9], we have

$$\text{system efficiency} = \frac{\text{the number of slots filled with one tag}}{\text{current frame size}} = \frac{a_1^{N,n}}{N} \tag{4}$$

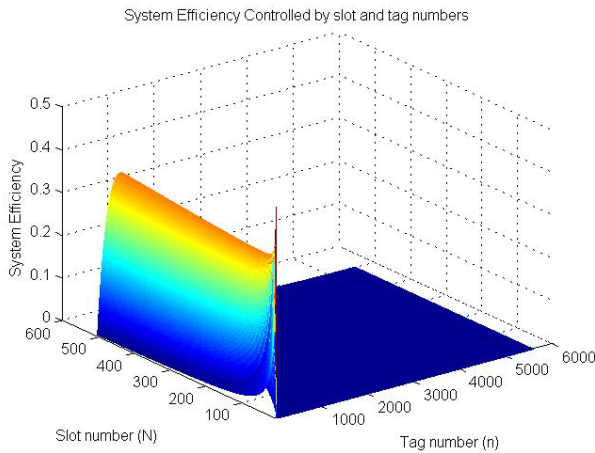
To find the optimum condition for the unread tag number,  $n$ , in a system, we may let  $d(a_k^{N,n})/(dn) = 0$ , and from equations (3) and (4) we obtained:

$$n = -\frac{2 \ln 2}{\ln W}, \text{ with } W = 1 - \frac{1}{N} \tag{5}$$

Figure 1 shows the relation between unread tag number  $n$  and slot number  $N$  in a frame. If we submit the  $N = 256$ , as the number presented in [9], we have  $n = 354$ ,



**Fig. 1.** The relationship between the unread tag number and slot number in a frame



**Fig. 2.** How the system efficiency controlled by unread tag number and slot number

which comes the equation (12) in [9]. Figure 2 shows how the system efficiency controlled by those two parameters,  $N$  and  $n$ .

From the Figure 1, it is clearly to see that the size 256 slots per frame may be suggested ([9] did not mention why 256 was taken) if the tag numbers are very large and the optimum system efficiency is about 35.5% for all the  $N$  and  $n$  from Figure 2 ([9] only showed 4 cases), therefore we may keep this for all the frames if the tag number is large enough to be divided into multi-frames. Now we have an improved enhanced dynamic framed slotted ALOHA (IEDFSA), namely we can use the previous information to estimate the unread tag number to make a decision in the beginning or we can use the optimum condition as the first reading and establishing read information for the next step. Then we keep the optimum condition as threshold to build a frame if there are very large tags (grouping the tags), if the tag number is not such large we

can back to the method described in [9], as [9] showed that if the system efficiency keep about high efficiency, the performance will be > 80% in comparison with other methods.

## 5 Conclusion

Following the definition of system efficiency presented in previous paper [9], a close form relation was obtained in this paper and the relations between unread tag number and frame size, in terms of slot number, were fully shown. In particularly how the system parameters, namely unread tag number and the slot number of a frame, control the defended system efficiency was clearly presented, and hence, the whole system can be always run in optimum condition.

## References

1. G. P. Huber, "Organizational Learning: Contributing Processes and the Literatures," *Org. Sci.*, vol. 2, no. 1 pp88-115, 1991.
2. V. Standford, "Pervasive Computing Goes the Last Hundred Feet with RFID Systems," *IEEE Perv. Comp.* April-May 2003.
3. R. Angeles, "RFID Technologies: Supply-Chain Applications and Implementation Issues," *Info. Sys. Mgmt.*, Winter pp51-65, 2005.
4. C. A. Thompson, "Radio Frequency Tags for Identifying Legitimate Drug Products Discussed by Tech Industry," *Amer. J. Health-Sys. Pharm.*, 61, 14. July 15, pp1430-1431, 2004.
5. G. Yang and S. I. Jarvenpaa, "Trust and Radio Frequency Identification (RFID) Adoption within an Alliance," In R. Sprague (Ed.), *Proc 38<sup>th</sup> Hawaii Intl. Conf. Sys. Sci.*, Big Island, HI, January, pp855-864, IEEE Comp. Soc. Press. Los. Alamitos, CA, USA. 2005.
6. P. Singer, "A new Approach to Low Cost RFID Tags," *Semiconductor Intl.*, Feb. 2005. Available at [www.reed-electronics.com/semiconductor/article/CA499653?pubdate=2%2F1%2F05](http://www.reed-electronics.com/semiconductor/article/CA499653?pubdate=2%2F1%2F05)
7. E. K. Clemons and Y. Wang, "Special Issue: Tchnology Strategy for Electronic Marketplaces," *J. Mgmt. Info. Sys.* vol 17, no.2, pp.5, 2000.
8. K. Finkenzeller. *RFID handbook*. 2<sup>nd</sup> edition. John Wiley & Sons, 2003.
9. Su-Ryun Lee, Sung-Don Joo, and Chae-Woo Lee, "An Enhanced Dynamic Framed Slotted ALOHA Algorithm for RFID Tag Identification," *The Proceeding of the 2<sup>nd</sup> Annual International Conference on Mobile and Ubiquitous Systems*, 2005.
10. R. Rom and M. Sidi, "Multiple Access Protocols/Performance and Analysis." Springer-Verlag, pp47-77, 1990.
11. H. Vogt. Efficient, "Object Identification with Passive RFID Tags," *Proc. Pervasive* pp.98-113, 2002..
12. J. E. Wieselthier, A. Ephermides, and L. A. Michels, "An exact Analysis and Performance Evaluation of Framed ALOHA with Capture," *IEEE Tans. on Communications.* 37(2), pp125-137 Feb. 1989.
13. H. Vogt., "Multiple Object Identification with Passive RFID Tags," *2002 IEEE International Conference on Systems, Man and Cybernetics*. 2002.



# A Dynamic Threshold Technique for XML Data Transmission on Networks

Xu Huang, Alexander Ridgewell, and Dharmendra Sharma

School of Information Sciences and Engineering, University of Canberra, ACT, 2601  
Australia

{Xu.Huang, Alexander.Ridgewell,  
Dharmendra.Sharma}@canberra.edu.au

**Abstract.** XML is increasingly being used to transmit data on networks but it is a verbose format. One may employ a middleware to enhance performance by minimizing the impact of transmission time [1, 2]. Normally, to reduce the amount of data sent the XML documents are converted to a binary format using a compression routine such as XMill [3]. However while this would reduce the amount of data, it results in an increase in the CPU time. We present a technique to decide if it would be transmitting the XML document as a compressed document or not depending on a threshold that we first establish. Experimental results show our method is superior to the NAM method [1]. The simulation results shows that for an example of a 4.5 MB XML file in our method will make the CPU time decreasing 22.69% and total transition time will save 4.61% in comparison with the method described in [1].

## 1 Introduction

XML has become an increasingly important data standard for use in organizations as a way to transmit data [4, 5, 6, 7]. Additionally it is being used to enable web services and similar, often custom, RPC functionality to allow greater access to data across multiple systems within an organization and allowing the possibility of future systems to be created from collections of such RPC functionality.

XML is a verbose, text based format with strict requirements on structure and is often criticized for its large space requirements. This large size can be particularly problematic for use in transmission across a network, where network bandwidth restrictions can cause significant delays in receiving the transmission.

One solution to this problem is to look at reducing the size of these transmissions by rendering them in a binary format, such as by using XMill to compress an XML document. However such methods can take longer as compressing and decompressing may take more time than what is saved transmitting the smaller XML document.

One solution to this problem may be the Network Adaptable Middleware (NAM) raised by Ghandeharizadeh et al [8], even though there are some ways to directly compress, such as column-wise compression and row-wise compression for large message sizes [9]. This solution estimates the time it will take to compress, transmit in binary format and decompress a document compared to an estimate of how long it would take to transmit the document as uncompressed text. The estimates are based on a persistent collection of information on how the system has performed in the past

and provides an accurate estimate on whether it would be faster to compress the document before transmission or not.

We have looked at another way of determining when to compress an XML document before transmitting it in our *One Pass Technique* (OPT). In this technique we determine a threshold size value for the network. Any XML document size smaller than this threshold it will be sent uncompressed while any XML document size larger it will be compressed before it is sent.

## 2 Establishing Threshold and One Pass Technique (OPT)

In contrast to the five network factors that contribute to the latency time of delivering a query output [1] based on the analysis of the one gigabyte TPC-H benchmark [10], our method presented here is utilizing an established “threshold” for the current working status and then to have “one-pass” transmission. We defined a threshold value for the network such that the transmitted time, for XML documents size are compressed (such as via XMill) and uncompressed, will be comparable. To determine what this value could be, we first need to determine the networks characteristics. As the networks characteristics will evolve with time the threshold value needs to dynamically change with the network.

Before OPT can be used on a network we need to determine the threshold value by making a number of XML transfers of different sizes across the network. The transmissions need to be made both with the document compressed, using XMill as an example, (and decompressed where it is received) and by transmitting the document without compression. An estimate of how long it takes to transmit a document of a given size can then be determined by curve fitting to these results. The threshold value is set to be the size when the estimated time to transmit it without compression is equal to the estimated time to transmit it with compression. In some situations this may result in a threshold value that will require compression of all documents or one that will never require compression of a document.

There are a number of factors that can prevent OPT from yielding the best result for all cases. The threshold value will only be valid for the network bandwidth it is calculated for, so if that bandwidth changes a threshold value will give an inaccurate result and a new threshold value will need to be determined.

The compression and decompression times are dependent on the CPU load. If the load on a CPU is heavier (or lighter) than it was when calculating the threshold value it may not make the appropriate decision on whether or not to use compression on the XML document. Similarly the technique works best with a homogenous set of CPUs. Different CPUs will take different time periods to compress and decompress the XML documents. The compression/decompression time of two low end CPUs on a network will be different to the compression/decompression time of two high end CPUs on the same network using the same threshold value. This can also lead to the OPT making a wrong decision on whether or not to compress the document.

OPT can also be affected by changes in the networks traffic density. If the network is under a heavier load than it was when the threshold value was calculated the technique is more likely to transmit an uncompressed XML document when a compressed document would have been faster, and with a lighter network load compressed XML transmissions are more likely to occur when an uncompressed transmission would

have been faster. OPT is best used in a homogenous environment where the network bandwidth is well known and network traffic is reasonably stable.

### 3 Experimental Results: Some Examples

A number of XML documents were gathered to test using a time based threshold to decide on when to compress a document and when not to. These files were of different sizes. An application program was written to transmit these documents a number of times across a network using a threshold value. Any XML document with a size greater than the threshold value is transmitted compressed while all other XML documents are sent uncompressed. The algorithm used is:

*If*  $Size_{Document} > Size_{Threshold}$  *Then* transmit\_compressed, *Else* transmit\_uncompressed

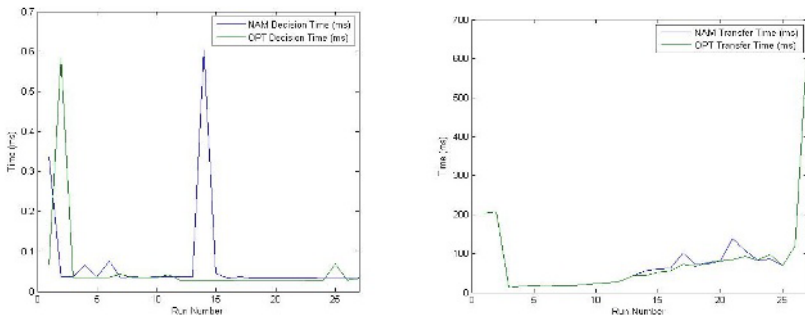
A similar application was set up to transfer the documents using the NAM methodology (Ghandehazrizadeh, 2003). NAM uses measured network and computer characteristics to compare estimates on how long it would take to transmit an uncompressed document against an estimate of how long it would take to transmit a compressed document. The algorithm used is:

*If*  $Time_{Uncompressed\ Transmission} > Time_{Document\ Compression} + Time_{Compressed\ Transmission} + Time_{Document\ Decompression}$  *Then* transmit\_compressed, *Else* transmit\_uncompressed.

The experiment was conducted using a client PC ( 754pin Athlon64 3200+@2.05GHz with 1GB RAM), one Server PC ( Celeron D 2.8@2.79GHz with 512MB RAM) connected by a Router (Billion BIPAC 7402G) over a 100MBit Ethernet connection.

A set of twenty-seven runs were carried out to determine the characteristics of the network before the applications were run against it, solving the quadratic equations used to get the time and size estimates NAM uses in it decision algorithm and determining the threshold value for the current network traffic load for the OPT. The threshold value was found to be 425KB.

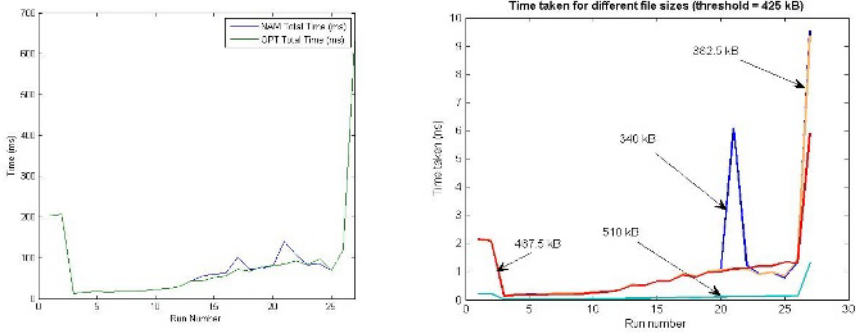
Figure 1 (a) shows the comparison between the time it took NAM to decide whether to send the XML document compressed or uncompressed and the time it took the STT to do the same.



**Fig. 1.** (a) NAM decision time vs. OPT decision time; (b) Total NAM transmission time vs. total OPT transmission time

Figure 1 (b) shows the total transmission time in each run using NAM with the total transmission time in the same run using the OPT.

Figure 2 (a) shows the combined total transmission and decision times for NAM with the total transmission and decision times for the OPT.



**Fig. 2.** (a) NAM total time vs. OPT total time; (b) Times taken for the files changed to the threshold values changed  $\pm 10\%$  and  $\pm 20\%$  and then (using five parameters) transmitting

In these experiments we see the OPT performing slightly better than NAM, completing the 27 runs 149.33734 ms faster than with NAM, 148.90656 ms in the total transmission time and 0.43078 ms faster than NAM in the decision making time.

In order to confirm our results, we take the file size of  $\pm 10\%$  and  $\pm 20\%$  of the threshold value to see how the file size to affect the transition time. The results are shown in Figure 2 (b). It is clearly shown that the lower size files are constantly taking more time in comparison with large size files. This shows that those files that the sizes are above the threshold should be compressed before any transmission and those files that the sizes are below the threshold should be uncompressed. The results of the example show that the worst case could make the time taken as large as  $> 10$  times longer than it normally takes.

## 4 Conclusion

We have examined the possibility of using the OPT to control when an XML document should be compressed before being transmitted over a network. We compared this technique to another control technique, the Network Adaptable Middleware (NAM), and found that for a stable network of known characteristics and a random selection of XML documents the OPT is able to out perform the NAM technique. While we suspect that the NAM technique would be able to match the transmission times of the OPT after enough data has been collected to refine its estimates the lower CPU decision making time required for the OPT means that it is a better choice for situations where a network is relatively stable in bandwidth, CPU load and network traffic density.

The experimental results demonstrate that our OPT method is superior to the method offered by [1] for fixed size, for example for 4.5 MB XML file our method

will make the CPU time decreasing 22.69% and total transition time will save 4.61% in comparison with method [1].

We also show the results that the time taken for the cases that if the file's size changed  $\pm 10\%$  and  $\pm 20\%$  of the threshold value. The final results strongly support our method. If we use five-parameter method (or NAM) to transmit those files, the results show that the worst case could make the time taken as large as  $> 10$  times what it normally takes.

## References

- [1] S. Ghandeharizadeh, C. Papadopoulos, M. Cai, and K. K. Chintalapudi, "Performance of Networked XML-Driven Cooperative Applications", In Proceedings of the Second International Workshop on Cooperative Internet Computing Hong Kong, China, August 2002.
- [2] Alexander Ridgewell, Xu Huang, and Dharmendra Sharma, "Evaluating the Size of the SOAP for Integration in B2B", the Ninth International Conference on Knowledge-Based Intelligent Information & Engineering Systems Melbourne, Australia, September, 2005. Part IV, pp.29.
- [3] H. Liefke and D. Suci. XMill: An efficient Compressor for XML Data. Technical Report MSCIS-99-26, University of Pennsylvania, 1999.
- [4] Curbera, F. Duftler, M. Khalaf, R. Nagy, W. Mukhi, N and Weerawarana, S.: Unraveling the web services web: An introduction to SOAP, WSDL, UDDI. IEEE Internet Computing, 6(2): 86-93, March-April 2002.
- [5] Fan, M. Stallaert, J. and Whinston, A. B.: The internet and the future of financial markets, Communications of the ACM, 43(11):83-88, November 2000.
- [6] Rabhi, F.A. and Benatallah, B.: An integrated service architecture for managing capital market systems. IEEE Network, 16(1):15-19, 2002.
- [7] Kohloff, Christopher and Steele, Robert: Evaluating SOAP for High Performance Business Applications: Real-Time Trading Systems, 2003, <http://www2003.org/cdrom/papers/alternate/P872/p872\kohloff.html>, accessed 22 March 2005.
- [8] S. Ghandeharizadeh, C. Papadopoulos, M. Cai, R. Zhou, P. Pol, NAM: A Network Adaptive Middleware to Enhance Response Time of Web Services, 2003, MASCOTS 2003: 136
- [9] R.R. Iyer and D. Wilhite. "Data Compression Support in Databases." In Proceedings of the 20<sup>th</sup> International Conference on Very Large Dasta Bases, 1994
- [10] M. Poess and C. Floyd. "New TPC Benchmarks for Decision Support and Web Commerce." ACM SIGMOD Record, 29(4), Dec 2000.

# Distributed Face Recognition: A Multiagent Approach

Girija Chetty and Dharmendra Sharma

School of Information Sciences and Engineering, University of Canberra, Australia  
{Girija.Chetty, Dharmendra.Sharma}@canberra.edu.au

**Abstract.** In this paper we present an application of agent technology to the problem of face recognition. With a new composite model consisting of multiple layers, the system can achieve high performance in terms of robustness and recognition in complex visual environmental conditions. The robustness of the complex face recognition system is enhanced due to integration with agent based paradigm, with more than 95% accuracy achieved under illumination, pose and expression variations of faces in images with multiple faces, and background objects. The results of preliminary findings are promising, suggesting further investigations in intelligent agent methodology for multimodal biometrics using fusion of face, gait, gesture, and voice biometric traits to person identity recognition problem in distributed scenarios such as video surveillance, health informatics, and crime investigation applications.

## 1 Introduction

In recent years applications of biometric technologies are not just limited to high security border control and national security scenarios, but in day-day civilian and e-commerce applications [1]. Many different biometric traits are available for person recognition such as fingerprint, face, voice, gait, retina, iris, hand geometry and vein patterns. However, recognition based on any one of these modalities may not be sufficiently robust or else may not be acceptable to a particular user group or in a particular situation or instance.

Current approaches to the use of single biometrics in person identity recognition are therefore limited, principally because no single biometric is generally considered both sufficiently accurate and user-acceptable for universal application. Multimodal biometrics can provide more robust solutions to security and convenience requirements of many applications such as video surveillance, crime investigation, and health informatics scenarios, where there is a need to recognize the identity from insufficient biometric sensor data. For example in face recognition systems in video surveillance scenarios, low resolution images in cluttered background, difficulty to obtain frontal face images, bad lighting, occlusions, pose and expression variations, and multiple persons in the scenes, are typical challenges which the face recognition algorithms are confronted with. Moreover, for distributed implementations, the different stages of face recognition, such as the acquisition stage of capturing face biometric information, the feature extraction stage, the template acquisition and classification stage are spatially and functionally distributed, with complex hierarchies of security levels and interacting user/provider requirements. The face recognition systems deployed in such distributed environments require that the system

is adaptive and flexible in configuration, for which an approach based on innovative multi-agent based paradigm [2-4] can be very promising.

## 2 Agent Based Methodology for Multi-modal Biometrics

In considering the use of multi-modal biometrics in another scenario for example, a realistic distributed civilian environment, such as a public system for regulated access to healthcare records, it is clear that there are several inter-related sources of variability which are likely to affect the required performance of the authentication system. These sources include, for example, environmental conditions, users' physiological/behavioural characteristics, users' preferences, variability of the communication channels, and so on. Thus, there is a clear requirement for the system to be able to adapt to user needs and conditions and, especially, to be able to determine and maintain an acceptable balance between confidence and convenience for its users through negotiations between information users and providers.

The use of intelligent agent methodology allows efficient management of complexity introduced by the use of multi-biometrics for remote access. Intelligent autonomous agents [3], and multi-agent systems form a vibrant and rapidly expanding research field [5]. Agents can be defined as computer sub-systems that interact with some environment, and are capable of autonomous action. In addition they are flexible in responding to their environment, pro-active in exploiting opportunities and seeking goals and "social" in their interactions with other agents where appropriate. In addition they may have other valuable properties such as adaptability or mobility.

A novel agent architecture, MARS multi-agent system [2] was proposed comprising a group of several interacting agents, and is well suited to situations where multiple perspectives of a problem-solving situation exist. Types of interaction that may best be suited to biometric security involve co-operation, co-ordination and negotiation between agents. The needs of the information provider for establishing sufficient trust in the user may have to be balanced with the confidentiality of the user's biometric information and his/her ease of use of the system. A balance may need to be struck for each service, transaction or session and may even be dynamically modified during use. Tasks for the agent systems include for example, handling of multiple authorisation levels, location of data across several repositories, and user interface and performance modification as required by the user or necessitated by the environment.

In this paper we report the proof of concept experiments based on innovative multi-agent architecture to solve the problem of distributed face recognition in complex environments. The architecture is based on adapting the MARSE, a multi-agent systems framework, proposed by Intelligent Systems Group at the University of Canberra [2] for distributed face recognition task. The architecture consists of a fusion of multi-layered structural and functional models in a network-oriented distributed environment. The system uses agent oriented implementation of multiple biometric modalities to check and verify identity, and the information from multiple biometric traits can be combined using a series of novel data fusion techniques to find an optimum degree of reliability in authentication. The fusion technique used adapts to environmental and person-specific variations in which a system is accessed, and

includes several features such as significance, confidentiality and cost of data, capture environment and recognition success rate histories of individual biometric traits for the person.

Here we describe face detection, facial feature extraction and face matching stages for implementing face recognition in a distributed fashion. The details of the other biometric traits used, and the complete multi-layer structural and functional model with different fusion algorithms is described elsewhere, [6]. The paper is organized as follows. Section 2 reviews previous approaches in related agent based computing paradigm, and Section 3 discusses the proposed system architecture, with highlighting its three-layer structural model and three-layer functional model. Section 4 reports the experimental results, with some conclusions presented in Section 5.

### **3 Previous Approaches**

There has been a limited effort in research literature until now in addressing distributed multimodal biometric recognition problems using agent-oriented methodology, though there has been some work in single mode biometrics. Aslandogan and Yu [4] developed a image search agent named Diogenes, which takes advantage of the text/HTML structure of web pages as well as the visual analysis of the images for personal image search and identification. Diogenes works well on internet that contain a facial image accompanied by a body of text that contains textural identification of the image. Another system combining intelligent software agents with face recognition engines has now been developed by the ANSER team [5]. The Missing Children Locator Agent [5] is an example of one of ANSER's systems. The software agents continuously search and retrieve facial images and relevant information from web sites on the Internet. A back-end face recognition engine analyses each image file to detect and match each face to a set of stored face images of missing children. The user can also search for an input image from the database for a match. Successful implementation of the system demonstrates its ability to speed up processing-intensive tasks and automate the processes required for face recognition. However, in the above two systems, the intelligent agent does not offer robustness in distributed environment. It can not migrate from host to host in a heterogeneous environment. This disadvantage makes it difficult to incorporate it with legacy systems when the system has to access image databases in different formats. An active tracking vision system was reported by Aoki et al. [7], that can detect human faces from sequential images. By integrating multiple information sensing from real-time and semi real-time agents, their system can track people in a room reliably and efficiently. objective faces in real time.

### **4 System Architecture**

The multi-layer structural and functional model can be viewed in two ways: one at structural level, and the other at the functional or operational level. The four-layers of the system at structural level are:



- The interface layer: A communication channel between the input/output device and the application server.
- The central controller layer: To control all inner biometric recognition schemes as well as implementing an intelligent detection strategy.
- The external assistant layer: Includes a group of agents to help implementation of feature extraction schemes.
- The remote application layer: An outer application part of the system which is used for face matching by connecting remote legacy databases consisting of biometric templates.

At functional level , it consist of three layers:

- A central agent host layer implementing an intelligent detection scheme on a single PC.
- A neighboring agent host layer implementing a feature extraction scheme in a parallel-computing environment.
- A remote agent host layer implementing a remote face-matching scheme in a distributed computing environment.

The details of each layer in the structural level can be summarized as:

- *The interface layer* – This layer allows an easy access to the system by establishing a point-to-point input/output layer explicitly connecting the input/output device and the application server. All recognition schemes are hidden from the end user. A user interface is used to accept the input image and return recognition results.
- *The central controller layer* – This layer is the main part of the whole system and connects all sub-systems. A novel face detection scheme based on fusion of multiple colour spaces is implemented in this layer. Three stages of face detection that is colour segmentation, the skin-region detection, and face fusion is done by three agents in this layer:
  - *Skin segmentation agent* - The facial region is localized by this agent based on statistical skin colour distribution and thresholding. Then the knowledge about facial patterns (distribution of non-skin sub-regions) is used to determine instances of face within such regions. The agent uses morphological operators to divide different convex objects, removing regions that are too small, and recovering regions' sizes while keeping the same topological structure.
  - *Skin region detection agent* : This agent localizes the face region by matching the skin regions with a template, and eliminates those skin regions which do not correspond to face region such as hands.
  - *Face Fusion agent*: This agent performs fusion of face segmentation and face region detector based on multiple colour spaces such hue-saturation(HSV) colour space, chrominance-luminance colour space (YCrCb) colour space, and RGB colour space.

Use of multiple colour spaces allowed a face detection accuracy of more than 90% under complex visual background conditions as shown in Figure 1.

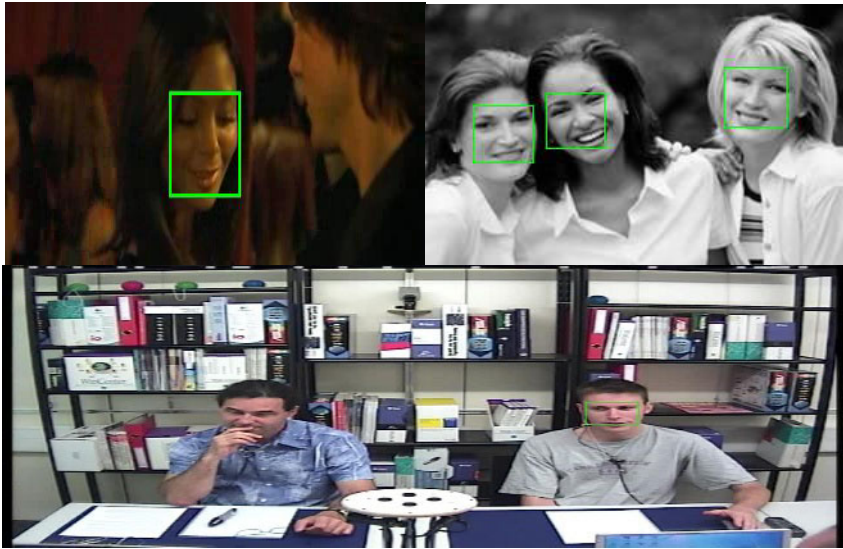


Fig. 1. Face detection in complex visual background

#### 4.1 The External Assistant Layer

The external assistant layer extracts features from face images, including both local and global features. First, global features are extracted by principal component analysis (PCA) or Eigenface approach, while local features are obtained by active appearance model (AAM) and Gabor wavelet transform (GWT) [8-11]. Extracting global and local features is computational intensive and hence we use a group of external computation agents and adopt the "Divide and Conquer" strategy in this layer. Figure 2 shows the extraction of global features based on Eigenface approach.



Fig. 2. Extraction of global features with eigen face approach

#### 4.2 The Remote Application Layer

Remote face-matching scheme is implemented in this layer. Based on the fact that large image databases with different formats might be located in different places over

the network, it makes more sense if the agent moves to the remote data source for searching and matching, rather than transferring large volumes of data over the network for processing. In this layer, we explicitly create a matching agent, initializing it with matching algorithms and dispatching it to the Internet. Upon reaching a new host, the matching agent interacts with remote agents and communicates with the backend databases for searching and matching.

A two-step-matching scheme is performed: geometric-based coarse matching and dynamic-link-architecture based accurate matching [12]. Geometric, feature-based face recognition is among the earlier algorithms proposed [10]. As the image database becomes larger, however, it turned out to be impossible to perform accurate recognition by simply using this scheme. However, we embed this algorithm into our proposed matching agent as an effective pre-filtering for the second-step for accurate matching.

In the second step, the overall geometrical configuration of the face features is described with a vector of numerical data, representing the position and size of the main facial features, e.g. eyes, nose and mouth, and supplemented by the shape of the face outline [11]. For both steps, a nearest neighbourhood classifier is used for the proof of concept experiments, and a Hidden Markov Model (HMM) classifier would have given better classification performance. Its performance is a function of the number of classes to be discriminated (people to be recognized) and of the number of examples per class.

## 5 Face Recognition Experiments

To implement and evaluate the system performance, we used the IBM ABLE environment [8,9]. ABLE is a Java framework, component library, and productivity tool kit for building intelligent agents using machine learning and reasoning. We conducted experiments at different levels of framework. Three sets of experiments are conducted: Face Detection Experiments (FDE) in the layer I, Feature Extraction Experiments (FEE) in the layer II, and remote Face-matching Experiments (FME) in the layer III. The FDE Test aims at evaluating the algorithm's robustness in the proposed detection agents under different visual environment variations. Then the FEE test evaluates the accuracy with which features are extracted and further used by remote face matching agent. Finally, the FME test examines the correctness and effectiveness of remote face matching scheme in a distributed environment.

### Face Detection Experiments

In order to obtain a stable colour segmentation with noisy face images, more than 1000 face images with variation in pose and illumination and different types of skin are chosen as the training samples for the colour segmentation agent. More than 100 face images were used for the FEE test. The experimental results illustrated in Table 1 show that the improvements achieved for FME tests in terms of robustness to the variations in pose, lighting, multiple faces, moderate tilt of faces and partial occlusion.

**Table 1.** Face Recognition accuracy (%)

Visual Environmental Conditions	FDE-I	FDE-II	FEE-I	FEE-II	FDE-II+FEE-II
Unevenness of lighting	84.56	86.57	89.79	91.46	96.23
Multiple faces in the scene	85.55	88.48	91.16	90.57	93.47
Moderate tilt of faces	87.59	90.29	92.85	89.66	94.18
Pose and expression var.	89.48	89.44	91.11	92.38	91.33

FDE-I :Face detection based on single colour space; FDE-II: Face detection based on fusion of colour spaces; E-I: feature extraction based on global PCA features; FEE-II based on fusion of PCA+GWT

## 6 Conclusion

Multimodal biometrics provide a practically viable approach for overcoming the performance and acceptability barriers to the widespread adoption of biometric systems. However, it is essential that the resulting complexities are managed in a seamless and effective way. The paper introduces a multi agent systems approach for distributed face recognition problem. The combination of powerful fusion algorithms with multi-layer structural and functional model allows higher face recognition accuracy even under adverse environmental conditions.

In contrast to current face recognition models, which suffer from slow performance and platform dependence, the proposed multi-layered system model with several improved algorithms has been tested through experiments, demonstrating its feasibility and effectiveness. Coupled with other supporting schemes, the system is potentially useful in a wide range of distributed face recognition services such as remote video surveillance, health informatics and criminal identity verification.

## References

- [1] F Deravi and M Lockie, "Biometric Industry Report - Market and Technology Forecasts to 2003", Elsevier Advanced Technology, December 2000.
- [2] Dharmendra Sharma, "Proposal for a Multi-Agent Reasoning System Environment, School of ISE, University of Canberra.
- [3] N R Jennings, K Sycara, M Wooldridge, "A Roadmap of Agent Research and Development", Autonomous Agents and Multi-Agent Systems, Kulwer Academic Publishers, Vol 1, pp 275-306,1998.
- [4] Y. A. Aslandogan and C. T. Yu, "Diogenes: A Web Search Agent for Content Based Indexing of Personal Images," in Proceedings of ACM SIGIR 2000, Athens, Greece, July 2000.

- [5] H. Wisniewski, "Face Recognition and Intelligent Software Agents - an Integration System," Prepared statement for the U.S. Senate Committee on Commerce, Science and Transportation, May 12 1999.
- [6] G.Chetty, and D.Sharma, "Multimodal biometric fusion based on multi-agent architecture", 2007 WSEAS Int. Conf. on COMPUTER ENGINEERING and APPLICATIONS (CEA'07)(under review).
- [7] Y. Aoki, K. Hisatomi and S. Hashimoto, "Robust and Active Human Face Tracking Vision Using Multiple Information," Proceedings of SCI'99 (World Multiconference on Systems, Cybernetics and Informatics), Vol.5, pp. 28-33, Aug. 1999, Orlando.
- [8] J. Kiniry and D. Zimmerman, "Special Feature: A Hands-on Look at Java Mobile agents
- [9] D.B Lange and M. Oshima, Programming and Deploying Java Mobile Agents with Aglets, Addison-Wesley, 1998.
- [10] Cai, A. Goshtasby and C.Yu, "Detecting Human Faces in Color Images," Image and Vision Computing, 18(1): 63-75, 2000. .
- [11] T. Kanade, "Picture Processing by Computer Complex and Recognition of Human Faces," Technical report, Kyoto University, Dept. of Information Science, 1973.
- [12] M. Lades, J. C. Vorbruggen and J. Buhmann, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," IEEE Transactions on Computers, Vol. 42, No. 3, Mar. 1993.

# Using Windows Printer Drivers for Solaris Applications – An Application of Multiagent System

Wanli Ma, Dat Tran, Dharmendra Sharma, and Abhishek Mathur

School of Information Sciences and Engineering  
University of Canberra  
{Wanli.Ma, Dat.Tran, Dharmendra.Sharma,  
Abhishek.Mathur}@canberra.edu.au

**Abstract.** This paper proposes using multiagent system technology to solve the problem of printing across heterogeneous operating systems without re-implementing printer drivers. The printing problem comes from a real world application, where we have to print from Solaris operating system applications to printers which only have Windows operating system drivers. Multiple intelligent agents are distributed on both Windows and Solaris platforms to integrate services running on these platforms. They are responsible for printing task interception, operating system spooling, printing format conversion, load balancing, and intelligent data routing. This approach avoids re-implementing Windows printer drivers on Solaris, yet provides seamlessly printing for Solaris applications to those printers.

**Keywords:** Multiagent, MAS, operating systems, printer drivers.

## 1 Introduction

One of the major problems facing today's IT operation is the interoperability among heterogeneous operating systems. Popular daily used operating systems, such as Windows family, Linux and Unix family, MacOS family, and mainframe operating systems etc., are not compatible with each other. Some applications only run on certain operating systems. For any sizable organization, it is very likely there is coexistence of heterogeneous operating systems in operation. The situation will stay as it is for a long time yet to come according to ComputerWorld [1].

Significant effort has been invested by the research communities and IT industries to improve the interoperability among heterogeneous operating systems. For example, middleware solutions, notably, Microsoft .NET environment [2], CORBA specification [3], and J2EE specification [4], and distributed operating system approaches, such as [5], are proposed and implemented, yet we are still not be able to easily integrate applications on different platforms. In real life, we are still battling the platform issues, such as how to run application X on platform Y.

The problem we encountered is to provide Solaris Sun Ray users with printing capability. It is from a real life project which aims at providing the community with basic function, almost nil maintenance, and low cost computers and Internet access. Sun Ray thin client technology is chosen as the delivery infrastructure. When coming

to printing, most home user level printers on the market are not PostScript compliant. Manufactures have their own proprietary printer drivers. All of the printers have Windows drivers, but none of them has Solaris drivers. The challenge is to make the printers work for Solaris, and therefore, provide Sun Ray users with printing capability to the printers of their choices from the market. Existing technology fails to provide a neat solution.

Multiagent system (MAS) technology has its root from distributed artificial intelligence [6]; however, the continuous and autonomous nature of agents and the communication among these agents make them a good candidate for distributed computing, and even just general software applications, for example, agent based software engineering [7], autonomic computing [8, 9], and many other distributed applications [10-14].

This paper proposes using multiagent system technology to seamlessly integrate services running on heterogeneous operating systems and thus solve the printing problem. Our proposal is alone the line with the idea of the autonomic computing system proposed by Tesauro et al [8, 9], but our primary focus is different. At this stage, our main task is to integrate Windows printer drivers into Solaris operating system.

The rest of the paper is organized as follows. Section 2 describes the problem in details, and Section 3 presents the multiagent solution. In Section 4, we discuss our prototype implementation and future work. Section 5 concludes the paper.

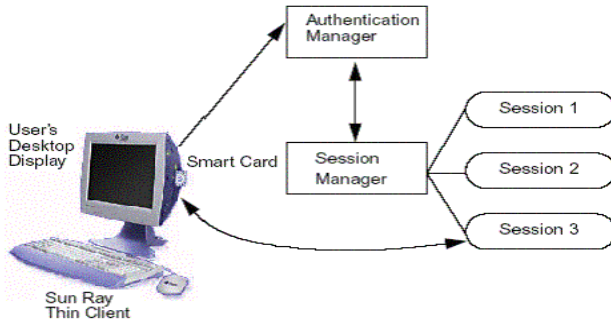
## 2 The Problem

The problem is, in essence, printing to low end ink jet printers from Solaris operating system. To describe the problem clearly, we will first briefly explain Sun Ray thin client technology and then the ink jet printers available on the market. Afterwards, the problem becomes clear.

### 2.1 Sun Ray Thin Client Technology

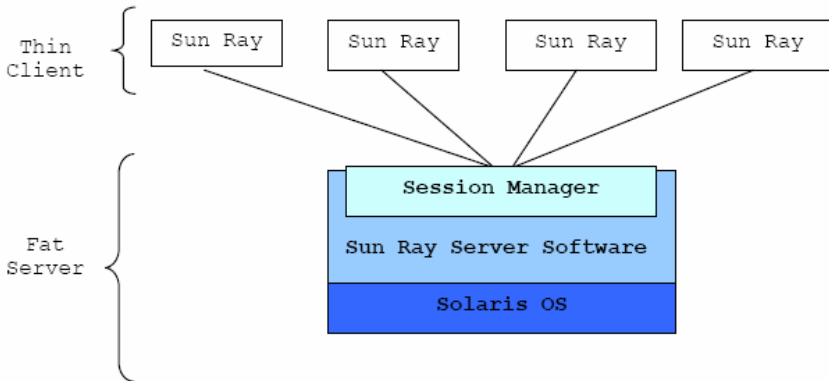
A Sun Ray thin client unit consists of a monitor, a built-in smart card reader, a network port, an embedded audio speaker, and 4 USB ports, which support keyboard, mouse, and other peripherals. The unit itself has little CPU power and memory capacity. It only needs to recognise keystrokes and mouse events and also to display pixel data received from the server. It is actually just an I/O device: taking keyboard and mouse actions and transmitting them across the network to the server. The server then transmits back the screen pixels, including mouse movements and the characters typed. The actual computing is performed on one or more remote servers, which run Sun Ray Server Software (SRSS) on Solaris operating system platforms. No computing is performed locally.

The absence of application files, operating system, and storage media on the local client defines the stateless nature – which also means that the desktop, i.e., the thin client unit, does not have to be administered locally. Thin client users are administered remotely at the server end. Administrators are freed from maintaining individual clients and performing repetitive task of setting up each client unit.



**Fig. 1.** Sun Ray Session Management

Remote processing requires all sessions to be run at a central server or servers, which maintain and authenticate sessions. Smart card allows hot swapping of sessions between terminals. Session Manager simply switches the session from one terminal to another and provides mobility to users (Fig. 1).



**Fig. 2.** SRSS and Solaris

The key of this thin client architecture lies in the Sun Ray Server Software (SRSS). It runs on the Solaris OS and is responsible for receiving and transmitting screen refreshes, keyboard strokes, and mouse movements between a client and its session (Fig. 2). It also manages user applications – starts and terminates applications on user behalf.

## 2.2 The Available Printers

Most industrial strength printers are PostScript compliant and also have their own network cards. They do not have problems with Solaris printing. However, the low end ink jet printers for home users are not PostScript compliant. These printers have



very little intelligence themselves. They heavily rely on the printer drivers running on the host computers to do all the necessary operation, including data format conversion. This type of printers is very popular with home users and small businesses. They are cheap to buy, can print on different media, and also can print in colors. A laser printer with the same ability costs a lot more.

Due to their low data processing power, printer drivers become the critical part of operating these printers. All ink jet printers have Windows drivers, but most of them do not have Solaris drivers. It is impossible to buy a randomly chosen printer and plug it into a Solaris machine and hope it will work smoothly. To make things worse, the printers on the market updates very fast just like any other IT products. It basically means that the printer drivers have also to be updated as fast.

### 2.3 The Problem Encountered

To be able to print, from an ordinary Sun Ray user's point of view, it would be the same as with other personal computers: buying a USB printer (almost certain, cheap ink jet type) and connecting it to the USB port of the Sun Ray client unit. Afterwards, the user would naturally expect printing starts.

In fact, on the contrary, the system has great difficulties to work with the printer. Sun Ray applications are running on Solaris operating system and controlled by Sun Ray Server Software (SRSS), as discussed in the previous section. The printing problem becomes the problem of printing from Solaris operating system to the printer. SRSS can recognise a new USB device being plugged; however, it won't be able to control it without proper driver.

We checked all available ink jet printers on Australian market in the middle of 2005 and failed to find any one providing a printer driver for Solaris. The problem further manifests itself into multiple problems:

- The fundamental problem is that there is no proper printer driver for Solaris operating system.
- If there were a printer driver from a particular printer, there is no guarantee of drivers from other models or brands or manufacturers. A Sun Ray user may buy a printer which does not have the driver for Solaris operating system.
- If there were a driver for the current printer, there is no guarantee of a driver for the next updated model.
- The available printer models and the associated printer drivers on the market keep changing.

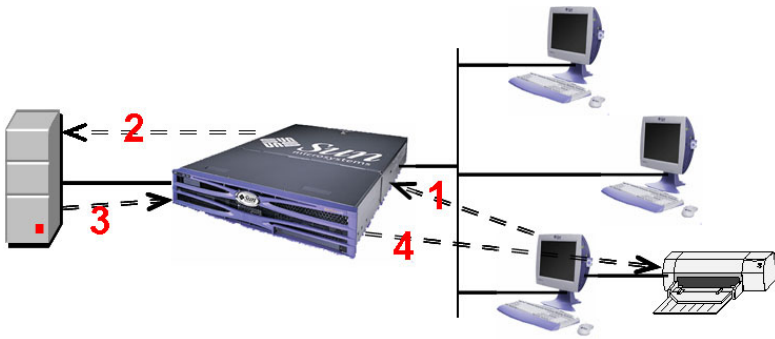
The lacking of printer drivers is the critical problem we have to solve in order to provide Sun Ray users with printing capability.

## 3 The Proposed Solution

In essence, to make printing from Solaris possible, we have to find a way to drive the printers. There are 2 possible solutions to this problem. One is to redevelop or port printer drivers for Solaris. The other way is to find an alternative by pass.

The former is not a real solution. There are a number of serious drawbacks with the approach:

- Full cooperation from printer manufacturers is needed for redeveloping the drivers; however, it is almost impossible. It is impractical to develop a printer driver without the detailed internal knowledge about the printer's hardware and firmware. Manufactures always treat this knowledge as trade secret and won't make it public.
- Without the cooperation, the redevelopment work has to rely on reverse engineering. Reverse engineering is tedious and time consuming. It may also have legal ramifications.
- Reverse engineering cannot keep the pace with the market, given the fact of many models, frequent model updates, and many manufacturers on the market, let alone new manufacturers entering into the market.



**Fig. 3.** Printing from Sun Ray to its printer via Solaris and Windows

To avoid all these difficulties, we propose an innovative approach: using Windows printer drivers for Solaris. The idea is simple: a Windows box is connected to a Solaris box (or several Solaris boxes) via network and is working as conversion channels. The Solaris box sends a printing job to a conversion channel, which corresponds to a particular printer. The conversion channel then translates the printing job into the proprietary format, which the printer understands, via its Windows printer driver. The printing job is then sent back to the Solaris box and then to the attached printer (Fig. 3).

Under the proposal, printing now becomes a very complicated task involving multiple operating systems and communication. To coordinate this complicated task, multiagent system is needed. A number of intelligent agents are distributed on both Solaris and Windows platforms, intercepting printing task, synchronizing operating system spooling, instructing printing format conversion, and also balancing the load and routing the data.

Solaris resident agents are responsible for capturing printing requests and probing for printer device descriptions. They are also responsible for sending the printing requests over to Windows based agents, which perform the tasks of converting printing data to the forms which can be understood by the printers connected to the Sun Ray clients (on Solaris platform). After the conversion, the data is sent back to the Solaris agents in and then to the printer, Fig. 4.

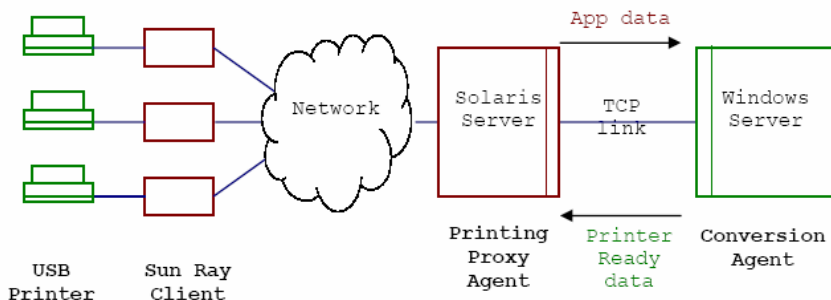


Fig. 4. The Agent Architecture

A printing request is now handled by the agents in the following steps:

- **Intercepting the printing requests from a Sun Ray client**

On the Solaris end, there are many *printing proxy agents*. An agent will intercept the printing request whenever a user performs a printing operation. A virtual print spooler and its associated directories are setup on Solaris to help the interception. The intercepted data will then be sent to a conversion agent on the Windows end. The printing proxy agent has the knowledge of the current user, the current application, and the printer model etc. related information. The knowledge is used to decide to which conversion agent the task will be sent.

- **Data conversion**

On the Windows end, there are many *conversion agents*. Upon receiving a message from a printing proxy agent, a conversion agent will convert the received data into printing format by simulating printing operations of human users on Windows operating system. Of course, the printer driver of this particular printer has to be installed on the Windows machine, but there is no need for a physical printer. With the help of the printer driver, the converted data is in the very format that printer can understand.

- **Transmitting the data back to Solaris and then the printer connected to the Sun Ray client**

The conversion agent is then transmits the converted data back to the printing proxy agent. The data is further forwarded to the printer connected to the Sun Ray client.

## 4 Prototype Implementation and Future Work

On the surface, the implementation is straightforward – only 2 types of agents. Actually, it is rather complicated. There are a number of factors which complicate the situation. For example, there might be more than 1 Solaris hosts and more than 1 Windows hosts. The agents on these hosts have to negotiate with each the best pair of printing proxy agents and conversion agents, with respect of the current working load of each host and the printer types. In addition, intercepting printing request on Solaris

side and mimicking human printing operation on Windows side require programming at the deep hearts of both operating systems.

To prove the concept, at the very beginning, manual interception and redirection of data were performed, and printing was successfully done from Solaris applications to an HP inkjet printer connected to a Sun Ray client. Solaris application data was converted to the printer understandable format via a Windows machine as conversion channel.

Up to date, a prototype implementation is very well under the way. So far, the implementation of the agents has been done. The reminding work is the communication among the agents.

Implementing the communication among the agents does not only bring us to the full implementation of the proposed system, it also prompts us to think about the future work.

In the near future, related to the printing problem itself, we need to decide how to do load balance and also if we allow mobile agents. Being able to understand the load of each involved host is critical for real life applications. We haven't yet decided how we are going to tackle the problem in our proposal. It might be helpful to introduce another type of agents – traffic control agents – to coordinate the interactions among the printing proxy agents and conversion agents. More investigation is needed. As to the communication among the agents, both broadcast and peer-to-peer are employed.

To allow mobile agents or not is not so critical at this stage. It is more about agent research than solving the printing problem per se. However, we do see the benefit of having mobile agents, especially for load balance purpose and also in ad hoc network environment.

Further down the track, we'd like to test this proxy and conversion agent idea on other applications, such as (i) bringing Windows multimedia formats (e.g., Windows media files) to Solaris and other operating system environments and, in more general sense, (ii) providing Windows like GUI to other operating systems without much programming of redevelopment.

## 5 Conclusion

The paper proposed to use multiagent technology seamlessly integrating the printing services on Windows and Solaris operating systems. Printing proxy agents, which reside on the Solaris operating system side, are responsible for capturing printing requests and then pass the captured data and other related information to conversion agents, which reside on the Windows operating system side. A conversion agent is responsible for converting the printing data into the right format the final printer understands. With the effort of printing proxy agents and conversion agents, not only is it possible for Solaris to print to the printers which only have Windows printer driver; but also the whole printing operation is transparent to the end users.

In essence, our proposal is about making applications on one operating system as the pluggable services to the other operating system. Multiagent technology makes the integration smoothly and seamlessly with minimum effort. In the future, we expect to apply this proposal to wider areas of software application integration across heterogeneous computer systems and distributed computing.

**Acknowledgments.** The research work is supported by The Completion Grant of Division of Business, Law and Information Sciences, University of Canberra.

## References

1. Thibodeau, P. *Mixed IT environments remain king with large users*. 2005 [cited 2006 January]; Available from: [http://www.computerworld.com/hardwaretopics/hardware/story/0,10801,106776,00.html?source=NLT\\_AM&nid=106776](http://www.computerworld.com/hardwaretopics/hardware/story/0,10801,106776,00.html?source=NLT_AM&nid=106776).
2. Microsoft. *NET homepage*. 2006 [cited 2005 December]; Available from: <http://www.microsoft.com/net/default.msp>.
3. CORBA. *CORBA homepage*. [cited 2005 December]; Available from: <http://www.corba.org/>.
4. Sun. *J2EE homepage*. [cited 2005 November]; Available from: <http://java.sun.com/javae/index.jsp>.
5. Kon, F., et al. *2K: A Distributed Operating System for Dynamic Heterogeneous Environments*. in *9th IEEE International Symposium on High Performance Distributed Computing (HPDC'00)*. 2000. Pittsburgh, USA: IEEE Press.
6. Stone, P. and M.M. Veloso. *Multiagent Systems: A Survey from a Machine Learning Perspective*. *Autonomous Robots*, 2000. **8**(3): p. 345-383.
7. Jennings, N.R., *On agent-based software engineering*. *Artificial Intelligence*, 2000. **117**: p. 277-296.
8. Kephart, J.O. and D.M. Chess, *The vision of autonomic computing*. *Computer*, 2003. **36**(1): p. 41-52.
9. Tesauro, G., et al. *A Multi-agent Systems Approach to Autonomic Computing*. in *Third International Conference on Autonomous Agents & Multiagent Systems (AAMAS'04)*. 2004. New York, New York, USA: ACM press.
10. Honavar, V., L. Miller, and J. Wong. *Distributed Knowledge Networks*. in *IEEE Information Technology Conference*. 1998.
11. Pan, L., et al. *Mobile Agents - The Right Vehicle for Distributed Sequential Computing*. in *9th International Conference on High Performance Computing - HiPC2002*. 2002. Bangalore, India: Springer-Verlag.
12. Martin, D., A. Cheyer, and D. Moran, *The Open Agent Architecture: a framework for building distributed software systems*. *Applied Artificial Intelligence*, 1999. **13**(1/2): p. 91-128.
13. Drashansky, T., et al., *Networked agents for scientific computing*. *Communications of the ACM*, 1999. **42**(3): p. 48-53.
14. Zhang, Z., et al. *Multiagent system solutions for distributed computing, communications, and data integration needs in the power industry*. in *General Meeting of the IEEE Power Engineering Society*. 2004: IEEE Press.

# A Novel Approach to Programming: Agent Based Software Engineering

Dharmendra Sharma, Wanli Ma, Dat Tran, and Mary Anderson

School of Information Sciences and Engineering  
University of Canberra  
{Dharmendra.Sharma, Wanli.Ma, Dat.Tran,  
Mary.Anderson}@canberra.edu.au

**Abstract.** Agent (and multiagent system) technology originates from artificial intelligence. The continuous and autonomous nature of agents and the communication among these agents also make them an excellent candidate for distributed computing and even general software applications. Agent based software engineering advocates using the agent technology for typical software development. Evolving from object oriented software engineering to agent based software engineering may be as significant as evolving from the procedure oriented concept to the object oriented concept. In this paper, we examine agent based software engineering in relation to the three main steps of software development: system analysis and modeling, design and implementation, and validation.

**Keywords:** Multiagent Systems, MAS, Agent Based Software Engineering, ABSE, Software Engineering.

## 1 Introduction

Agent-based software engineering (ABSE) is a new area in software development. The basic idea of ABSE is to use agents as the building blocks for a software system, the same way that objects are the building blocks in object-oriented software engineering. ABSE promises a simplified and enhanced approach to construct complicated software systems, stemming from the continuous and autonomous nature of agents and also the high level of abstraction and communication amongst the agents. Since its debut, agent-based software engineering has attracted a high level of interest in the research community.

Software engineering “*is an engineering discipline which is concerned with all aspects of software production from the early stages of system specification through to maintaining the system after it has gone into use.*” [1, p. 6]. There are three main steps in any software development project: system analysis and modeling, design and implementation, and software validation [1]:

- **System analysis and modeling** are used to understand and document the real world. Many routine procedures and modeling methodologies have been developed for traditional software engineering.

- **Design and implementation** are about software system architecture design, component partitioning, relationships among different components, and programming.
- **Software validation** is responsible for system testing and debugging.

In addition to these practical approaches, the research community has always dreamed of specifying and verifying software by rigid mathematical formulae; thus, formal specification and verification have always been research topics of software engineering. ABSE is no exception; it may provide a way to develop real life, usable, formal specification and verification.

In this paper, we will exam ABSE along the lines of the three steps of software development, including formal specification and verification. There are two reasons for this paper. The first is straightforward: we'd like to reflect the new developments in agent based software engineering. Due to page limit, we will only list the new developments in the last a couple of years. Even though, our selection of references is by no mean close to exclusive. Secondly, instead of surveying ABSE papers and grouping them into different categories by themselves, we'd like to put the development in this area into the three main steps of software development. By doing so, it is easier to see, from a practitioners' point of view, what has been achieved and what is still missing.

The rest of the paper is organized as follows: in Section 2, we will briefly study software agent and agent-based software engineering; Sections 3, 4, and 5 examine ABSE technology and tools for system analysis and modeling, design and implementation, and software validation; Section 6 examines formal specification and verification; finally, Section 7 concludes the paper with a summary.

## 2 Software Agents and Agent-Based Software Engineering

Agent technology originated from artificial intelligence research and can be traced back to the Actor model by Hewitt [2] of 1970'. The agent concept might be the most diverse topic among research communities. According to Bradshaw, a software agent is "*a software entity which functions continuously and autonomously in a particular environment*" [3]. Wooldridge and Jennings define agents by their characteristics. If a piece of software has the characteristics of *autonomy*, *social ability*, *reactivity*, and *pro-activity*, it is an agent in weak notion. On top of these characteristics, if the software has further characteristics of *adaptability*, *mobility*, *veracity*, and *rationality*, it becomes a stronger agent. The combination of these characteristics is what makes agents good candidates for general software applications. Nwana suggests that "agent" is an umbrella term, under which many different types of agents exist [4]. Almost every agent system consists of multiple agents. A single agent system may exist, but its ability as a system is in doubt. Multiagent system (MAS) technology is basically another way of talking about agents, with emphasis on multiple, perhaps distributed, agents and the communication amongst them. Stone wrote a good survey paper on agent communication to achieve the common goal [5].

Using agents as the basic building blocks to construct complicated software systems was first suggested by Shoham in 1993 [6]. The phase "agent-oriented

software engineering” was coined by Jennings and Wooldridge [7, 8]. In 2000, Charles Petrie suggested to use “agent based software engineering” as a “*refinement of some aspect of AOSE, based upon our practice experience in agent building*” [9]. We feel that agent based software engineering provides a better description of the discipline – using agents as the basic building blocks to construct software systems.

Jennings argued systemically on the feasibility and benefits of ABSE. His argument was based on the observations of building a large and complex software system. The essential activities in tackling the complexity of a software system are decomposition, abstraction, and organization:

- **Decomposition**, the so called divide-and-conquer strategy, is the oldest and most effective way for human beings to handle any large system, not just in software development.
- **Abstraction** is looking at the system as a whole. Abstraction makes complicated systems easier to understand and manage.
- **Organization** is identifying and organizing the relationships among subsystems, which are the result of decomposition.

Agent technology can handle all three activities very well; hence it is a good candidate for software development. The methodology and utilities developed to facilitate agent based software development constitute ABSE.

There are survey papers on Agent Based Software Engineering [7, 10]. In this paper, we attempt to put the technology developed for Agent Based Software Engineering to the cycle of software development, namely, system analysis and modeling, design and implementation, and software validation. Specification and verification are also a focus of this paper.

### 3 System Analysis and Modeling

The purpose of the system analysis phase is to identify an agent’s functionalities, regarded as roles. Steegmans *et al* [11] proposed to use the role model as a high level agent model and to use the role diagram, action diagram and commitment schema to support the role model design.

In the modeling phase, there are two main areas: agent modeling and formal modeling [12]. Agent modeling represents the system’s static structure and shows the agents involved in the system, the relationship between the agents and the attributes characterizing the agents. Formal modeling is used to describe the functionalities and operations of the Multi-Agent system and to study the consistency that the system should exhibit before its implementation. The formal model is constituted by the Agents Meta Language (AML), which is based on belief temporal logic [12]. Shehory [13] suggested that an agent modeling technique should have the following criteria: *preciseness, accessibility, expressiveness, modularity, complexity management, executability, testability, refinability, analyzability and openness*. On the other hand, an agent-based system modeling technique should have the following characteristics: *autonomy, complexity, adaptability, concurrency, distribution, and communication richness*.



Typical modeling techniques are AOM (Agent-Oriented Methodology), ADEPT (Advanced Decision Environment for Process Tasks) [14], and DESIRE (Design and Specification of Interacting Reasoning).

Shehory performed an evaluation on the three modeling techniques and concluded that the modeling techniques already provide a wide array of features advantageous for agent modeling but there is still a need for agent-based system developers for further exploration of the issues mentioned above [13].

## 4 Design and Implementation

In the agent design phase, we need to identify and determine the agents involved in the problem, the agents' organization scheme and the activities they carry out inside the system [12]. Steegmans *et al* [11] proposed a state chart modeling language to support the design of roles for situated agents.

An interesting approach has been proposed by Sanchis [15]. Normally, software agent designers first define the characteristics of the application and then define the agents that will implement it [16]. This methodology is called APG (APplication first, aGent after) and has a drawback which is to limit the expression of the properties of the agents to a particular realization of the target application. Sanchis [15] proposed a reverse methodology which is called AGP (AGent first, aPplication after). Given a target application, designers will select some specific agents from existing agents and adapt their expression of the properties (qualities and attributes).

## 5 Software Validation

It seems that the research in this area is far behind the development in the other areas of ABSE. Not so many reports can be found. Hall [17] proposed to use the following techniques for validation:

- *Conflict detection*: to analyse the model definition to determine whether any inconsistent state transitions could possibly occur. Inconsistencies can arise.
- *Scenario simulation*: to express formal concrete behaviour scenarios, each of which is an interleaved sequence of input events and expected output events resulting from them.
- *Scenario coverage*: to measure and report the coverage of a given set of scenarios.
- *Scenario maintenance*: to keep scenarios correct when the specification is evolved to meet new or changed requirements.
- *Interactive theorem proving*: to prove correctness properties of the model

## 6 Specifications and Verification

Formal specification and verification has been a long pursued goal of research communities and the IT industry. There is no exception in ABSE. A multiagent based software system can be described by the following related but different aspects: the

internal operation of an agent, the interaction of agents, the beliefs of an agent, the goal of the system or an individual agent, and the role of an agent etc.

A multiagent system can be described at two different levels: the interactions among agents and the internal operations of the agents to achieve their individual goals. Hilaire et al [18] suggested to use Object Z to describe agent internal operations and use Statecharts to describe agent interactions. The combination of Object Z and Statecharts is called OZS notation.

Agent based software development starts with system analysis, which results in agent organizational structure, i.e., agent roles and interactions. Based on the structure, agent types are designed, and then the agents are implemented in a programming language (or languages). In paper [19], Dastani et al concentrate on defining formal semantics, using the first order logic, of agent roles, especially in an open MAS, where the roles of an agent change from time to time depending on its context. The authors defined four types of role operations an agent may have, *enact/deact* and *activate/deactivate*.

Although there is no consensus on what an agent is, autonomy is a universally agreed upon attribute of an agent. The meaning of “autonomy” is somewhat vague and open to different interpretation. Weiss et al attempted to provide formal semantics to autonomy [20]. In their paper, a formal language ASL (Autonomy Specification Language) is introduced. The role of an agent consists of a set of activities. For every activity norms and sanctions are attached to regulate the agent behaviors. ASL defines three types of norms – *permissions*, *obligations*, and *interdictions* – and two types of sanctions – *reward* and *punishment*.

An important aspect of MAS is the goals of the agents, which make up the Multiagent System. Simon and Flouret [21] define agent goals by temporal logic of actions (TLA) [22]. The focus of this approach is the goal decomposition tree (GDT). The authors proposed a tree structured goal decomposition, where the system goal is decomposed into sub-goals, and each of the sub-goals is further decomposed into yet smaller sub-sub-goals.

It has long been debated what the better way is to achieve formal semantics – declarative or procedural. The former is better for system verification (a proof system), and the later is better for implementation (code transformation). Winikoff et al take an interesting approach. They combine both semantics into a single system. In [23], they argued that both declarative and procedural are important. They mapped both declarative and procedural semantics to the goal of an agent – “Goals have two aspects: declarative, where a goal is a description  $s$  of the state of the world which is sought ( $Env \models s$ ); and procedural, where a goal is a set of procedures  $P$  which is executed (in an attempt) to achieve the goal”. A goal is defined as **Goal( $s$ ,  $P$ ,  $f$ )**, which means achieving the goal  $s$  by using the procedure  $P$ , and abort whenever  $f$  becomes true.

Agent interaction plays an important role in achieving system goals of MAS. Artikis et al use causal theory to model and validate the specification of the social laws of agents in electronic societies [24]. For an open society of agents, the internal architecture of an agent is unknown to other agents of the same society. Therefore, its behaviors can only be regulated by social constraints.

Cabac et al suggested using colored Petri net to specify agent interaction [25]. In paper [25], the authors tried to convert agent interaction protocol (AIP) of AUML into

a type of colored Petri net, which is called reference net. Petri net provides AIP with formal operational semantics.

Beliefs are an important aspect of an agent, for example, in BDI model (believed-desire-intention) [26], where believed, desire, and intention, the three mental attitudes, decide the MAS behaviors. In [27], Benerecetti and Cimatti proposed to use propositional branching-time temporal logic (CTL) [28] as the formal language to describe the beliefs of an agent. A model check system NuMAS (NuSMV for Multi-Agent Systems), which is based on from NuSMV [29], is developed for model checking the temporal logic specification.

While complete formal specification and verification have always had difficulties in dealing with real life application, Perini et al [30] advocate a mixed approach, which interleaves formal and informal specification. The approach is to “*allow for lightweight usage of formal verification techniques, that are used as services in an ‘informal’ development methodology*”. The authors suggested keeping informal visual modeling languages, but interpreting them formally. The authors tested the idea on the visual modeling language provided by Tropos agent-oriented software development methodology [31], where a Tropos model is translated into linear time temporal logic specification and then verified by the NuSMV model checker [29].

## 7 Summaries

In this paper, we presented the recent advances in agent-based software engineering. ABSE was discussed in relation to the three main steps of software development: system analysis and modeling, design and implementation, and validation. ABSE was also examined in terms of specification and verification, in the pursuit of fault free software. The purpose of this structure was to provide a collaborative visual of the where ABSE development currently stands and point out the areas that need further work.

The authors of this paper is building a multiagent based IT security system MAITS [32-34]. The system use Chemical Abstract Machine as the underlying computation and communication model and temporal logic as formal versification tool [35]. We expect more first hand experience in systematically testing against agent based software development cycles.

ABSE could be the next step forward from object-oriented programming. The basic idea of ABSE is to use agents as building blocks for the development of software systems. Agents can have the basic characteristics of autonomy, social ability, reactivity, and pro-activity, and also the more advanced characteristics of adaptability, mobility, veracity, and rationality. Agents are also adept at the three essential activities in tackling complex systems: decomposition, abstraction, and organization. The combination of these characteristics and abilities is what makes agents good candidates for the development of general software applications. ABSE also promises a simplified and enhanced approach to construct complicated software systems.

**Acknowledgments.** The research work is supported by The Completion Grant of Division of Business, Law and Information Sciences, University of Canberra and the Multidisciplinary Research Grant of University of Canberra.

## References

1. Sommerville, I., *Software Engineering*. 6th ed. 2001: Addison Wesley.
2. Hewitt, C., *Viewing Control Structures as Patterns of Passing Messages*. Artificial Intelligence, 1977. **8**(3): p. 323-364.
3. Bradshaw, J.M., *An Introduction to Software Agents*, in *Software Agents*, J.M. Bradshaw, Editor. 1997, AAAI Press/The MIT Press. p. 3-46.
4. Nwana, H.S., *Software Agents: An Overview*. Knowledge Engineering Review, 1996. **11**(3): p. 1-40.
5. Stone, P. and M.M. Veloso, *Multiagent Systems: A Survey from a Machine Learning Perspective*. Autonomous Robots, 2000. **8**(3): p. 345-383.
6. Shoham, Y., *Agent-oriented programming*. Artificial Intelligence, 1993. **60**(1): p. 51-92.
7. Wooldridge, M. and P. Ciancarini. *Agent-Oriented Software Engineering: The State of the Art*. in *The First International Workshop on agent-oriented software engineering (AOSE2000)*, LNCS 1957. 2001: Springer-Verlag.
8. Jennings, N.R., *On agent-based software engineering*. Artificial Intelligence, 2000. **117**: p. 277-296.
9. Petrie, C. *Agent-Based Software Engineering*. in *The First International Workshop on agent-oriented software engineering (AOSE2000)*, LNCS 1957. 2001: Springer-Verlag.
10. Tveit, A. *A Survey of Agent-Oriented Software Engineering*. in *the First NTNU CS/GS Conference*. 2001.
11. Steegmans, E., Weyns, D., Holvoet, T. and Berbers, Y., *A Design Process for Adaptive Behavior of Situated Agents*. AOSE2004, LNCS 3382: p. 109-125.
12. Ramos, F.F. *Methodology for Analysis and Design of Systems*. in *Proceedings of the Second IEEE Workshop on Software Technologies for Future Embedded and Ubiquitous Systems (WSTFEUS'04)*. May 2004.
13. Shehory, O., and Sturm, A. *Evaluation of modeling techniques for agent-based systems*. in *Proceedings of the Fifth International Conference on Autonomous Agents*. May 2001: ACM Press.
14. Jennings, N., and Wooldridge, M., *Software Agents*. IEEE Review, January 1996: p. 17-20.
15. Sanchis, E. *Designing new Agent Based Applications Architectures with the AGP Methodology*. in *Proceedings of the Twelfth International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*. June 2003.
16. Green, S., Hurst, L., Nangle, B., Cunningham, P., Somers, F. and Evans, R., *Software agents: A review, Technical Report*. 1997, Department of Computer Science, Trinity College: Dublin, Ireland.
17. Hall, R.J., Automated Software Engineering, 2002. **Springer 9**: p. 233-261.
18. Hilaire, V., et al. *A Formal Approach to Design and Reuse Agent and Multiagent Models*. in *5th International Workshop on agent-oriented software engineering, AOSE 2004*, LNCS 3382. 2004. New York, NY, USA: Springer.
19. Dastani, M., et al. *Enacting and Deacting Roles in Agent Programming*. in *5th International Workshop on agent-oriented software engineering, AOSE 2004*, LNCS 3382. 2004. New York, NY, USA: Springer.
20. Weiss, G., et al. *Operational Modeling of Agent Autonomy*. in *6th International Workshop on agent-oriented software engineering, AOSE 2005*, LNCS (to be published). 2005. Utrecht, The Netherlands: Springer.
21. Simon, G. and M. Flouret. *Implementing Validated Agent Behaviors with Automata based on Goal-Decomposition Trees*. in *6th International Workshop on agent-oriented software engineering, AOSE 2005*, LNCS (to be published). 2005. Utrecht, The Netherlands: Springer.

22. Lamport, L., *The temporal logic of actions*. ACM transactions on Programming Language and Systems, 1994. **16**(3): p. 872-923.
23. Winikoff, M., et al. *Declarative & Procedural Goals in Intelligent Agent Systems*. in *Proceedings of the 8th International Conference on Principles and Knowledge Representation and Reasoning (KR-02)*. 2002. Toulouse, France: Morgan Kaufmann.
24. Artikis, A., M.J. Sergot, and J. Pitt. *Specifying Electronic Societies with the Causal Calculator*. in *The 3rd International Workshop, AOSE 2002, LNCS 2585*. 2002. Bologna, Italy: Springer.
25. Cabac, L. and D. Moldt. *Formal Semantics for AUML Agent Interaction Protocol Diagrams*. in *5th International Workshop on agent-oriented software engineering, AOSE 2004, LNCS 3382*. 2004. New York, NY, USA: Springer.
26. Rao, A.S. and M.P. Georgeff. *BDI-agents: from theory to practice*. in *Proceedings of the First International Conference on Multiagent Systems (ICMAS-95)*. 1995. San Francisco, USA: The MIT Press.
27. Benerecetti, M. and A. Cimatti. *Validation of Multiagent Systems by Symbolic Model Checking*. in *The 3rd International Workshop, AOSE 2002, LNCS 2585*. 2002. Bologna, Italy: Springer.
28. Emerson, E.A., *Temporal and modal logic*, in *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics*, J.v. Leeuwen, Editor. 1990, Elsevier. p. 996-1072.
29. Cimatti, A., et al. *NuSMV 2: An OpenSource Tool for Symbolic Model Checking*. in *Computer Aided Verification, CAV 2002, LNCS 2404*. 2002. Copenhagen, Denmark: Springer.
30. Perini, A., et al. *Agent-Oriented Modeling by Interleaving Formal and Informal Specification*. in *4th International Workshop on agent-oriented software engineering, AOSE 2003, LNCS 2935*. 2003. Melbourne, Australia: Springer.
31. Bresciani, P., et al., *Tropos: An Agent-Oriented Software Development Methodology*. Autonomous Agents and Multi-Agent Systems, 2004. **8**(3): p. 203 - 236.
32. Ma, W. and D. Sharma. *A Multiple Agents Based Intrusion Detection System*. in *Ninth International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES2005)*. 2005. Melbourne, Australia: Springer-Verlag.
33. Sharma, D., W. Ma, and D. Tran. *On an IT Security Framework*. in *Ninth International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES2005)*. 2005. Melbourne, Australia: Springer-Verlag.
34. Sharma, D., Ma, W., Tran, D., Liu, S., and Anderson M. *MAITS: A Multiagent Based IT Security Approach*. Accepted for publication as a chapter in the book *Architectural Design of Multi-Agent Systems: Technologies and Techniques*.
35. Ma, W., Tran, D., Sharma D., Lin H., and Anderson M. *Concurrent Programming with Multiagents and the Chemical Abstract Machine*. Accepted for publication as a chapter in the book *Architectural Design of Multi-Agent Systems: Technologies and Techniques*.

# Personalisation of Web Search: An Agent Based Approach

G.L. Ligon, M.B. Balachandran, and D. Sharma

School of Information Sciences and Engineering  
Division of Business, Law and Information Sciences  
University of Canberra  
ACT 2601, Australia  
g.ligon@student.canberra.edu.au,  
{bala.balachandran, dharmendra.sharma}@canberra.edu.au

**Abstract.** Web based information searching is increasingly becoming an essential day-to-day activity. Currently search engines are the only dominant tools used for searching information over the Internet. Most search engines are powerful tools in finding information and cataloguing them, however they are not much efficient in meeting the personalised information needs of users. This paper presents research work done for personalising web search practice by customizing search engine results via an agent called Personalised Search Agent (PSA). In this paper, we will describe the architecture, design and implementation of a PSA system. Finally, we show several experimental results obtained by comparing the performance of the PSA against traditional search engines.

## 1 Introduction

Internet is constantly growing bigger and bigger. As a result, retrieval of *relevant information* from the Internet becomes a time-consuming and laborious task. since this activity involves the analysis and separation of interesting pages from a great set of candidate pages (Convey, 1992). Today's search engines are the most widely used tools for searching and retrieving information from the Internet. Users provide a set of key words to a search engine which in turn returns a set of *links to WebPages* that are related to those key words. The simpler the keyword is the more general will be the results returned by the search engines. This causes users to spend considerable amount of time in searching and retrieving quality information. In terms of today's business world, this means waste of human resources, computing resource, energy and time – making it a pullback force in productivity. Recently, web search personalisation has been discussed from various viewpoints (Cooke, 1999; Amandi, 2000; Ligon et al, 2005).

In this paper, we present an agent-based approach which attempts to personalise the information retrieval process from the Internet. Our definition of a software agent is a computer program that acts on behalf of its user to accomplish a specified task. Issues involved in designing and developing agent systems have been described by many researchers (Padgham and Winikoff, 2004; Luck, Ashri and d'Inverno, 2004). We utilise a Personalised Search Agent (PSA) which acts between users and search

engines, and applies sophisticated techniques to mine and exploit relevant and personalised information for users. Some key characteristics considered in the design of the agent include utilisation concept categories, learning from user actions and using meta-search engines. The overall architecture of the PSA system is based on a multi-layered client/server architecture as described later in the paper. The client handles personalising information regarding one user whereas the server effectively combines the personalising information of all the clients (i.e. its users) to generate a global profile. Both client and server apply the category concept where user selected URLs are mapped against categories. The PSA learns the *user relevant URLs* both by requesting explicit feedback and by implicitly capturing user actions (for instance the active time spent by the user on a URL). The PSA, is a novel architecture and contributes to the domain of knowledge *web information searching*, by delivering new ideas such as active time based user relevancy calculations and automatic generation of sensible search keyword combinations. Moreover, the PSA implementation is based on a multi-layer agent architecture which has high potential for future extensions. The utilisation of data mining techniques which employ case-based reasoning makes the PSA a more responsive, more accurate and more effective tool for personalised information searching.

## 2 The PSA Framework

The general framework of the Personalised Search Agent (PSA) system is shown in Figure 1.

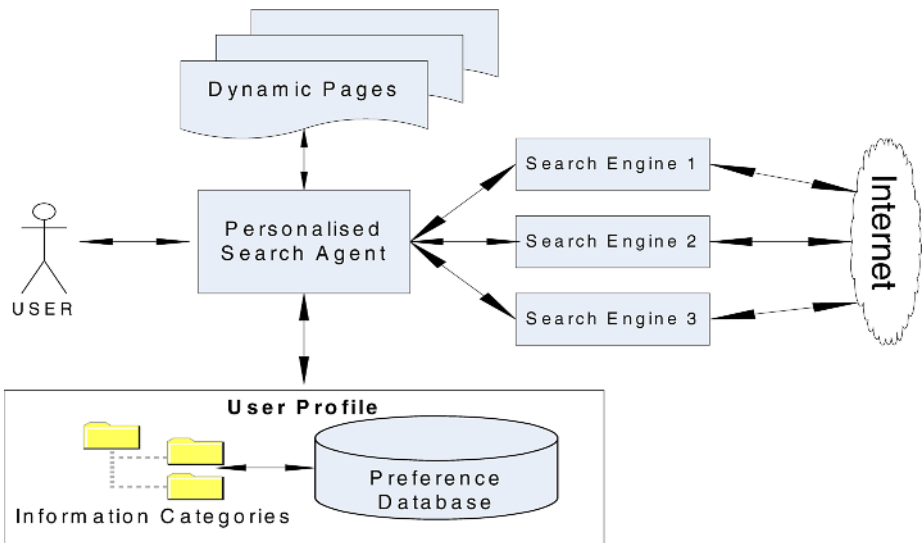


Fig. 1. A general framework of the PSA

## 2.1 PSA Components

The most important modules are described below:

Information categories:- The concept of information categories is a technique used to organise information in a structured hierarchical order (Giuseppe, 1999). In the PSA these categories are useful in two ways. Firstly, the categories help to enhance the user-entered keywords, by effectively combining the keywords with category values. The second use of categories in PSA is that categories can be used to remember the user preference rather than asking the user to recall the preference.

Preference databases :- In PSA, preference databases are the actual personalisation database. All the search details are stored in this database. A successful search will always have at least the following details:

- User keyword
- Preferred category value
- Generated keyword
- URLs (links) opened
- Search engines from which each link was retrieved.

Again a successful search is determined by positive user feedback. The mapping of the keywords with the category is done in the preference database and stored with other relevant information.

User profile:- The term user profile is more a conceptual or abstract term used in PSA. User profile is represented by the effective combination of the categories (General Subject Directories, 2005) and a preference database. After a certain period of use the database and the categories will become quite uniquely customised to the user. The user profile form is a data entry form designed to capture an initial *Local Category* structure with a minimal set of category items and values.

## 2.2 PSA Functionalities

The key functionalities are briefly described below:

Generating keywords: - During conventional web search, we normally shuffle the keywords by changing the order or replacing one keyword with another one to achieve the desired result (Lawrence,1999). The keyword generation feature of the PSA automates this process. Basically, the PSA accepts keyword/s from the user along with a category name. The PSA then makes different combinations of 'keywords' and 'category values' based on a keyword-generating algorithm. The actual search (with the search engines) is done using all the 'keyword sets' generated and presented to the user in a predefined order until the user receives positive feedback.

Managing the Preference Database: - Preference Database contains the base data for all the ranking calculations carried out both on the client and on the server side. They normally contain all the possible attributes of a successful search such as the keyword pattern, the category item, the category value and the search engine ranking for that search, as well as the name and location of the associated dynamic page. Almost all the non-derivable information of a successful search is stored in the preference database.



### 2.3 Rankings and Algorithms

To determine the relevant information from the captured user details several ranking approaches have been used (Haveliwala, 2002). In the PSA development, we have chosen the following algorithms:

The Keyword Generating Algorithm : - It is an effort to simply assist the user in generating different keyword combinations based on a simple algorithm explained below. By doing this the PSA not only eases user effort but also substantially saves user time. If 'A' be the category of preference,  $B_1$  to  $B_n$  be the words in the user keyword combination then the sequence would be:

“A + (B<sub>1</sub>...B<sub>n</sub>)”  
 A + “(B<sub>1</sub>...B<sub>n</sub>)”  
 A + B<sub>1</sub> + “(B<sub>2</sub>...B<sub>n</sub>)”, .....

Ideally, this sequence will stop either when the user gives positive feedback or when all the probable combinations of words are exhausted. However, in actual practice, proper limits are applied in the number of probable combinations.

The Link Ranking Algorithm: - Link ranking is a new approach to find personalisation at its grass roots level. Basically, link ranking works by calculating the active time the user spends on each URL in a successful search result. A successful result is defined by user feedback. User actions will be continuously monitored once the results are submitted but only considered for processing, if the user gives positive feedback on the result. The link ranking algorithm can simply be explained as follows:

Let 't' be the active time spent by the user on a link, 's' be the search engine ranking for that user and 'n' be the number of times the link (or URL) is viewed. Then

$$\text{Link Rank} = (t + s + n)$$

The Search Engine Ranking :- The search engine ranking is given to each dynamic page both on the client side and on the server side. Ranking is calculated by grouping all the URLs in the page by the search engines (which returned the link) and then taking the sum of the link rankings of the URLs in each group. If two or more search engine have the same ranking, then their relative ranking is recalculated by counting the number of URLs returned by each, and furthermore with the LIFO algorithm.

The Life Saving Rank : - Let the link ranking for the URL be LR and T be the total time elapsed after storing that URL to the dynamic page, then the life saving rank is = T/LR.

## 3 PSA Development and Implementation

In this section, we describe the implementation of the PSA model in terms of its architecture, design and implementation tools. Using a client-server architecture was found to be a simple and effective solution that provides a multi-user support (Edelstein, 1994). In our client-server model, both client and server are online web

applications / web services. On the client side the preference database and the dynamic pages would be specific to a particular user whereas on the server side they are more general. The server side profile database and the categories are the union of all the clients (users). Each successful search that a client makes therefore will have an impact not only in updating the local details but also in affecting the server details. This results in the server acting as a personalised information source or a global database of personalised structured information after a period of use. All the clients request the information from the server; the server then checks its global profile for relevant information and in the absence of any, conducts a real search. Once the server has enough links in its global profile, the search request from a client can be returned almost instantaneously. A client-server based multi-layer architecture of the PSA is illustrated in Figure 2.

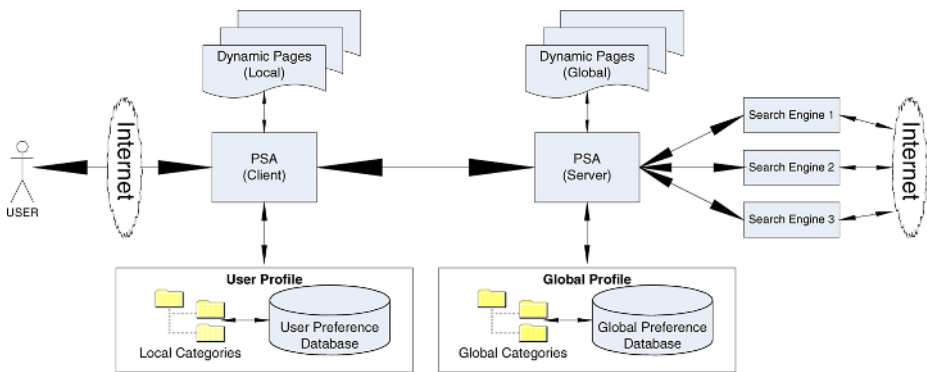


Fig. 2. Multi-layer architecture of the PSA

The key components are described below:

**Global profile:** - Like the user profile, the global profile is an effective combination of the global preference database and the global categories and is rather an abstract term to represent the ‘generalised personalisation’ of all the clients in the PSA. The global profile can be useful in predicting general trends and in determining search engine efficiency with respect to the category and the types of user.

**Dynamic pages:** - Dynamic pages are the result pages created by the PSA based on the cached relevant results from the database. On the client side they are real WebPages whereas, on the server side they are a package of results generated as per the client request and passed on to the client after creation. The client side dynamic pages can be a well-structured report having a small screen shot of each URL, the related category and other retrieval details.

### 3.1 PSA Implementation Tools

A prototype system based on the concepts and techniques described in the previous sections was developed and implemented as a personalised search agent. The

prototype system has been implemented on Windows platforms using the .NET technologies. The programming languages used were ASP.NET, ADO.NET and C#.

## 4 Experimental Result

In order to verify the validity of the proposed method, we conducted some experiments. A series of tests was carried out to investigate the effectiveness of the web search provided by the PSA. We used three factors, namely efficiency, relevancy and responsiveness to compare our results with those produced by traditional search engines.

**Relevancy:** - relevancy is highly personal and it is therefore difficult to define standards for its measurement. The best method is a user survey. However, in this research, relevancy was measured based on a set of assumptions as follows:

The search domain is set as the University Of Canberra website, [www.canberra.edu.au](http://www.canberra.edu.au).

Relevancy is measured considering the user as a research student and by pre defining some of the preferences of such a user, which are listed in Figure 3.

Keyword (pre condition)	Preference assumed for a Research Student (post condition)
Job	Jobs in relation to research field
Fee	Fees for research students
Thesis	Guidelines for writing a thesis.
Scholarships	Scholarships for international research students.

**Fig. 3.** Assumed preferences against test keywords

The candidate systems selected for comparison are [www.google.com](http://www.google.com), [www.yahoo.com](http://www.yahoo.com) and [www.altavista.com](http://www.altavista.com). How the results are relevant to the assumption is judged from a research student's perspective. However, the results can be verified with a wider assumption and with greater variety in groups of users. This is just a matter of time. A good database entry has been made for the selected search keywords, as real use of the system comes with time and how long it has been used; hence how much data is available for that user.

**Efficiency:** - efficiency unveils how effectively the system can remember user preferences rather than asking users to recall their preferences each time. Apparently, it tests how the system learns by capturing user input and feedback and uses that information to increase the efficiency of current search practice.

**Responsiveness:** - responsiveness is not how fast the results are displayed to the user, but how fast the user receives relevant results.

Bench marking of the test results for the PSA and candidate search engines was done based on the result obtained by carrying out the above test criteria. Firstly, the average relevancy, termed the Relevancy Index, was for each candidate system as depicted in Figure 4.

Keyword		PSA	Google	AltaVista	Yahoo
Relevancy Averages	Job	2.80	2.00	0.6	1.40
	Fee	3.00	1.75	1.5	1.25
	Thesis	4.00	0.00	3.5	3.00
	Scholarship	3.00	1.80	1.5	2.20
<b>Total</b>		<b>12.8</b>	<b>5.5</b>	<b>7.1</b>	<b>7.85</b>
<b>Average (Relevancy index)</b>		<b>3.2</b>	<b>1.4</b>	<b>1.8</b>	<b>2</b>

Fig. 4. Relevancy index of the PSA and the candidate search engines from a research student perspective

The results data tables under section two of this chapter also contain the time taken to display the search results. Figure 5 shows the compilation of the time averages for the PSA and the candidate search engines.

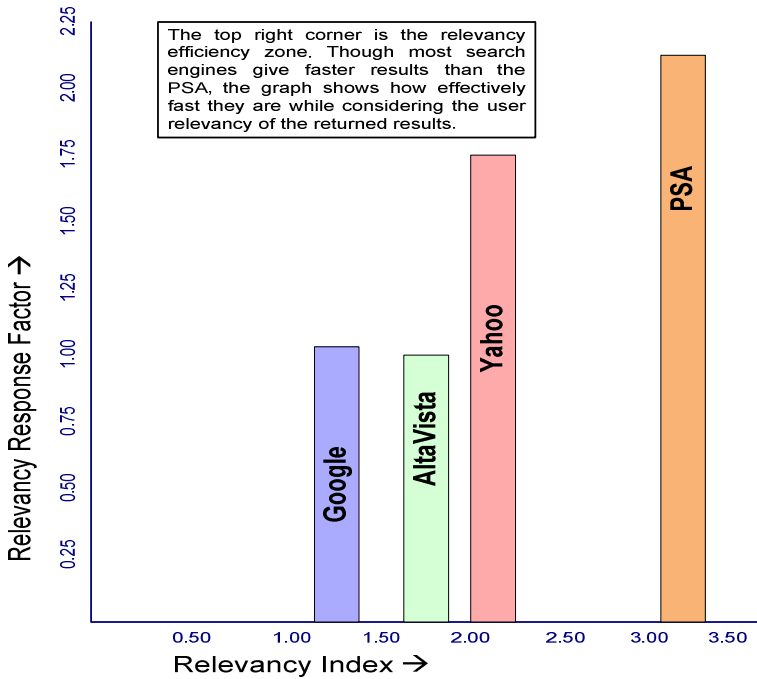
Keyword		PSA	Google	AltaVista	Yahoo
Time Averages	Job	1.00	0.55	1.00	0.31
	Fee	1.50	0.25	0.28	0.24
	Thesis	1.00	0.18	0.74	0.19
	Scholarship	0.50	0.25	0.85	0.07
<b>Total (in Seconds)</b>		<b>4.00</b>	<b>1.23</b>	<b>2.87</b>	<b>0.81</b>
<b>Average</b>		<b>1</b>	<b>0.31</b>	<b>0.72</b>	<b>0.20</b>

Fig. 5. Averages of the time taken by the PSA and the candidate search engines under various search criteria

A relevancy response factor (time taken for delivering the relevant results not just the results) is just an indication of relevant results and can be taken by processing the data in compilation figures 4 and 5 as follows.

$$\text{Relevancy Response Factor} = \text{Relevancy Index} - \text{Average Response Time}$$

It is important to note that the above line is not a mathematically valid equation as it contains an abstract variable of the relevancy index. Instead this can be considered as an effort to test the relevancy in a more accountable way. Applying the above equation, the relevancy response factor of the PSA and the candidate search engines are calculated. Figure 6 illustrates the comparisons graphically.



**Fig. 6.** Graph depicting the *relevancy of the results* and the *response time* for delivering relevant results comparing the PSA with other candidate search engines

## 5 Summary and Conclusions

We proposed an agent-based approach for personalising web search. The key characteristics of the proposed method are as follows:

- Uses the category or directory approach of organising similar information in a hierarchy
- Learns from experience and user feedback
- Uses multiple search engines (meta search approach)
- Creates dynamic pages containing the links (URLs) that the user has visited during a search
- Ranks the links based on the time spent on and other parameters
- Ranks the search engines based on the successful links returned to the user

The PSA is capable in capturing feedback from the user, either by explicitly requesting feedback from the user or by learning the user's habits. In the former method, the PSA requests feedback on the relevancy of a link presented to the user. However, the PSA also receives many implicit inputs, for example, active time, when the user gives feedback on an explicit request from the user. These inputs are used to make various calculations regarding the relevancy of the link. Moreover, the PSA effectively reduces the user inputs through features such as keyword generating

algorithms. It can also automatically switch to search mode with different combinations if the user is unhappy with the presented results.

The effectiveness of the proposed method was verified by extensive testing. The test results have shown that the PSA is successful in delivering personalised results to the user. With the client-server architecture, the PSA server, is potentially a good resource for a personalised information database. Each and every link in the server database is reviewed and accepted by at least one user, which confirms that it is relevant to that particular category. This is particularly important, considering that, information search services like the MSN Search are more likely to present human-powered listings (How search engines work, 2002), whereas the PSA has all its cached URLs reviewed and ranked during the search process itself. Furthermore, the experimental result shows the high flexibility of the proposed method.

## References

- Amandi, D.G.a.A.(2000), *PersonalSearcher: An Intelligent Agent for Searching Web Pages*.  
 Convey, J. (1992). *Online Information Retrieval*. London: Library Association Publishing.  
 Cooke, A. (1999). *A guide to finding quality information on the Internet*. London: Library Association Publishing.  
 Giuseppe,A. & Umberto, S.(1999). *User profile modelling and applications to digital libraries*. Italy: C.N.R.  
 Edelstein, H(1994). Unraveling Client/Server Architecture. *DBMS*, Vol.34, p.7.  
 General Subject Directories: Table of Features. (2005). Retrieved 2 January 2005, from <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SubjDirectories.html#Guides>  
 Haveliwala, T.H. (2002). *Topic-sensitive PageRank: Proceedings of the 11th International World Wide Web Conference*. USA: Hawaii.  
 How search engines work: hybrid search engines. (2002). Retrieved 12 April, 2005, from <http://searchenginewatch.com/webmasters/article.php/2168031>  
 Ligon GL., Balachandran, M.B. and Sharma, D. (2005). Personalised search on electronic information (KES'2005), Melbourne, Australia, Part IV:pp 35-41.  
 Luck, M, Ashri, R and d'Inverno, M. (2004). *Agent-Based Software Development*, Artech House, Inc. Norwood, MA.  
 Padgham, L and Winikoff, M. (2004). *Developing Intelligent Agent Systems, A Practical Guide*. John Wiley & Sons.

# Autonomy and Intelligence – Opportunistic Service Delivery in Mobile Computing

Jiangyan Chen<sup>1</sup>, M.J. O'Grady<sup>2</sup>, and G.M.P. O'Hare<sup>2</sup>

<sup>1</sup> School of Computer Science & Informatics, University College Dublin (UCD), Belfield, Dublin 4, Ireland

jiangyan.chen@ucd.ie

<sup>2</sup> Adaptive Information Cluster (AIC), School of Computer of Computer Science & Informatics, University College Dublin (UCD), Belfield, Dublin 4, Ireland

{michael.j.ogrady, gregory.ohare}@ucd.ie

**Abstract.** Significant opportunities exist for mobile network operators to enhance and refine those services they offer their subscribers. Developments in mobile devices have rendered it feasible to deploy solutions that incorporate an intelligent agent component. The use of agents is particularly apt in those situations that facilitate what may be termed opportunistic services provision. Such services are suited to certain domains that facilitate unobtrusive observance of the subscriber. As an illustration of the issues involved, a mobile blogging application is described.

## 1 Introduction

Mobile computing is now both a usage paradigm and research discipline in its own right. However, a number of different strategies may be adopted for deploying applications and services for mobile users. For example, ubiquitous computing [1] envisages a world saturated with embedded computational technology that can be accessed in an anywhere, anytime basis. A more conventional view of how mobile users might access services would be that of via a so called smartphone. In either case however, it is acknowledged that the use of intuitive user interfaces and interaction modalities are essential, resulting in the Ambient Intelligence (AmI) [2] initiative. Quite how the intelligence requirement is fulfilled remains to be seen.

As developments on mobile hardware and associated software mature, the possibility of deploying applications and services that incorporate Artificial Intelligence (AI) techniques becomes more realistic. One promising approach for achieving this concerns the modeling and implementation of individual services as Multi-Agent Systems (MAS). As well as delivering an intelligence quotient, traditional characteristics of agents make them particularly suited to modeling the inherently dynamic context of the average mobile user. In this paper, these issues are explored further and illustrated through the description of mobile blogging (mblogging) application for tourists. In brief: blogging is the maintenance of an online diary in which people record their observations on various subjects. Though normally a task that occurs at a workstation, it can also take place while in a mobile context.

This paper is structured as follows: Section 2 presents some ongoing research concerning agent deployment on mobile devices as well as some ongoing developments in the mobile blogging area. In Section 3, various aspects pertaining to the delivery of services to mobile users using intelligent agents are examined. To demonstrate the feasibility of the approach, a brief overview of a mobile blogging application is presented in Section 4 after which the paper is concluded.

## 2 Related Research

A number of researchers have recognized the potential offered by the deployment of intelligent agents on mobile devices. Given the ubiquity of the Java language, most have implemented their systems such that they are J2ME compliant. Practically all are extensions of well-documented workstation implementations. For example, 3APL-M [3] is based on the Artificial Autonomous Agents Programming Language (3APL) [4]. The Light Extensible Agent Platform (LEAP) [5] has evolved from the JADE [6] platform. Likewise, microFIPA-OS [7] has its foundations in the open source platform FIPA-OS [8]. KSACI [9], an extension of SACI [10], is essentially an inter-agent communications infrastructure. In contrast, some agent platforms support a more sophisticated reasoning ability and conform to the BDI model [11]. JACK [12] and, more recently Agent Factory [13], being just two examples.

Blogging is a popular activity on the internet. However, blogging can also have a spatial-temporal component when undertaken by mobile users, and has thus attracted the interest of some researchers. GTWeb [14] investigates issues relating to the presentation of geotemporal data, and in particular, the GPS and digital photography issues necessary to create and maintain trip diaries. GeoNotes [15] is primarily concerned with the issue of information overload, as users proceed to annotate space in a massive scale, and identifies some content-based filtering techniques to address this. Urban Tapestries [16] focuses both on accessing and publishing location-specific multimedia. BRAINS [17] is concerned with the provision of ubiquitous access to personal blogs on the WWW. From a mobile device perspective, Beale considers the case of a smart phone for social interaction, and illustrates his case with a description of SmartBlog [18].

## 3 Opportunistic Service Provision

Practically all services and applications available to the mobile computing community are user driven. The user must initiate the service, formulate their request and await a response, much the same as in a traditional workstation setting. However, a more opportunistic approach is feasible in certain domains where a model of both the user and their environment is available. A classic example is that of route guidance where drivers are prompted to follow certain routes as they make their way to their destinations. In a more traditional mobile computing scenario, it can be easily envisaged how certain activities are particularly suited to opportunistic service provision. Tourism is one example which will be discussed in Section 4. Education is another where, for example, as a student explores an archaeological site perhaps, their



attention is directed to certain artifacts and information that is pertinent to both their personal situation and their location is provided. Another example concerns mobile commerce where a shopper is directed to shops that stock items on their shopping list.

Before services can be opportunistically delivered to mobile users, a number of prerequisites must be established.

1. **User & Domain Model:** A model of the user is essential if a service is to be personalized to meet their requirements. Essential aspects of a user model include age, sex and language amongst others. However, this core model must be augmented with sub-models pertinent to certain application domains. In the case of tourists, a model of their cultural interests would be essential.
2. **Environmental Model:** A model of the environment in which the mobile user is operating is essential. However, the nature of the model will reflect the application domain in question. In the case of a tourist, details of various tourist attractions must be included in the model.
3. **Content Model:** Any content will be domain specific and may well be dynamic in nature. In a mobile commerce scenario, the content model would contain details of products maintained by various shops in the environment. In the tourist case, multimedia presentations concerning the various tourist attractions could be stored.

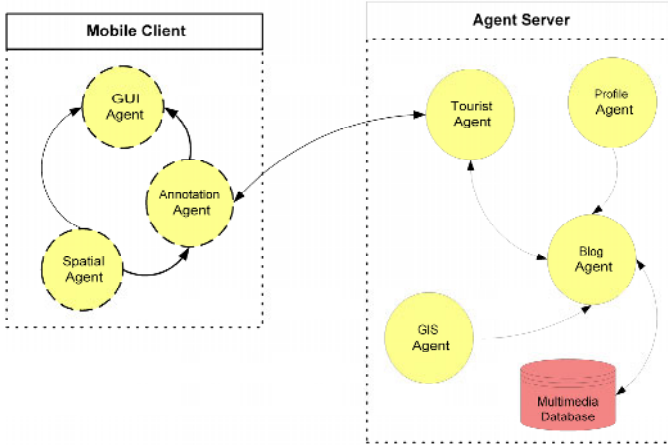
Having established the prerequisites for opportunistic service provision, the distinguishing attributes of a necessary software solution begin to crystallize. Given the mobility of the user, a degree of autonomy is essential. However, it is in reconciling the various models such that the user's needs are addressed that poses the greatest challenge. It is here a capability for intelligent reasoning is essential. Indeed, a degree of proactiveness is also necessary as opportunities for presenting information to the user must be identified quickly. Thus the aptness of the intelligent agent paradigm for opportunistic service provision can be seen. Their inherent characteristics of autonomy, proactiveness and reactivity are well known and, depending on the implementation in question, a sophisticated reasoning facility may be reasonably anticipated. Such characteristics offer an excellent opportunity to augment the traditional blogging scenario, as illustrated by the systems outlined in Section 2, with an additional level of dynamism, ultimately enhancing the user experience. In the next section, a prototypical mobile blogging application modeled on the intelligent agent paradigm is described.

## 4 A Mobile Blogging Scenario

As a tourist explores some area of cultural significance, they may wish to leave a record of their observations so that others can access them. Alternatively, they may wish to know what others think of the attraction in question. Blogging, a popular activity on the internet, offers a mechanism around which a solution may be derived.

An architecture for a mobile blogging service is presented in Fig. 1. Given the ubiquity of the mobile phone, it is realistic to assume that a tourist will possess one, and that they would envisage using it to access the service. A central repository is essential to maintain all blog entries, thus a wireless data connection to a fixed server is also essential. A prudent strategy is to disperse the agents between the tourist's host

device and the server. An advantage of this approach is that those agents that need to autonomously monitor the tourist's behavior may reside on their device while those that require a strong reasoning component should operate on the fixed server. A brief description of each agent is now provided.



**Fig. 1.** Agent Architecture for mobile blogging

- **GUI Agent:** This agent monitors all tourists' interaction with their device and controls the user interface.
- **Spatial Agent:** Given that the tourist's geolocation or spatial context is of fundamental importance, the Spatial Agent must autonomously capture and interpret it, before proceeding to inform the other agents.
- **Annotation Agent:** This agent handles all mblogging-related functionality on the device. It collaborates with the GUI Agent to enable it to capture an mblog entry (fig. 2a). The tourist may select what type of mblog entry they wish to create. Simple text, photo, audio and video combinations are theoretically possible, though phone capability and tourist preference will determine what constitutes the final mblog entry. The Spatial Agent is interrogated to ascertain the tourist's spatial context. After tagging some necessary metadata, the mblog entry is dispatched to the Blog Agent on the server. The second task that the Annotation Agent provides concerns annotation retrieval (fig 2b). The tourist may select some initial criteria for determining which mblog entries they wish to view. Type of annotation, position (relative to their current location) and date are three criteria they can select. Again, the request is sent to the Blog Agent which returns those entries that conforms to the selected criteria. The GUI Agent renders them for the tourist to browse.
- **Tourist Agent:** This is essentially the tourist's proxy agent on the server. All tourists are assigned a Tourist Agent on commencing a session. As well as making system design intuitive, such an approach also increases the net scalability of the system.

- Profile Agent: This agent manages user's profiles and supplies pertinent aspects of the profile to other agents on request.
- GIS Agent: A model of the environment is maintained by the GIS Agent. Sub-models are made available to other agents on request.



Fig. 2. How the tourist (a) records and (b) retrieves mblog entries

- Blog Agent: An initial task of this agent is to classify and store mblog entries in the database, in response to incoming upload requests from the Annotation Agent. Before being stored in the database, the entry must be semantically indexed. The appropriate tags are identified by querying the Profile Agent for pertinent aspects of the tourist profile, language being just one obvious example. Likewise, the GIS Agent is queried in an attempt to reconcile the tourist's physical geographical position with the nearest landmark. Finally, a complete record is constructed for the mblog entry in the multimedia database. This represents an augmentation of the core system content and may be viewed as a non-censored content overlay. A second task undertaken by the Blog Agent concerns the extraction of entries that conform to the tourist's request, and to dispatch them to the Annotation Agent on the mobile device. Again it must negotiate with the Profile and GIS Agents to identify (using appropriate profile and GIS tags) those entries of most importance.

From an implementation perspective, the client has been realized using the J2ME wireless toolkit. To enable communications between J2ME and J2SE environments,

an object transformation process has been implemented. A number of Java Specification Requests (JSRs) must be supported by any mobile phone aspiring to support this mblogging service. Bluetooth support (JSR 82) is necessary for accessing a GPS device. The Mobile Multimedia API (JSR 135) is necessary for audio and video rendering. Advanced camera operation may require JSR 234. Finally, if data must be cached on the device, file system access must be available through JSR 75.

Agent Factory (AF) [19], including its mobile component [13], has been employed for implementing the constituent agents. All AF agents conform to the BDI model as well as encapsulating those characteristics normally associated with agents. Autonomy and reactivity are exhibited by those agents residing on the mobile device. More sophisticated behavior can be realized on the network where agents collaborate (using a FIPA compliant ACL) and maintain sophisticated mental models, implemented as *beliefs* and *commitment rules*. Essential behaviors are modeled as *desires* and *intentions* respectively. A distinguishing characteristic of AF concerns its support for *Agent Cloning*. This is essential for load balancing so when an agent's workload passes a certain threshold, the agent is automatically cloned. In this way, acceptable performance can be maintained.

## 5 Future Work and Conclusion

Research is ongoing towards realizing a fully functional prototype after which it is planned to conduct extensive evaluations. However, an immediate priority is to utilize the proactive nature of the deployed agents to opportunistically draw the tourist's attention to mblog entries that may be of interest to them. A second priority is to increase the adaptivity exhibited considerably, particularly by increasing the kind of phones the system could be deployed on. A third aspect that needs further consideration concerns the refinement of the content filters, given that popular tourist attractions may attract considerable comment.

Finally, the vision of enabling people to dynamically record their observations in a multimedia fashion is one that researchers have sought to realize in recent years. Given the momentum behind the blogging phenomenon, it may well be that this concept will provide the scaffolding by which this vision is ultimately realized.

## Acknowledgements

This material is based upon works supported by the Science Foundation Ireland under Grant No. 03/IN.3/1361.

## References

1. Vasilakos, A., Pedrycz, W., Ambient Intelligence, Wireless Networking, Ubiquitous Computing, Artec House, 2006.
2. Aarts, E, Marzano, S. (editors), The New Everyday: Views on Ambient Intelligence, 010 Publishers, Rotterdam, The Netherlands, 2003.
3. 3APL-M: Platform for Lightweight Deliberative Agents: <http://www.cs.uu.nl/3apl-m/>

4. Birna van Riemsdijk, M.D., Dignum, F., Meyer, J.J., A Programming Language for Cognitive Agents: Goal Directed 3APL. Proceedings of the First Workshop on Programming Multiagent Systems: Languages, frameworks, techniques, and tools (ProMAS), Melbourne, 2003.
5. Berger, M., Rusitschka, S., Toropov, D., Watzke, M., Schichte, M., Porting Distributed Agent-Middleware to Small Mobile Devices: Proceedings of the Workshop on Ubiquitous Agents on embedded, wearable, and mobile devices held in conjunction with the joint conference on Autonomous Agents and Multi-Agent Systems (AAMAS), Bologna, 2002.
6. Bellifemine, F., Rimassa, G., Poggi, A., JADE - A FIPA compliant Agent Framework, Proceedings of the 4th International Conference and Exhibition on the Practical Application of Intelligent Agents and Multi-Agents, London, 1999.
7. Tarkoma, S., Laukkanen, M., Supporting Software Agents on Small Devices, Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-2002), Bologna, Italy, July 2002.
8. Foundation for Intelligent Physical Agents (FIPA), <http://www.fipa.org>.
9. R. Albuquerque, J. Hubner, G. Eliano de Paula, J. Sichman, and G. Ramalho, KSACI: A Handheld Device Infrastructure for Agents Communication. In Proceedings of the ATAL conference, Seattle, USA, 2001.
10. SACI – Simple Agent Communications Infrastructure - <http://www.lti.pcs.usp.br/saci/>.
11. Rao, A.S., Georgeff, M.P., Modelling Rational Agents within a BDI Architecture. In: Principles of Knowledge Representation. & Reasoning, San Mateo, CA. 1991.
12. JACK - The Agent Oriented Software Group, <http://www.agent-software.com>.
13. Muldoon, C., O Hare, G.M.P., Collier, R.W., O Grady, M.J., Agent Factory Micro Edition: A Framework for Ambient Applications, Proceedings of: Intelligent Agents in Computing Systems, Reading, UK, May, 2006.
14. Spinellis, D. D., Position-Annotated Photographs: A Geotemporal Web, Pervasive Computing, 2 (2), (2003), 72-79.
15. Espinoza, F., Persson, P., Sandin, A., Nystrom, H., Cacciatore, E., Bylund, M., GeoNotes: Social and Navigational Aspects of Location-Based Information Systems, Proceedings of the 3rd international conference on Ubiquitous Computing, Atlanta, Georgia, USA, 2001.
16. Lane, G. Urban Tapestries: Wireless networking, public authoring and social knowledge, personal & Ubiquitous Computing, 7 (2003), 169-175.
17. Leu, J-S, Chi, Y-P., Shih, W-K., Design and implementation of Blog Rendering and Accessing INSTANTLY system (BRAINS), Journal of Network and Computer Applications, 2006 (in press).
18. Beale, R., Supporting Social Interaction with Smart Phones, IEEE Pervasive Computing, 4 (2) (2005), 35-41.
19. O'Hare G.M.P., Agent Factory: An Environment for the Fabrication of Multi-Agent Systems, in Foundations of Distributed Artificial Intelligence (G.M.P. O'Hare and N. Jennings eds) pp. 449-484, John Wiley and Sons, Inc., 1996.

# Acoustic Parameter Extraction from Occupied Rooms Utilizing Blind Source Separation

Yonggang Zhang<sup>1</sup>, Jonathon A. Chambers<sup>1</sup>, Paul Kendrick<sup>2</sup>, Trevor J. Cox<sup>2</sup>,  
and Francis F. Li<sup>3</sup>

<sup>1</sup> The Centre of Digital Signal Processing, Cardiff School of Engineering,  
Cardiff University, Cardiff CF24 0YF, UK

(zhangy15, chambersJ)@cf.ac.uk

<sup>2</sup> School of Acoustics and Electronic Engineering  
University of Salford, Salford M5 4WT, UK

<sup>3</sup> Department of Computing and Mathematics  
Manchester Metropolitan University, Manchester M1 5GD, UK

**Abstract.** Room acoustic parameters such as reverberation time (RT) can be extracted from passively received speech signals by some ‘blind’ methods, which mitigates the need for good controlled excitation signals or prior information of the room geometry. However, noise will degrade such methods greatly. In this paper a new framework is proposed to extend these methods for room parameter extraction from noise-free cases to more realistic noise environment, such as occupied rooms, where noises are generated by occupants. In this proposed framework, blind source separation (BSS) is combined with an adaptive noise canceller (ANC) to remove the noise from the passively received reverberant speech signal. Room acoustic parameters can then be extracted from the output of the ANC with existing ‘blind’ methods. As a demonstration we will utilize this framework combined with a maximum-likelihood (ML) based method to estimate the RT of a simulated occupied room. Simulation results show that the proposed framework provides a good estimate of the RT in such a simulated occupied room.

## 1 Introduction

The performance of an acoustic environment can be formulated by a set of acoustic parameters. For example, room RT is an important parameter which is defined as the time taken by a sound to decay 60dB below its initial level after it has been switched off [1]. Many methods have been proposed to estimate the RT [1][2][3]; ‘blind’ methods which utilize the passively received speech signals are particularly attractive in certain environments as good controlled excitation signals are unnecessary [4][5]. In [4] an artificial neural network (ANN) is trained to extract the RT from passively received speech utterances. In [5], an exponentially damped Gaussian white noise model is used to describe the reverberation tail of the received speech signal, and an ML estimation method is then performed on segments of the speech signal to extract the RT time. As

shown by the authors, both of these algorithms provide reliable RT estimates in a noise free environment. To estimate the RT in noisy environments, such as occupied rooms, the results of these methods will be contaminated or severely biased. Therefore both methods are limited by the noise level and unsuitable for occupied rooms.

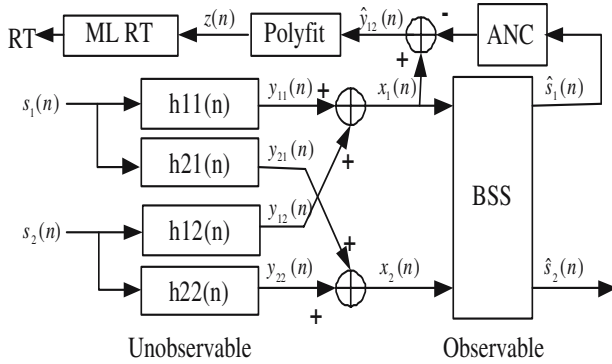


Fig. 1. Proposed blind RT estimation framework for occupied rooms

To make the RT estimation methods more robust and accurate, an intuitive way is to remove the unknown noise signal from the received speech signal as much as possible before the RT estimation. A powerful tool for extracting some noise interference signal from a mixture of signals is the convolutive BSS method [6]. Naturally, given two spatially distinct observations, BSS can attempt to separate the mixed signals to yield two independent signals. One of these two signals mainly consists of the excitation speech signal plus residue of noise and the other signal contains mostly of the noise signal. Using this estimated noise signal as a reference signal, the noise contained in the received speech signal can then be removed by an ANC. Our new framework is motivated by BSS and ANC. Different stages of this framework in an occupied room are shown in Fig.1. The signal  $s_1(n)$ , which is assumed to be the noise signal in this work, is independent of the excitation speech signal  $s_2(n)$ . The passively received signals  $x_1(n)$  and  $x_2(n)$  are modelled as convolutive mixtures of  $s_1(n)$  and  $s_2(n)$ . The room impulse response  $h_{ji}(n)$  is the impulse response from source  $i$  to microphone  $j$ . BSS is used firstly to obtain the estimated excitation speech signal  $\hat{s}_2(n)$  and the estimated noise signal  $\hat{s}_1(n)$ . The estimated noise signal  $\hat{s}_1(n)$  then serves as the reference signal for the ANC to remove the noise component from  $x_1(n)$ . The output of the ANC  $\hat{y}_{12}(n)$  is an estimation of the noise free reverberant speech signal  $y_{12}(n)$ . As compared with  $x_1(n)$ , it **crucially** retains the reverberant structure of the speech signal and has a low level of noise, therefore it is more suitable to estimate the RT of the occupied room. As a demonstration, the ML estimation method is performed to extract the RT with the output signal of the ANC. As shown by our later simulations, the RT extracted by this framework

in a simulated occupied room is reasonable and similar to the results obtained in a noise free case.

The following section introduces the BSS process. The ANC is described in Section 3. Section 4 describes the ML method. Simulation results are given in Section 5. Section 6 summarizes the paper.

## 2 Blind Source Separation

As shown by Fig.1, the goal of BSS is to extract the estimated noise signal  $\hat{s}_1(n)$  from received mixture signals  $x_1(n)$  and  $x_2(n)$ . If we assume that the room environment is time invariant, the received mixtures  $x_1(n)$  and  $x_2(n)$  can be modeled as weighted sums of convolutions of the source signals  $s_1(n)$  and  $s_2(n)$ . Assume that  $N$  sources are recorded by  $M$  microphones (here  $M=N=2$ ) the equation that describes this convolved mixing process is:

$$x_j(n) = \sum_{i=1}^N \sum_{p=0}^{P-1} s_i(n-p)h_{ji}(p) \tag{1}$$

where  $s_i(n)$  is the source signal from a source  $i$ ,  $x_j(n)$  is the received signal by a microphone  $j$ , and  $h_{ji}(n)$  is the P-point response from source  $i$  to microphone  $j$ . Using a T-point windowed discrete Fourier transformation (DFT), time domain signal  $x_j(n)$  can be converted into the time-frequency domain signal  $X_j(\omega, n)$  where  $\omega$  is a frequency index and  $n$  is a time index. For each frequency bin we have

$$\mathbf{X}(\omega, n) = \mathbf{H}(\omega)\mathbf{S}(\omega, n) \tag{2}$$

where  $\mathbf{S}(\omega, n) = [S_1(\omega, n), \dots, S_N(\omega, n)]^T$  and  $\mathbf{X}(\omega, n) = [X_1(\omega, n), \dots, X_M(\omega, n)]^T$  are the time-frequency representations of the source signals and the observed signals respectively and  $(\cdot)^T$  denotes vector transpose. The separation can be completed by the unmixing matrix  $\mathbf{W}(\omega)$  of a frequency bin  $\omega$

$$\hat{\mathbf{S}}(\omega, n) = \mathbf{W}(\omega)\mathbf{X}(\omega, n) \tag{3}$$

where  $\hat{\mathbf{S}}(\omega, n) = [\hat{S}_1(\omega, n), \dots, \hat{S}_N(\omega, n)]^T$  is the time-frequency representations of the estimated source signals and  $\mathbf{W}(\omega)$  is the frequency representation of the unmixing matrix.  $\mathbf{W}(\omega)$  is determined so that  $\hat{S}_1(\omega, n), \dots, \hat{S}_N(\omega, n)$  become mutually independent. Exploiting the nonstationary of the speech signal we define the cost function as follows:

$$J(\mathbf{W}(\omega)) = \arg \min \sum_{w=1}^T \sum_{k=1}^K F(\mathbf{W})(\omega, k), \tag{4}$$

where  $K$  is the number of signal segments and  $F(\mathbf{W})(\omega, k)$  is defined as

$$F(\mathbf{W})(\omega, k) = \|\mathbf{R}_{\hat{\mathbf{S}}}(\omega, k) - \text{diag}[\mathbf{R}_{\hat{\mathbf{S}}}(\omega, k)]\|_F^2 \tag{5}$$



where  $R_{\hat{s}}(\omega, k)$  is the autocorrelation matrix of the separated signals and  $\|\cdot\|_F^2$  denotes the squared Frobenius norm, and  $k$  is the block index. The separation problem is then converted into a joint diagonalization problem. Obviously, the solution  $\mathbf{W}(\omega) = 0$  will lead to the minimization of  $F(\mathbf{W})(\omega, k)$ . To avoid this some constraints should be added to the unmixing matrix. In [6] a penalty function is added to convert the constrained optimization problem into an unconstrained optimization problem. The cost function of penalty function based joint diagonalization is as follows:

$$J(\mathbf{W}(\omega)) = \arg \min_w \sum_{w=1}^T \sum_{k=1}^K F(\mathbf{W})(\omega, k) + \lambda g(\mathbf{W})(\omega, k) \quad (6)$$

where  $\lambda$  is the penalty weight factor and  $g(\mathbf{W})(\omega, k)$  is a form of penalty function based on a constraint of the unmixing matrix. With a gradient-based descent method we can calculate the unmixing matrix after several iterations from equation (6). The separated signals  $\hat{s}_1(n)$  and  $\hat{s}_2(n)$  can then be obtained from (3) after applying an inverse DFT.

### 3 Adaptive Noise Canceller

After BSS we obtain the estimated noise signal  $\hat{s}_1(n)$ . This signal is then used as a reference in the ANC stage signal to remove the noise component from the received signal  $x_1(n)$ . A new variable step size LMS algorithm which is suitable for speech processing is used in the ANC. The updates of the step size can be formulated as follows:

$$e(n) = x_1(n) - \hat{\mathbf{s}}_1^T(n) \mathbf{w}(n) \quad (7)$$

$$\mathbf{g}(n) = \frac{e(n) \hat{\mathbf{s}}_1(n)}{\sqrt{L[\hat{\sigma}_e^2(n) + \hat{\sigma}_s^2(n)]}} \quad (8)$$

$$\mathbf{p}(n) = \beta \mathbf{p}(n-1) + (1-\beta) \mathbf{g}(n) \quad (9)$$

$$\mu(n+1) = \alpha \mu(n) + \gamma \|\mathbf{p}(n)\|_F^2 \quad (10)$$

where  $\mu(n)$  is the variable step size,  $\hat{\mathbf{s}}_1(n) = [\hat{s}_1(n), \dots, \hat{s}_1(n-L+1)]^T$ ,  $\mathbf{w}(n)$  is the weight vector of the adaptive filter,  $L$  is the filter length,  $\hat{\sigma}_e^2(n)$  and  $\hat{\sigma}_s^2(n)$  are estimations of the temporal error energy and the temporal input energy,  $0 < \alpha < 1$ ,  $0 < \beta < 1$ ,  $\gamma > 0$ ,  $\mathbf{g}(n)$  is the square root normalized gradient vector,  $\mathbf{p}(n)$  is a smoothed version of  $\mathbf{g}(n)$ . The recursion of the filter weight vector is as follows

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu(n) \frac{e(n) \hat{\mathbf{s}}_1(n)}{L[\hat{\sigma}_e^2(n) + \hat{\sigma}_s^2(n)]} \quad (11)$$

The square root normalized gradient vector  $\mathbf{g}(n)$  in (8) is used to obtain a robust measure of the adaptive process. The first-order filter based averaging operation

in (9) removes the disturbance brought by the target signal. The variable step size  $\mu(n)$  in (10) is adapted to obtain a fast convergence rate during the early adaptive process and a small misadjustment after the algorithm converges. The adaptation of the weight vector in (11) is based on the sum method in [7] which is designed to minimize the steady state mean square error. Equations (7)(8)(9)(10)(11) provide a new variable step size LMS algorithm for the ANC stage. The output signal of the ANC  $\hat{y}_{12}(n)$  should then be a good estimation of the noise free reverberant speech signal  $y_{12}(n)$ . Next, the estimation of the RT from  $\hat{y}_{12}(n)$  must be considered.

### 4 ML RT Estimation Method

In the ML RT estimation method, the fine structure of the reverberant tail of the output signal  $\hat{y}_{12}(n)$  is overlap segmented by a window with a width of  $N$  [5]. At each segment we obtain an observed vector  $\mathbf{y}_N$ . This observed vector is then modelled as a product of two sequence

$$\mathbf{y}_N(i) = \mathbf{x}_N(i)\mathbf{a}_N(i), i = 1 \dots N \tag{12}$$

where  $\mathbf{x}_N$  is a vector whose elements are drawn from a random white Gaussian sequence  $x(n) \sim (0, \sigma^2)$  and  $\mathbf{a}_N$  is an exponentially damped sequence whose elements are determined by  $\mathbf{a}_N(i) = a^i, i = 1 \dots N$  where  $a = 1/\exp(-\tau)$ ,  $\tau$  is a constant which describes the damping rate. It is easy to see that  $\tau$  actually describes the damping rate of sequence  $\mathbf{a}_N(i)$ , which is used to model the envelope of the reverberant speech signal. According to the definition the RT can be obtained from this decay rate [5]:

$$T_{60} = \frac{-3\tau}{\log_{10}(\exp(-1))} = 6.91\tau \tag{13}$$

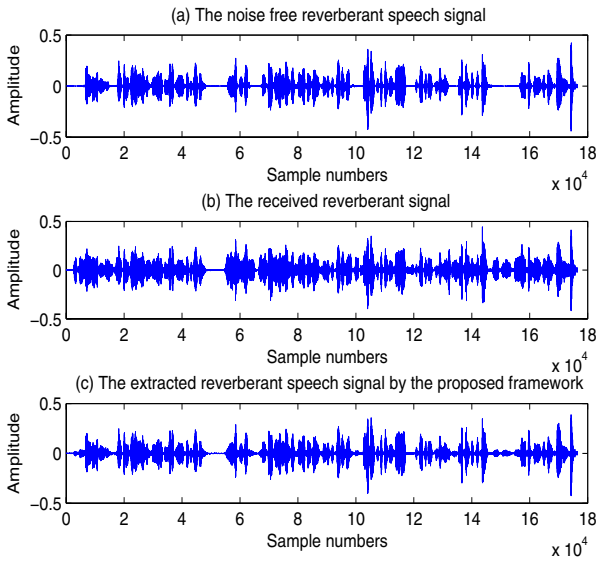
The logarithm likelihood function of the observed samples of  $\mathbf{y}_N$  with respect to the parameters  $a, \sigma$  is then obtained

$$E\{L(\mathbf{y}_N; a, \sigma)\} = -\frac{N(N-1)}{2} \ln(a) - \frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N a^{-2n} \mathbf{y}_N^2(n) \tag{14}$$

With this function the parameters  $a$  and  $\sigma$  can be obtained by the ML approach [8]. With the ML estimate of parameter  $a$ , the decay parameter  $\tau$  and the RT can also be obtained. From each segment we can obtain an estimate of RT, and a series of estimates of RT can be obtained with the total output signal  $\hat{y}_{12}(n)$ . These estimates can then be used to identify the most likely RT of the room by using an order-statistic filter [5]. As shown by the authors, it provides reliable RT estimates in noise free environments [5]. The simulation in the next section will show that combined with the proposed framework it can also extract a good estimate of the RT of a simulated occupied room.

## 5 Simulation

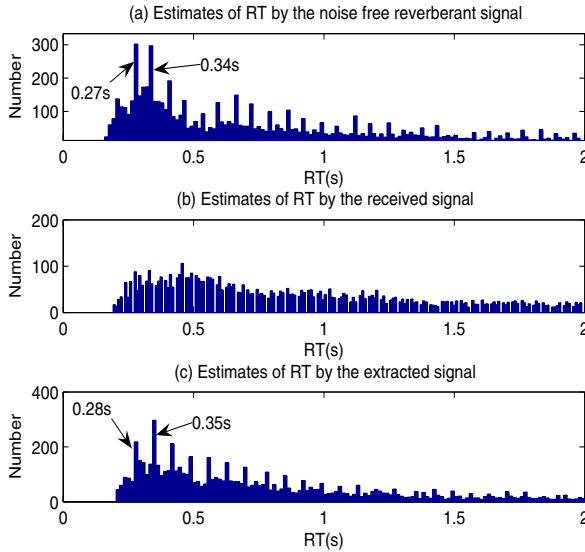
In this section we examine the performance of the proposed framework. The flow chart of the simulation is shown in Fig.1. The two source signals are two anechoic 40 seconds male speech signals with a sampling frequency of 8kHz. One of the speech signal is modelled as the excitation speech signal, and the other speech signal is modelled as the noise signal. The size of the simulated occupied room is set to be 10m\*10m\*5m where ‘m’ is meter. The positions of these two sources are set to be [1m 3m 1.5m] and [3.5m 2m 1.5m]. The positions of the two microphones are set to be [2.45m 4.5m 1.5m] and [2.55m 4.5m 1.5m] respectively. The occupied room and its impulse response  $h_{ji}$  between source  $i$  and microphone  $j$  are simulated by an image room model [9] with a reflection coefficient of 0.7, which is in rough correspondence with the actual room. The RT of this room measured by Schroeder’s method [2] is 0.27s.



**Fig. 2.** The noise free signal  $y_{12}(n)$ , the received signal  $x_1(n)$  and the extracted signal  $\hat{y}_{12}(n)$

As shown by Fig.1, BSS is performed firstly to extract the estimated noise signal  $\hat{s}_1$ . This signal contains mostly the noise signal and a low level of the desired excitation speech signal. To evaluate the BSS performance we use a noise to signal ratio (NSR) which is the energy ratio defined between the component of the noise signal and the component of the excitation speech signal contained in  $\hat{s}_1$ . The NSR of  $\hat{s}_1$  in this simulation is 38dB, therefore it has a strong correlation with the noise signal  $s_1$  and a slight correlation with the excitation speech signal. This signal is then used in the ANC model as a reference signal. The filter length of the ANC is set to be 500 and the parameters  $\alpha$ ,  $\beta$ ,  $\gamma$  are set to be 0.99, 0.9999,

200 respectively in the simulation. The last 1000 samples of the filter coefficients are used to measure the steady-state performance. The output signal of the ANC  $\hat{y}_{12}$  contains two components: the reverberant excitation speech signal and the residue of the noise signal. The signal to noise ratio (SNR) between these two components is 43dB.



**Fig. 3.** The RT estimates obtained by the noise free signal  $y_{12}(n)$ , the received signal  $x_1(n)$  and the extracted signal  $\hat{y}_{12}(n)$

We plot the first approximately 20s of the noise free reverberant speech signal  $y_{12}(n)$ , the contaminated reverberant speech signal  $x_1(n)$  and the output signal of ANC  $\hat{y}_{12}$  in Fig.2(a), Fig.2(b) and Fig.2(c) respectively. It is easy to see that after BSS and ANC the noise contained in  $x_1(n)$  is reduced greatly, and  $\hat{y}_{12}(n)$  is a good estimate of  $y_{12}(n)$ , which is suitable for RT estimation.

Then the ML method is used to extract the RT by using a window width of 1,200, which is chosen to obtain a good performance. To show the advantage of the proposed framework, we estimate the RT from both signals  $x_1(n)$  and  $\hat{y}_{12}$ , and compare the results with the estimates obtained by the noise free signal  $y_{12}$ . It is clearly to see in Fig.3 that the estimated RTs obtained with the extracted signal  $\hat{y}_{12}$  are similar to the results estimated with the noise free signal  $y_{12}$ , and it is difficult to identify the correct RT from the RT estimates with the received signal  $x_1(n)$ . Note that both the estimates in Fig.3(a) and Fig.3(c) have another peak which is larger than the correct RT 0.27s due to the lack of sharp transients in the clean speech [5]. Although presented for only one scenario due to the space limitation, such improvement is consistently found in our experimental studies.

The performance of the whole framework highly depends on the performance of the BSS algorithm. Although limited by the performance of frequency domain

BSS, this framework is only suitable for the room whose RT is less than 0.3s and the noise is directional, it can be potentially developed for more real cases with further enhancement of the BSS algorithm. As shown by our simulations above, nonetheless, reliable RT can be extracted using this framework within a highly noisy occupied room, something that has not previously been possible.

## 6 Conclusion

This paper proposes a new framework for room acoustic parameter extraction in occupied rooms. Simulation results show that the noise is removed greatly by the proposed framework from the reverberant speech signal and the performance of this framework is good in a simulated low RT occupied room environment. Due to the motivation of our framework BSS and ANC can be potentially used in many acoustic parameter estimation methods as a preprocessing. This framework provides a new way to overcome the noise disturbance in RT estimation. Future work will focus on the theoretic analysis of this blind RT estimate framework and the improvement of its stages, especially the improvement of convolutive BSS in long reverberation environments.

## References

1. H. Kuttruff: Room Acoustics 4th ed. Spon Press. (2000) 315–333
2. M. R. Schroeder: New method for measuring reverberation time. *J. Acoust. Soc. Am.* **37** (1965) 409–412
3. ISO 3382: Acoustics-measurement of the reverberation time of rooms with reference to other acoustical parameters. International Organization for Standardization. (1965)
4. T. J. Cox, F. Li and P. Darlington: Extracting room reverberation time from speech using artificial neural networks. *J. Audio. Eng. Soc* **49** (2001) 219–230
5. R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr., C. R. Lansing and A. S. Feng: Blind estimation of reverberation time. *J. Acoust. Soc. Amer.* **114** (2003) 2877–2892
6. W. Wang, S. Sanei and J. A. Chambers: Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources. *IEEE Tran. Signal Processing* **53** (2005) 1654–1669
7. J. E. Greenberg: Modified LMS algorithm for speech processing with an adaptive noise canceller. *IEEE Trans. Signal Processing* **6** (1998) 338–351
8. R. Ratnam, D. L. Jones and W. D. O'Brien Jr.: Fast algorithms for blind estimation of reverberation time. *J. Acoust. Soc. Amer.* **11** (2004) 537–540
9. J. B. Allen and D. A. Berkley: Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Amer.* **65** (1979) 943–950

# An Online Method for Detecting Nonlinearity Within a Signal

Beth Jelfs<sup>1</sup>, Phebe Vayanos<sup>1</sup>, Mo Chen<sup>1</sup>, Su Lee Goh<sup>1</sup>, Christos Boukis<sup>2</sup>,  
Temujin Gautama<sup>3</sup>, Tomasz Rutkowski<sup>4</sup>, Tony Kuh<sup>5</sup>, and Danilo Mandic<sup>1</sup>

<sup>1</sup> Imperial College London, UK

<sup>2</sup> AIT, Greece

<sup>3</sup> Phillips Leuven, Belgium

<sup>4</sup> BSI RIKEN, Japan

<sup>5</sup> University of Hawaii

**Abstract.** A novel method for online analysis of the changes in signal modality is proposed. This is achieved by tracking the dynamics of the mixing parameter within a hybrid filter rather than the actual filter performance. An implementation of the proposed hybrid filter using a combination of the Least Mean Square (LMS) and the Generalised Normalised Gradient Descent (GNGD) algorithms is analysed and the potential of such a scheme for tracking signal nonlinearity is highlighted. Simulations on linear and nonlinear signals in a prediction configuration support the analysis. Biological applications of the approach have been illustrated on EEG data of epileptic patients.

## 1 Introduction

Signal modality characterisation is becoming an increasingly important area of multi-disciplinary research and large effort has been put into devising efficient algorithms for this purpose. The idea is that the changes in the signal nature between linear and nonlinear and deterministic and stochastic can reveal information (knowledge) which is critical in certain applications (e.g. health conditions). Our focus is on the detection of linear/nonlinear changes in signal modality. The existing algorithms in this area are based on hypothesis testing [1, 2, 3] and describe the signal changes in a statistical manner. However, there are very few online algorithms which are suitable for this purpose. The purpose of the new approach described in this paper is to show the possibility of an online algorithm which can be used not only to identify the nature of the signal, but also to track changes in the nature of the signal (signal modality detection).

One intuitive method to determine the nature of a signal has been to present the signal as input to two adaptive filters with different characteristics, one nonlinear and the other linear. By comparing the responses of each filter this can be used to identify whether the input signal is linear or not. Whilst this is a very useful simple test for signal nonlinearity, it does not provide an online solution. There are ambiguities due to the need to choose many parameters of the corresponding filters and this approach does not rely on the “synergy” between the filters considered.

### 1.1 Existing Approaches

In [4] an online approach is considered which successfully tracks the degree of nonlinearity of a signal using adaptive algorithms, but relies on a parametric model to effectively model the system in order to provide a true indication of the degree of nonlinearity. Figure 1 shows an implementation of this method using a third order Volterra system expansion as the system model and the normalised LMS (NLMS) algorithm with a step size  $\mu = 0.008$  to update the system parameters. The system input and output can be described by

$$u[n] = \sum_{i=0}^I a_i x[n - i] \text{ where } I = 2 \text{ and } a_0 = 0.5, a_1 = 0.25, a_2 = 0.125 \quad (1)$$

$$y[n] = F(u[n]; n) + \eta[n] \quad (2)$$

where  $x[n]$  are i.i.d uniformly distributed over the range  $[-0.5, 0.5]$  and  $\eta[n] \sim \mathcal{N}(0, 0.0026)$ . The function  $F(u[n]; n)$  varies with  $n$ ,  $F(u[n]; n) = u^3[n]$  for  $10000 < n \leq 20000$ ,  $F(u[n]; n) = u^2[n]$  for  $30000 < n \leq 40000$  and  $F(u[n]; n) = u[n]$  at all other times. The output  $y[n]$  can be seen in the first trace of Fig. 1 the second and third traces show the residual estimation errors of the optimal linear system and Volterra system respectively, the final trace is the estimated degree of system nonlinearity. Whilst these results show that this approach can detect changes in nonlinearity and is not affected by the presence of noise this may be largely due to nature of the input signal being particularly suited to the Volterra model.

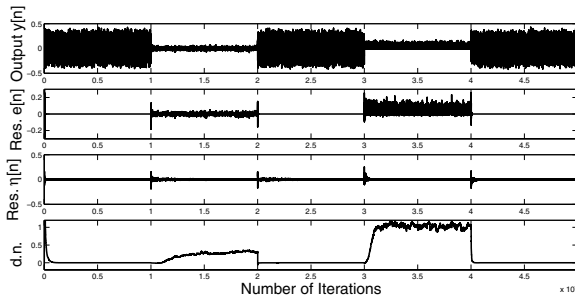


Fig. 1. NLMS with Volterra series

### 1.2 The Proposed Approach

The proposed approach develops on the tracking capabilities of adaptive filters by using combinations of adaptive filters in a more natural way to produce a single hybrid filter without the need for any underlying signal generation models. Hybrid filters consist of multiple individual adaptive subfilters operating in parallel and all feeding into a mixing algorithm which produces the single output of the filter [5, 6]. This mixing algorithm is also adaptive and combines the outputs

of the subfilters based on their current performance on the input signal. Previous applications of hybrid filters have focused mainly on the improved performance they can offer over the individual constituent filters. One effect of this mixing algorithm is that it can give an indication of which filter is currently responding to the input signal most effectively. Therefore by again selecting algorithms which are suitable for either linear or nonlinear signals, it is possible to cause this mixing algorithm to adapt according to fundamental properties of the input signal. A simple form of mixing algorithm for two adaptive filters is a convex combination. Convexity can be described as [7]

$$\lambda x + (1 - \lambda)y \text{ where } \lambda \in [0, 1] \tag{3}$$

and  $x$  and  $y$  are two points on a line, the resultant of this will lie on the same line between  $x$  and  $y$ . By observing the mixing parameter  $\lambda$  and which of the subfilters is dominating, conclusions can then be drawn about the nature of the input signal.

The proposed approach uses the least mean square (LMS) algorithm [8] to train one of the subfilters and the generalised normalised gradient descent (GNGD) algorithm [9] for the other, with the purpose of distinguishing the linearity/non-linearity of a signal and is illustrated in Fig. 2. The LMS algorithm was chosen as it is widely used, known for its robustness and excellent steady state properties whereas the GNGD algorithm has a faster convergence speed and much better tracking capabilities. We aim at exploiting these properties and set out to show that due to the synergy and simultaneous mode of operation, our proposed hybrid filter has excellent tracking capabilities for signals with extrema in their inherent linearity and nonlinearity characteristics.

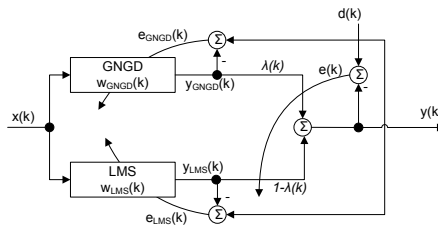


Fig. 2. Convex combination of adaptive filters

## 2 Derivation of the Proposed Approach

Unlike the existing approaches to hybrid adaptive filters which focus on the quantitative performance of such filters, our aim is to design a hybrid filter which combines the characteristics of two adaptive filters with the aim of making the value of the “mixing” parameter  $\lambda$  adapt according to the dynamics of the input signal. This is achieved following the approach from Fig. 2 where two adaptive filters are combined, one trained by the GNGD algorithm and the other by the



LMS algorithm, in a convex manner. The output of the hybrid filter from Fig. 2  $y(k)$  is an adaptive convex combination of the output of the LMS trained subfilter  $y_{LMS}$  and the output of the GNGD trained subfilter  $y_{GNGD}$  and is given by

$$y(k) = \lambda(k)y_{GNGD}(k) + (1 - \lambda(k))y_{LMS}(k) \tag{4}$$

where  $y_{LMS}$  is generated from [8]

$$\begin{aligned} y_{LMS}(k) &= \mathbf{x}^T(k)\mathbf{w}_{LMS}(k) \\ e_{LMS}(k) &= d(k) - y_{LMS}(k) \\ \mathbf{w}_{LMS}(k + 1) &= \mathbf{w}_{LMS}(k) + \mu_{LMS}e_{LMS}(k)\mathbf{x}(k) \end{aligned} \tag{5}$$

and  $y_{GNGD}$  is the corresponding output of the GNGD trained subfilter given by [9]

$$\begin{aligned} y_{GNGD}(k) &= \mathbf{x}^T(k)\mathbf{w}_{GNGD}(k) \\ e_{GNGD}(k) &= d(k) - y_{GNGD}(k) \\ \mathbf{w}_{GNGD}(k + 1) &= \mathbf{w}_{GNGD}(k) + \frac{\mu_{GNGD}}{\|\mathbf{x}(k)\|_2^2 + \varepsilon(k)}e_{GNGD}(k)\mathbf{x}(k) \\ \varepsilon(k + 1) &= \varepsilon(k) - \rho\mu_{GNGD}\frac{e_{GNGD}(k)e_{GNGD}(k - 1)\mathbf{x}^T(k)\mathbf{x}(k - 1)}{(\|\mathbf{x}(k - 1)\|_2^2 + \varepsilon(k - 1))^2} \end{aligned} \tag{6}$$

where  $e_{GNGD}(k)$  and  $e_{LMS}(k)$  are the individual output errors of the subfilters at time instant  $k$ ,  $d(k)$  is the desired signal,  $\mathbf{x}(k) = [x(k - 1), \dots, x(k - N)]^T$  is the input signal vector,  $N$  is the length of the filter and  $\mathbf{w}_{LMS}(k) = [w_{LMS_1}(k), \dots, w_{LMS_N}(k)]^T$  and  $\mathbf{w}_{GNGD}$  are the filter weight vectors corresponding to the LMS and GNGD trained subfilters. The step-size parameters of the filters are  $\mu_{LMS}$  and  $\mu_{GNGD}$ , and in the case of the GNGD  $\rho$  is the step-size adaptation parameter and  $\varepsilon$  the regularisation term.

To preserve the inherent characteristics of the subfilters, which are the basis of our approach, the constituent subfilters are each updated by their own errors  $e_{LMS}(k)$  and  $e_{GNGD}(k)$ , whereas the parameter  $\lambda$  is updated based on the overall error  $e(k)$ . The convex mixing parameter  $\lambda(k)$  is updated using the following gradient adaptation

$$\lambda(k + 1) = \lambda(k) - \mu_\lambda \nabla_\lambda E(k)|_{\lambda=\lambda(k)} \tag{7}$$

where  $\mu_\lambda$  is the adaptation step-size. From (4) and (7), the  $\lambda$  update can be shown to be

$$\lambda(k + 1) = \lambda(k) - \frac{\mu_\lambda}{2} \frac{\partial e^2(k)}{\partial \lambda(k)} = \lambda(k) + \mu_\lambda e(k)(y_{GNGD}(k) - y_{LMS}(k)) \tag{8}$$

To ensure the combination of adaptive filters remains a convex function it is critical  $\lambda$  remains within the range  $0 \leq \lambda(k) \leq 1$ . To ensure this, in [5] the authors used a sigmoid function as a post-nonlinearity to bound  $\lambda(k)$ . Since, in order to determine the changes in the modality of a signal (linear, nonlinear) we

are not interested in the overall performance of the filter but in the variable  $\lambda$  the use of a sigmoid function would interfere with true values of  $\lambda(k)$  and was therefore not appropriate. A hard limit on the set of allowed values for  $\lambda(k)$  was therefore implemented.

### 3 Simulations

For all simulations the proposed approach was evaluated in an adaptive one step ahead prediction setting with the length of the adaptive filters set to  $N = 10$ .

Several experiments were conducted in order to illustrate the ability of the proposed approach to track the modality changes within a signal of interest. In the first experiment the behaviour of  $\lambda$  was investigated for benchmark linear and nonlinear inputs. Values of  $\lambda$  were averaged over a set of 1000 independent simulation runs. The inputs used in the simulations were a stable linear AR(4) process given by:

$$x(k) = 1.79x(k-1) - 1.85x(k-2) + 1.27x(k-3) - 0.41x(k-4) + n(k) \quad (9)$$

and a benchmark nonlinear signal [10] given by:

$$x(k+1) = \frac{x(k)}{1+x^2(k)} + n^3(k) \quad (10)$$

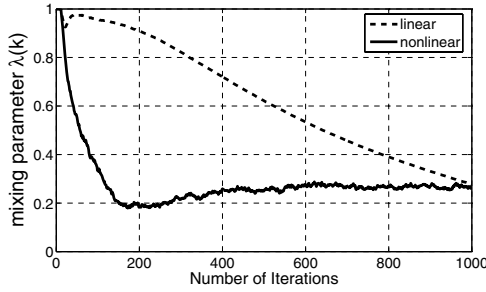
where  $n(k)$  is a zero mean, unit variance white Gaussian process.

The values of the step-sizes used were  $\mu_{LMS} = 0.01$  and  $\mu_{GNGD} = 0.6$ . For the GNGD filter  $\rho = 0.15$  and the initial value of the regularisation parameter was  $\varepsilon(0) = 0.1$ . Within the convex combination of the filters the step-size for the adaptation of  $\lambda(k)$  was  $\mu_\lambda = 0.05$  and the initial value of  $\lambda(0) = 1^1$ . From the curves shown in Fig. 3 the value of  $\lambda(k)$  for both inputs move towards zero as the adaptation progresses. As expected the output of the proposed convex combination of adaptive filters approaches the output of the LMS filter  $y_{LMS}$  predominately due to the better steady state properties of the LMS filter when compared to the GNGD filter. In the early stages of adaptation the nonlinear input (10) adapts to be dominated by the LMS filter much faster than the linear input and quickly converges whereas the linear input (9) has a much more gradual change between the two filters<sup>2</sup>.

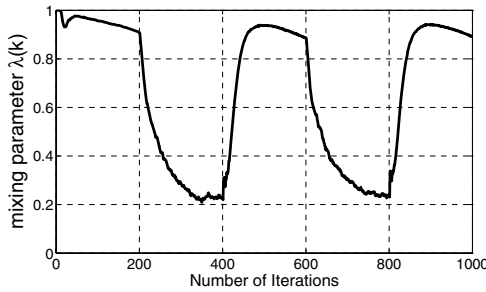
In the second experiment, we investigate whether we can use the changes in  $\lambda$  along the adaptation to track the changes in signal modality. Since the behaviour of  $\lambda$  as a response to the different inputs is clearly distinct, especially in the earliest stages of adaptation, the proposed convex combination was presented with an input signal which alternated between linear and nonlinear. The input signal was alternated every 200 samples and the dynamics of the mixing

<sup>1</sup> Since GNGD exhibits much faster convergence than LMS it is natural to start the adaptation with  $\lambda(0) = 1$  this way, we avoid possible artefacts that may arise due to the slow initial response to the changes in signal modality.

<sup>2</sup> Both filters perform well on a linear input and are competing along the adaptation.



**Fig. 3.** Comparison of  $\lambda$  for linear and nonlinear inputs



**Fig. 4.** Mixing parameter  $\lambda$  for input nature alternating every 200 samples

parameter  $\lambda(k)$  are shown in Fig. 4. Figure 4 shows how the value of  $\lambda(k)$  adapts in a way which ensures that the output of the convex combination is dominated by the filter most appropriate for the input signal characteristics.

The third experiment was to investigate the fastest speed with which the proposed approach can adapt to the alternating signal. Figure 5 shows the response of  $\lambda(k)$  to the input signal alternating every 100 and 50 samples respectively. There is a small anomaly in the values of  $\lambda$  immediately following the change in input signal from nonlinear to linear, which can be clearly seen in Fig. 5b around sample numbers  $100i, i = 1, 2, \dots$  where the value of  $\lambda$  exhibits a small dip before it increases. This is due to the fact that the input to both the current AR process 9 and the tap inputs to both filters use previous nonlinear samples where we are in fact predicting the first few “linear” samples. This does not become an issue when alternations between the input signals occur less regularly or if there is a more natural progression from “linear” to “nonlinear” in the the input signal.

Finally, to examine the usefulness of the proposed approach for the processing of real world signals a set of EEG signals has been analysed. Following the standard practice, the EEG sensor signals were averaged across all the channels and any trends in the data were removed. Figure 6 shows the response of  $\lambda$  when applied to two different sets of EEG data from epileptic patients, both

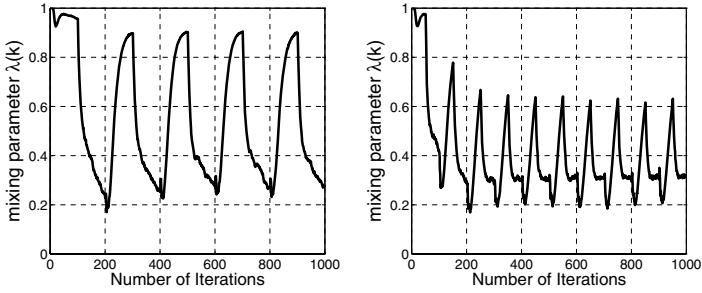


Fig. 5. Mixing parameter  $\lambda$  for input nature alternating every a)100 b)50 samples

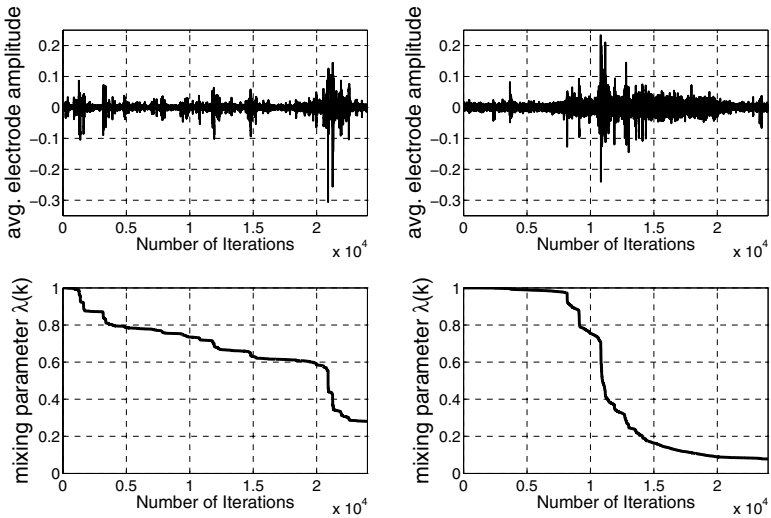


Fig. 6. EEG signals for two patients showing epileptic seizures and corresponding  $\lambda$

showing the onset of a seizure. These results show that the proposed approach can effectively detect changes in the nature of the EEG signals which can be very difficult to achieve otherwise. It would be interesting to ascertain whether by performing multiple step ahead prediction would help detect a change in signal nature before the actual onset of that change.

### 4 Conclusions

We have proposed a novel approach to identify changes in the modality of a signal. This is achieved by a convex combination of two adaptive filters for which the transient responses are significantly different. By training the two filters with different algorithms, in this case the least mean square (LMS) and generalised normalised gradient descent (GNGD) algorithms, it is possible to exploit the

different performance capabilities of each. The evolution of the adaptive convex mixing parameter  $\lambda$ , helps determine which filter is more suited to the current input signal dynamics, and thereby gain information about the nature of the signal. The analysis and simulations illustrate that there is significant potential for the use of this method for online tracking of some fundamental properties of the input signal.

## References

1. Schreiber, T., Schmitz, A.: Discrimination power of measures for nonlinearity in a time series. *Physical Review E* **55** (1997) 5443–5447
2. Gautama, T., Mandic, D., Hulle, M.V.: The delay vector variance method for detecting determinism and nonlinearity in time series. *Physica D* **190** (2004) 167–176
3. Gautama, T., Mandic, D., Hulle, M.V.: Signal nonlinearity in fmri: A comparison between bold and mion. *IEEE Transactions on Medical Imaging* **22** (2003) 636–644
4. Mizuta, H., Jibu, M., Yana, K.: Adaptive estimation of the degree of system nonlinearity. In: *Proceedings IEEE 2000 Adaptive Systems for Signal Processing and Control Symposium (AS-SPCC)*. (2000) 352–356
5. Figueras-Vidal, A., Arenas-Garcia, J., Sayed, A.: Steady state performance of convex combinations of adaptive filters. In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*. Volume 4. (2005) 33–36
6. Kozat, S.S., Singer, A.C.: Multi-stage adaptive signal processing algorithms. In: *Proceedings IEEE Sensor Array Multichannel Signal Processing workshop*. (2000) 380–384
7. Cichocki, A., Unbehauen, R.: *Neural Networks for Optimisation and Signal Processing*. Wiley (1993)
8. Widrow, B., Stearns, S.: *Adaptive Signal Processing*. Prentice-Hall (1985)
9. Mandic, D.: A generalized normalized gradient descent algorithm. *IEEE Signal Processing Letters* **11** (2004) 115–118
10. Narendra, K., Parthasarathy, K.: Identification and control of dynamical systems using neural networks. *IEEE Transactions on Neural Networks* **1** (1990) 4–27

# Using Hierarchical Filters to Detect Sparseness in Unknown Channels

C. Boukis and L.C. Polymenakos

Athens Information Technology, Greece

**Abstract.** A novel algorithm particularly suited for determining and identifying the sparseness and the associated features within an unknown channel model is proposed. This is achieved by virtue of hierarchical filters which exhibit relatively steady state characteristics, when more than one of their constitutive sub-filters have to identify non-zero parts of the unknown channel. This way, sparseness is detected by comparing, in an on-line fashion, the performance of a hierarchical filter with that of the standard finite impulse response (FIR) filter trained by the least mean square (LMS) algorithm. In addition, the character and type of sparseness can be detected based on the architecture of the underlying hierarchical filter. The analysis is supported by simulations evaluated in a rigorous statistical framework.

## 1 Introduction

Communication channels usually have most of their energy concentrated in a just a few locations, resulting in sparse impulse responses. Typical examples are the echo paths in communication channels and underwater acoustic channels [1]. The determination of the degree and the type of sparsity of a channel is important since it might reveal significant information relative to its nature. Given the sparseness of a channel and the approximate locations of the non-zero samples of its impulse response more efficient identification can be performed. Moreover, significant reduction of the computational complexity can be succeeded since only the non-zero parts need to be identified. Contrary to existing approaches [2,6] we are not directly interested on the identification of the unknown channel; our major concern is the determination of the sparsity of the unknown channel, if any.

The proposed system consists of a hierarchical and an FIR filter performing in parallel. Their coefficients are updated according to the hierarchical least mean square (HLMS) [7] and the least mean square (LMS) [8] algorithms respectively. Comparing the error signals of the two adaptive structures can reveal whether the unknown channel is sparse or not. Moreover, by grouping appropriately the samples of the tap-delay line of the input layer into sub-filters the area of the impulse response of the unknown channel where the energy is located can be determined.

This paper is organised as follows: A brief presentation of hierarchical filters and the HLMS algorithm is given in section 2. The proposed - sparseness detection - structure is introduced in section 3. Simulation results are given in section 4. Finally section 5 concludes the paper.

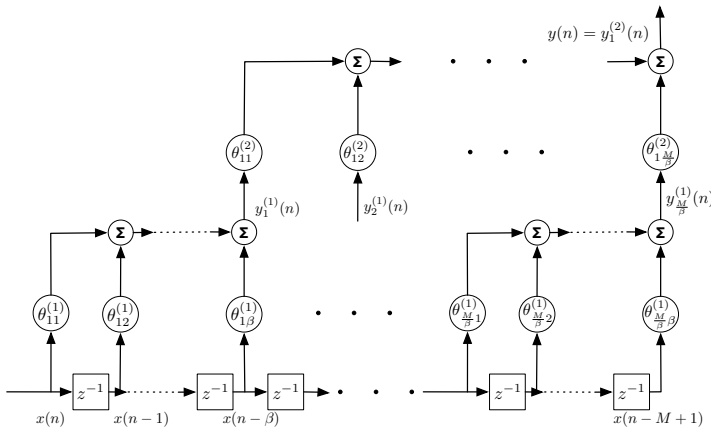
## 2 Background: Hierarchical Structures

The most common approach in system identification is the use of finite impulse response filters (FIR), also known as tap-delay lines, whose coefficients are updated with the LMS algorithm [8]. The output of such a structure is a finite sum of delayed and weighted versions of their input, i.e.  $y_{LMS}(n) = \sum_{i=0}^{M-1} \theta_i x(n-i)$ . Using the LMS algorithm to adapt its coefficients  $\boldsymbol{\theta} = [\theta_0, \theta_1, \dots, \theta_{M-1}]^t$  yields  $\boldsymbol{\theta}(n+1) = \boldsymbol{\theta}(n) + \mu e_{LMS}(n) \mathbf{x}(n)$  where  $\mathbf{x}(n) = [x(n), x(n-1), \dots, x(n-M+1)]$  the regressor vector,  $\mu$  the step size and

$$e_{LMS}(n) = d(n) - y_{LMS}(n) \tag{1}$$

the adaptation error. By  $d(n)$  the desired response is denoted. The convergence behaviour of these adaptive structures, which has been thoroughly analysed in the past [8], depends strongly on their length  $M$ . The larger this is the slower the filter converges and the higher the steady state error is.

Hierarchical filters, contrary to FIR, are multilayer structures [7]. By combining the outputs of several transversal filters of small length they attempt to identify systems of higher order. A typical example of a hierarchical filter consisting of two layers given in Fig. 2. The coefficients of its lower layer (input layer) are grouped into  $M/\beta$  sub-filters, of order  $\beta-1$  each. The outputs of these sub-filters are then used as input in the second layer (output layer) to produce a global output. So the input layer produces local estimates of the desired response while the output layer fuses them. The number of delays in the input layer determines the order of the hierarchical filter.



**Fig. 1.** A hierarchical filter with 2 layers, consisting of  $M/\beta$  sub-filters of length  $\beta$  each at the input layer and 1 sub-filter of length  $M/\beta$  at the output layer

The output of the  $i$ -th sub-filter of the input layer of the structure of Fig. 2 is given by

$$y_i^{(1)}(n) = \sum_{j=0}^{\beta-1} \theta_{i,j}^{(1)} x(n - i\beta - j) \tag{2}$$

while the global output is

$$y_{HLMS}(n) = y_1^{(2)}(n) = \sum_{i=0}^{M/\beta-1} \theta_{1,i}^{(2)} y_i^{(1)}(n) \tag{3}$$

Combining (2) and (3) yields

$$y_{HLMS}(n) = \sum_{i=0}^{M/\beta-1} \theta_{1,i}^{(2)} \sum_{j=0}^{\beta-1} \theta_{i,j}^{(1)} x(n - i\beta - j) \tag{4}$$

From (4) it is observed that hierarchical filters have: (a) linear input-output equations and (b) more degrees of freedom than common FIR filters, since they have more coefficients to adjust. For example the structure of Fig. 2 has  $M + M/\beta$  coefficients while an FIR filter of the same order has  $M$ . Notice that the output of a hierarchical filter of order  $M$  consisting of two layers and having sub-filter length  $\beta$  is equivalent to that of an FIR filter of the same order whose  $k$ -th coefficient is given by  $\theta_k^{FIR} = \theta_{1,i}^{(2)} \theta_{i,j}^{(1)}$ , where  $i = \text{floor}(k/\beta)$  and  $j = \text{mod}(k, \beta)$ . The HLMS algorithm applies a stochastic gradient descent technique for the optimisation of the coefficients of each sub-filter which attempts to minimise the square of the local errors defined as

$$e_i^{(1)}(n) = d(n) - y_i^{(1)}(n) \tag{5}$$

$$e_1^{(2)}(n) = e_{HLMS}(n) = d(n) - y_1^{(2)}(n) \tag{6}$$

for the input and the output layer respectively ( $i = 0, 1, \dots, \beta - 1$ ). This results in the following set of updating equations

$$\theta_i^{(1)}(n+1) = \theta_i^{(1)}(n) + \mu_i^{(1)} e_i^{(1)}(n) \phi_i^{(1)}(n) \tag{7}$$

$$\theta_1^{(2)}(n+1) = \theta_1^{(2)}(n) + \mu_1^{(2)} e_1^{(2)}(n) \phi_1^{(2)}(n) \tag{8}$$

where  $\theta_i^{(1)} = [\theta_{i,0}^{(1)}, \theta_{i,1}^{(1)}, \dots, \theta_{i,\beta-1}^{(1)}]^t$  the coefficient's vector of the  $i$ -th subfilter of the input layer,  $\theta_1^{(2)} = [\theta_{1,0}^{(2)}, \theta_{1,1}^{(2)}, \dots, \theta_{1,M/\beta-1}^{(2)}]^t$  the coefficients of the output layer. The corresponding regressor vectors are  $\phi_i^{(1)}(n) = [x(n - i\beta), x(n - i\beta - 1), \dots, x(n - (i+1)\beta - 1)]^t$  and  $\phi_1^{(2)}(n) = [y_0^{(1)}(n), y_1^{(1)}(n), \dots, y_{M/\beta-1}^{(1)}(n)]^t$ . The adaptation is controlled by the learning rates  $\mu_i^{(1)}$  and  $\mu_1^{(2)}$ .

Assuming that the desired response is produced by a LTI channel with coefficients  $\theta^* = [\theta_0^*, \theta_1^*, \dots, \theta_{M-1}^*]^t$ , i.e.  $d(n) = \sum_{i=0}^{M-1} \theta_i^* x(n - i)$  and exploiting (4) the adaptation errors given in (5) and (6) become



$$e_i^{(1)}(n) = \sum_{j=0}^{\beta-1} \left( \theta_{i\beta+j}^* - \theta_{i,j}^{(1)} \right) x(n - i\beta - j) + \sum_{k=0}^{i\beta-1} \theta_k^* x(n - k) + \sum_{k=(i+1)\beta}^M \theta_k^* x(n - k)$$

$$e_1^{(2)}(n) = e_{HLMS}(n) = \sum_{i=0}^{M/\beta-1} \sum_{j=0}^{\beta-1} \left( \theta_{i\beta+j}^* - \theta_{1,i}^{(2)} \theta_{i,j}^{(1)} \right) x(n - i\beta - j)$$

Evaluating  $e_i^{(1)}(n)$  and  $e_1^{(2)}(n)$  for specific cases (i.e.  $\theta_{i\beta+j}^* = \theta_{1,i_o}^{(2)} \theta_{i_o,j}^{(1)}$  or  $\theta_{i\beta+j}^* = \theta_{i_o,j}^{(1)}$ ) shows that the adaptation errors cannot be driven to zero simultaneously. Hence HLMS has increased steady state error compared to the LMS [5]. A rigorous mathematical analysis is provided in [4]. Only when

$$\theta_k^* = \theta_{i\beta+j}^* = \begin{cases} \theta_{i_o\beta+j}^* \neq 0 & i = i_o \text{ and } j = 0, \dots, \beta \\ 0 & i = 0, \dots, i_o - 1, i_o + 1, \dots, M/\beta \text{ and } j = 0, \dots, \beta \end{cases}$$

the HLMS and the LMS algorithm have the same steady state error [3]. In this case the HLMS converges faster than the standard LMS. This is illustrated in Fig. 2 where the mean square error (MSE) curves of the HLMS and the LMS algorithms are compared for a sparse channels of length 256 with 8 nonzero coefficients. The performance of the HLMS was evaluated for three distinct values of the sub-filter length  $\beta$ : 2, 4 and 8. The values nonzero samples of the impulse response that was used for the derivation of these graphs were given by

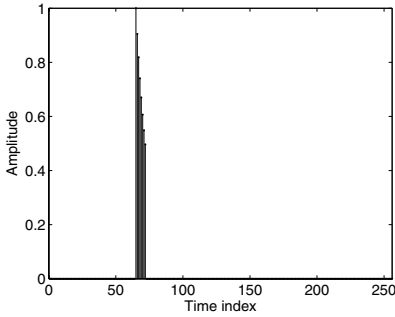
$$\theta_k^* = \exp(-(k - k_0)/\tau), \text{ where } k = k_0, k_0 + 1, \dots, k_0 + d \quad (9)$$

where  $k_0$  is the location of the first nonzero sample,  $d$  is the number of nonzero samples and  $\tau$  a time constant that controls the decaying rate of the envelope of this impulse response.

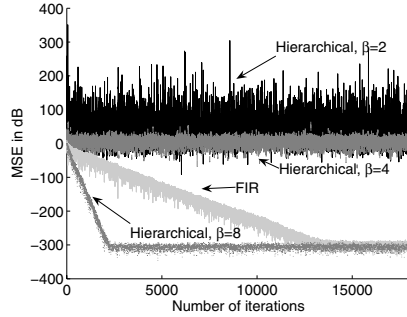
### 3 Sparseness Detection

The block diagram of the system that was developed for sparseness detection in unknown channels is depicted in Fig. 3. Its functioning can be described as follows: Initially a proper order  $M - 1$  for the adaptive filters is chosen so as to avoid under-modelling of the unknown channel. A value for  $n_\infty$ , which denotes the number of iterations required by the LMS algorithm to reach steady state is also chosen. Notice that  $n_\infty$  depends on the order of the adaptive filters and the employed learning rate. The hierarchical filter has 2 layers. The sub-filter length is initially set to  $\beta = M/2$ . At time instant  $n = 0$  the adaptation of the coefficients of the FIR and the hierarchical filter commences. The quantity that is observed so as to make a decision concerning the sparsity of the channel is

$$\Delta e^2(n) = e_{HLMS}^2(n) - e_{LMS}^2(n) \quad (10)$$

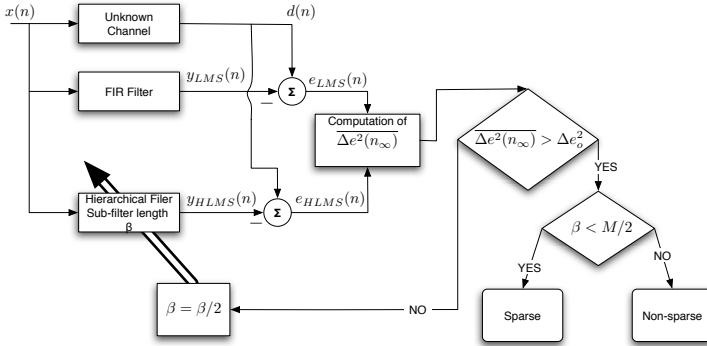


(a) Channel with 8 nonzero samples



(b) Hierarchical and FIR filter performance for channel (b)

**Fig. 2.** Comparison of the MSE curves of the HLMS and the LMS algorithms for two sparse channels

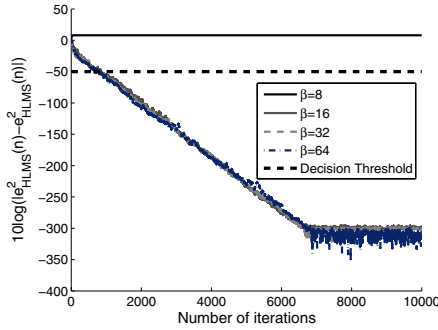


**Fig. 3.** Diagram of the experimental set-up used for sparsity detection

In order to reduce the "randomness" of this signal and to be able to get more robust decisions a momentum term was introduced resulting in

$$\overline{\Delta e^2(n)} = (1 - \lambda)\overline{\Delta e^2(n - 1)} + \lambda \Delta e^2(n) \tag{11}$$

When the value of  $\overline{\Delta e^2(n)}$  tends to zero as  $n$  increases, then the unknown channel is sparse. On the contrary, when there is no sparseness  $\overline{\Delta e^2(n)}$  progressively increases. Thus, observing the value of  $\overline{\Delta e^2(n)}$  at a time instant  $n_\infty$  after the commencement of the adaptation process and comparing it to a hard-bound  $\Delta e_o^2$  a decision can be made on the sparsity of the channel. If  $\overline{\Delta e^2(n)} < \Delta e_o^2$  the sub-filter length is decreased and the whole procedure is repeated for this new value. When  $\overline{\Delta e^2(n_\infty)}$  is found to have a large positive value then the final length of the sub-filter  $\beta_f$  is checked. If this is smaller than  $M/2$ , i.e. its initial value, then the channel is sparse. The nonzero coefficients in this case are located in a neighbourhood of width  $2\beta_f$  in the impulse response of the unknown channel. Their exact location can be found by checking which one of the coefficients of the



**Fig. 4.** The value of  $\Delta e^2(n)$  as a function of time for several phases of the proposed sparsity detection method

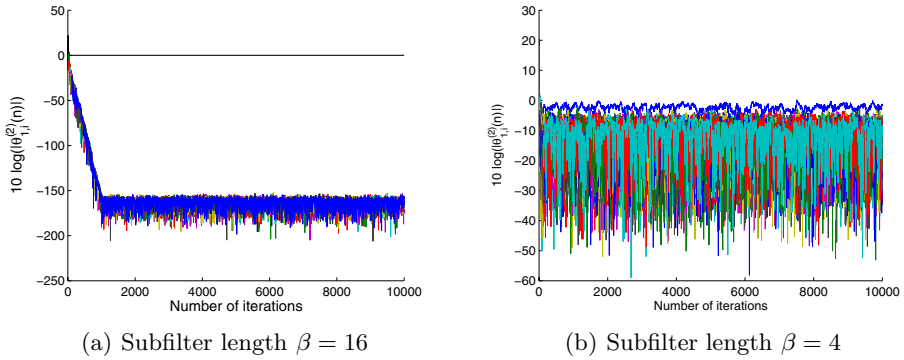
output layer converges to a nonzero value when  $\beta = 2\beta_f$  (Fig. 5(a)). Notice that the employed hierarchical filter can detect sparseness of the form (2) in channel impulse responses.

### 4 Simulation Results

To evaluate the performance of proposed method, the determination of the sparseness of an unknown channel of length 256 was attempted. Its impulse response had only 8 nonzero samples located at positions  $k = 85, 86, \dots, 92$  and their values were given by (9). This channel is depicted in Fig. 2(a). To enhance adaptation the learning rate of the LMS algorithm was normalised with respect to the power of its input vector. For the HLMS algorithm the step-size of every sub-filter was normalised with respect to the power of its regressor vector. The order of both adaptive filters was set to 256 as well, to avoid under-modelling. For this length a proper value for  $n_\infty$  was  $3 \times 10^3$ . The decision on the sparsity of the unknown channel was based on the comparison of  $\overline{\Delta e^2(n_\infty)}$  to the threshold  $\Delta e_o^2 = 10^{-5}$ . The input signal was white noise of zero mean and unit variance and for the computation of  $\overline{\Delta e^2(n_\infty)}$  the momentum term  $\lambda$  was set to 0.8.

Initially the sub-filter length was set to  $\beta = M/2$ , i.e. 64. Since  $\Delta e^2(n)$  fell below the predefined threshold within a few iterations the sub-filter length was reduced to  $\beta = M/4 = 32$ . The value of  $\Delta e^2(n)$  became again smaller than  $\Delta e_o^2$  and the  $\beta$  was reduced again. The same procedure was repeated several times. After some sub-filter order reductions a value  $\beta_f$  was found for which  $\overline{\Delta e^2(n_\infty)} > \Delta e_o^2$ . This value in our case was  $\beta_f = 8$ . Since  $\beta_f < M/2$ , it was concluded that the unknown channel was sparse (Tab. 4). Moreover, observing that  $\theta_{1,5}^{(2)} \gg \theta_{1,i}^{(2)}$  for  $i = 1, \dots, 4, 6, \dots, 8$  when  $\beta = 16$  (Fig. 5(a)) it was found that the energy of the impulse response of the unknown channel was concentrated in the coefficients  $[\theta_{i_o,\beta}^* \dots \theta_{(i_o+1)\beta-1}^*] = [\theta_{80}^* \dots \theta_{95}^*]$ .

The squared error difference  $\overline{\Delta e^2(n)}$  for the different stages of the proposed method is illustrated in Fig. 4 as a function of time. In Fig. 5(a) the coefficients



**Fig. 5.** Time evolution of the coefficients of the second layer for the cases of sub-filter length  $\beta = 16$  and  $\beta = 4$ . In this layer the channel estimates of the sub-filters of the input layer are fused.

**Table 1.** Decisions of the consecutive steps proposed algorithm for sparsity detection. The decision threshold  $\Delta e_o^2$  was set to  $10^{-5}$ .

Step	$\beta$	$10 \log(\Delta e^2(n_\infty))$	$\Delta e^2(n_\infty) < \Delta e_o^2$
1	64	-283.3	Yes
2	32	-285.44	Yes
<b>3</b>	<b>16</b>	<b>-287.36</b>	<b>Yes</b>
4	8	10.43	No

of the second layer of the hierarchical filter sub-filter are illustrated for  $\beta = 16$ . It is observed that when  $\overline{\Delta e^2(n)}$  is decreasing with time the coefficient  $\theta_{1,i_o}^{(2)}$  of the second layer that converges to 1 indicates that the nonzero coefficients  $\theta_k^*$ , of the unknown channel, have indexes  $k = i_o\beta, \dots, (i_o + 1)\beta - 1$ . This does not hold when  $\overline{\Delta e^2(n_\infty)} > \Delta e_o^2$  as can be observed from Fig. 5(b) where  $\beta = 4$ .

## 5 Conclusions

A novel technique for the detection of sparseness in unknown channels has been proposed in this paper. This approach exploits the fact that the HLMS has significantly higher steady state error than the LMS algorithm unless the unknown channel is sparse and its nonzero coefficients lie in the vicinity spanned by a single sub-filter of the input layer. The core of the proposed system is an FIR and a hierarchical filter that perform in parallel. Depending on the design of the input layer of the hierarchical filter more efficient sparsity detection can be achieved. Simulations for the case of a sparse channel having an impulse response with only a few consecutive nonzero elements have been conducted to support the analysis.

## Acknowledgements

The authors would like to thank Dr Danilo Mandic, Dr Wai Leong, Mr Mo Chen and Ms Beth Jelfs for fruitful discussions and suggestions during Dr Boukis' visit to Imperial College, which has been instrumental towards communicating this idea.

## References

1. W.S. Burdic, *Underwater Systems Analysis*, Prentice-Hall: Englewood Cliffs, NJ, 1984.
2. J. Balakrishnan, W.A. Sethares and C.R. Johnson, Jr, *Approximate Channel Identification via  $\delta$ -Signed Correlation*, *Int. J. Adapt. Control Signal Proc.*, **16**:309-323, February 2002.
3. V. Nascimento, *Analysis of the Hierarchical LMS Algorithm*, *IEEE Signal Proc. Let.*, **10**(3):78-81, March 2003.
4. P. Stoica, M. Agrawal and P. Ahgren, *On the Hierarchical Least-Squares Algorithm*, *IEEE Communication Let.*, **6**(4):153-155, April 2002.
5. M. Macleod *Performance of the Hierarchical LMS Algorithm*, *IEEE Communication Let.*, **9**(12):436-437, December 2002.
6. R.K. Martin, W.A. Sethares, R.C. Williamson and C.R. Johnson, Jr, *Exploiting Sparsity in Adaptive Filters*, *IEEE Trans. Signal Proc.*, **50**(8):1883-1894, August 2002.
7. T.-K. Woo, *Fast Hierarchical Least Mean Square Algorithm*, *IEEE Signal Proc. Let.*, **8**(11):289-291, November 2001.
8. B. Widrow and S. Stearns. *Adaptive Signal Processing*. Prentice Hall, 1985.

# Auditory Feedback for Brain Computer Interface Management – An EEG Data Sonification Approach

Tomasz M. Rutkowski<sup>1</sup>, Francois Vialatte<sup>1</sup>, Andrzej Cichocki<sup>1</sup>,  
Danilo P. Mandic<sup>2</sup>, and Allan Kardec Barros<sup>3</sup>

<sup>1</sup> Laboratory for Advanced Brain Signal Processing  
Brain Science Institute RIKEN, Japan  
[tomek@brain.riken.jp](mailto:tomek@brain.riken.jp)

<http://www.bsp.brain.riken.jp/>

<sup>2</sup> Department of Electrical and Electronic Engineering  
Imperial College London, United Kingdom  
[d.mandic@imperial.ac.uk](mailto:d.mandic@imperial.ac.uk)

<http://www.commsp.ee.ic.ac.uk/>

<sup>3</sup> Laboratory for Biological Information Processing  
Universidade Federal do Maranhão, Brazil  
[allan@ufma.br](mailto:allan@ufma.br)

<http://pib.dee.ufma.br/>

**Abstract.** An auditory feedback for Brain Computer Interface (BCI) applications is proposed. This is achieved based on the so-called sonification of the mental states of humans, captured by Electro-Encephalogram (EEG) recordings. Two time-frequency signal decomposition techniques, the Bump Modelling and Empirical Mode Decomposition (EMD), are used to map the EEG recordings onto musical scores. This auditory feedback proves to have extremely high potential in the development of on-line BCI interfaces. Examples based on the responses from visual stimuli support the analysis.

## 1 Introduction

Brain Computer Interface (BCI) techniques have received much attention recently, owing to the exciting possibility of computer-aided communication with the outside world. This is achieved in a non-invasive manner, which poses several important and difficult challenges. In terms of signal processing these include the detection, estimation and interpretation of brain signals, and cross-user transparency [1]. It comes as no surprise, therefore, that this technology is envisaged to be at the core of future “intelligent” prosthetics, and is particularly suited to the needs of the handicapped and paralyzed. Other industries which would benefit greatly from the development of BCI include the entertainment, computer games, and automotive industries, where the control and navigation in a computer-aided application is achieved without resorting to using hands, or gestures. Instead, the onset of “planning an action” recorded from the head (scalp)

surface, and the relevant information is “decoded” from this information carrier. Notice this is notoriously difficult due to the “blind” nature of the problem, lack of any sort of feedback, and overwhelming noise presence within the signal. Apart from purely signal conditioning problems, in most BCI experiments other issues such as user training and adaptation, which inevitably causes difficulties and limits in wide spread of BCI technology due to the lack of “generality” caused by cross-user differences [2]. To help mitigate some of these issues, we propose to make use of an auditory feedback during BCI training or utilization which will inform the user about the “goodness” of brain activities. In our approach, experiments based on visual stimuli were conducted within the so-called Steady State Visual Evoked Potential (SSVEP) mode. Within this framework, the subjects are asked to focus their attention on simple flashing stimuli, whose frequency is known to cause a physiologically stable response present in EEG [3,4]. In the next step EEG signals are mapped into the auditory domain using two signal decomposition techniques one wavelet based and the other based on a decomposition onto a set of AM-FM basis functions. This way, the proposed multimodal BCI scheme uses the EEGs captured with several electrodes, subsequently preprocessed, and transformed into informative and pleasant artificial music, in order for the user to efficiently control the states of their mind (neurofeedback).

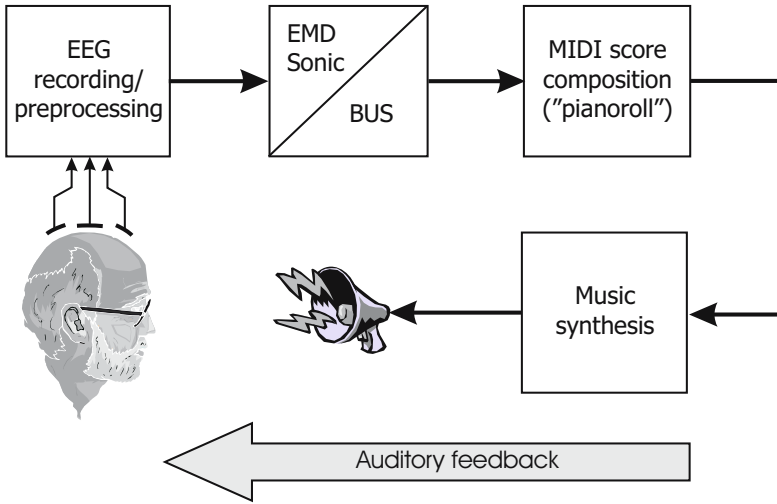
## 2 Methods

Sonification of EEG signals is a procedure in which electrical brain activity captured from human scalp is transformed into an auditory representation [5]. This paper proposes two novel approaches based on multimodal information fusion, whereby the so introduced “audio” modality provides perceptual feedback, but for which there is no unique generation procedure (see a conceptual diagram of the proposed approach in Figure 1).

We aim at looking at the level of detail (richness of information source) obtainable, and compare the usefulness of two signal decomposition approaches in this context. The first approach is based on a sparse signal representation in the Time-Frequency (TF) domain by standard wavelet transformations and is referred to as the Bump Modeling Sonification (BUS) technique. The second one is referred to as the Empirical Mode Decomposition Sonification (EMDSonic) technique, which rests on the identification of signal’s non-stationary and non-linear features, these also represent different “mind” modalities captured by the sensors. This novel method, as shown later, allows us to create slowly modulated tones representing changing brain activities.

### 2.1 Bump Modelling for Sparse EEG Sonification

A direct mapping of a signal or its wavelet TF representation onto a music file would produce a highly noisy result, this is due to the nature of EEG signals generation and recording techniques [6]. To reduce such artifacts, and provide a



**Fig. 1.** Block diagram of the proposed EEG sonification scheme for brain computer interfaces. An auditory feedback i.e. a mapping from EEG features onto a discrete set of musical variables, provides a convenient insight into the dynamics and patterns of EEG events.

simplified, yet rich in information representation, the bump modeling technique [7,8] has been proposed, which extracts interesting patterns from the TF maps. The main idea of this method is to approximate a TF map with a set of pre-defined elementary parameterized functions called bumps, whereby the map is represented by the set of parameters of the bumps. This provides a very sparse encoding of the map, resulting in information compression rates that range from a hundred to a thousand (further details about bump modeling are given in [7,8]). Prior to bump modeling, we compute the so-called “z-score” from the TF map [7]. This way, high normalized amplitude values represent “significant” patterns. Therefore, the bump modeling will extract “interesting” patterns of activity from the background containing the most relevant information in the EEG signal. The algorithm can be outlined in the following steps:

- (i) partition the map to define the zones to be modeled (those windows form a set of overlapping areas of the map);
- (ii) find the window that contains the maximum amount of energy;
- (iii) adapt a bump  $b$  to the selected zone, and withdraw it from the original map;
- (iv) should the amount of information modeled by the bumps reaches a threshold, stop; otherwise return to (iii).

To isolate “islands” of significant EEG activities, we used half-ellipsoid functions, defined by [8]:

$$\varphi_b(f, t) = \begin{cases} a\sqrt{1 - \nu} & \text{for } 0 \leq \nu \leq 1 \\ 0 & \text{for } \nu > 1 \end{cases} \quad (1)$$



where  $\nu = \left( e_f^2 + e_t^2 \right)$  with  $e_f = (f - \mu_f) / l_f$  and  $e_t = (t - \mu_t) / l_t$ . The variables  $\mu_f$  and  $\mu_t$  are the coordinates of the center of the ellipsoid,  $l_f$  and  $l_t$  are the half-lengths of the principal axes,  $a$  is the amplitude of the function,  $t$  and  $f$  the time and frequency index.

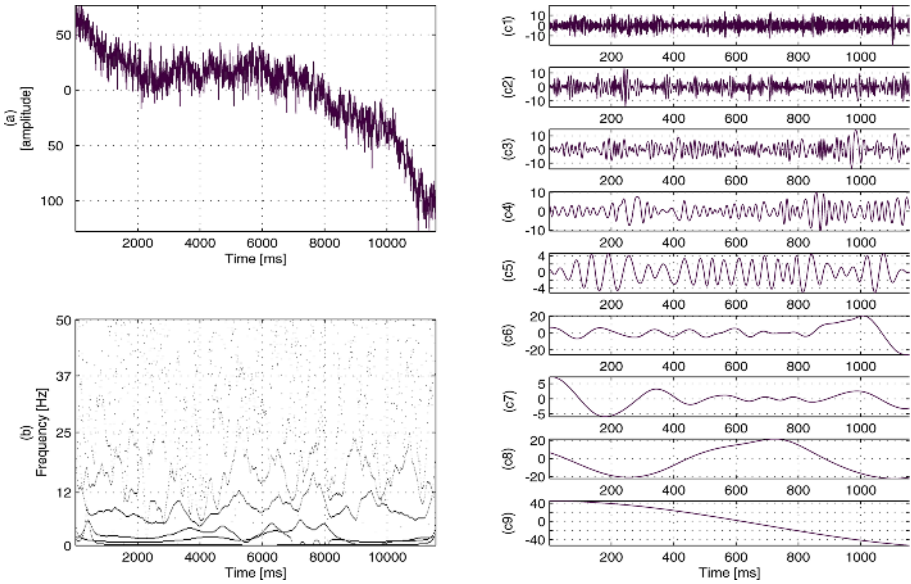
## 2.2 Empirical Mode Decomposition for EEG Sonification

Empirical Mode Decomposition (EMD) [10] utilizes empirical knowledge of oscillations intrinsic to a time series in order to represent them as a into superposition of components with well defined instantaneous frequencies. These components are called Intrinsic Mode Functions (IMF). IMFs, which should approximately obey the requirements of (i) completeness; (ii) orthogonality; (iii) locality; and (iv) adaptiveness. To obtain an IMF it is necessary to remove local riding waves and asymmetries, which are estimated from local envelope of minima and maxima of the waveform. The Hilbert spectrum for a particular IMF allows us to represent in the amplitude - instantaneous frequency - time plane. An IMF satisfies thus the two conditions: (i) in the whole data set, the number of extrema and the number of zero crossings should be equal or differ at most by one; (ii) at any point of IMF the mean value of the envelope defined by the local maxima and the envelope defined by the local minima should be zero. The technique of finding IMFs corresponds thus to finding limited-band signals. It also corresponds to eliminating riding-waves from the signal, which ensures that the instantaneous frequency will not have fluctuations caused by an asymmetric wave form. IMF in each cycle is defined by the zero crossings. Every IMF involves only one mode of oscillation, no complex riding waves are thus allowed. Notice that an IMF is not limited to be a narrow band signal, as it would be in traditional Fourier or wavelets decomposition, in fact, it can be both amplitude and frequency modulated at once, and also non-stationary or non-linear. The procedure to obtain IMF components from a signal, called sifting [10] and consists of the following steps:

- Identify the extrema of the signal waveform  $x(t)$ ;
- Generate “signal envelopes” by connecting local maxima by a cubic spline. Connect signal minima by another cubic spline;
- Determine the local mean,  $m_1$ , by averaging the two spline envelopes;
- Subtract the mean from the data to obtain:

$$h_i = x(t) - m_i; \tag{2}$$

- Repeat as necessary until there are no more possible IMF to extract.
- Proper IMF is a first component containing the finest temporal scale in the signal;
- The residue  $r_i$  should be generated by subtracting out proper IMF found from the data;
- The residue contains information about longer periods which will be further resifted to find additional IMFs.

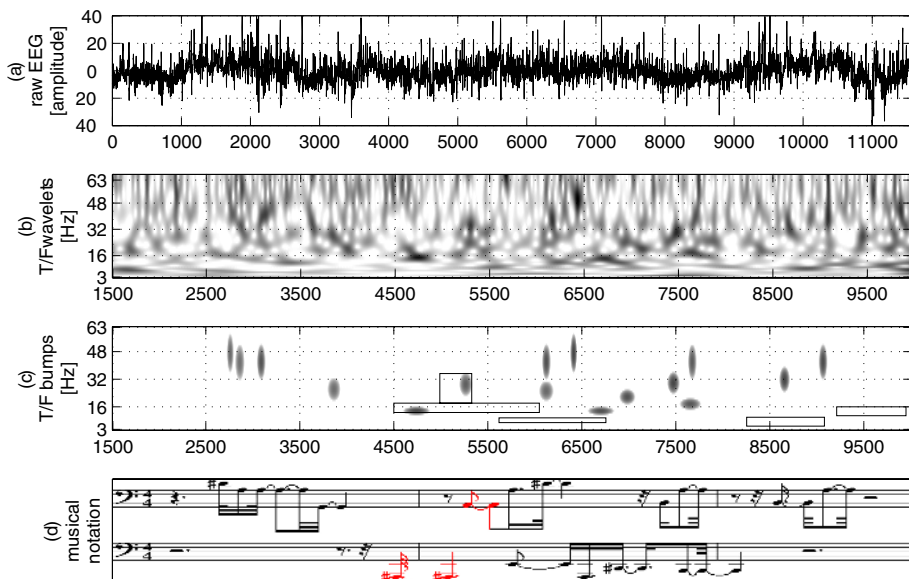


**Fig. 2.** An example of EEG signal decomposition using EMD technique. Clockwise from top left panel: (a) Raw EEG signal recorded from single frontal (Fp1) electrode. The panels (c1)-(c9) present nine IMF components extracted sequentially. IMFs represent oscillations occurring in EEG from higher to lower frequencies. (b) The oscillatory components ranging from 1Hz to 50Hz are combined in form of Hilber-Huang spectrum [9].

An example of EEG data decomposition using the above procedure is illustrated in Figure 2.2 where single channel EEG signal in panel (a) was decomposed into eight oscillatory components as in panel (b) of that figure. It is easy to spot a very low frequency component which represents very slow amplitude drift caused by amplifiers or a loosely connected reference electrode. The higher frequency oscillations are ordered into ascending components. Using the above procedure, EEG signals from chosen electrodes were decomposed separately forming subsets of IMF functions, from which low frequency drifts and high frequency spikes were further removed. From the obtained IMFs corresponding spectrograms were produced by applying the Hilbert transform to each component, as first suggested in [10]. The Hilbert transform allows us to depict the variable amplitude and the instantaneous frequency in the form of functions of frequency and time (in contrast to Fourier expansion, for example, where frequencies and amplitudes are fixed for its bases). Such an approach is very suitable for the non-stationary EEG and subsequent sonification results have slow changing frequency components indicating drifts in phase and amplitudes.

### 2.3 From Time-Frequency Representation into MIDI

The TF representations of EEG introduced in the above sections can be transformed into musical scores using MIDI procedures [11]. In both cases of EMDsonic or

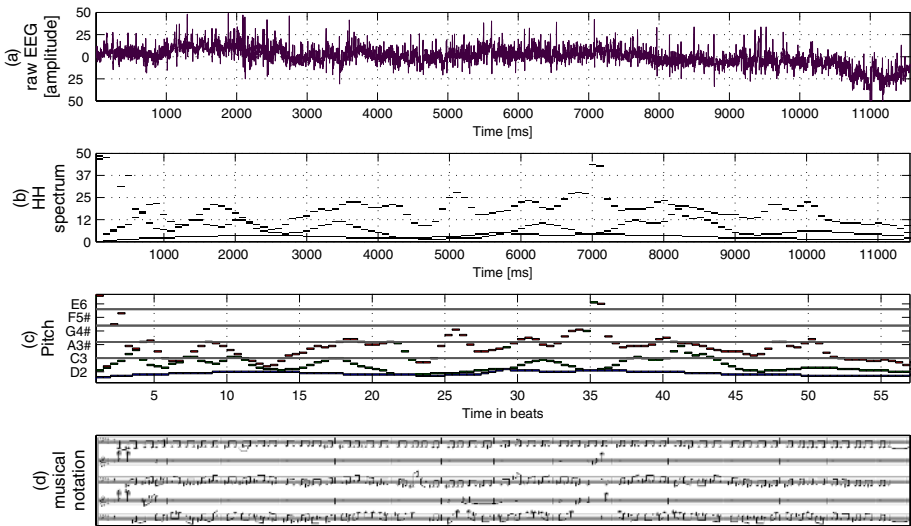


**Fig. 3.** An example of TF domain bumps modeling for EEG sonification. The above diagrams present: (a) The original raw EEG signal captured during a steady state visual evoked paradigm experiment; (b) wavelets representation the above EEG signal; (c) sparse bump modeling procedure of the above wavelets; (d) is a translation of obtained MIDI representation into classical musical sheet.

BUS TF maps, musical tones were obtained via so called “pianorolls” [11], which directly represent musical scores. In the case of EMDSonic technique, this mapping is performed as shown in Figure 4. (b) and (c), where the Hilbert-Huang spectrum of the original EEG signal is directly mapped by searching for ridges [12]. The location and duration of the ridges is transformed into musical notes with appropriate duration. In case of BUS analysis, the transformation into musical scores is performed as presented in Figure 3: (i) the velocity was obtained from the amplitude  $a$ ; (ii) the note pitch was obtained following a pentatonic scale (the pentatonic scale is based on five pitch values, for instance:  $60 - 63 - 65 - 67 - 70\text{Hz}$ ) from  $\mu_f$  (for instance, here 3Hz represents pitch 20Hz and 50Hz pitch 70Hz); (iii) the onset of the note was obtained from  $\mu_t$  using  $l_t/2$ ; (iv) the duration was computed from  $l_t$ .

### 3 Experiments

EEG sonification experiments were conducted for subjects performing SSVEP based BCI management. Subjects were asked to try to concentrate on a single flashing chessboard whose frequency was later recognized by an separate procedure. The EMDSonic or BUS algorithms were used to inform the user via auditory feedback about the level of concentration in every single trial, which



**Fig. 4.** An example of EMD domain oscillatory components modeling for EEG sonification. The above diagrams present: (a) The original raw EEG signal captured during a steady state visual evoked paradigm experiment; (b) Huang-Hilbert spectrum of the above EEG signal; (c) so called “piano-roll” composed from the Huang-Hilbert spectrogram (similarity of both diagrams in (b) and (c) shows the accuracy of presented EEG to midi sound transformation); (d) is a translation of obtained MIDI representation into classical musical sheet.

is necessary to accurately classify attentionality enhanced frequency of flashing stimuli. Results of EEG sonification using both procedures are depicted in Figure 3 and 4, where same EEG channel and trial was transformed into music. From the figures the differences between the two approaches become apparent.

## 4 Conclusions

We have presented two approaches to sonify EEG data for direct application in BCI environments. EMDSonic have shown novel and very interesting natural response in auditory domain due to very powerful ability to track slowly varying oscillations in EEG. In online application this approach also introduces delay related to data window analysis and simple decomposition, which is not destructive for monitoring slow cortical potentials in EEG [3]. For EMDSonic it was also easy to segregate MIDI scores into separate channels, later assigned to different instruments, due to filter banks alike EMD decomposition (see middle panel in Figure 2.2). On the other hand more traditional approach using wavelets together with still emerging bumps decomposition allowed us to create very sparse musical scores. Due to a very high computational cost, bump modeling is suitable only for offline EEG sonification. However this limit can be overcome by

an extraction of a TF masks, computed from several significant (e.g. training) trials. Both approaches are somehow complementary due to focus on different components in EEG and they provide insights into brain waves visualization and auditory feedback for BCI.

## References

1. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Braincomputer interfaces for communication and control. *Clinical Neurophysiology* **113** (2002) 767–791
2. Wolpaw, J.R., McFarland, D.J.: Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proceedings of National Academy of Sciences of the United States America* **101**(51) (2004) 17849–17854
3. Niedermeyer, E., Da Silva, F.L., eds.: *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. 5 edn. Lippincott Williams & Wilkins (2004)
4. Kelly, S.P., Lalor, E.C., Finucane, C., McDarby, G., Reilly, R.B.: Visual spatial attention control in an independent brain-computer interface. *IEEE Transactions on Biomedical Engineering* **52**(9) (2005) 1588–1596
5. Miranda, E., Brouse, A.: Interfacing the brain directly with musical systems: On developing systems for making music with brain signals. *LEONARDO* **38**(4) (2005) 331–336
6. Jovanov, E., Starcevic, D., Samardzic, A., Marsh, A., Obrenovic, Z.: EEG analysis in a telemedical virtual world. *Future Generation Computer Systems* **15** (1999) 255–263
7. Vialatte, F.: *Modelisation en bosses pour l'analyse des motifs oscillatoires reproductibles dans l'activite de populations neuronales : applications a l'apprentissage olfactif chez l'animal et a la detection precoce de la maladie d'Alzheimer*. PhD thesis, Paris VI University, Paris (2005) [http://www.neurones.espci.fr/Theses\\_PS/VIALATTE\\_F.pdf](http://www.neurones.espci.fr/Theses_PS/VIALATTE_F.pdf)
8. Vialatte, F., Cichocki, A., Dreyfus, G., Musha, T., Rutkowski, T., Gervais, R.: Blind source separation and sparse bump modelling of time frequency representation of EEG signals: New tools for early detection of Alzheimer's disease. In: *Proceedings of the 2005 IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, Mystic CT, USA, IEEE* (2005) 27–32
9. Rilling, G., Flandrin, P., Goncalves, P.: On empirical mode decomposition and its algorithms. In: *Proceedings of IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, NSIP-03, IEEE* (2003)
10. Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N.C., Tung, C., Liu, H.: The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **454**(1971) (1998) 903–995
11. Eerola, T., Toiviainen, P.: MIR in Matlab: The MIDI toolbox. In: *Proceedings of The 5th International Conference on Music Information Retrieval, ISMIR2004, Barcelona, Spain, Audiovisual Institute, Universitat Pompeu Fabra* (2004) 22–27
12. Przybyszewski, A., Rutkowski, T.: Processing of the incomplete representation of the visual world. In: *Proceedings of the First Warsaw International Seminar on Intelligent Systems, WISIS'04, Warsaw, Poland* (2004)

# Analysis of the Quasi-Brain-Death EEG Data Based on a Robust ICA Approach

Jianting Cao<sup>1,2</sup>

<sup>1</sup> Department of Electronic Engineering, Saitama Institute of Technology,  
1690 Fusaiji, Fukaya-shi, Saitama 369-0293, Japan  
cao@sit.ac.jp

<sup>2</sup> The Lab. for Advanced Brain Signal Processing, Brain Science Institute, RIKEN,  
2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan

**Abstract.** The brain-death is defined as the cessation and irreversibility of all brain and brain-stem function. A brain-death diagnosis is made according to precise criteria and in a well-defined process. Since the process of brain-death determination usually takes a longer time and with a risk (e.g. shortly remove the breath machine in a spontaneous respiration test), therefore, a practical, safety and rapid method is expected to be developed in the pre-test of the quasi-brain-death patient. This paper presents a practical EEG examination method associated with a robust data analysis method for the pre-testing of a quasi-brain-death patient. The developed EEG examination method is applied in the bedside of patient using a small number of electrodes. The developed single-trial data analysis method is used to reduce the power of additive noise and to decompose the overlapped brain and interference signals.

## 1 Introduction

Electroencephalogram (EEG) technique is used to a confirmatory test in the determination of brain death since it can evaluate the function of the cerebral cortex [4]. This paper presents a practical EEG examination method to pre-test of a quasi-brain-death patient. Since EEG measurement is applied in the bedside of patient in our method, the accuracy of EEG is influenced by environmental noise.

In this paper, we presents a robust data analysis method [1-3] to decompose the single-trial EEG raw data recorded from some quasi-brain-death patients. The developed method has two procedures. In the first stage, a robust pre-whitening technique with an additive noise reduction technique is presented. In the second stage, a robust nonlinear function derived by the parameterized  $t$ -distribution model is applied to decompose the mixtures of sub-Gaussian and super-Gaussian source components. The experimental results illustrate that the effectiveness of this robust data analysis method.

## 2 The Model and Method of EEG Data Analysis

### 2.1 A Practical Model of EEG Data Analysis

The model based on a practical EEG measurement can be formulated by

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \boldsymbol{\xi}(t), \quad t = 1, 2, \dots, \tag{1}$$

where  $\mathbf{x}(t) = [x_1(t), \dots, x_m(t)]^T$  represent the transpose of  $m$  electrodes at time  $t$ . Each electrode signal  $x_i(t)$  contains  $n$  common components (e.g. brain activities, interferences, etc.) represented by the vector  $\mathbf{s}(t) = [s_1(t), \dots, s_n(t)]^T$  and a unique component (sensor noise from electrode) which is a element in the vector  $\boldsymbol{\xi}(t) = [\xi_1(t), \dots, \xi_m(t)]^T$ . Since the source components are overlapped, and transferred rapidly to the electrodes, an element of the numerical matrix  $\mathbf{A} \in \mathbf{R}^{m \times n} = (a_{ij})$  can be consider as a quantity related to the physical distance between  $i$ -th sensor and  $j$ -th source. Based on this definition of Eq. (1), we noted that a source component  $s_i$  at least contributes to more than two sensors, and a noise component  $\xi_i$  contributes at most to only one sensor.

There are two kinds of noises have to be reduced or discarded in the EEG data analysis. The first kind of noise is called additive noise (unique component) which is generated from the individual sensor. The standard Independent Component Analysis (ICA) approach is usually failed to reduce such kind of noise. Therefore, we will apply the robust pre-whitening technique in the pre-processing stage to reduce the power of additive noise. The second kind of noise is a common component such as an interference. This kind of noise can be discarded after the independent source decomposition.

### 2.2 The Method of EEG Data Analysis

In this subsection, we will first apply the robust pre-whitening technique to reduce the power of additive noise. Next, we will apply the developed ICA algorithm based on the parameterized  $t$ -distribution density model to separate the mixture of sub-Gaussian and super-Gaussian signals[3].

Let us rewrite Eq. (1) in a data matrix form as

$$\mathbf{X}_{(m \times N)} = \mathbf{A}_{(m \times n)}\mathbf{S}_{(n \times N)} + \boldsymbol{\Xi}_{(m \times N)}, \tag{2}$$

where  $N$  denotes data samples. When the sample size  $N$  is sufficiently large, the covariance matrix of the observed data can be written as

$$\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T + \boldsymbol{\Psi}, \tag{3}$$

where  $\boldsymbol{\Sigma} = \mathbf{X}\mathbf{X}^T/N$ , and  $\boldsymbol{\Psi} = \boldsymbol{\Xi}\boldsymbol{\Xi}^T/N$  is a diagonal matrix. For convenience, we assume that  $\mathbf{X}$  has been divided by  $\sqrt{N}$  so that the covariance matrix can be given by  $\mathbf{C} = \mathbf{X}\mathbf{X}^T$ .

When  $\boldsymbol{\Psi}$  cannot be ignored in the model, we employ the cost function as

$$L(\mathbf{A}, \boldsymbol{\Psi}) = tr[\mathbf{A}\mathbf{A}^T - (\mathbf{C} - \boldsymbol{\Psi})][\mathbf{A}\mathbf{A}^T - (\mathbf{C} - \boldsymbol{\Psi})]^T. \tag{4}$$

Minimizing the cost function, we obtain the estimate

$$\hat{\Psi} = \text{diag}(\mathbf{C} - \hat{\mathbf{A}}\hat{\mathbf{A}}^T). \quad (5)$$

The estimate  $\hat{\mathbf{A}}$  can be obtained by  $\frac{\partial L(\mathbf{A}, \Psi)}{\partial \mathbf{A}} = 0$ . Here, we employ the eigenvalue decomposition

$$\hat{\mathbf{A}} = \mathbf{U}_n \mathbf{\Lambda}_n^{1/2}, \quad (6)$$

where  $\mathbf{\Lambda}_n$  is a diagonal matrix whose elements are the  $n$  largest eigenvalues of  $\mathbf{C}$ . The columns of  $\mathbf{U}_n$  are the corresponding eigenvectors.

Once the estimates  $\hat{\mathbf{A}}$  and  $\hat{\Psi}$  converge to stable values, we can finally compute the score matrix by using

$$\mathbf{Q} = [\hat{\mathbf{A}}^T \hat{\Psi}^{-1} \hat{\mathbf{A}}]^{-1} \hat{\mathbf{A}}^T \hat{\Psi}^{-1}. \quad (7)$$

Using the above result, the new transformation data can be obtained by  $\mathbf{z} = \mathbf{Q}\mathbf{x}$ .

It should be noted that it is necessary to use the robust pre-whitening technique to reduce the power of the noise with the decorrelation procedure. It is insufficient to obtain the independent components, since an orthogonal matrix in general contains additional degrees of freedom. Therefore, the remaining parameters need to be estimated by ICA method.

There several BSS/ICA methods can be used in EEG data analysis. For example, applying the natural gradient based approach [1], we obtain an updating rule as

$$\Delta \mathbf{W}(t) = \eta [\mathbf{I} - \varphi(\mathbf{y}(t))\mathbf{y}^T(t)] \mathbf{W}(t), \quad (8)$$

where  $\eta > 0$  is a learning rate,  $\mathbf{y}(t) = \mathbf{W}\mathbf{z}(t)$  is a vector of the decomposed component,  $\mathbf{W}$  is the demixing matrix. The activation functions derived by using the  $\mathbf{t}$ -distribution density model and the light-tailed distribution density model are[3]

$$\varphi_i(y_i) = \frac{(1 + \beta)y_i}{y_i^2 + \frac{\beta}{\lambda_\beta^2}}, \quad k_\beta = \hat{k}_i > 0, \quad (9)$$

$$\varphi_i(y_i) = \alpha \lambda_\alpha \text{sgn}(y_i) |\lambda_\alpha y_i|^{\alpha-1}, \quad k_\alpha = \hat{k}_i \leq 0, \quad (10)$$

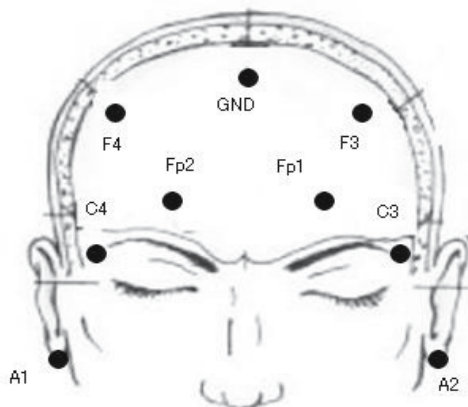
where  $\alpha$  and  $\beta$  are the parameters which control the shape of a family of distributions (sub-Gaussian or super-Gaussian),  $\lambda_\alpha$  and  $\lambda_\beta$  are scaling factors. The estimate of the kurtosis is defined by  $\hat{k}_i = \hat{m}_4 / \hat{m}_2^2 - 3$ , and it can be obtained by estimating the 2nd- and 4th-order moments as  $\hat{m}_j(t) = [1 - \eta] \hat{m}_j(t-1) + \eta y_i^j(t)$ ,  $j = 2, 4$ .

### 3 Experimental Conditions and Results

The EEG measurement was done in the Shanghai Huashan Hospital in affiliation with Shanghai Fudan University (P. R. China). The EEG data was directly recorded in the bedside of patient where the level of environmental noise was



higher. The EEG recording machine was a portable NEUROSCAN ESI system. In the system, a total of nine electrodes were planted on the scalp (see Fig. 1). The reference was set by using A1 + A2. The sampling rate of EEG data was 1000 Hz, and the resistances of electrodes were set under  $8\text{ k}\Omega$ .



**Fig. 1.** The layout of electrodes

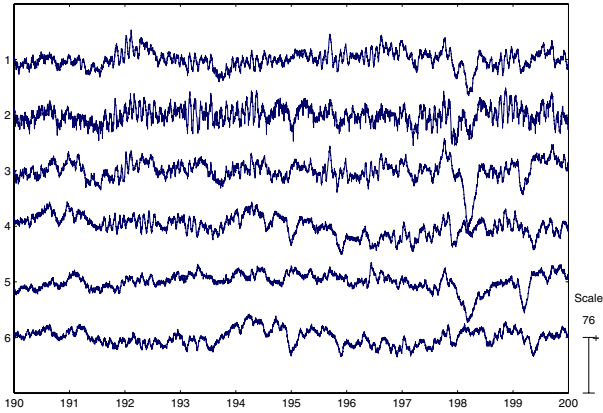
The total of 36 quasi-brain-death patients have been examined by using EEG from June 2004 to March 2006. In this paper, we focus on two typical cases of quasi-brain-death patients. The patients in both cases were in a deep-coma state at first. One patient was going to awake after the treatment, however another patient was going to brain death.

### 3.1 A Patient from Deep-Coma to Awake

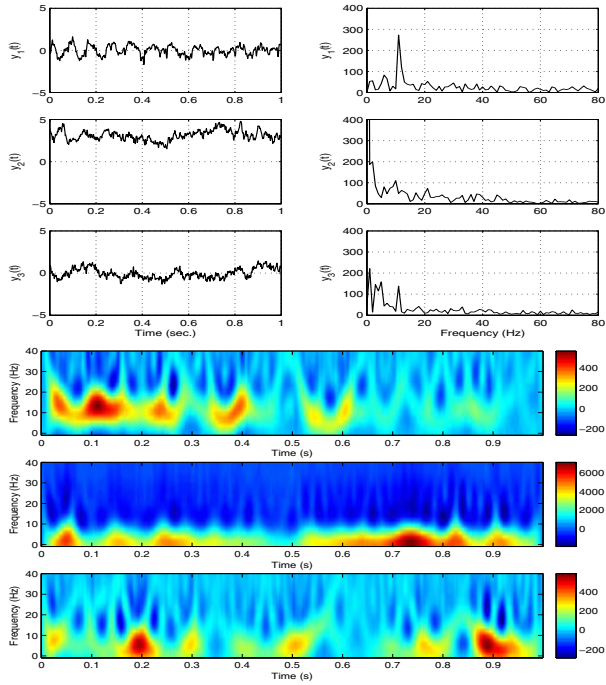
In the first example, a 18-year-old male patient with a primary cerebral disease was admitted to the hospital. After one month hospitalization, on June 22, 2004, the patient lost his consciousness and in a deep-coma state. His pupil was dilated, and the breath machine was used. The patient was completely unresponsive to external visual, auditory, and tactile stimuli and was incapable of communication. The symptom was very similar to a brain-death case.

At the same day, the EEG examination was applied in the bedside. The recorded EEG data in the first time was in five minutes (total in three times). As an example, a small time window of 10 seconds EEG is shown in Fig. 2 (a). Applying the developed data analysis method described in SECTION 2 to the recorded data, we obtained the result shown in Fig. 2 (b). In this result, a typical alpha-wave component with 12.5 Hz was decomposed successfully. When we assumed the number of the decomposed components is equal to the number of sensors, the similar result has been obtained.

Without loss of generality, we applied the same method to a random time window of recorded EEG data. The similar result to this example has been obtained.



(a) Recorded EEG data for the first patient



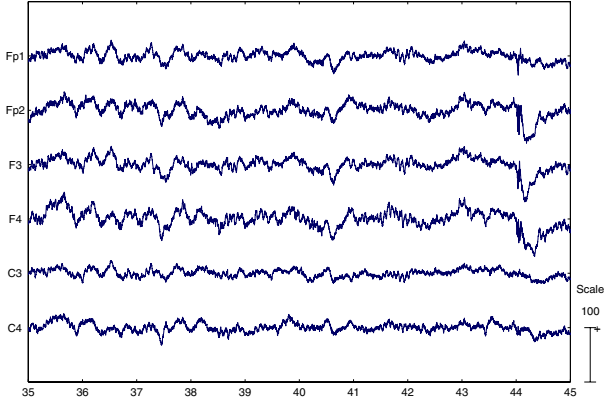
(b) Decomposed alpha-wave component in the time and frequency domains

**Fig. 2.** Recorded EEG data and analyzed results

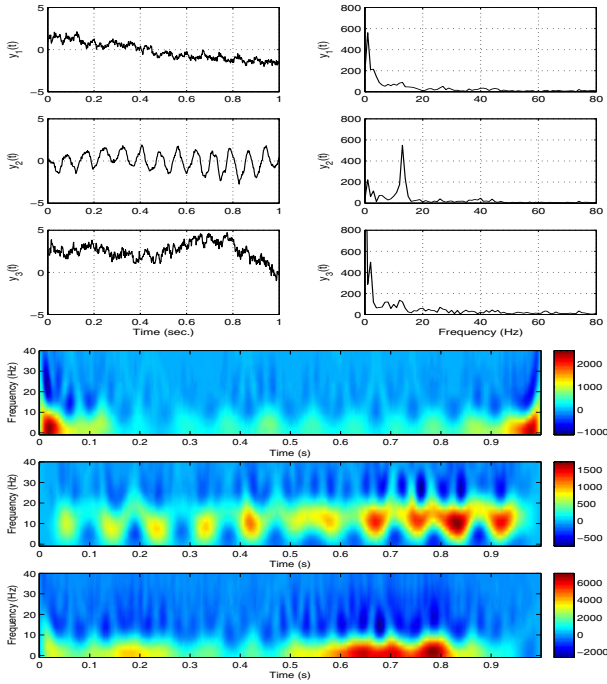
Although the symptom of patient was very similar to a brain-death case, our data analyzed result indicate that the patient had physiological brain activity. In fact, the patient came to consciousness little by little after that. On August 31 2004, the patient was able to respond to simple questions, and left the hospital.

### 3.2 A Patient from Deep-Coma to Brain-Death

In the second example, the patient was a 17-years-old female with a virus encephalitis. On March 16, 2005, the patient was in a deep-coma state, the pupil was dilated but had a very weak visual response, the breath machine was used.



(a) Recorded data for the second patient (first time EEG examination)

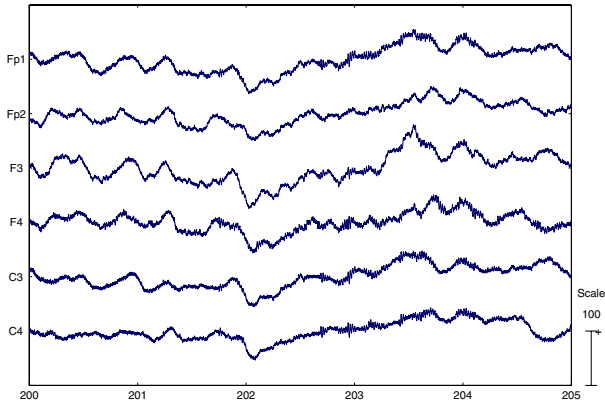


(b) Decomposed alpha-wave component in the time and frequency domains

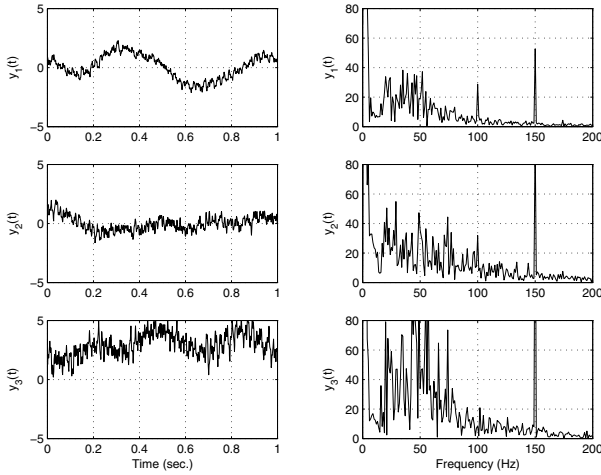
**Fig. 3.** EEG measured data and its analyzed results

The EEG examination in the bedside of patient was done in three times at same day. Each EEG recording was in five minutes. As an example, a time window of ten seconds EEG signals is plotted in Fig. 3 (a). Applying the developed data analysis method described in SECTION 2 to this recorded data, we obtained the result shown in Fig.3 (b). In this result, a typical alpha-wave component with 12.5 Hz was decomposed successfully.

Without loss of generality, we have applied the same method to the other recorded EEG data. The similar result (an alpha-wave component with 12.5 Hz) has been obtained as shown in Fig. 3. This result indicate that the patient has physiological brain activity which is identical to the clinical diagnosis (a very weak visual response appears in a few time).



(a) Recorded EEG data in the second time examination



(b) The decomposed interference components in the time or frequency domain

**Fig. 4.** EEG measured data and analyzed results

The second time EEG examination was done on March 22, 2005 (8 days after first examination). The clinical examination was that the patient is completely unresponsive to external visual, auditory, and tactile stimuli. Therefore, the diagnosis was given as a quasi-brain-death patient. The second time EEG examination was applied to the same patient at the same day. The recorded EEG data was in five minutes. As an example, a time window of five seconds EEG signals were shown in Fig. 4(a). Applying the data analysis method described in SECTION 2 to the recorded EEG data, we obtained the result shown in Fig.4 (b). In this result, three high-frequency components (interference components) were extracted, but no more alpha-wave component was recovered by analyzed result.

This data analysis result is identical to the clinical diagnosis in which given the term of the quasi-brain-death patient is because of the standard brain-death diagnosis is necessary to be made according to a precise criteria.

## 4 Conclusions

We have proposed a practical EEG examination method associated with a robust data analysis method to a quasi-brain-death patient. Through the analysis of real-measured patient's EEG data by this method, an alive deep-coma patient or a brain-death patient can be distinguished. Moreover, since the proposed method can be applied to the bedside of patient with a real-time measurement and real-time data analysis, it is very useful to the pre-test of a quasi-brain-death patient. We expect this method enable us to support the clinical doctor in avoid of the misleading diagnosis of brain death.

## Acknowledgments

The author would like to acknowledge Mr. Liangyu Zhao at Saitama Institute of Technology, Dr. Zhen Hong, Dr. Guoxian Zhu and Yue Zhang at Shanghai Huashan Hospital, Prof. Yang Cao and Prof. Fanji Gu at Fudan University, China for the EEG experiment and useful comments. This research project is supported by the Japan Society for the Promotion Science (JSPS) and the National Natural Science Foundation of China (NSFC) in the Japan-China Research Cooperative Program.

## References

1. Amari, S., Cichocki, A., Yang, H. H.: A New Learning Algorithm for Blind Signal Separation. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (eds.): *Advances in Neural Information Processing System*, Vol. 8. The MIT press, (1996) 757-763
2. Cao, J., Murata, N., Amari, S., Cichocki, A., Takeda, T.: Independent Component Analysis for Unaveraged Single-Trial MEG Data Decomposition and Single-Dipole Source Localization. *Neurocomputing* 49 (2002) 255-277
3. Cao, J., Murata, N., Amari, S., Cichocki, A., Takeda, T.: A Robust Approach to Independent Component Analysis with High-Level Noise Measurements. *IEEE Trans. on Neural Networks* 14 (3) (2003) 631-645
4. Taylor, R. M.: Reexamining the Definition and Criteria of Death. *Seminars in Neurology* 17 (1997) 265-270

# A Flexible Method for Envelope Estimation in Empirical Mode Decomposition

Yoshikazu Washizawa<sup>1</sup>, Toshihisa Tanaka<sup>2,1</sup>,  
Danilo P. Mandic<sup>3</sup>, and Andrzej Cichocki<sup>1</sup>

<sup>1</sup> Brain Science Institute, RIKEN,  
2-1, Hirosawa, Wako-shi, Saitama, 351-0198, Japan  
{washizawa, cia}@brain.riken.jp

<sup>2</sup> Department of Electrical and Electronic Engineering, Tokyo University of  
Agriculture and Technology,  
2-24-16, Nakacho, Koganei-shi, Tokyo, 184-8588, Japan  
tanakat@cc.tuat.ac.jp

<sup>3</sup> Department of Electrical and Electronic Engineering, Imperial College London,  
SW7 2BT, United Kingdom  
d.mandic@imperial.ac.uk

**Abstract.** A flexible and efficient method for finding the envelope within the empirical mode decomposition (EMD) is introduced. Unlike the existing (deterministic) spline based strategy, the proposed envelope is a result of an optimisation process and sought as a minimum of a quadratic cost function. A closed form solution of this optimisation problem is obtained and it is shown that by choosing free parameters, we can fine-tune the frequency resolution or the number of intrinsic mode functions (IMFs) as well as the shape of the envelopes. Computer simulations on both the synthetic and real-world electro-encephalogram (EEG) data support the analysis.

## 1 Introduction

The empirical mode decomposition (EMD) proposed by Huang et. al. in 1998 [1] is a technique for the analysis of non-linear and non-stationary signals in the time-frequency domain and its applications are manifold. EMD decomposes a signal  $x(t)$  into its components called intrinsic mode functions (IMFs)  $c_i(t)$ ,  $i = 1, 2, \dots, n$ , and the residual  $r(t)$ , in the following way:

$$x(t) = \sum_{i=1}^n c_i(t) + r(t), \quad (1)$$

The idea behind this approach is that every IMF has a very narrow frequency band all time, which allows us to produce a time-frequency spectrum called Hilbert-Huang (HH) spectrum by using the Hilbert transform (for more details, see [2]). The spectrum is sharp and clear curves unlike the Wavelet or short time Fourier spectrum. As a result, to so defined spectrum exhibits well defined

spectral components. EMD needs no a priori knowledge on the signal and as such belongs to the class of Exploratory Data Analysis techniques [3].

EMD is totally characterized by so called “upper” and “lower” envelopes of an input signal. Despite the enormous interest in this technique by the ocean modeling and each sciences research communities, and its huge potential in signal processing, little attention has been paid to and “optimal” choice of these envelopes, and instead the originally proposed spline based techniques have been employed [1]. Well-known interpolations of maxima or minima such as cubic spline are often used [1]. The so-generated EMD with conventional interpolations automatically determine the number of IMFs, or “the resolution of frequency.” Therefore a practical problem is how to control this “frequency resolution.”

To solve this problem, we propose a class of envelopes for EMD which minimizes a quadratic penalty cost function involving cost parameters, which allows us to control the bandwidth of IMFs. Since these parameters are weight coefficients in the frequency-domain. This enables a very convenient adjustment of the frequency resolution. In practical terms, if we assume wider bandwidth, fewer IMFs are needed, whereas for narrower bandwidths, we obtain more IMFs, this way can control the number of IMFs and more suitable time-frequency spectrum depending on an application.

## 2 Empirical Mode Decomposition

EMD decomposes a given signal  $x$  into a number of IMFs, which have the following properties:

1. Along the signal, the number of extrema and the number of zero crossings must either be equal or differ at most by one;
2. At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

Huang et. al. showed [1] that such an IMF has a very narrow frequency band. Signals such as AM, FM, AM-FM are a natural candidate for an IMF.

For given signal  $x$ , EMD decomposes it as in eq. (1) by the so called sifting process given by:

1. Let  $\mathcal{I}$  be a index set of IMFs. Initialize  $\mathcal{I} = \phi$  (empty set).
2. **While**  $(x - \sum_{i \in \mathcal{I}} c_i)$  has extremum,
  - (a) Set  $h = x - \sum_{i \in \mathcal{I}} c_i$
  - (b) **While**  $h$  is not IMF.
    - Detect local maxima and minima. Interpolate them by cubic splines. Let  $u$  and  $l$  be respectively the upper and the lower envelope.
    - Set  $h \leftarrow h - \frac{1}{2}(u + l)$
  - (c) Add  $h$  to the set of IMF.

The stopping criterion of IMF in 2.(b) is a crucial problem which in [1] was solved by using the standard deviation (SD):

$$SD = \sum_t \frac{|h_{old}(t) - h_{new}(t)|^2}{h_{old}^2(t)}. \tag{2}$$

In [1] a typical heuristic value for this stopping criterion was set to a value between 0.2 and 0.3.

### 3 Definition of Envelopes by Quadratic Penalty Function

Let  $f$  be a signal to be decomposed,  $\{t_i^u\}$  and  $\{t_j^l\}$  ( $i = 1, \dots, N_u, j = 1, \dots, N_l$ ) be sets of time instances of maxima and minima, respectively.

Without loss in generality, we shall consider the upper envelope (the lower envelope can be analyzed the same way). The problem to find  $u$  can be formulated as a classical linear estimation problem from a set of samples  $\{f(t_i^u)\}$ . Specifically, we assume that the sampling of  $u$  was performed as,

$$\sum_{i=1}^{N_u} (e_i \otimes \overline{k(\cdot, t_i^u)})u = \begin{bmatrix} f(t_1^u) \\ \vdots \\ f(t_{N_u}^u) \end{bmatrix}, \tag{3}$$

where  $k(\cdot, \cdot)$  is the reproducing kernel of the signal space,  $(\cdot \otimes \bar{\cdot})$  is the Neumann-Shatten product defined by  $(a \otimes \bar{b})c = \langle c, b \rangle a$ , and  $e_i$  is the  $i$ -th natural basis in  $\mathbb{R}^{N_u}$ . Let  $A = \sum_{i=1}^{N_u} (e_i \otimes \overline{k(\cdot, t_i^u)}) = A$  and  $f_v = \begin{bmatrix} f(t_1^u) \\ \vdots \\ f(t_{N_u}^u) \end{bmatrix}$ . Then (3) is reduced to

$$Au = f_v. \text{ If } u \text{ is a discrete consisting of } N \text{ sample, then } A \in \mathbb{R}^{N_u \times N} \text{ is explicitly expressed as}$$

$$A = \sum_{i=1}^{N_u} (e_i \hat{e}_{t_i^u}^\top), \tag{4}$$

where  $\hat{e}_i$  is an  $i$ -th natural basis in  $\mathbb{R}^N$ .

The upper envelope is given by a solution of the above linear formula to yield

$$u = A^\dagger f_v + w, \quad w \in \mathcal{N}(A), \tag{5}$$

where  $A^\dagger$  is the Moore-Penrose generalized inverse of  $A$  and  $\mathcal{N}(A)$  is the null space of  $A$ . If  $AA^* = I$ ,  $A$  is a partial isometry matrix i.e.,  $A^\dagger = A^\top$ . We still have an ambiguity on  $u$ , that is  $w$  an arbitrary vector. The following discussion is devoted to determine the unique  $w$ .

Let  $\mathcal{F}$  be a Fourier transform operator and  $U$  be a Fourier transform of  $u$ , that is

$$U = \mathcal{F}u = \sum_{t=0}^{N_u-1} u(t) \overline{\exp(i\xi t)}. \tag{6}$$



We can now introduce the following cost function,

$$J = \sum_{i=1}^N U(\xi_i) \overline{LU(\xi_i)} = \langle U, LU \rangle, \tag{7}$$

where  $L$  is a self adjoint cost operator defined as,

$$L = \text{diag}(\boldsymbol{\rho}) = \begin{bmatrix} \rho_1 & & 0 \\ & \ddots & \\ 0 & & \rho_N \end{bmatrix}. \tag{8}$$

and  $\{\xi_i\}_{i=1}^N$  is a set of discretized frequencies such that  $\xi_i < \xi_j$  if  $i < j$ .

The underlying idea behind this cost function is that an envelope should be smooth, i.e., its frequency spectrum should be biased to its lower part. Different definitions of  $\rho_i$  result in different envelopes. For example,  $\boldsymbol{\rho}$  can be given as follows:

1.  $\boldsymbol{\rho} = (\underbrace{0 \dots 0}_n \underbrace{1 \dots 1}_{N-2n} \underbrace{0 \dots 0}_n)$
2.  $\boldsymbol{\rho} = (1^\alpha \ 2^\alpha \ \dots \ (N/2)^\alpha \ (N/2)^\alpha \ \dots \ 1^\alpha)$
3.  $\boldsymbol{\rho} = (e^\alpha e^{2\alpha} \dots e^{N\alpha/2} e^{N\alpha/2} \dots e^\alpha)$

Intuitively, the case 1 suggests that high frequency components of an envelope are mostly suppressed. The second and third cases imply that a higher components of an envelope cost the higher because the shape of frequency spectrum tends to decays polynomially or exponentially. Various  $\boldsymbol{\rho}$  can be alternated during a sifting process.

In summery, the optimization problem to be solved here is:

$$\begin{aligned} \min_u \quad & J = \langle \mathcal{F}u, L\mathcal{F}u \rangle, \tag{9} \\ \text{subject to} \quad & u = A^\dagger f_v + w, \quad w \in \mathcal{N}(A). \end{aligned}$$

**Theorem 1.** *Optimization problem (9) is minimized when*

$$\begin{aligned} u &= A^\dagger f_v \\ &- (I - A^\dagger A) ((I - A^\dagger A)^* \mathcal{F}^* L\mathcal{F} (I - A^\dagger A))^\dagger (I - A^\dagger A)^* \mathcal{F}^* L\mathcal{F} A^\dagger f_v. \end{aligned} \tag{10}$$

(Proof). We can change the constraint to

$$u = A^\dagger f_v + (I - A^\dagger A)w, \tag{11}$$

without loss of generality since  $(I - A^\dagger A)$  is the projection operator onto  $\mathcal{N}(A)$  i.e,  $\mathcal{N}(A) = \mathcal{R}(I - A^\dagger A)$ .

By subsisting  $u$  as in (11) to (9), we have

$$\begin{aligned} J &= \langle \mathcal{F}(A^\dagger f_v + (I - A^\dagger A)w), L\mathcal{F}(A^\dagger f_v + (I - A^\dagger A)w) \rangle \\ &= \langle \mathcal{F}A^\dagger f_v, L\mathcal{F}A^\dagger f_v \rangle + \langle w, (I - A^\dagger A)^* \mathcal{F}^* L\mathcal{F}A^\dagger f_v \rangle \\ &\quad + \langle (I - A^\dagger A)^* \mathcal{F}^* L\mathcal{F}A^\dagger f_v, w \rangle + \langle w, (I - A^\dagger A)^* \mathcal{F}^* L\mathcal{F}(I - A^\dagger A)w \rangle. \end{aligned}$$

Then the vector which has to be optimized is changed to  $w$ .

Since  $(I - A^\dagger A)^* \mathcal{F}^* L\mathcal{F}(I - A^\dagger A)$  is non-negative definite,  $w$  is given as

$$w = -(I - A^\dagger A)((I - A^\dagger A)^* \mathcal{F}^* L\mathcal{F}(I - A^\dagger A))^\dagger (I - A^\dagger A)^* \mathcal{F}^* L\mathcal{F}A^\dagger f_v. \tag{12}$$

□

### 4 Experiments

Two experimental results illustrate the effectiveness and flexibility of the proposed method. We used the publicly available MATLAB codes [4]-[6] to compare our proposed EMD with the conventional one.

First, to illustrate the behavior of the envelopes given in (10), we generated a synthetic signal,

$$x(t) = \sin\left(\frac{2}{40}\pi t\right) \sin\left(\frac{2}{400}\pi t\right), \tag{13}$$

where  $x(t)$  is discrete signal with  $t = 1, \dots, 800$ . We set  $\rho = (1^\alpha \ 2^\alpha \ \dots \ 400^\alpha \ 400^\alpha \ \dots \ 1^\alpha)$ . The envelopes obtained from the proposed method with variable  $\alpha$  are shown in Figure 1. It can be observed that a larger  $\alpha$  yields a smoother envelope. We can also observe from Figure 1 the various shapes of the mean of upper and lower envelopes. When  $\alpha = 1$ , variation in the mean signal is considerable, which implies that the sifting process removes high frequency components. In that case, the bandwidth of IMF is narrow, and the number of IMFs is large.

Next, we applied the proposed method to a real biomedical signal. We used a single channel of electroencephalogram (EEG). The experiments was conducted by Rutkowski et. al. within the so-called Steady State Visual Evoked Potential (SSVEP) mode [7]. Within this framework, the subjects are asked to focus their

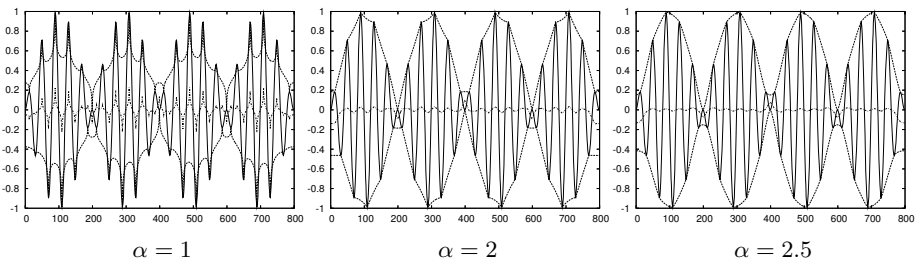


Fig. 1. Original signals, envelopes and means of envelopes

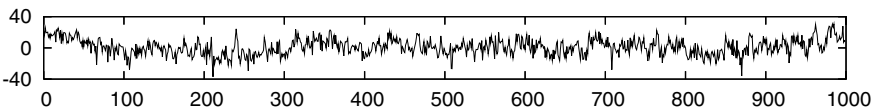
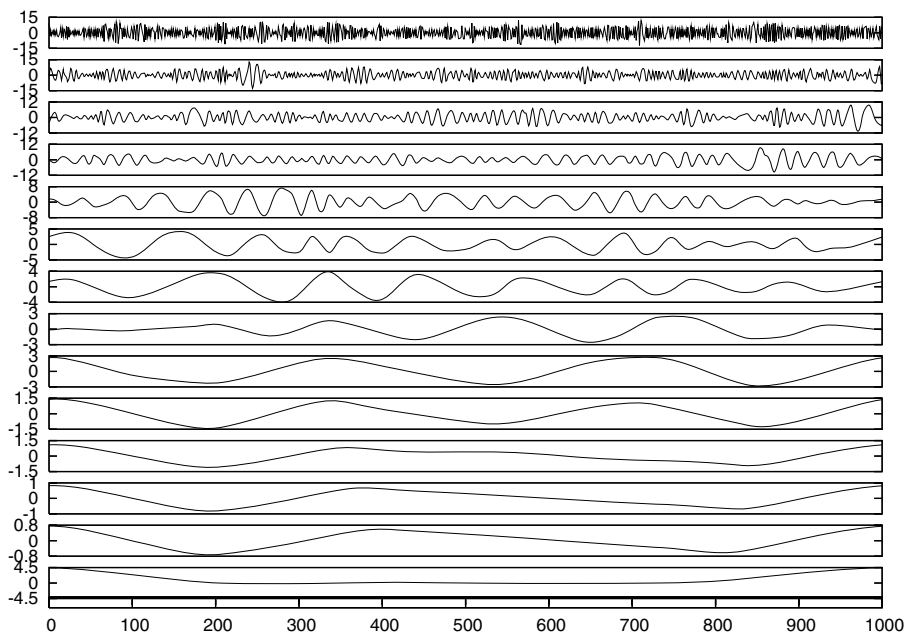
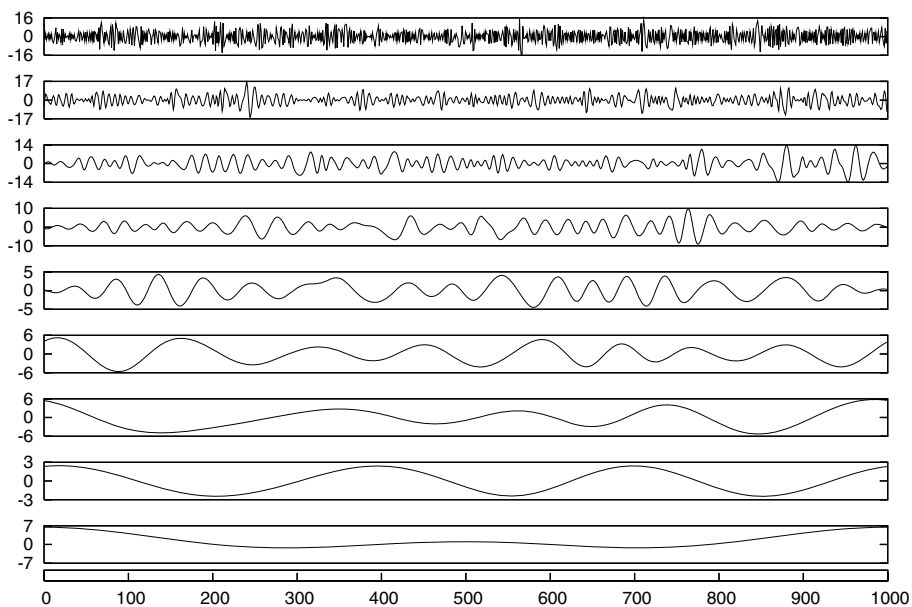


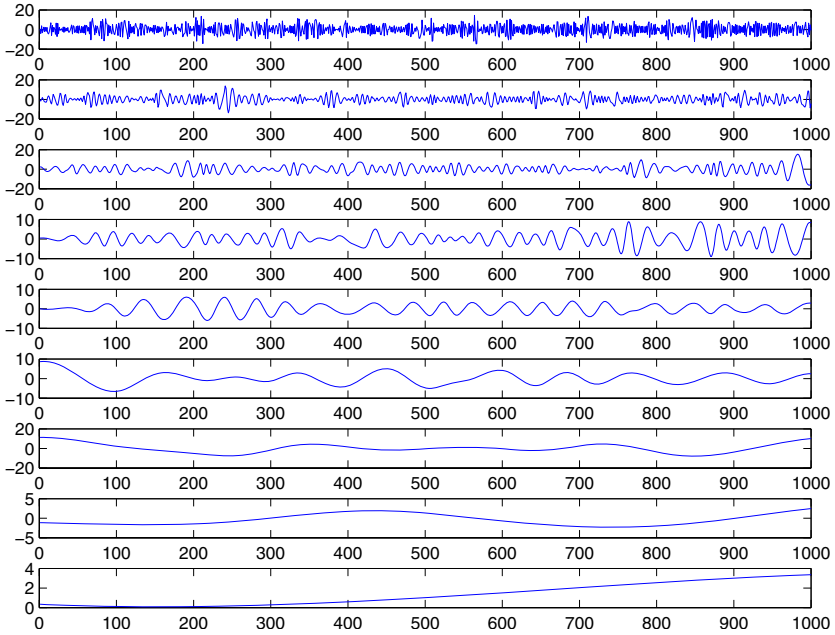
Fig. 2. A single channel of EEG signal



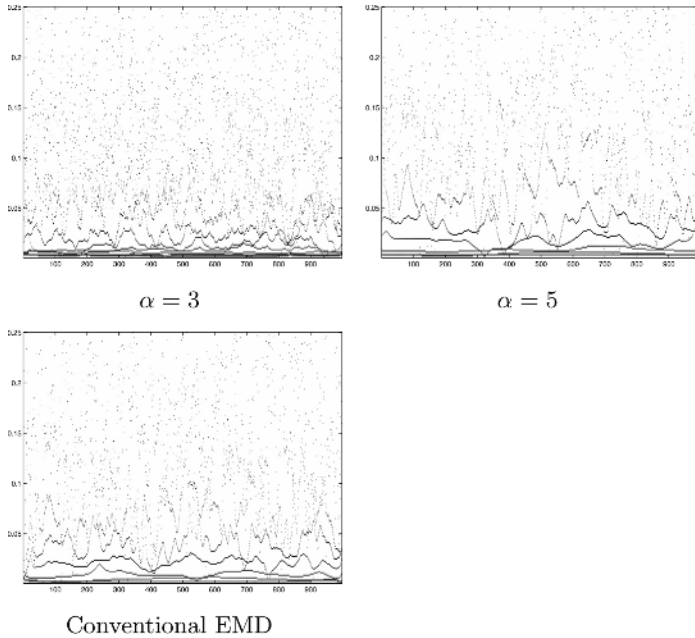
**Fig. 3.** IMFs of the EEG signal (Proposed method  $\alpha = 3$ )



**Fig. 4.** IMFs of the EEG signal (Proposed method  $\alpha = 5$ )



**Fig. 5.** IMFs of the EEG signal (Conventional method)



**Fig. 6.** Comparison of Hilbert-Huang spectra (Vertical axis: normalized frequency, horizontal axis: sample)

attention on simple flashing stimuli, whose frequency is 1/26Hz. The stimuli is given from one second after (205 sample). The EEG signal under this subject was recorded by an A/D converter with a sampling rate of 2048Hz and a bit depth of 24bits, followed by a down-sampler to 204.8Hz. Figure 2 shows this signal.

We tested three values of  $\alpha$ ,  $\alpha = 3, 5, 7$ , with  $\rho = (1^\alpha \ 2^\alpha \ \dots \ 500^\alpha \ 500^\alpha \ \dots \ 1^\alpha)$ . Figures 3, 4 and 5 compare the IMFs obtained by the proposed method with those from the conventional EMD with cubic spline. When  $\alpha = 3$ , the number of IMFs was 14, while when  $\alpha = 5$ , the number of IMFs was 9. The number of IMFs of conventional EMD was 8. These results imply that to tune parameter  $\rho$ , we can control the number of IMFs and the bandwidth of IMF.

Figure 6 shows the Hilbert-Huang spectrum of the EEG signal. For  $\alpha = 3$  observe a dense and higher resolution spectrum, while for  $\alpha = 5$ , the spectrum is rough and with lower resolution.

## 5 Discussion and Conclusions

A new approach for the derivation of the envelopes within the empirical mode decomposition (EMD) method has been proposed. It allows us to control the frequency bandwidth of IMFs, their number, and the resolution of the Hilbert-Huang spectrum.

In the proposed method, to obtain the solution, we have to compute a generalized inverse to  $\mathbb{R}^N$  of which dimension equals to the number of samples. It requires heavy computation for the inverse, which is a subject of our follow-up work. In our experiments, we used fixed cost parameter  $\rho$ . Indeed, it can be changed during sifting process. Various  $\rho$  will also give different and interesting results. For example, by changing  $\rho$ , higher resolution in low frequency parts in Hilbert-Huang spectrum could be obtained. Efficient and useful methods to alter  $\rho$  would be addressed in the near future.

## References

1. Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N-C., Tung, C. C., Liu, H. H.: Empirical Mode Decomposition Method and the Hilbert Spectrum for Non-stationary Time Series Analysis. Proc. Roy. Soc. London, A454, (1998), 903-995.
2. Hahn, S. L.: Hilbert Transforms in Signal Processing. Artech House Publishers, 1996.
3. Tukey, J. W.: Exploratory data analysis. Addison-Wesley, 1977.
4. The time-frequency toolbox. <http://tftb.nongnu.org/>
5. Rilling, G., Flandrin, P., Gonçalvès, P.: MATLAB codes for EMD. <http://perso.ens-lyon.fr/patrick.flandrin/emd.html>
6. Lambert, M., Engroff, A., Dyer, M., Byer, B.: MATLAB codes for EMD. <http://www.owl.net.rice.edu/elec301/Projects02/empiricalMode/code.html>
7. Rutkowski, T. M., Vialatte, F., Cichocki, A., Mandic, D. P., Barros, A. K.: Auditory Feedback for Brain Computer Interface Management An EEG Data Sonification Approach. Proc. of 10th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES2006) in this volume.

# The Performance of LVQ Based Automatic Relevance Determination Applied to Spontaneous Biosignals

Martin Golz and David Sommer

University of Applied Sciences Schmalkalden, 98574 Schmalkalden, Germany

**Abstract.** The issue of Automatic Relevance Determination (ARD) has attracted attention over the last decade for the sake of efficiency and accuracy of classifiers, and also to extract knowledge from discriminant functions adapted to a given data set. Based on Learning Vector Quantization (LVQ), we recently proposed an approach to ARD utilizing genetic algorithms. Another approach is the Generalized Relevance LVQ which has been shown to outperform other algorithms of the LVQ family. In the following we present a unique description of a number of LVQ algorithms and compare them concerning their classification accuracy and their efficacy. For this purpose a real world data set consisting of spontaneous EEG and EOG during overnight-driving is employed to detect so-called microsleep events. Results show that relevance learning can improve classification accuracies, but do not reach the performance of Support Vector Machines. The computational costs for the best performing classifiers are exceptionally high and exceed basic LVQ1 by a factor of  $10^4$ .

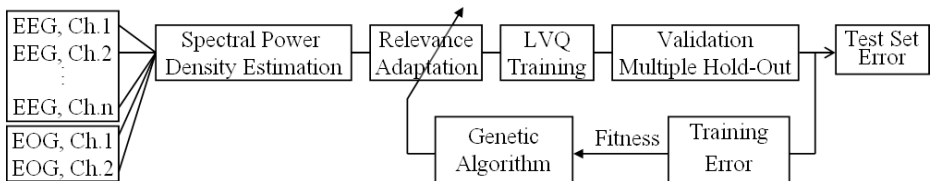
**Keywords:** Automatic Relevance Determination, Learning Vector Quantization, Support Vector Machines, Electroencephalogram.

## 1 Introduction

In many data fusion applications it was shown that combining heterogeneous sources is necessary in order to process much more information and to improve classification accuracy. In case of detecting so-called microsleep events (MSE), which are observed as attention lapses and prolonged eye lid closures during overnight driving simulations, we reported that fusion on the feature level of different sources contributes to improve the classification accuracy and the stability of the classifier [1]. Problems arise for large number of features, because all non-parametric local classification methods have fundamental problems due to the so-called "curse of dimensionality", i.e. the performance deteriorates when going to higher dimensions in the input space [2]. In this respect, simple local algorithms such as the nearest-neighbour classifier suffer more than non-local learning algorithms such as support vector machines (SVM). Note that for a given high-dimensional input vector the nearest neighbour is not much closer than other input vectors, or in other words, the ratio of the distance between the nearest and the farthest distance converges to one [2].

One way to overcome these difficulties is pruning of irrelevant features after relevances have been gained with respect to the classification task. Models based on Bayesian statistics were proposed by MacKay [3] and Neal [4] under the terminology of Automatic Relevance Determination (ARD). ARD is also attractive as it yields simpler and more interpretable models. It is a kind of knowledge extraction in applications where the importance of features is unknown. This is also the case of MSE detection where up to now no consistent expert knowledge is available. Most known facts are related to fatigue and are not appropriate for MSE detection. Based on Learning Vector Quantization (LVQ) [5], which is a widely used and very intuitive approach to classification, three methods of ARD has been introduced in the last decade, namely distinction sensitive LVQ (DSLQV) [6], relevance LVQ (RLVQ) [7] and generalized relevance LVQ (GR-LVQ) [8]. All methods define a diagonal metric in input space which is adapted during training according to plausible heuristics. Moreover, GR-LVQ benefits of a gradient dynamics on an appropriate error function. It generalizes RLVQ which is based on simple Hebbian learning and which showed worse and instable results on real world data [8].

The adaptation schemes which adjust weighting factors constitute a method for determining the intrinsic dimensionality of the data. The weighting factors can be regarded as relevance values. Dimensions with zero weight have no influence on the distances and are not relevant. Hence, the dimensions which possess the smallest relevance values are ranked as least important, i.e. they can be removed. In general, an input space dimension as small as possible is desirable for the above mentioned methods, for the sake of efficiency, accuracy, and simplicity of neural network processing.



**Fig. 1.** Automatic Relevance Detection scheme using LVQ1 classifiers and a genetic algorithm optimizing feature weight factors in order to minimize the mean training errors [1]

In the same line, we proposed an adaptive metric optimization approach (in the following labeled as 'GA-OLVQ1') based on the fast converging and robust OLVQ1 algorithm and a genetic algorithm (Fig. 1) [1]. Data of different sources, in our case several channels of electroencephalogram (EEG) and two channels of electrooculogram (EOG), are pre-processed and afterwards features are extracted using spectral power density estimation techniques. For each feature an individual weight factor is assigned. In the subsequent step of classification the weights are used in the distance calculation between input and prototype vectors

using weighted Euclidian metric. The classification accuracy, estimated by multiple hold-out validation of the trained networks, serves as fitness measure of a genetic algorithm. Consequently, test set errors are not used, directly and indirectly, for any step of optimization. The genetic algorithm generates populations of LVQ networks with different sets of feature weighting factors. At the end of this kind of optimization a population of well fitted LVQ networks remains. Over the ten best fitting individuals, ranked by their training errors, the weight factors are averaged. These are the final relevance values.

The purpose of this paper is to present a unique description of several algorithms of the LVQ family (section 2) and to compare their classification accuracy and the efficacy of ARD algorithms to algorithms with non-adaptive metric on a real world data set (section 4). The data set consists of spontaneous EEG and EOG during driving and is introduced in section 3. The paper is closing by some conclusions.

## 2 The Family of LVQ Algorithms

Given a finite training set of feature vectors  $x^i = (x_1, \dots, x_n)^T$  assigned to class labels  $y^i : S = \{(x^i, y^i) \in \mathbb{R}^n \times \{1, \dots, C\} | i = 1, \dots, m\}$  where  $C$  is the number of different classes. And given a set of randomly initialized prototype vectors  $w^i$  assigned to class labels  $c^i : W = \{(w^i, c^i) \in \mathbb{R}^n \times \{1, \dots, C\} | i = 1, \dots, p\}$ . The goal of all LVQ algorithms is to adapt the prototypes in order to yield good generalization, i.e. to make the probability of misclassification of undrawn items of a given population as small as possible.

Execute the following steps repetitively:

1. Select randomly  $(x^i, y^i) \in S$  and calculate the squared distances  $d_A = \|x^i - w^A\|_{r^o}^2$  and  $d_B = \|x^i - w^B\|_{r^o}^2$  of  $x^i$  to the nearest and second nearest prototype vector  $w^A$  and  $w^B$ , respectively, using a weighted Euclidian metric:  $\|x - w\|_{r^o} = \sqrt{\sum_{k=1}^n r_k^o |x_k - w_k|^2}$ , where  $r^o = (r_1, \dots, r_n)^T$  is a normalized weight vector containing the relevance values  $r_1, \dots, r_n$ .  
 With respect to the class labels  $c^A, c^B$  assigned to  $w^A, w^B$  resp., and the label  $y^i$  assigned to  $x^i$ , four different cases are possible: (a)  $c^A = y^i \wedge c^B \neq y^i$ , (b)  $c^A \neq y^i \wedge c^B = y^i$ , (c)  $c^A = y^i \wedge c^B = y^i$ , (d)  $c^A \neq y^i \wedge c^B \neq y^i$ .  
 Depending on these cases, a pair of factors  $(\nu_A, \nu_B)$  is to be chosen from table 1. They are needed in step (3) and modulate the step size and set the sense of direction of each step. For GLVQ [9] and GRLVQ the factors  $\nu_A$  and  $\nu_B$  are variable and are to be calculated in each iteration (Table 1).
2. Check if case (a) or case (b) is given and if  $x^i$  is in a window around the perpendicular bisector between  $w^A$  and  $w^B$ :  $\min(d_A/d_B, d_B/d_A) > (\frac{1-s}{1+s})^2$ , where  $s$  is the window size ( $s = 0.2, \dots, 0.3$ ). If these conditions are fulfilled go to step (3), otherwise do not execute update steps and go to step (1).
3. Update both prototype vectors:  $\Delta w^A = \nu_A \eta(t) (x^i - w^A)$  and  $\Delta w^B = \nu_B \eta(t) (x^i - w^B)$ . The step size  $\eta(t)$  controls the rate of convergence of the algorithm and depends on the iteration index  $t$ ;  $\eta(t)$  is a monotonically decreasing function.



**Table 1.** Iteration steps and factors ( $\nu_A, \nu_B$ ) for different LVQ algorithms. For GLVQ and GRLVQ  $\kappa_A = sgd' \left( \frac{d_A - d_B}{d_A + d_B} \right) \left[ \frac{d_B}{(d_A + d_B)^2} \right]$ ,  $\kappa_B = sgd' \left( \frac{d_A - d_B}{d_A + d_B} \right) \left[ \frac{d_A}{(d_A + d_B)^2} \right]$  is to be calculated where  $sgd'$  is the first derivative of the sigmoid function. For LVQ 3, OLVQ 3 and DSLVQ the parameter  $\kappa$  is fixed and should be in the range  $0.1, \dots, 0.5$ .

Method	Steps				Pairs of factors ( $\nu_A, \nu_B$ )			
					case (a)	case (b)	case (c)	case (d)
LVQ 1; OLVQ 1	(1)	(2)	(3)	(1; 0)	(-1; 0)	(1; 0)	(-1; 0)	
LVQ 2	(1)	(2)	(3)	(1; -1)	(0; 0)	(0; 0)	(0; 0)	
LVQ 2.1	(1)	(2)	(3)	(1; -1)	(-1; 1)	(0; 0)	(0; 0)	
LVQ 3; OLVQ 3	(1)	(2)	(3)	(1; -1)	(-1; 1)	( $\kappa; \kappa$ )	(0; 0)	
RLVQ	(1)	(2)	(3)	(4)	(1; 0)	(-1; 0)	(1; 0)	(-1; 0)
DSLVQ	(1)	(2)	(3)	(4)	(1; -1)	(-1; 1)	( $\kappa; \kappa$ )	(0; 0)
GLVQ	(1)	(2)	(3)	(4)	( $\kappa_A; -\kappa_B$ )	( $-\kappa_A; \kappa_B$ )	(0; 0)	(0; 0)
GRLVQ	(1)	(2)	(3)	(4)	( $\kappa_A; -\kappa_B$ )	( $-\kappa_A; \kappa_B$ )	(0; 0)	(0; 0)

4. Update the weight vector  $r^o: \forall k = 1, \dots, n$  do the following three steps: (i) update, (ii) threshold, (iii) normalization.

RLVQ: (i)  $r_k = r_k^o - \nu_A \eta_r (x_k^i - w_k^A)^2$ , (ii)  $r_k = \max(r_k, 0)$ , (iii)  $r^o = \frac{1}{\|r\|_2} r$

GRLVQ: (ii)  $r_k = \max(r_k, 0)$ , (iii)  $r^o = \frac{1}{\|r\|_2} r$

(i)  $r_k = r_k^o - \eta_r sgd' \left( \zeta \frac{d_A - d_B}{d_A + d_B} \right) \left[ \zeta \frac{d_B \cdot (x_k^i - w_k^A)^2 - d_A \cdot (x_k^i - w_k^B)^2}{(d_A + d_B)^2} \right]$ ,  $\zeta = \begin{cases} 1, \text{case(a)} \\ -1, \text{case(b)} \\ 0, \text{otherwise} \end{cases}$

DSLVQ: (i)  $r = r^o + \eta_r (h^o - r^o)$ ,  $h_k = \zeta \frac{|x_k^i - w_k^B| - |x_k^i - w_k^A|}{\max(|x_k^i - w_k^B|, |x_k^i - w_k^A|)}$ ,  $\zeta = \begin{cases} 1, \text{case(a)} \\ -1, \text{case(b)} \\ 0, \text{otherwise} \end{cases}$

$h^o = \frac{1}{\|h\|_1} h$ , (ii)  $r_k = \begin{cases} 1, & r_k \geq 1 \\ 10^{-4}, & r_k \leq 10^{-4} \\ r_k, & \text{otherwise} \end{cases}$ , (iii)  $r^o = \frac{1}{\|r\|_1} r$

Note, DSLVQ executes step(3) but not step(4) in case(c) which leads to  $\zeta=0$  [6].  $R_k, x_k^i, w_k^A, w_k^B$  are the  $k$ -th components of  $r, x^i, w^A, w^B$ , respectively.

5. Go to step (1) until an abortion criterion is fulfilled.

The final weight vector  $r^o$  contains the relevances for each input space dimension.

### 3 Experimental Data Set

Experiments were conducted in our real car driving simulation lab. Seven EEG channels from different scalp positions (C3, C4, Cz, O1, O2, A1, A2) and two EOG-signals (vertical, horizontal) were recorded from 23 young adults during driving sessions lasting 35 minutes. These sessions were repeated every hour between 1 a.m. and 8 a.m. This way, the likelihood of the occurrence of MSE was gradually increasing due to at least 16 hours without sleep prior to the experiment.

MSE are typically characterized by driving errors, prolonged eye lid closures or nodding-off. Towards automatic detection, two experts performed the initial MSE scoring, whereby three video cameras were utilized to record i) drivers head and upper part of the body, ii) right eye region and iii) driving scene. For further processing, only clear-cut cases, where all the experts agreed on the MSE, were taken into account. Despite providing enough test data to tune our algorithms, the human experts could not detect some of the typical attention lapses, such as the one with open eyes and stare gaze. The number of MSE varied amongst subjects and was increasing with time of day for all subjects. In all 3,573 MSE (per subject: mean number  $162 \pm 91$ , range 11-399) and 6,409 non-MSE (per subject: mean number  $291 \pm 89$ , range 45-442) were collected. Non-MSE are periods between MSE where the subject is drowsy but shows no clear or unclear MSE. This clearly highlights the need for an automated data fusion based MSE detection system, which would not only detect the MSE also recognized by human experts, but would also offer a possibility to detect the critical MSE cases which are not recognizable by human experts.

Features were extracted of 8 sec long EEG and EOG segments during MSE or non-MSE by power spectral density estimation and subsequent logarithmic scaling and averaging in frequency bands in the range from 0.5 to 35.5 Hz and a width of 1 Hz.

## 4 Results

In the following we want to compare between several algorithms within the LVQ family and with other classification methods applied to our real world two-class problem. The main question is addressed to classification accuracies which we estimate by computing test errors in a cross validation scheme. There is no indication that the chosen method of multiple hold-out has a remarkable estimation bias compared to the leave one-out method which is an always unbiased estimator of the true classification error [10]. In addition to the original proposed LVQ variants (LVQ1, LVQ2.1, LVQ3, OLVQ1) [5] we examine four further variants additionally executing relevance detection (DSLQV, RLVQ, GR-LVQ, GA-OLVQ1) as mentioned above. Furthermore, we compare them also to the well-known nearest neighbour (1-NN and k-NN) algorithm, to the linear discriminant analysis (LDA), the Error Backpropagation neural network (EBP) and to the Support Vector Machine (SVM). SVM is applied using four different kernel functions because it is not known a priori which matches best for the given problem: 1) linear kernel  $k(x^i, x^j) = \langle x^i, x^j \rangle$ , 2) polynomial kernel:  $k(x^i, x^j) = (\langle x^i, x^j \rangle + 1)^d$ , 3) sigmoidal kernel:  $k(x^i, x^j) = \tanh(\alpha \langle x^i, x^j \rangle + \Theta)$  and 4) RBF kernel:  $k(x^i, x^j) = \exp(-\gamma \|x^i - x^j\|^2)$  for all  $x^i, x^j \in \mathfrak{R}^n$ . Mean training errors and mean test errors are reported in order to quantify the ability to adapt to and to generalize the given problem (Table 1). Training errors are differing largely. Some methods are able to adapt perfectly such as 1-NN, but are not protected against overfitting. 1-NN, a typical example of a local classifier, as well as LDA, a simple global classifier, are exceeded by all LVQ

**Table 2.** Results of multiple hold-out cross validation: mean and standard deviation of training and test errors. Different algorithms have been applied to spontaneous biosignals of the two classes "microsleep event" and "non-microsleep event". Parameters were optimized empirically. C is the regularization parameter of SVM.

Method	Parameter values	$E_{\text{TRAIN}} [\%]$	$E_{\text{TEST}} [\%]$
LVQ1	#neurons = 500	$10.2 \pm 0.2$	$15.7 \pm 0.3$
LVQ2.1	#neurons = 350	$9.6 \pm 0.1$	$15.5 \pm 0.4$
LVQ3	#neurons = 350	$10.3 \pm 0.1$	$15.6 \pm 0.4$
OLVQ1	#neurons = 500	$9.3 \pm 0.2$	$15.7 \pm 0.4$
RLVQ	#neurons = 500; $\eta_r = 0.01$	$15.8 \pm 0.4$	$19.5 \pm 0.4$
DSLQVQ	#neurons = 250; $\eta_r = 0.05$	$8.5 \pm 0.2$	$15.5 \pm 0.3$
GLVQ	#neurons = 400	$9.6 \pm 0.1$	$15.5 \pm 0.4$
GRLVQ	#neurons = 350; $\eta_r = 0.01$	$6.5 \pm 0.2$	$14.3 \pm 0.4$
GA-OLVQ1	#generat. = 200, #popul. = 128	$8.8 \pm 0.2$	$12.9 \pm 0.4$
SVM linear kernel	$C = 10^{-2.75}$	$15.5 \pm 0.1$	$16.9 \pm 0.2$
SVM polynomial k.	$C = 10^{-2.6}$ ; $d = 2$	$7.1 \pm 0.1$	$14.7 \pm 0.3$
SVM sigmoid k.	$C = 10^{+4.4}$ ; $\alpha = 10^{-2.3}$ ; $\vartheta = -1.6$	$7.9 \pm 0.1$	$12.9 \pm 0.3$
SVM Gaussian k.	$C = 10^{+0.31}$ ; $\gamma = 10^{-2.1}$	$0.1 \pm 0.0$	$10.1 \pm 0.4$
LDA	-	$15.6 \pm 0.1$	$17.4 \pm 0.3$
1-NN	-	$0.0 \pm 0.0$	$20.1 \pm 0.5$
k-NN	k = 11	$11.6 \pm 0.1$	$14.7 \pm 0.2$
EBP	#neurons = 8 (hidden layer)	$12.3 \pm 1.1$	$18.1 \pm 0.7$

variants. No important differences in the classification accuracy occurred within the LVQ family, despite the RLVQ which is by 4% inferior and GRLVQ which is by 1% slightly superior. GRLVQ is outperformed by our proposed approach (GA-OLVQ1). But SVM performs still better if a Gaussian kernel function has been utilized and if the hyperparameter and the regularization parameter have been optimized.

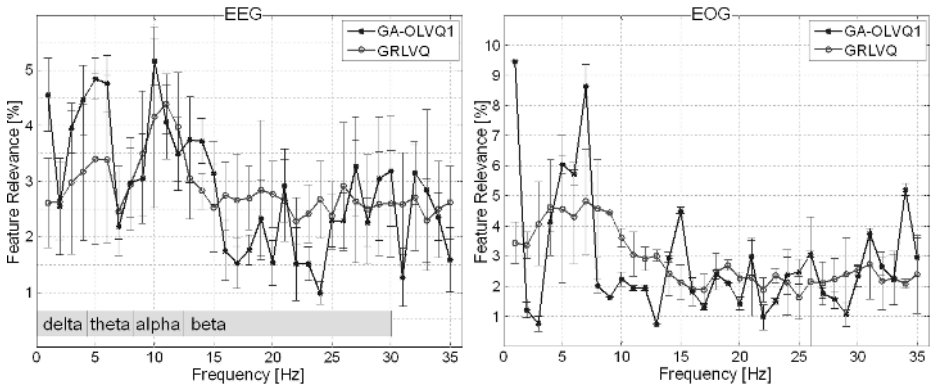
The parameters of all applied methods have been found empirically in order to minimize test errors (Table 1). We report here only the most important parameters and their optimal values for this data set. This optimization has been done on a single training / test partition and does not influence results of other partitions. Therefore, a separate validation set is not necessary.

The computational load of the compared methods is differing largely. OLVQ1 and LVQ1 are unproblematic w.r.t. to the choice of their parameters and they have lowest computational costs, which are in the region of  $10^4$  iterations. This takes about  $10^2$  sec on a modern personal computer. This is the main reason why we used OLVQ1 in our GA-OLVQ1 approach.

RLVQ exhibits large problems when parameters are not optimal. It is very sensitive to the step size  $\eta_r(t)$  (should be about  $10^{-3}$ ), otherwise RLVQ converges as fast as LVQ1. The window parameter is crucial when LVQ2, LVQ2.1, or LVQ3 is used (we have found out  $s = 0.1$  to be good). They need 10 times more iterations as LVQ1 and advantageously, they perform with a lower optimal number of prototypes. DSLVQ showed the same problems and the same need of iterations as LVQ2. It is not sensitive as much as RLVQ to the choice of the step size  $\eta_r(t)$ . Except for the computational load, GLVQ seem to have the same

properties as DSLVQ, though no metric adaptation is learned. GLVQ needs  $10^6$  iterations in the average and therefore, it causes 100 times more computational costs than LVQ1. GRLVQ is in the same shape as GLVQ. The same problems as with LVQ2 are occurring and it needs about 100 times longer than LVQ1.

Our GA-OLVQ1 approach surpasses the computational costs of all other methods. It takes about  $10^4$  times longer than LVQ1. Therefore, we distribute the population of LVQ1 networks in a pool of 32 top modern personal computers and achieve a temporal consumption of about one day. The same amount of computational cost is reached by SVM because scanning for optimal values of the hyperparameter and of the regularization parameter is necessary. A single run of SVM adaptation needs about 10 times longer as for LVQ1, except when the hyperparameter value is far from the optimum. In these cases a single run of SVM can take more than  $10^4$  times longer as for LVQ1. Lastly, we want to present the



**Fig. 2.** Relevance values of GA-OLVQ1 and of GRLVQ for each frequency band (range: 0.5, . . . , 35.5 Hz, 1 Hz width)

acquired relevance values (Fig. 2). To some extent, they are differing between both ARD methods. The relevance values of GA-OLVQ1 show higher dynamic and mostly lower standard deviations. EEG frequencies in the region between the delta and theta band and in the alpha band, but not in the beta band are important for MSE detection. (The just now mentioned bands are common in the EEG community.) These results are in line with them in fatigue research, but the often observed downshift from alpha to high theta is in contrast to our results. In this region low relevance values were found.

## 5 Conclusions

We have presented an overview of some methods of the LVQ family including four approaches to automatic relevance determination. Their classification accuracy has been compared on a biomedical data set consisting of about  $10^4$  items. It

turned out that two of the four ARD approaches performed better than the rest of the LVQ family. Therefore, the usefulness of the underlying global metric adaptation is corroborated.

Our approach which combines all features of all the recorded EEG and EOG channels and which adapts relevance values using genetic algorithms outperformed all other LVQ methods. But best results, with test errors down to 10%, were obtained by Support Vector Machines utilizing a Gaussian kernel function and neglecting the need of metric adaptation.

Unfortunately, the computational costs of both best performing methods are exceptionally high. These costs exceed LVQ1 by a factor of  $10^4$ .

The relevance values of the PSD features of the EEG were similar to findings of other authors in the adjacent field of fatigue research, but for a deeper understanding much more research is needed. Furthermore, there are large inter-individual differences of the EEG and EOG characteristic [1]. It would be interesting to investigate whether methods of local metric adaptation can handle this problem and therefore gaining better and more stable results than global adaptation schemes. Another future issue should be the extension to a greater variety of feature extraction methods which is also likely to improve and stabilize the MSE detection. These issues will be further steps on the long way to establish a reference measure needed for the development of video-based drowsiness warning systems.

## References

1. Sommer, D., Chen, M., Golz, M., Trutschel, U., Mandic, D.: Fusion of State Space and Frequency-Domain Features for Improved Microsleep Detection. *Int Conf Artificial Neural Networks (ICANN 2005)*; LNCS 3697, Springer, (2005) 753–759
2. Bengio, Y., Delalleau, O., Le Roux, N.: The Curse of Dimensionality for Local Kernel Machines. *Techn. Rep. 1258*, Université de Montréal, (2005)
3. MacKay, D. J. C.: Probable Networks and Plausible Predictions - a Review of Practical Bayesian Methods for Supervised Neural Networks. *Network: Computation in Neural Systems*, 6, (1995) 469–505
4. Neal, R. M.: Bayesian Learning for Neural Networks. PhD thesis, University of Toronto, Canada, LNS 118, Springer, Berlin, (1996)
5. Kohonen, T.: *Self-Organizing Maps*. Springer, Berlin, 3rd ed., (2001)
6. Pregonzer, M., Flotzinger, D., Pfurtscheller, G. Distinction Sensitive Learning Vector Quantization - A New Noise-Insensitive Classification Method. In *Proc Int Conf Neural Networks (ICNN-94)*, Orlando, (1994) 2890–2894
7. Bojer, T. et al.: Relevance Determination in Learning Vector Quantization. In: M. Verleysen (ed.), *Europ Symp Artif Neural Netw, D-facto public*, (2001) 271–276
8. Hammer, B., Villmann, T.: Generalized Relevance Learning Vector Quantization. *Neural Networks*, 15 (8-9), (2002) 1059–1068
9. Sato, A.S., Yamada, K.: Generalized Learning Vector Quantization. In: *Adv. Neural Inform. Process. Systems 8*, MIT Press, Cambridge, (1996) 423–429
10. Sommer, D., Golz, M.: A Comparison of Validation Methods for Learning Vector Quantization and for Support Vector Machines on Two Biomedical Data Sets. in M. Spiliopoulou et al. (eds.): *From Data and Information Analysis to Knowledge Engineering*. Springer, (2006) 150–157

# Alertness Assessment Using Data Fusion and Discrimination Ability of LVQ-Networks

Udo Trutschel<sup>2</sup>, David Sommer<sup>1</sup>, Acacia Aguirre<sup>2</sup>, Todd Dawson<sup>2</sup>,  
and Bill Sirois<sup>2</sup>

<sup>1</sup> University of Applied Sciences Schmalkalden, 98574 Schmalkalden, Germany

<sup>2</sup> Circadian Technologies, Inc., 2 Main Street, Suite 310 Stoneham, MA 02180  
utrutschel@circadian.com

**Abstract.** To track the alertness changes of 14 subjects during a night driving simulation study traditional alertness measures such Visual Analog Sleepiness Scale, Alpha Attenuation Test (AAT), and number of Microsleep events per driving session were used. The aim of the paper is to assess these traditional alertness measures regarding their mutual correlations, revise one of them (AAT) and introduce new more general methods to capture changes in human alertness without too many constraints attached. The applied methods are utilizing data fusion methods and data discrimination capabilities via Learning Vector Quantification networks. The advantage of using more general data analysis methods which allows one to assess the validity of proposed alertness measures and opens possibilities to get a more comprehensive knowledge of obtained results.

## 1 Introduction

Recent technical developments have produced a 24-hour, advanced society that continues to grow on a global scale. Consequently, the basic human circadian rhythm ("working during the day and sleeping at night") is under constant siege. Because of the long working hours that eat up people's sleeping time, a general deterioration of people's daytime alertness and an increase in driver drowsiness is seen. Especially, accidents caused by drowsy drivers have a high fatality rate and high costs. To prevent these accidents a reliable tool to accurately measure human alertness levels is needed.

The first attempts to quantify human alertness were subjective reports that consisted of documenting the individual's self-assessment. The main measures include the Stanford Sleepiness Scale (SSS), the Visual-Analog Scale (VAS), and the Epworth Sleepiness Scale (ESS). More objective measures of human alertness can be derived from electroencephalogram (EEG) and electrooculogram (EOG) data. For example, the Multiple Sleep Latency Test (MSLT) measures the time to fall asleep while lying in a quiet, dark bedroom on repeated opportunities at 2 hours intervals throughout the day using EEG for sleep onset determination. The Maintenance of Wakefulness Test (MWT) requires that subjects sit in chairs in a darkened room and remain awake for 40 minutes. After applying different

mathematical and statistical techniques, EEG-frequency bands (delta, theta, alpha, beta, etc.) were used to define a variety of parameters, such as slow-wave activity, alpha slow-wave index, the alpha quotient, and others to estimate alertness. More and more, such alertness parameters have been introduced, in general using the Power Spectral Density (PSD) and multiple combinations of EEG-bands. A good review of the subjective and objective alertness measure can be found in [5] and [6]. Because of the shortcomings of these methods, a new test was developed by Michimori et al. in [1], the Alpha Attenuation Test (AAT) using the occipital (O1) - auricular (A2) EEG-derivation. The AAT is defined as ratio of Eyes-Closed (EC) to Eyes Open (EO) PSD of the alpha band (8 Hz - 12 Hz).

However, there are several drawbacks to all these proposed alertness measures. First, they are based on the countless definitions involving the PSD of a variety of EEG bands. Second, the definition of the EEG-frequency bands introduces artificial boundaries for the data analysis. Third, the separate analysis of the EEG-channels and frequency bands often leads to inconsistent results. Therefore, more general methods with less predefined assumptions are needed for comprehensive human alertness estimation. In order to identify insufficient perceptual capabilities (e.g. prolonged eye closure) and no ability to process external information (e.g. microsleep) reduced alertness should be defined as a combination of brain (EEG) and eye function (EOG). There are modern concepts of data fusion to combine multiple EEG and EOG signals in a way that ensures optimal information gain. For example, a Feature Fusion (FF) approach in combination with Learning Vector Quantization (LVQ) networks was already successfully applied by Sommer et al. in [3] for improving the detection of Micro-Sleep Events (MSE). The FF approach and the ability of LVQ networks to classify and discriminate data with low error rates [4] was utilized for the definition of generalized alertness measures. The high sensitivity of LVQ networks to small and unknown changes in the data is exactly what is required to detect variations in human alertness which are hidden in EEG and EOG.

## 2 Study Design and Data Recorded

Fourteen young adults participated in night time driving study at the University of Schmalkalden. They arrived at the driving simulator facility in the evening, after a day of normal activity and at least 16 hours of continuous wakefulness, which was checked by wrist actigraph. After being wired up for EEG recordings, they started driving on a driving simulator at 1:00 A.M They had to complete seven driving sessions lasting 40 min, each followed by a 10 min period during which they estimated their subjective alertness using a VAS and performed a 5-minute AAT with five alternating 30 seconds Eyes-Open (EO) and Eyes-Closed (EC) episodes. Before the next driving session a 10-min break was scheduled. Experiments ended at 8:00 A.M. During the entire study seven EEG channels from different scalp positions (A1, A2, C3, C4, Cz, O1, O2) and two EOG-signals (vertical, horizontal) were recorded. The driving tasks were monotonous by design to induce drowsiness and Micro-Sleep Events (MSE). MSE are typically

characterized by driving errors, prolonged eye lid closures or nodding-off. Two experts performed an initial manual MSE scoring. Three video cameras were utilized to record (1) drivers face, (2) right eye region and (3) driving scenario. The number of MSE varied amongst subjects and increased during the night for all subjects indicating a clear deterioration in alertness.

The preprocessing of the all EEG and EOG data involved linear trend removal and applying the Hanning window to the data segments. Power Spectral Density (PSD) estimation was performed by the discrete Fourier transform. The so calculated PSD coefficients were averaged within 1.0 Hz wide bands. Further improvements in classification were achieved by applying a monotonic continuous transformation  $\log(x)$  to the PSD.

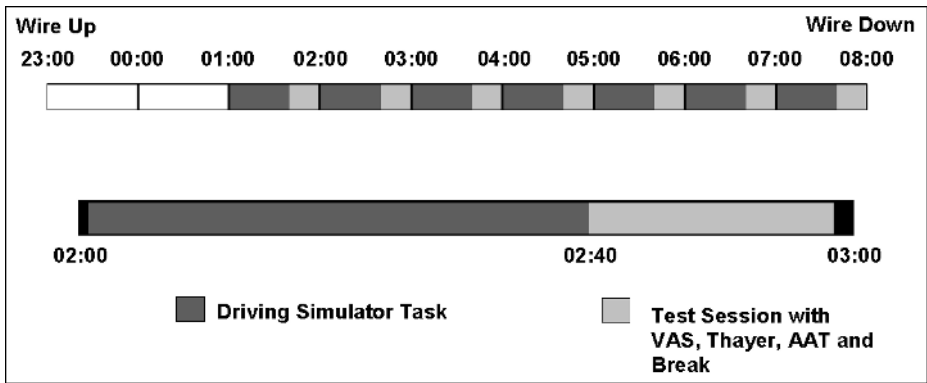


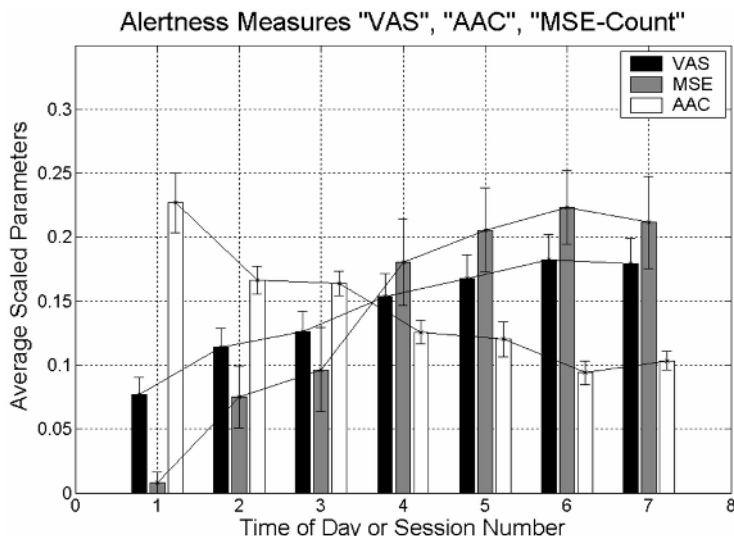
Fig. 1. Design of the driving simulator study

### 3 VAS, MSE and AAT – Correlation Results

First, we present results obtained from VAS, MSE and AAT per driving session. Because of the different nature of the parameters, only the relative changes are meaningful as alertness measures. VAS score and number of MSE increase with reduced alertness, whereas the value of the so-called Alpha Attenuation Coefficient (AAC) decreases. The measures used are not only different in their correlation to alertness; they are different in the time period they cover. VAS reflects a punctual subjective alertness estimation. The AAT provides an objective alertness measure for a 5 minute period and the number of scored MSE per driving session gives an alertness score for a 40 minute period. A reliable EEG based alertness measure on a time scale of five minutes would be extremely useful for many applications. Therefore, we will focus on the Alpha Attenuation Coefficient (AAC), which was introduced by Michimori in [1] and is defined as ratio between the PSD of the alpha band (8 Hz - 12 Hz) for EC and EO episodes, respectively.

Despite the different nature of VAS, MSE and AAC as alertness measure, the average scaled results show a remarkably similar alertness trend over the course of the night hours. Average results for all 14 subjects are shown in Fig. 2. This





**Fig. 2.** Averaged, scaled results with Standard Error for VAS, MSE and AAC. Correlations: 'VAS-MSE'= 0.99; 'VAS-AAC'= -0.99; 'MSE-AAC'= -0.98.

alertness trend is well-known from other studies and was thus expected [2]. The high VAS-MSE correlation will serve further as benchmark to evaluate the other alertness measures.

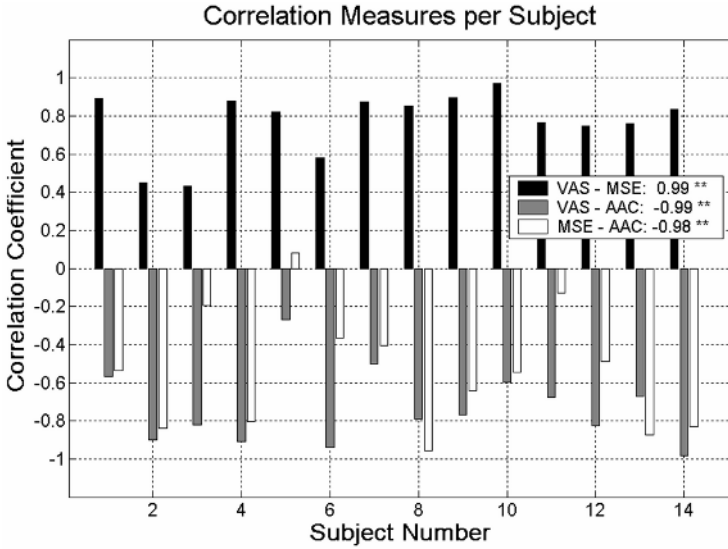
However, the good overall VAS-AAC and MSE-AAC correlation results hide the large individual variability and thus can not accurately estimate or predict individual changes in alertness. Therefore, individualized correlation results are shown in Fig. 3.

Considering the individual correlation results it appears that for subject '5' the AAC fails to reflect the alertness course established by VAS and MSE. This could be because this individual did not produce alpha waves and/or by the restrictive definition of the AAC. Nevertheless, it would be extremely beneficial to have an objective, general valid test to measure the alertness for a broad population and not only for alpha wave producing individuals.

#### 4 "Alpha Attenuation Test" (AAT) – Revised

The AAT was developed based on two assumptions. First, the PSD of alpha waves changes in correspondence to the alertness level of the individual and second, the rate of change occurs in opposite direction for the AAT sections 'EC' and 'EO'.

To test the fundamental assumptions behind the AAT, the following generalized hypothesis should be investigated. Alertness would be high when the EEG-PSD differs substantially between EC and EO episodes, resulting in good LVQ discrimination and a low LVQ test error. Alertness would be reduced when

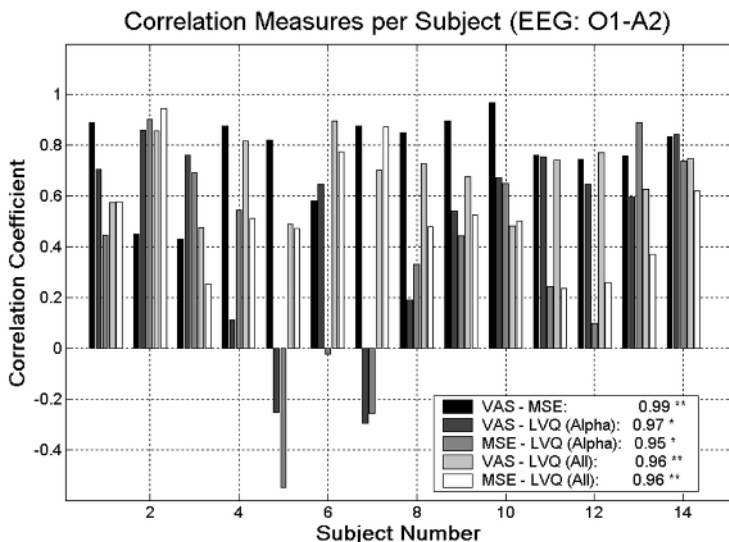


**Fig. 3.** Individualized correlation results between VAS, MSE and AAC. Legend shows the overall correlations of averaged results for 14 subjects, \*\* =  $p < 0.01$ .

**Table 1.** Overall Correlations of Alertness Measures. \*\* =  $p < 0.01$

Correlations	VAS	MSE	AAC	LVQ (Alpha)	LVQ
VAS	1	<b>0.99**</b>	<b>-0.99**</b>	<b>0.97**</b>	<b>0.96**</b>
MSE	<b>0.99**</b>	1	<b>-0.97**</b>	<b>0.95**</b>	<b>0.96**</b>
AAC	<b>-0.99**</b>	<b>-0.97**</b>	1	<b>-0.95**</b>	<b>0.96**</b>
LVQ (Alpha)	<b>0.97**</b>	<b>0.95**</b>	<b>-0.95**</b>	1	--
LVQ	<b>0.96**</b>	<b>0.96**</b>	<b>-0.96**</b>	--	1

the EEG-PSD during EC and EO are similar, resulting in low LVQ discrimination and a high LVQ test error. Therefore, the test error of the LVQ network should be directly correlated to the relative change in alertness. The EEG channel 'O1-A2' is used for the LVQ analysis allowing a direct comparison with the AAC. The overall LVQ correlation results (Table 1) are very close to the correlation for the AAC. Unfortunately, on the subject level there is disagreement. In addition to subject '5' now subject '7' shows no correlation any more to VAS and MSE (Fig. 4), if only the PSD of the alpha band is used (LVQ-A). The situation improves slightly if the band concept is abandoned and the PSD of the integer frequencies are used as features (LVQ). Still, there are troublesome indicators that the selected method is not able to correctly capture the alertness



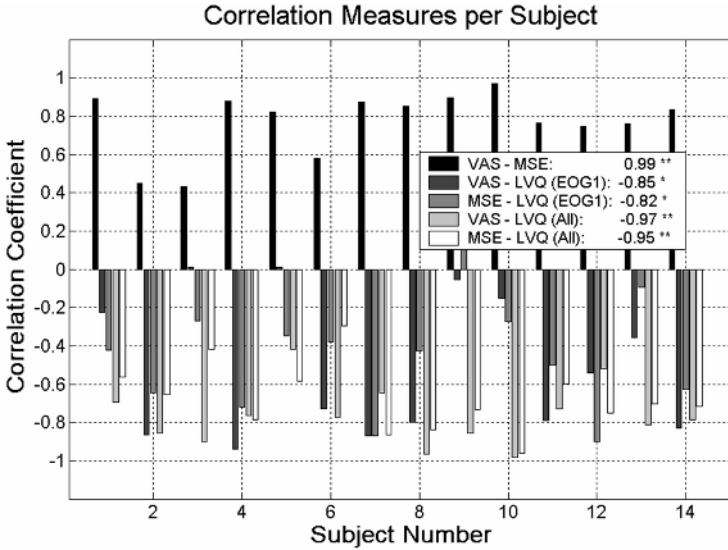
**Fig. 4.** Individualized correlation results between VAS, MSE and LVQ (EO v EC). Legend shows the overall correlations of averaged results for 14 subjects, \*\* =  $p < 0.01$ .

changes during the night. Adding additional EEG channels to the data set and using the FF results in further deterioration of the correlation to the VAS and MSE measures. For example, involving all EEG and EOG channels in the LVQ analysis reduces the correlations for VAS-LVQ and for MSE-LVQ to 0.65 and 0.62, respectively. This should not happen to a generalized method. We concluded that using the differences and/or ratios of EEG-PSD between EO and EC episodes is not the most efficient way for detecting alertness changes.

## 5 “Alpha Attenuation Test” (AAT) – Modified

From the results showed in the previous section it became clear that any ratios between EC and EO are not useful to extract information about alertness from the EEG data. As a modified approach, we propose a further simplification of the relative alertness estimation. We assume that EEG and EOG data during the first AAT probably reflect the highest alertness. Thus, the EEG and EOG of all other AAT sessions during the night will be compared by means of LVQ networks to the first AAT-session which is used as reference.

If the EEG and EOG data sets during a given AAT session are extremely different from the EEG and EOG data sets during the first AAT, then the LVQ network discriminates well, resulting in a low test error. A substantial change in alertness between different AAT sessions has then occurred. On the other hand, if the EEG and EOG data sets during a given AAT are similar to the EEG and EOG data sets during the first AAT, then the LVQ network can not discriminate well, resulting in a high test error. No significant change in alertness between



**Fig. 5.** Individualized correlation results between VAS, MSE and LVQ (S1 v S2 . . . S7). Legend shows the overall correlations of averaged results for 14 subjects, \* =  $p < 0.05$ .

different AAT sessions has then occurred. With this simple pairwise comparison of EEG and EOG for different AAT-times, we were able to obtain changes in alertness over the night which is correlating to a high degree with VAS and MSE (Fig. 5) on a subject by subject basis. The fact that a remarkable improvement in the correlation results was achieved by fusing the data from all EEG and EOG signals is encouraging as well.

Whereas the correlation results using only eye data (VAS-LVQ-EOG1, MSE-LVQ-EOG1) for most subjects are good (except for subject '9', '13'), a substantial improvement can be achieved using all EEG and EOG data (VAS-LVQ-All, MSE-LVQ-All). The new proposed alertness measure based on LVQ pairwise discrimination AAT-session 1 (S1) versus AAT-session 2 (S2). to AAT-session 7 (S7), correlates well with VAS and MSE for all subjects.

Nevertheless, this method of tracking alertness changes based on data discrimination using LVQ, presented here for the first time, is not perfect. Our assumption that the point of the highest alertness is at the beginning of the night may be not valid for every subject. The potential of the approach would be further improved if a second marker for the lowest alertness could be added.

## 6 Conclusions

The newly introduced methods of estimating alertness are utilizing the advantages of Feature Fusion (FF) and the capabilities of LVQ neural networks to classify data with low error rates and good discrimination sensitivity. High sensitivity of LVQ network to small and unknown changes in the PSD of the EEG

and EOG data was applied to trace subtle alertness changes. These proposed methods do not make any assumptions that certain EEG signals and predefined frequency bands has to be used in order to get a reliable alertness measure.

Our first approach relies only on the general assumption that the characteristics of the EEG-PSD between EO and EC episodes changes fundamentally when the alertness of an individual is changing over the course of time. Unfortunately, there are strong indications that the approach suffers from the same weaknesses than the AAT. The utilization of ratio between EO and EC episodes for EEG-PSD bands diminishes the sensitivity to alertness related changes in the signals.

Our second approach relies on pairwise comparison of the PSD from well defined EEG and EOG data at different moments in time. This approach shows great potential and should be evaluated further using a second reference point of low alertness.

## References

1. Michimori A, Stone P, Aguirre A, Stampi C., Analysis of the alpha attenuation test. *Sleep Research*, 23, (1994) 454
2. Trutschel, U., Guttkuhn, R., Ramsthaler, C., Golz, M., Moore-Ede, M.: Automatic Detection of Microsleep Events Using a Neuro-Fuzzy Hybrid System. Proc 6th European Congress of Intelligent Techniques Soft Computing (EUFIT98), Vol 3, Aachen (ISBN 3-89653-5005-5) (1998) 1762–66
3. Sommer, D., Chen, M., Golz, M., Trutschel, U., Mandic, D.: Fusion of State Space and Frequency-Domain Features for Improved Microsleep Detection. in W. Duch et al. (Eds.): International Conf Artificial Neural Networks (ICANN 2005); LNCS 3697, Springer, Berlin (2005) 753–759
4. Kohonen, T.: *Self-Organizing Maps*. Springer, Berlin, 3rd ed., (2001)
5. Sforza, E., Grandin, S., Jouny, C., Rochat, T., Ibanez, V.: Is waking electroencephalographic activity a predictor of daytimesleepiness in sleep-related breathing disorders? *Eur Respir J* (2002) 645–652
6. Leproult, R., Coleccia, E.F., Berardi, A.M., Stickgold, R., Kosslyn, S.M., Van Cauter, E.: Individual differences in subjective and objective alertness during sleep deprivation are stable and unrelated. *Am J Physiol* 284, (2002) R280–R290

# Signal Reconstruction by Projection Filter with Preservation of Preferential Components

Akira Hirabayashi<sup>1</sup> and Takeshi Naito<sup>2</sup>

<sup>1</sup> Yamaguchi University, Ube Yamaguchi 755-8611, Japan  
a-hira@yamaguchi-u.ac.jp

<http://www.ir.csse.yamaguchi-u.ac.jp>

<sup>2</sup> Omron Corporation, Social Systems Solutions Business Company,  
2-2-1, Nishikusatsu, 525-0035, Kusatsu, Japan

**Abstract.** Signal reconstruction is one of the knowledge extraction problems. Especially in this problem, knowledge means underlying signal or image, which is extracted from the downsampled one. In this paper, we propose a novel reconstruction filter which perfectly reconstructs predetermined preferential components, and makes a reconstructed signal/image agree with the oblique projection of an original one. It enables us to get rid of artifacts which arise in reconstructed signals by the conventional partial projection filter when the number of samples is small compared with the dimension of the approximation subspace. By simulations, we show that the proposed filter performs better than the conventional methods.

## 1 Introduction

Signal/image reconstruction is a problem of estimating an original signal/image from a downsampled one. So far, many reconstruction filters have been proposed. The most popular one is the Wiener filter [1]. It reconstructs the optimal approximation which minimizes the mean squared error between a reconstructed image and an original image. The ‘mean’ implies average with respect to both image and noise ensembles. Hence, the criterion does not guarantee that the reconstructed image is the best approximation to the individual original image. In order to solve the problem, the projection filter was proposed [2,3]. It reconstructs the optimal approximation which minimizes the squared error between a reconstructed image and an original image in the sense of average with respect to only the noise ensemble. The partial projection filter is an extension of the projection filter to the case where original images belong to a certain subspace [4].

The partial projection filter provides good reconstruction results if downsampled images have enough pixels compared with the dimension of the subspace used in the filter. If not the case, a reconstructed image is degraded. Precisely, degree of degradation differs depending on image components. If heavy degradation happens to low frequency components in images, especially for the constant component, then strong artifacts arise in total reconstructed images.

In order to solve the problem, we propose a novel projection filter. The key idea is to explicitly specify important components to be perfectly reconstructed. A reconstructed image agrees with an oblique projection of an original image. Hence, we call the filter the *Perfect Preferential component reconstruction Projection filter (the PPP filter)*.

This paper is organized as follows. We formulate the image reconstruction problem by using the so-called inverse problem in Section 2. In Section 3, we briefly review the partial projection filter. In Section 4, we propose the novel filter, and derive its closed formula. In Section 5, the proposed filter is applied to the image magnification problem, and it is shown that the proposed filter performs better than the partial projection filter and the bicubic interpolation. Section 6 concludes the paper.

## 2 Formulation of Image Reconstruction Problem

Let us start with a formulation of image reconstruction problem in the continuous-discrete model. The discrete-discrete model can also be formulated in a similar way. In the model, original images are defined on a continuous domain, while observed images are defined on a discrete domain. We assume that an original image  $f(x, y)$  belongs to a Hilbert space  $H$ . The inner product in  $H$  is denoted by  $\langle f, g \rangle$  for elements  $f$  and  $g$  in  $H$ .

An observed image has a finite number,  $N_x \times N_y$ , of pixels. Each pixel value  $d_{n_x, n_y}$  ( $n_x = 1, 2, \dots, N_x, n_y = 1, 2, \dots, N_y$ ) can be expressed as the inner product between  $f$  and a sampling function  $\psi_{n_x, n_y}$ :

$$d_{n_x, n_y} = \langle f, \psi_{n_x, n_y} \rangle. \quad (1)$$

Let  $\mathbf{d}$  be a vector in which  $d_{n_x, n_y}$  is the  $(n_x + (n_y - 1)N_x)$ -th element. Let us define an operator  $A$  by

$$Af = \mathbf{d}. \quad (2)$$

An approximation of  $f$  is denoted by  $\tilde{f}$ , which also belongs to  $H$ . Let  $X$  be an operator that maps  $\mathbf{d}$  into  $\tilde{f}$ :

$$X\mathbf{d} = \tilde{f}. \quad (3)$$

Image reconstruction problem becomes a problem of finding  $X$  that maps  $\mathbf{d}$  into an optimal  $\tilde{f}$  under some criterion.

The following notations are used in this paper. The adjoint operator of an operator  $T$  is denoted by  $T^*$ . The range and the nullspace of  $T$  are denoted by  $R(T)$  and  $N(T)$ , respectively. Let  $T^\dagger$  denote the Moore-Penrose generalized inverse of  $T$ .

## 3 Partial Projection Filter

Let us briefly review the partial projection filter for the noiseless case. Eqs. (2) and (3) yield

$$\tilde{f} = XAf. \quad (4)$$

The basic idea is to make  $XAf$  agree with the best approximation in a certain closed subspace in  $H$  when  $f$  is assumed to belong to a closed subspace  $S$  in  $H$ . Based on the assumption,  $R(P_S A^*)$  becomes the subspace in which the best approximation is obtained, and the partial projection filter is defined as follows.

**Definition 1 (Partial projection filter [4]).** *An operator  $X$  is called a partial projection filter if  $X$  satisfies*

$$XAP_S = P_{R(P_S A^*)}, \tag{5}$$

where  $P_{R(P_S A^*)}$  is the orthogonal projection operator onto  $R(P_S A^*)$ .

The reconstructed image by the partial projection filter for an image  $f$  in  $S$  is the orthogonal projection onto  $R(P_S A^*)$ . Note that  $R(P_S A^*)$  is the subspace spanned by orthogonal projections of the sampling functions  $\{\psi_{n_x, n_y}\}_{n_x=1}^{N_x} \{n_y=1}^{N_y}$  onto  $S$ . Such projections usually oscillate frequently. If we have only a small number of pixels, the oscillation appears in the reconstructed image. Then, strong artifacts arise in the reconstructed images.

### 4 Novel Projection Filter

In order to prevent the problem mentioned above, we propose a novel reconstruction filter. Let  $S$  be a closed subspace in  $H$ . For simplicity, we assume that

$$S + N(A) = H, \tag{6}$$

where the symbol  $+$  denotes the sum of subspaces. In this case,  $S$  and  $N(A)$  have an intersection in general. Similarly,  $\dot{+}$  and  $\oplus$  denotes the direct sum and the orthogonal direct sum, respectively.

In contrast to the partial projection filter, we do not assume that  $f$  belongs to  $S$ . We use  $S$  as an approximation subspace. That is, we approximate an original image  $f$  in  $H$  by using  $\tilde{f}$  which we restrict within  $S$ . If an original image  $f$  belongs to  $S$ , then we want to perfectly reconstruct the image. However,  $Af = 0$  for any  $f$  in  $S \cap N(A)$ . Then, perfect reconstruction is possible only for  $f$  in a complement of  $S \cap N(A)$  in  $S$ . We are free to choose such a complement. Along this context, the orthogonal complement of  $S \cap N(A)$  in  $S$  is chosen in the partial projection filter. This is the reason why the important components in images are not always reconstructed well by the partial projection filter.

We propose another choice of the complement. Let  $\{\phi_m\}_{m=1}^M$  be the important components in images that belong to  $S$ , but not to  $S \cap N(A)$ . The subspace spanned by  $\{\phi_m\}_{m=1}^M$  is denoted by  $V_i$ . The suffix  $i$  means ‘interest.’ If

$$V_i \dot{+} \{S \cap N(A)\} = S, \tag{7}$$

then the complement is uniquely determined as  $V_i$ . However, Eq. (7) is not true in general. Then, we have to determine the rest of the complement. For the purpose, we use the orthogonal complement,  $V_c$ , of  $V_i \dot{+} \{S \cap N(A)\}$  in  $S$  in order to be



consistent with the partial projection filter. The suffix  $c$  means ‘complement.’ By using these two subspaces  $V_i$  and  $V_c$ , we finally choose the complement as

$$L = V_i \oplus V_c. \tag{8}$$

The subspace  $S$  is decomposed into

$$S = L \dot{+} \{S \cap N(A)\}. \tag{9}$$

Eqs. (6) and (9) imply that

$$L \dot{+} N(A) = H. \tag{10}$$

Hence, we can define the projection onto  $L$  along  $N(A)$ . The projection operator is denoted by  $P_{L,N(A)}$ . Note that this projection is *oblique* in general.

**Definition 2.** *An operator  $X$  is called the Perfect Preferential component reconstruction Projection filter (the PPP filter) if  $X$  satisfies*

$$XA = P_{L,N(A)}. \tag{11}$$

In the left-hand side of Eq. (11), the orthogonal projection operator  $P_S$  does not appear. This implies that an original image  $f$  is not restricted in  $S$ . On the other hand,  $P_S$  appears in the left-hand side of Eq. (5), because  $f$  is restricted in  $S$  in the partial projection filter.

A closed form of the PPP filter is given as follows. By using the orthogonal projection operators  $P_{V_i}$  and  $P_{V_c}$  onto  $V_i$  and  $V_c$ , respectively, we define an operator  $P_L$  by

$$P_L = P_{V_i} + P_{V_c}. \tag{12}$$

Since  $V_i$  and  $V_c$  are perpendicular to each other as is shown in Eq. (8),  $P_L$  is the orthogonal projection operator onto  $L$ . The detail of how to construct  $P_L$  can be found in [7].

**Theorem 1.** *The PPP filter always exists, and its general form is given by*

$$X = (AP_L)^\dagger + Y(I - AA^\dagger), \tag{13}$$

where  $Y$  is an arbitrary operator from  $\mathbf{C}^N$  to  $H$ .

A proof of Theorem 1 is deferred the appendix. For any  $f$  in  $H$ , a reconstructed image  $\tilde{f}$  by the PPP filter agrees with the oblique projection of  $f$  onto  $L$  along  $N(A)$ . In order to compare the PPP filter with the partial projection filter, let us assume that  $f$  belongs to  $S$ . Then,  $f$  can be decomposed into two components as

$$f = f_i + f_j, \tag{14}$$

where  $f_i$  and  $f_j$  lie in  $V_i$  and  $\{V_c \oplus (S \cap N(A))\}$ , respectively. The first component  $f_i$  is perfectly reconstructed because  $f_i$  lies in  $V_i$ . On the other hand,  $f_j$  is not

perfectly reconstructed, but projected onto  $V_c$ . The projection is denoted by  $f_c$ . Then, the reconstruction for  $f$  is given by

$$\tilde{f} = f_i + f_c. \tag{15}$$

Note that, since  $V_c$  is perpendicular to  $S \cap N(A)$ ,  $f_c$  is the minimum error approximation of  $f_j$  in  $V_c$ . This guarantees that the PPP filter reconstructs a good approximation of  $f$ .

### 5 Application Example

In this section, we apply the PPP filter to image magnification, and compare its performance to the partial projection filter and the bicubic interpolation. Figure 1 shows the original entire image. Figure 2 shows the target image in the simulation corresponding to a  $20 \times 20$  pixels area indicated by the box in Figure 1. Figure 3 shows a  $10 \times 10$  pixels image synthesized from Figure 2 by the bilinear interpolation. It is used to estimate Figure 2.

The width and the height of the image are denoted by  $l_x$  and  $l_y$ , respectively, which are equal to  $N_x$  and  $N_y$ . The coordinate of the center of each pixel is denoted by  $(x_{n_x}, y_{n_y})$  ( $n_x = 1, 2, \dots, N_x, n_y = 1, 2, \dots, N_y$ ). Assume that  $f$  belongs to the space  $H$  of square-integrable functions on the area  $[0, l_x] \times [0, l_y]$ . The inner product  $\langle f, g \rangle$  in  $H$  is defined by

$$\langle f, g \rangle = \frac{1}{l_x l_y} \int_0^{l_x} \int_0^{l_y} f(x, y)g(x, y) dx dy.$$

For positive real numbers  $r$  and  $p$ , let  $\psi$  be a function defined by

$$\psi(x) = \begin{cases} p & (|x| \leq r), \\ 0 & (|x| > r). \end{cases} \tag{16}$$

The sampling function  $\psi_{n_x, n_y}(x, y)$  is assumed to be

$$\psi_{n_x, n_y}(x, y) = \psi(x - x_{n_x})\psi(y - y_{n_y}). \tag{17}$$

Let  $S$  be a subspace spanned by functions

$$\varphi_{k_x, k_y}(x, y) = \varphi_{k_x}(x)\varphi_{k_y}(y) \tag{18}$$

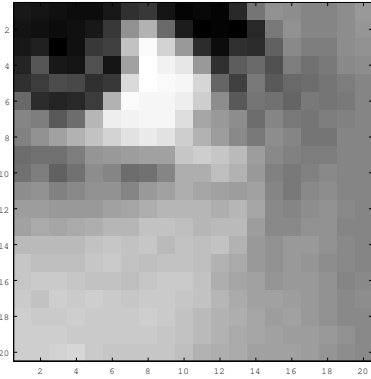
for  $k_x = 1, 2, \dots, K_x$  and  $k_y = 1, 2, \dots, K_y$ , where

$$\varphi_{k_x}(x) = \begin{cases} 1 & (k_x = 1), \\ \sqrt{2} \cos \frac{(k_x - 1)\pi x}{l_x} & (k_x = 2, 3, \dots, K_x), \end{cases}$$

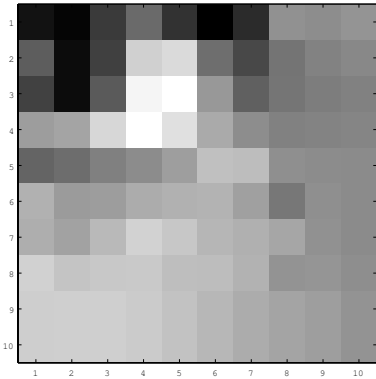
$$\varphi_{k_y}(y) = \begin{cases} 1 & (k_y = 1), \\ \sqrt{2} \cos \frac{(k_y - 1)\pi y}{l_y} & (k_y = 2, 3, \dots, K_y). \end{cases}$$



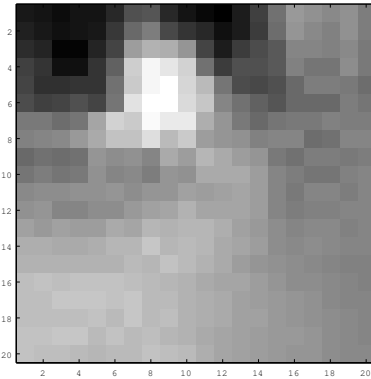
**Fig. 1.** The original entire image



**Fig. 2.** The target image in the simulation corresponding to a  $20 \times 20$  pixels area indicated by the box in Figure 1



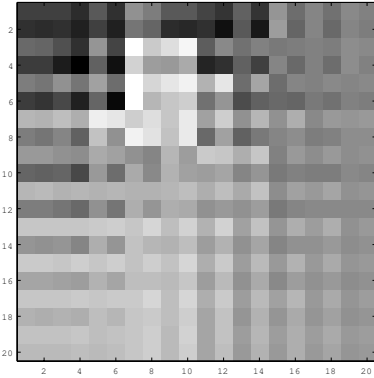
**Fig. 3.** Downsampled image of  $10 \times 10$  pixels synthesized from Figure 2 by the bilinear interpolation. This image is used to estimate Figure 2.



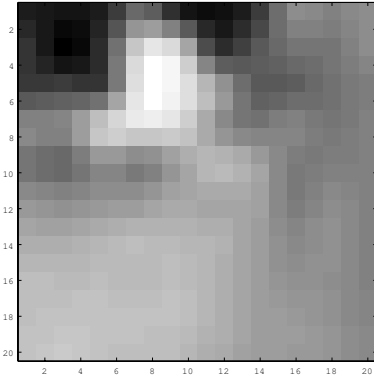
**Fig. 4.** Reconstructed image by the proposed filter. PSNR = 27.87dB.

In the simulations,  $K_x = K_y = 20$ , which is the same as the number of pixels in each axis in Figure 2.

Let the preferential components  $\{\phi_m\}_{m=1}^M$  be 36 functions  $\{\varphi_{k_x, k_y}\}_{k_x=1}^6 \{\varphi_{k_x, k_y}\}_{k_y=1}^6$ . Figure 4 and Figure 5 show reconstructed images by the PPP filter and the partial projection filter, respectively. We can see that artifacts like a lattice arise in Figure 5, while not in Figure 4. The PSNRs of Figure 4 and Figure 5 are 27.87dB and 20.52dB, respectively. From both points of view, the PPP filter performs much better than the partial projection filter. Figure 6 shows an image obtained by the bicubic interpolation. We do not see big difference between Figure 4 and Figure 6. The PSNR is, however, 27.20dB which means the PPP filter gives still better result than the bicubic interpolation.



**Fig. 5.** Reconstructed image by the partial projection filter. PSNR = 20.52dB.



**Fig. 6.** Reconstructed image by the bicubic interpolation. PSNR = 27.20dB.

The preferential components  $\{\phi_m\}_{m=1}^M$  in the simulation are chosen from low frequency components of the basis  $\{\varphi_{k_x, k_y}\}_{k_x=1}^{K_x} \{\varphi_{k_x, k_y}\}_{k_y=1}^{K_y}$  in  $S$ . We also used  $\{\varphi_{k_x, k_y}\}_{k_x=1}^1 \{\varphi_{k_x, k_y}\}_{k_y=1}^1$ ,  $\{\varphi_{k_x, k_y}\}_{k_x=1}^2 \{\varphi_{k_x, k_y}\}_{k_y=1}^2$ ,  $\dots$ ,  $\{\varphi_{k_x, k_y}\}_{k_x=1}^7 \{\varphi_{k_x, k_y}\}_{k_y=1}^7$ . Among them, the best result was obtained by  $\{\varphi_{k_x, k_y}\}_{k_x=1}^6 \{\varphi_{k_x, k_y}\}_{k_y=1}^6$ . Some other efficient way to chose  $\{\phi_m\}_{m=1}^I$  may further improve the result.

## 6 Conclusion

We proposed a novel reconstruction filter which perfectly reconstructs predetermined preferential components, and makes a reconstructed image agree with the oblique projection of an original image. It was named the Perfect Preferential component reconstruction Projection filter (the PPP filter). It enabled us to get rid of artifacts which arise in reconstructed images by the partial projection filter when the number of pixels is small compared with the dimension of the approximation subspace. By simulations, we showed that the PPP filter performed better than the partial projection filter and the conventional bicubic interpolation.

## References

1. H.C. Andrew and R.B. Hunt, *Digital Image Restoration*. New York: Prentice Hall, 1977.
2. H. Ogawa and N. Nakamura, "Projection filter restoration of degraded images," in *Proc. Int. Conference on Pattern Recognition 1984*, Montreal Canada, July 30-August 2. 1984, pp. 601–603.
3. N. Nakamura and H. Ogawa, "Optimum digital image restoration under additive noises," *The Transactions of IEICE D*, vol. J67-D, pp. 563–570, Jan. 1984

(in Japanese). Its English translation appeared in *Systems-Computers-Controls*, Scripta Technica, Inc., USA, vol. 15, pp. 73–82, 1984.

4. H. Ogawa and S. Hara, “Partial projection filter for image restoration,” in *Proc. 6th Int. Scandinavian Conference on Image Analysis*, Oulu, Finland, June 19–22, 1989, pp. 270–277.
5. A. Albert, *Regression and the Moore-Penrose Pseudoinverse*. New York: Academic Press, 1972.
6. Y. C. Eldar, “Sampling and reconstruction in arbitrary spaces and oblique dual frame vectors,” *The Journal of Fourier Analysis and Applications*, vol. 1, no. 9, pp. 77–96, 2003.
7. A. Hirabayashi and M. Unser, “An extension of oblique projection sampling theorem,” in *Proc. the 2005 International Conference on Sampling Theory and Applications (SampTA2005)*, Samsun, Turkey, July 10–14, 2005. pp. 1–6 (CD-ROM).

## A Proof of Theorem 1

In order to prove Theorem 1, we use the following lemmas.

**Lemma 1.** [5] *For any operator  $T$  and the orthogonal projection operator  $P$  onto a closed subspace in  $H$ , the following two relations hold:*

$$P(TP)^\dagger = (TP)^\dagger, \quad (19)$$

$$(PT)^\dagger P = (PT)^\dagger. \quad (20)$$

**Lemma 2.** *The oblique projection operator  $P_{L,N(A)}$  is expressed as*

$$P_{L,N(A)} = (AP_L)^\dagger A. \quad (21)$$

(Proof). Since  $P_L$  is the orthogonal projection operator onto  $L$ , Proposition 1 in [6] implies that  $P_L(AP_L)^\dagger A$  is the oblique projection operator onto  $L$  along  $N(A)$ . It follows from Eq. (19) that

$$P_{L,N(A)} = P_L(AP_L)^\dagger A = (AP_L)^\dagger A, \quad (22)$$

which implies Eq. (21).  $\square$

**(Proof of Theorem 1).** Eq. (21) implies that Eq. (11) yields

$$XA = (AP_L)^\dagger A. \quad (23)$$

Since  $N(A) \subset N((AP_L)^\dagger A)$ , Theorem (3.13.4) in [5] guarantees Eq. (23) has a solution  $X$ , and its general form is give by

$$X = (AP_L)^\dagger AA^\dagger + Y(I - AA^\dagger). \quad (24)$$

It follows from Eq. (20) that

$$(AP_L)^\dagger AA^\dagger = (P_{R(A)}AP_L)^\dagger P_{R(A)} = (P_{R(A)}AP_L)^\dagger = (AP_L)^\dagger.$$

That is,

$$(AP_L)^\dagger AA^\dagger = (AP_L)^\dagger. \quad (25)$$

Hence, Eqs. (24) and (25) imply Eq. (13).  $\square$

# Sensor Network Localization Using Least Squares Kernel Regression

Anthony Kuh<sup>1</sup>, Chaopin Zhu<sup>1</sup>, and Danilo Mandic<sup>2</sup>

<sup>1</sup> University of Hawaii

<sup>2</sup> Imperial College London

**Abstract.** This paper considers the sensor network localization problem using signal strength. Unlike range-based methods signal strength information is stored in a kernel matrix. Least squares regression methods are then used to get an estimate of the location of unknown sensors. Locations are represented as complex numbers with the estimate function consisting of a linear weighted sum of kernel entries. The regression estimates have similar performance to previous localization methods using kernel classification methods, but at reduced complexity. Simulations are conducted to test the performance of the least squares kernel regression algorithm. Finally, the paper discusses on-line implementations of the algorithm, methods to improve the performance of the regression algorithm, and using kernels to extract other information from distributed sensor networks.

## 1 Introduction

Information gathering is relying more on distributed communication and distributed networking systems. Ad hoc sensor networks are being deployed in a variety of applications from environmental sensing to security and intrusion detection to medical monitoring [1]. These sensor networks are becoming increasingly more complex with sensors responsible for different tasks. We will consider a sensor network consisting of two different types of sensors: base sensors where the locations are known (the base sensors could have GPS) and simple sensors called motes where the locations of the motes are unknown. Assuming all sensors are capable of transmitting information via signal strength we use kernel least squares regression methods to estimate the location of the motes. The kernel regression method performs similarly to kernel classification methods [9] at much reduced complexity. We also discuss an on-line implementation of the algorithm to perform tracking of mobile sensors and ways of extracting other information from sensors using kernel regression methods.

When signal strength is available a common method to perform sensor localization is using ranging information as discussed in [2,6]. Range-based methods commonly use a two step approach when performing localization: first signal distance is estimated between pairs of devices from signal strength and then a localization algorithm is used based on these estimated distances. In wireless radio in ideal space, the signal attenuation  $s$  satisfies,

$$s \propto Pd^{-\eta}, \quad (1)$$

where  $d$  is the distance between transmitter and receiver  $\eta > 2$  is a constant and  $P$  is the transmitting power [13]. Because of rich scattering in the real world, the received signal strength is quite noisy. This makes it more difficult for range-based methods such as [2,6] to get accurate readings of the distance between different devices. Statistical methods such as Maximum Likelihood Estimation (MLE) [10], EM algorithm [12], and Bayesian networks [3] were used to alleviate the scattering effect. These methods usually have high computing complexity.

We use an alternative approach first established in [9] where signal strength information is stored in a kernel matrix. In [9,16] a classification problem is solved to determine whether a sensor lies in a given region  $\mathcal{A}$  or not. The classification problem uses training data from base sensors (whose location is known) to learn the parameters  $\alpha(i)$  and threshold value  $b$  given by the following equation

$$f(x) = \text{sgn}\left(\sum_{i=1}^l \alpha(i)y_i K(x, x_i) + b\right) \quad (2)$$

where  $f(x)$  is the decision function to determine whether  $x \in \mathcal{A}$  or not and  $x_i$  are input locations and  $y_i = 1$  if  $x_i \in \mathcal{A}$  otherwise  $y_i = -1$ . In [9] the classification problem is solved using the Support Vector Machine (SVM) (with hinge loss function) [8,15]. In [16] the classification problem is solved using the Least Squares SVM (LS-SVM) (with quadratic loss function with equality constraints) [14,7]. Fine localization is achieved by performing the classification problem several times with different overlapping regions. A mote's location is estimated from the average of the centroids of each region it belongs to.

This method gives reasonably accurate estimates of locations of motes, but is computationally expensive as the classification problem must be solved for each region considered. This paper estimate locations of motes using one complex least squares kernel regression problem. Sensors are located on a two-dimensional grid with their location represented by a complex number. The parameters  $\alpha(i)$  are also complex and the kernels are real. The regression method saves computational costs while giving similar performance to the fine localization algorithm.

The paper is organized as follows. In Section 2 we discuss the ad hoc sensor network model. Section 3 gives a discussion of the least squares kernel regression algorithm, an on-line recursive version of the algorithm, and how signal strength is incorporated into the kernels. Section 4 simulates a sensor network model where sensors are randomly place on a two-dimensional grid. Finally, Section 5 discusses extensions of this work including and on-line implementation of the algorithm so mobile sensors can be tracked, improvements to the kernel regression algorithm, and references to how kernel regression methods are used to extract other sensor information. We also discuss learning in a distributed environment subject to communication and power constraints.

## 2 Sensor Network Model

Assume that an ad hoc network of size  $N$  is deployed in a two-dimensional geographical area  $\mathcal{T}$ . An integer from 1 to  $N$  represents each node (sensor) as its ID. Denote the set of all nodes in the network by  $\mathcal{N} = (1, \dots, N)$ . The location of node  $i \in \mathcal{N}$  is denoted by  $x_i$ . Assume that the first  $m$  nodes are *base sensors*. These nodes have more computational capabilities than other nodes and their location is known (i.e.  $x_i, 1 \leq i \leq M$  are known) as these nodes may be positioned or the nodes may have GPS. A goal is to determine the location of the other  $N - M$  nodes called *notes* (i.e. estimate  $x_{M+1}, \dots, x_N$ ).

Each node is capable of transmitting signal strength to each of the other nodes. As mentioned in the previous section signal strength is often related to distance by equation (1). The signal strength is also contaminated by additive noise and may be affected by terrain conditions. For this paper we assume the same signal strength model as [9,16] given by

$$s(x_i, x_j) = \exp\left\{-\frac{\|x_i - x_j\|^2}{\Sigma} + V\right\} \quad (3)$$

where  $V$  is a zero mean Gaussian random variable with standard deviation  $\tau$ .

The information about all signal strengths is stored in a kernel matrix  $K$ . Kernel entries are a function of signal strength and are constructed so that  $K$  is symmetric and positive semi-definite.

## 3 Least Squares Subspace Kernel Regression Algorithm

This algorithm is described in more detail in [7]. We describe the basic algorithm followed by a discussion of implementing a recursive on-line version of the algorithm. This section concludes by discussing the kernel regression localization algorithm.

### 3.1 Least Squares Kernel Subspace Algorithm

We are given  $m$  training examples or observations drawn from input  $X \in \mathbf{R}^n$  and output  $Y \in \mathcal{R}$ . The observations at each time  $(x_i, y_i), 1 \leq i \leq m$  are independent and can be represented compactly as  $(\mathbf{x}, \mathbf{y})$  where  $\mathbf{x} = [x_1 | \dots | x_m]$  and  $\mathbf{y} = [y_1, \dots, y_m]^T$ . The inputs are transformed from input space to feature space via kernel functions  $\phi(x)$  that map inputs from  $\mathcal{R}^n$  to feature space  $\mathcal{R}^d$ . Let  $\Phi(\mathbf{x}) = [\phi(x_1) | \dots | \phi(x_m)]$ . The estimate is given by  $\hat{Y}(x) = w^T \phi(x) + \mathbf{1}b$  where  $w \in \mathcal{R}^d$  is the weight vector,  $\mathbf{1}$  is an  $m$  vector of 1s, and  $b$  is the scalar threshold value.

The weight vector can be expressed as a linear combination of each of the feature vectors. For the standard LS-SVM, [14] the weight vector depends on all training feature vectors. This can be expressed as  $w = \Phi(\mathbf{x})\alpha$  where  $\alpha$  is an  $m$  vector. Each of the training examples  $x_i$  associated with a nonzero  $\alpha(i)$  is called a support vector. For the standard SVM only a fraction of the training examples



are support vectors as an  $\epsilon$  - insensitive cost function is used that removes training inputs as support vectors that are close to the zero error solution. For the LS-SVM an external procedure needs to be established to reduce the number of training examples. Methods to intelligently choose training examples are discussed briefly in the next subsection with more detail in [7]. Here we assume that the subset is chosen and we will denote it by  $\mathbf{x}_S$  which is a matrix containing  $l \leq m$  columns from  $\mathbf{x}$ . Denote the intelligent feature vectors by  $\Phi(\mathbf{x}_S)$ .

The optimization problem is a quadratic programming problem with equality constraints and is shown below:

$$\min J(w, b) = \min_{w, b} \frac{1}{2} \|w\|^2 + \frac{\gamma}{2} \|e\|^2 \tag{4}$$

subject to

$$e = \mathbf{y} - \Phi(\mathbf{x})w - \mathbf{1}b. \tag{5}$$

and

$$w = \Phi(\mathbf{x}_S)\alpha \tag{6}$$

where now  $\alpha$  is an  $l$  vector describing the weighting of the training feature vectors. Defining  $K_{SS} = \Phi^T(\mathbf{x}_S)\Phi(\mathbf{x}_S)$  and  $K_S = \Phi^T(\mathbf{x}_S)\Phi(\mathbf{x})$  and substituting equation (5,6) into (4) we have that

$$\min Q(\alpha, b) = \min \frac{1}{2} \alpha^T K_{SS} \alpha + \frac{\gamma}{2} \|Y - K_S^T \alpha - \mathbf{1}b\|^2. \tag{7}$$

This problem is optimized by finding the solution to the following set of linear equations.

$$\begin{bmatrix} l & \mathbf{1}^T K_S^T \\ K_S \mathbf{1} & K_{SS}/\gamma + K_S K_S^T \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T Y \\ K_S Y \end{bmatrix}. \tag{8}$$

Assuming  $A = K_{SS}/\gamma + K_S K_S^T$  is invertible, then by elimination we get that

$$b = \frac{\mathbf{1}^T Y - \mathbf{1}^T K_S^T A^{-1} K_S Y}{l - \mathbf{1}^T K_S^T A^{-1} K_S \mathbf{1}} \tag{9}$$

and

$$\alpha = A^{-1} K_S (Y - \mathbf{1}b) \tag{10}$$

### 3.2 Recursive Kernel Subspace Least Squares Algorithm

Here we discuss a recursive on-line procedure for updating the algorithm presented in the last subsection. At each update we update the estimate using a windowed recursive least squares algorithm. The basic algorithm can be described as follows:

1. Train parameters on initial set of data using batch or online methods.
2. Get new training data and add to information set.

3. If new data satisfies specified criteria add as a support vector.
4. Selectively prune support vectors.
5. Selectively prune information data.
6. If there is more data go to 2.

Here we consider a fixed window size where If we decide to add a support vector from training data we must also delete a support vector. A simple criteria is used to add and delete support vectors. To add a support vector, the new feature data is tested to see if it can reduce the training error data. The vector is added as a support vector when the training error can be reduced by a prescribed amount. This can be implemented as follows: evaluate the kernel vector between the newest data point and all other  $l_W$  data points. Compare to the training error vector  $e$ . We normalize each of the vectors to having magnitude one and compute the inner product between the two vector which is the same as computing the cosine of the angle between the two vectors. If the magnitude of the inner products is above a specified threshold value, then the new training data is added as a support vector. This criteria is also used in [4]. During the deletion process we delete the support vector that makes the least contribution to the weight vector.

### 3.3 Kernel Localization Algorithm

In order to implement the algorithms described in the previous subsections we need to form a kernel matrix from signal strength information. Signal strength information is not symmetric due to additive noise and other impairments. In [9] several different kernel functions are formed. Here we use the following kernel function described by

$$K(x_i, x_j) = \exp\left\{-\frac{\sum_{t=1}^M (s(x_i, x_t) - s(x_j, x_t))^2}{2\sigma^2}\right\} \quad (11)$$

Here the targets  $Y$  are complex numbers representing the locations of the base sensors. We pick the base sensor locations  $i_1, \dots, i_l$  that we want as support vectors and then use equation (8) to solve for  $\alpha$  and  $b$  which are complex valued. We then can estimate the location of mote  $j$  by using the following estimate

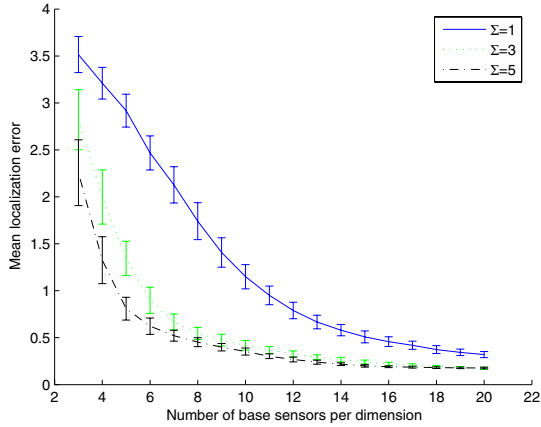
$$f(x_j) = \sum_{n=1}^l \alpha(i_n)K(x_j, x_{i_n}) + b, \quad M + 1 \leq j \leq N. \quad (12)$$

There is considerable savings by using the complex kernel regression algorithm as only one regression algorithm is performed to get estimates of the mote locations as opposed to the many kernel classification algorithms necessary to get fine localization in [9,16].

## 4 Simulations

Several simulations were conducted to test the performance of the kernel regression localization algorithm. Sensors were placed on a 10 by 10 grid. There were

$M = k^2$  base sensors placed on a grid and then the locations were perturbed by additive Gaussian noise. There were 400 motes placed randomly on the grid. Signal strength had additive noise with deviation  $\tau = 0.2$ . The subspace algorithm was used with  $l = 3k$  support vectors. The algorithm worked well with the regularization parameter set at  $\gamma = 60$  and the width of the Gaussian kernels was set at  $\sigma = 2.7$ . Fig. 1 shows how the mean localization error varies as the number of base sensors increases and  $\Sigma$  is varied. Simulations were conducted 100 times for each setting with average curves and standard deviations shown.



**Fig. 1.** Mean localization error versus number of base sensors

The plots show that the localization estimation error goes down as the number of base sensors increases. The results are similar to the results for fine localization shown in [9,16]. When  $\Sigma = 3, 5$  signal strength decreases rapidly as distance and the error rate remains roughly constant at a little below .5 when more than 100 base sensors are used.

## 5 Discussion and Extensions

This paper shows that sensor localization can be successfully performed using a least squares kernel subspace regression algorithm. This algorithm has good performance and has computational savings over kernel classification algorithms. There are several further directions to this work.

The sensor localization error is dependent on the number of base sensors. If there are too few base sensors the localization error will be high. In real sensor applications the number of base sensor may be limited. Additional information needs to be considered for the sensor localization algorithm. Note that signal strength information is not used between base sensors and motes in determining the regression estimates. If this information could be incorporated into the kernels this could improve estimation error. In practical applications the exact

location of motes may not be known, but the location may be modelled by a random distribution. This knowledge could be combined with signal strength information to get more accurate estimates of location using kernel methods.

Another consideration are power and communication constraints placed on sensors. If there are constraints on transmission of signal strength information how does this affect localization estimation. If signal strength is weak only nearby base sensors may receive the signal strength. Can good estimation procedures be established when sensors have power and communication constraints.

In Section 3.2 we discussed a recursive on-line least squares kernel subspace algorithm. The kernel localization algorithm can be modified to perform tracking capabilities when sensors are mobile. This is discussed in more detail in [17].

Finally, other information can be gathered from sensors networks using kernel regression algorithms. In [5,11] distributed kernel regression algorithms are used to approximate sensor information such as temperature information. Distributed kernel learning algorithms are proposed to learn this information. Distributed learning could also be applied to develop computationally efficient learning algorithms for sensor localization.

## References

1. I. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, A survey on sensor networks, *IEEE Communications Magazine* 102-114, Aug. 2002.
2. N. Bulusu, J. Heidemann, D. Estrin. "GPS-less low cost outdoor localization for very small devices", Tech. Rep. 00-0729, Computer Science Dept., Univ. of Southern California, 2000.
3. P. Castro, P. Chiu, T. Kremenek, and R. Muntz, "A probabilistic room location service for wireless networked environments", ACM Ubicomp 2001, Atlanta, GA., 2001.
4. B. de Kruif. *Function approximation for learning control, a key sample based approach*, Ph.D. thesis, University of Twente, Netherland, 2004.
5. C. Guestrin, P. Bodik, R. Thibaux, M. Paskin, S. Madden. "Distributed regression; an efficient framework for modeling sensor network data", *Information Processing in Sensor Networks 2004*, Berkeley, CA, Apr. 2004.
6. J. Hightower, G. Borriello, "Real-time error in location modeling for ubiquitous computing, in Location, Modeling for Ubiquitous Computing, 21-27, Ubicomp 2001 Workshop Proceedings, 2001.
7. A. Kuh, "Intelligent recursive kernel subspace estimation algorithms", the 39th Annual Conference of Information Sciences and Systems (CISS 2005), 216-221, Baltimore, MD., 2005.
8. K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf. An Introduction to Kernel-Based Learning Algorithms. *IEEE Trans. on Neural Networks*, Vol. 12, #2, 181-202, 2001.
9. X. Nguyen, M. Jordan, and B. Sinopoli. A kernel-based learning approach to ad hoc sensor network localization, *ACM Transactions on Sensor Networks*, Vol 1(1), pages 134-152, 2005.
10. N. Patwari, A. Hero, M. Perkins, N. Correat, R. O'Dea. "relative location estimation in wireless sensor networks", *IEEE Transaction on Signal Processing*, vol. 51, no. 8, pp. 2137-2148, Aug. 2003.

11. J. B. Predd, S.R. Kulkarni, H. V. Poor. Distributed Regression in Sensor Networks: Training Distributively with Alternating Projections, *Proceedings of the SPIE Conference and Advanced Signal Processing Algorithms, Architectures, and Implementations XV*, San Diego, CA, Aug., 2005.
12. T. Roos, P. Myllymaki, H. Tirri, "A statistical modeling approach to location estimation", *IEEE Transactions on Mobile Computing*, vol. 1, no. 1, pp, 59-69, Jan.-Mar. 2002.
13. S. Seidel, T. Rappaport (1992), "914MHz path loss prediction models for indoor wireless communications in multifloored buildings", *IEEE Transactions on Antennas and Propagation*, vol. 40, no. 2, pp. 207-217, Feb. 1992.
14. J. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least squares support vector machines*, World Scientific Publishing Co., Singapore, 2002.
15. V. Vapnik. *Statistical learning theory*. John Wiley, New York City, NY, 1998.
16. C. Zhu and A. Kuh. "Sensor network localization using pattern recognition and least squares kernel methods", in *Proceedings of 2005 Hawaii, IEICE and SITA Joint Conference on Information Theory, (HISC 2005)*, Honolulu, HI, May 25-27, 2005.
17. C. Zhu, A. Kuh. "Dynamic Ad Hoc network localization using online least squares kernel subspace methods", submitted to *2006 IEEE International Symposium on Information Theory*, Seattle WA., July 2006.

# Author Index

- Abe, Akinori III-22  
Abe, Jair Minoro II-844, II-851,  
II-858, II-871  
Abe, Norihiro II-620  
Abe, Noriyuki II-628  
Abraham, Ajith II-500, III-398,  
III-677, III-1128  
Adachi, Naoto III-166  
Adachi, Yoshinori II-401, II-1170,  
II-1176  
Aftarczuk, Kamila III-805  
Agell, Nria II-425  
Aguilar-Ruiz, Jess S. II-1264,  
II-1272  
Aguirre, Acacia III-1264  
Ahn, Byung-Ha I-122, I-130  
Ahn, Chan-Min III-84  
Ahn, Tae-Chon I-52, III-219  
Akamatsu, Norio III-692  
Akashi, Takuya III-692  
Alepis, Efthymios I-435  
Alexopoulos, Angelos II-1152  
Alfpio, Pedro I-1083  
Amaral, Jos Franco M. III-307  
Amorim, Pedro II-1248  
An, Jiyuan II-1305  
Anastasopoulos, Dionysios III-633  
Anderson, Mary III-1184  
Anguita, Davide II-442  
Angulo, Cecilio II-425  
Anshin, Peter III-988  
Aoe, Jun-ichi II-275, II-303, II-317,  
II-325  
Aoki, Terumasa II-371  
Aoyama, Atsushi II-553  
Apeh, Edward Tersoo I-1216  
Aquino, Andreia C. de I-268  
Araki, Takayoshi I-506  
Araki, Yutaka II-212  
Arita, Daisaku II-212  
Ariton, Viorel II-569  
Aritsugi, Masayoshi II-1216  
Arroyo-Figueroa, Gustavo I-943  
Atkinson, John I-1190  
Atlam, El-Sayed II-275, II-303, II-317,  
II-325  
Augusto, Juan Carlos II-171  
Awcock, Graeme J. I-1226  
Azzini, Antonia III-1111  
Baba, Norio III-663  
Babiak, Emilia III-797  
Back, Barbro I-720  
Bae, Dae Jin I-401  
Bae, Hyeon-Deok I-808, I-817  
Bae, Ihn-Han I-483  
Baek, Jonghun I-1011  
Baek, Joong Hwan I-866  
Baek, Yong Sun III-248  
Bagis, Aytekin I-94  
Baik, Ran III-284  
Baik, Sung Wook III-284  
Bajaj, Preeti I-46  
Balachandran, M. Bala III-1192  
Balfanz, Dirk I-753  
Balvig, Jens J. II-603  
Bannore, Vivek II-36  
Bao, Yongguang II-393  
Barbancho, Antonio I-475  
Barbancho, Julio I-475  
Bardis, Georgios I-425  
Barriga, Angel II-348  
Barros, Allan Kardec III-1232  
Baruque, Bruno III-432  
Bashar, Md. Rezaul I-532, III-116,  
III-124  
Bařtrk, Alper II-331  
Batten, Adam III-349  
Baturone, Iluminada II-363  
Becker, Matthias I-706  
Beer, Martin D. II-1240  
Beligiannis, Grigorios I-968  
Bellas, F. III-292  
Bellik, Yacine II-154  
Beřdok, Erkan I-606  
Bi, Jun II-678  
Bi, Leo I-220  
Bian, Zhengzhong III-507

- Bianco, Adriana C. I-268  
 Bien, Zeungnam III-248  
 Bilgen, Bilge I-37  
 Bilgin, Mehmet Zeki I-1075  
 Bingul, Zafer I-138  
 Blázquez-del-Toro, José M. III-580  
 Bolat, Emine Dođru I-841  
 Borzemski, Leszek II-195, III-789  
 Bouhafis, Lyamine I-409  
 Boukis, Christos III-1216, III-1224  
 Bourda, Yolaine II-154  
 Bradley, Jeremy III-366, III-374  
 Brdiczka, Oliver II-162  
 Brown, David J. I-639, I-825, I-1198  
 Brox, Piedad II-363  
 Brunner, Levin III-614  
 Bu, Jiajun I-417  
 Bugarín, Alberto III-623  
 Burguillo-Rial, J.C. II-659, III-263  
 Burša, Miroslav II-409  
 Byun, Yong Ki III-907
- Cai, Keke I-417  
 Calbi, Alessandro II-179  
 Cao, Jianting III-1240  
 Caponetti, Laura II-340  
 Cardoso, Luciana S. I-268  
 Carvalho, Paulo I-1083  
 Castells, Pablo III-598  
 Castiello, Ciro II-340  
 Ceravolo, Paolo III-1111  
 Cerekovic, Aleksandra II-220  
 Cha, Chang-II I-443  
 Cha, Jae-Sang III-883  
 Cha, Jaehyuk I-1043  
 Chambers, Jonathon A. III-1208  
 Chamorro-Martínez, Jesús II-355  
 Chan, Chee Seng I-639  
 Chan, Eddie Chun Lun II-652  
 Chang, Byoungchol I-1043  
 Chang, Elizabeth I-728, III-1119  
 Chang, Hoon III-541  
 Chang, Jae-Woo I-1067, III-76  
 Chang, KyungHi III-457  
 Chang, Sekchin III-449  
 Chang, Tae-Gyu II-500, III-677,  
 III-1128  
 Chang, Ying-Chen II-1  
 Chao, Pei-Ju II-1
- Chen, Chen-Wen II-888  
 Chen, Chun I-417  
 Chen, Guobin I-631  
 Chen, Jiangyan III-1201  
 Chen, Kuang-Ku I-300  
 Chen, Liming II-171  
 Chen, Mo III-1216  
 Chen, Yen-Wei II-55, II-63  
 Chen, Yi-Ping Phoebe II-1305  
 Chen, Yuehui III-398  
 Chen, Yumei I-145  
 Chetty, Girija III-1168  
 Chi, Sheng-Chai I-1  
 Chis, Monica III-677  
 Cho, Dongyoung I-110  
 Cho, Jae-Soo I-1011  
 Cho, Jungwon I-392  
 Cho, Ming-Yuan I-179  
 Cho, Nam-deok II-819  
 Choi, Byung-Uk I-559  
 Choi, Chang-Seok III-108  
 Choi, Dae-Young I-490, I-984  
 Choi, Dong You I-1242  
 Choi, Eui-in III-1058  
 Choi, Hun I-808, I-817  
 Choi, Sang-Yule III-883  
 Choi, Soo-Mi I-753  
 Choi, Woo-kyung III-101  
 Choi, Yoo-Joo I-753  
 Chong, Zhang I-459  
 Chou, Ming-Tao II-902  
 Chou, Pao-Hua I-300  
 Chun, Seok-Ju III-84  
 Chung, Jaehak III-449, III-473, III-480  
 Chung, Yongwha I-906  
 Cichocki, Andrzej III-1232, III-1248  
 Cigale, Boris III-515  
 Cisternino, Virginia III-1102  
 Çiviciođlu, Pmar I-606  
 Consoli, Angela III-497  
 Corallo, Angelo III-1083, III-1092,  
 III-1102  
 Corchado, Emilio II-433, III-432  
 Corcho, Oscar III-588  
 Corella, Miguel Ángel III-598  
 Costa-Montenegro, E. II-659  
 Couce, Beatriz III-300  
 Cox, Robert J. III-1143  
 Cox, Trevor J. III-1208  
 Crowley, James L. II-162

- Crowther, Patricia S. III-1143  
 Crua, Cyril I-1179  
 Cutler, Philip II-479  
  
 Daggard, Grant I-976  
 Datta, Avijit I-825  
 Dawson, Todd III-1264  
 Debenham, John I-228  
 Deep, Kusum I-951  
 Degemmis, Marco III-606  
 Deshmukh, Amol I-46  
 Dias, Douglas Mota III-307  
 Dillon, Tharam S. I-728, III-1119  
 Djaiz, Chaker I-687  
 Dosil-Outes, I. III-263  
 Drapała, Jarosław III-1012  
 Druszcz, Adam III-789  
 Duanmu, C.J. II-11  
 Duro, Richard J. III-292  
  
 Edwards, Graeme C. III-349  
 Egri-Nagy, Attila III-333  
 Endo, Mamoru II-1045  
 Eto, Kaoru II-977  
 Eto, Tsutomu III-166  
  
 Falcón, Juanjo I-679  
 Fan, Shaofeng II-1144  
 Fanelli, Anna Maria II-340  
 Fang, Fu-Min II-1  
 Fang, Hsin-Hsiung II-922  
 Fang, Zhijun I-631  
 Feng, Honghai I-145, I-498, I-714,  
 I-1029  
 Feng, Jun I-1037, I-1059, II-1095,  
 II-1191, II-1199  
 Fernández-García, Norberto III-580  
 Fernández-Hernández, Felipe II-363  
 Ferreira, Mauricio G.V. I-268  
 Figueroa, Alejandro I-1190  
 Finn, Anthony II-537  
 Finnie, Gavin I-1115  
 Fisteus, Jesús Arias III-580  
 Fitch, Phillip II-523  
 Flórez-Revuelta, Francisco III-424  
 Freeman, Michael III-415  
 Fuchino, Tetsuo II-553  
 Fuente, Raúl de la III-300  
 Fujii, Kunihiko I-1021  
 Fujii, Kunikazu III-205  
  
 Fujita, Yoshikatsu II-281  
 Fukazawa, Yusuke I-1021  
 Fuketa, Masao II-275, II-303, II-325  
 Fukue, Yoshinori II-289  
 Fukui, Ken-ichi II-929  
 Fukumi, Minoru III-692  
 Furumura, Takashi III-150, III-189  
 Fuwa, Yasushi II-994  
 Fyfe, Colin II-472, II-508  
  
 Gabrys, Bogdan I-1216, III-432  
 Galán-Perales, Elena II-355  
 Gallego, Josune III-277  
 García-Almanza, Alma Lilia III-30  
 García Chamizo, Juan Manuel III-424  
 García Rodríguez, José III-424  
 García-Sebastian, M. Teresa III-277  
 Gau, Shih-Wei I-179  
 Gautama, Temujin III-1216  
 Geem, Zong Woo I-86  
 Georgoulas, George I-515  
 Gerasimov, Vadim III-315  
 Ghada, Elmarhomy II-303, II-317,  
 II-325  
 Gil, Joon-Min I-1147  
 Gil-Castiñeira, F. III-263  
 Giráldez, Raúl II-1264, II-1272  
 Go, Kentaro II-1027  
 Goel, Piyush I-825  
 Goh, Su Lee III-1216  
 Golz, Martin III-1256  
 Gómez-Pérez, Asunción III-588  
 Gong, Peng III-441  
 González-Castaño, Francisco J.  
 II-659, III-263  
 Górecki, Przemyslaw II-340  
 Goto, Masato II-1079  
 Graña, Manuel III-277  
 Griffith, Josephine III-766  
 Grosan, Crina III-677, III-1128  
 Groumpos, Peter I-515  
 Gu, Jinguang I-738  
 Gu, Qin I-1106  
 Guo, Lei II-19  
 Gutiérrez-Ríos, Julio II-363  
 Guzmán, Giovanni I-550, I-614  
  
 Ha, Sang-Hyung III-101  
 Håkansson, Anne I-342, I-352



- Hadzilacos, Thanasis II-1136  
 Hajjam, Amir I-409  
 Hall, Richard I-102, I-220  
 Hamaguchi, Takashi II-553, II-579,  
 II-587  
 Han, Chang-Wook I-850  
 Han, Dongsoo I-260  
 Han, Hee-Seop I-746  
 Han, Xuming I-21  
 Harada, Jun II-275  
 Harada, Kouji II-115  
 Harashina, Naoki II-1119  
 Hartung, Ronald I-342, I-352  
 Hasegawa, Tsutomu II-212  
 Hashimoto, Setsuo III-684  
 Hashimoto, Yoshihiro II-587  
 Hassan, Nashaat M. Hussein II-348  
 Hattori, Akira II-1079  
 Hatzilygeroudis, Ioannis I-968,  
 II-1313  
 He, LiYun I-145, I-498, I-714,  
 I-1029  
 Healy, Gerry III-315  
 Heo, Joo III-457  
 Hernández, Carmen III-277  
 Hernandez, Yasmin I-943  
 Herrero, Álvaro II-433  
 Higuchi, Yuki II-1027  
 Hiissa, Marketta I-720  
 Hill, Richard II-1240  
 Hirabayashi, Akira III-1272  
 Hiraishi, Wataru II-275  
 Hochin, Teruhisa II-1182  
 Hong, Chao-Fu III-1, III-46  
 Hong, Gyo Young I-29  
 Hong, Jun Sik III-84  
 Hong, Kwang-Seok I-788, I-798  
 Hong, Minsuk II-545  
 Hong, Sungeon I-203  
 Hong, Yeh Sun I-866  
 Höppne, Frank II-70  
 Hori, Satoshi II-611, II-620, II-628  
 Horie, Kenichi III-38  
 Hoschke, Nigel III-349  
 Hoshio, Kenji III-212  
 Hou, Yimin II-19  
 Howlett, Robert J. I-1179, I-1206,  
 I-1226  
 Hsu, Chia-Ling III-1  
 Hsu, Mu-Hsiu III-938  
 Hu, Hong I-976  
 Hu, Meng II-587  
 Huang, Eng-Yen II-1  
 Huang, Hung-Hsuan II-220  
 Huang, Peng I-417  
 Huang, Xinyin II-55  
 Huang, Xu III-1150, III-1157, III-1163  
 Hussain, Farookh Khadeer III-1119  
 Hussain, Omar Khadeer III-1119  
 Hwang, Chong-Sun I-1124, I-1139,  
 I-1155  
 Hwang, Kyu-Jeong I-443  
 Hwang, Seung-won II-1281  
 Hwang, Sun-myoung II-745  
 Hwang, Suntae III-248  
 Iannone, Luigi III-606  
 Ichalkaranje, Nikhil II-450, II-486  
 Ichikawa, Sachiyoishi II-387  
 Ichimura, Takumi III-742, III-749  
 Igarashi, Emi II-1035  
 Ikehara, Satoru III-715  
 Ikezaki, Masakazu II-1224, II-1232  
 Im, SeokJin I-1124, I-1139, I-1155  
 In, Jang-uk III-1058  
 Inokuchi, Ryo II-78  
 Inuzuka, Nobuhiro II-1162  
 Iribe, Yurie II-1010  
 Irie, Masayuki III-205  
 Iritani, Takeshi II-953  
 Isern, David II-1256, III-758  
 Ishibuchi, Hisao II-86  
 Ishida, Yoshiteru II-123, II-131, II-139,  
 II-146  
 Ishii, Naohiro II-379, II-387, II-393,  
 II-1170  
 Ishikawa, Norihiro III-159, III-166  
 Isokawa, Teijiro III-699  
 Isomoto, Yukuo II-985  
 Ito, Hideaki II-1208  
 Ito, Kei III-813  
 Ito, Maiko II-1035  
 Ito, Masahiro II-953  
 Itoh, Hidenori II-401, II-961  
 Itoh, Kohji II-1071  
 Itoh, Toshiaki II-587  
 Itou, Junko III-212  
 Ivancevic, Vladimir II-537  
 Iwahori, Yuji II-401, II-1176  
 Iwamura, Norikazu III-835

- Iwata, Masayuki III-647  
 Iwazaki, Kumiko II-1045  
  
 Jacquet, Christophe II-154  
 Jain, Lakhmi C. II-110, II-450,  
 II-458, II-472, II-531, II-537,  
 III-497, III-504  
 Jamali, Mohammed A. I-252  
 Jang, Eun Sill I-992  
 Jang, Ik-Jin I-1011  
 Jang, Jae-Hyuk II-777  
 Jang, Kyung-Won III-219  
 Jang, Min-Soo I-590  
 Jang, SangHyun I-1163  
 Jarvis, Bevan II-458  
 Jayasooriya, Thimal III-415  
 Jee, KyengWhan II-492  
 Jelfs, Beth III-1216  
 Jelonek, Jacek III-341  
 Jeng, Don Jyh-Fu III-922, III-964  
 Jeng, LiDer II-879  
 Jeng, MuDer II-879  
 Jeon, Hong-Tae III-101  
 Jeon, Jae Wook I-401  
 Jeon, Jaewook II-545  
 Jeon, Jin-Hong II-812  
 Jeon, M.G. I-130  
 Jeong, Hong I-1090  
 Jeong, Moon Seok III-284  
 Jeong, Seungdo I-559  
 Jeong, Seung-Hee II-829  
 Jevtic, Dragan I-284  
 Jezic, Gordan I-236  
 Ji, Younggun III-473  
 Jie, Min Seok I-29, I-858, I-866  
 Jifeng, He I-459  
 Jimbo, Takashi II-387  
 Jin, Chunming III-197  
 Jin, SongGuo III-124  
 Jo, Sun-Moon I-451  
 Jones, Jason I-212  
 Joo, Hyun Jea III-707  
 Jung, Eun Sung I-68, I-78, III-124  
 Jung, Kyung-Yong I-163, I-310  
 Juszczyszyn, Krzysztof II-243,  
 III-1020  
  
 Kabasawa, Yasuo II-977  
 Kabassi, Katerina II-1289  
 Kakegawa, Jun'ichi II-1071  
  
 Kakehi, Masahide II-1062  
 Kalles, Dimitris II-1136  
 Kamoshita, Yasuhiro II-1002  
 Kanagawa, Koji III-725  
 Kaneyama, Takashi III-150  
 Kang, Dazhou I-647, I-655  
 Kang, Eui-young I-392  
 Kang, Joonhyuk III-465  
 Kang, Mingyun III-1075  
 Kang, Sanggil I-260  
 Kang, Sang-Won I-1124, I-1139, I-1155  
 Kang, Seo Il II-793  
 Kang, Sukhoon II-769, III-248  
 Kang, Yeon Gu I-582, III-707  
 Kang, YunHee I-203, I-1163  
 Karel, Filip I-195  
 Karras, Dimitrios A. I-9  
 Karsten, Helena I-720  
 Karungaru, Stephen III-692  
 Kashihara, Akihiro II-1002  
 Katarzyniak, Radosław Piotr III-1027  
 Kato, Kei III-827  
 Kato, Shohei II-961  
 Kato, Toshikazu III-16  
 Kato, Yoshikiyo III-8  
 Katsurada, Kouichi II-1010  
 Kawaguchi, Masashi II-387  
 Kawanaka, Haruki II-401  
 Kawaoka, Tsukasa I-506, I-882, I-1002  
 Kawasaki, Takashi II-289  
 Kazienko, Przemysław II-417  
 Keegan, Stephen II-686  
 Kendrick, Paul III-1208  
 Kerre, Etienne I-623  
 Keskar, A.G. I-46  
 Khwan-on, Sudarat I-833  
 Kil, Min Wook II-701, III-1075  
 Kim, ChangHwan I-874  
 Kim, Dong Seong I-935  
 Kim, Dongwon III-255, III-670  
 Kim, Dong Wook III-859  
 Kim, Duk Kyung III-441  
 Kim, Haeng-Kon II-751, II-760  
 Kim, Hak-Man II-812, III-900  
 Kim, Hanil I-392  
 Kim, Hong Sok III-541  
 Kim, Howon I-924, I-1250  
 Kim, Hwangrae III-1068  
 Kim, Hyeoncheol I-746  
 Kim, Hyun Deok I-276, I-1011

- Kim, Hyung-Jun I-443  
 Kim, Ikno III-922, III-964  
 Kim, In-Cheol I-244  
 Kim, Jaehwan III-465  
 Kim, Jeom Goo III-564  
 Kim, Jeong Hyun III-556  
 Kim, Jin-Geol III-233  
 Kim, Jong Tae I-401, III-907, III-914  
 Kim, Jong-Yul III-900  
 Kim, Jongwan I-1124, I-1139, I-1155  
 Kim, Joohee III-480  
 Kim, Jung-Hyun I-788, I-798  
 Kim, Jungyeop I-203  
 Kim, Sang Tae I-276  
 Kim, Sang-Wook I-443, I-1043  
 Kim, Sangkyun I-1234, I-1259  
 Kim, Seokhyun III-473  
 Kim, Seoksoo II-694, II-701, II-718,  
 III-1075  
 Kim, Seong-Joo I-924, III-101  
 Kim, SungBu I-890  
 Kim, Sung-Hwan I-935  
 Kim, Taehee III-556  
 Kim, Tai Hoon II-701, II-745, II-836  
 Kim, Wankyung II-709, III-1068  
 Kim, Woong-Sik I-898  
 Kim, Yong I-1090  
 Kim, Yong-Guk I-590  
 Kim, Yong-Ki I-1067, III-76  
 Kim, Yong Soo III-248  
 Kim, Young-Chang I-1067  
 Kim, Young-Joong III-225  
 Kim, Youngsup III-1042  
 Kimura, Masahiro II-929, II-937  
 Kitajima, Teiji II-553  
 Kitakoshi, Daisuke II-969  
 Kitazawa, Kenichi III-189  
 Klawonn, Frank II-70  
 Ko, Il Seok III-1035, III-1050  
 Ko, Min Jung I-292  
 Ko, SangJun III-457  
 Kobayashi, Kanami II-55  
 Kogure, Kiyoshi III-22  
 Koh, Byoung-Soo II-777  
 Koh, Eun Jin I-368, I-569  
 Koh, Jae Young III-572  
 Kojima, Fumio III-684  
 Kojima, Masanori III-189  
 Kojiri, Tomoko I-771, II-1053,  
 II-1062  
 Kokol, Peter II-1297  
 Kolaczek, Grzegorz II-243  
 Komedani, Akira I-771  
 Komiya, Kaori III-16  
 Komosinski, Maciej III-341  
 Kompatsiaris, Yiannis III-633  
 Kondo, Michiro II-851, II-858, II-871  
 Kong, Jung-Shik III-233  
 Konishi, Osamu III-813  
 Koo, Jung Doo III-548  
 Koo, Jung Sook III-548  
 Koshimizu, Hiroyasu II-1208  
 Koukam, Abder I-409  
 Koutsojannis, Constantinos I-968,  
 II-1313  
 Kowalczuk, Zdzisław I-671  
 Kozakura, Shigeki II-620  
 Koziarkiewicz, Adrianna III-805  
 Król, Dariusz II-259, III-774  
 Kubota, Naoyuki III-684  
 Kucuk, Serdar I-138  
 Kuh, Anthony III-1280  
 Kuh, Tony III-1216  
 Kukkurainen, Paavo III-383  
 Kukla, Grzegorz Stanisław  
 II-259, III-774  
 Kulikowski, Juliusz L. II-235  
 Kunchev, Voemir II-537  
 Kunifuji, Susumu II-1019, III-851,  
 III-859, III-867  
 Kunimune, Hisayoshi II-994  
 Kunstic, Marijan I-284  
 Kurakake, Shoji I-1021  
 Kurazume, Ryo II-212  
 Kurimoto, Ikusaburo III-742  
 Kuroda, Chiaki II-561  
 Kusek, Mario I-236  
 Kusumoto, Yoshiki III-205  
 Kuwahara, Jungo III-159  
 Kuwahara, Noriaki III-22  
 Kuwajima, Isao II-86  
 Kuwata, Tomoyuki II-102  
 Kwak, Hoon Sung I-542  
 Kwak, Jin I-924  
 Kwak, Jong Min III-1050  
 Kwon, Taekyoung I-916  
 Kwon, Yangsoo III-480  
 Kye, Bokyung I-1163

- Laflaquière, Julien I-1171  
 Lam, Toby H.W. II-637, II-644  
 Lama, Manuel III-623  
 Lampropoulos, Aristomenis S. I-376,  
 I-384  
 Lampropoulou, Paraskevi S. I-384  
 Lasota, Tadeusz III-774  
 Lee, Boo-Hyung III-135  
 Lee, Byoung-Kuk III-875  
 Lee, Chang-Hwan I-187  
 Lee, Chil-Woo I-598  
 Lee, Chungwon III-556  
 Lee, Deok-Gyu II-793  
 Lee, Do Hyeon III-564  
 Lee, Dong Chun III-548  
 Lee, Eun-ser II-819  
 Lee, Geuk II-701, III-1042, III-1075  
 Lee, Hanho III-108  
 Lee, Hong Joo I-1234, I-1259  
 Lee, Hongwon III-473  
 Lee, Hsuan-Shih II-896, II-902, II-910,  
 II-917, II-922  
 Lee, Huey-Ming III-938  
 Lee, Hwa-Ju I-483  
 Lee, Hyun-Gu I-590  
 Lee, Hyun-Jae II-829  
 Lee, Im-Yeong II-793  
 Lee, JangMyung I-890  
 Lee, Jee Hyung I-401  
 Lee, Jeong-On III-225  
 Lee, Jong Hyuk Park Sangjin II-777  
 Lee, JooYoung I-1250  
 Lee, Joon-Sung II-829  
 Lee, Ju-Hong III-84  
 Lee, Junkyu III-465  
 Lee, Kang Woong I-29, I-858, I-866  
 Lee, Keon Myung I-401  
 Lee, Kyoung Jun I-1267  
 Lee, Kyung-Sook I-483  
 Lee, Moo-hun III-1058  
 Lee, Raymond S.T. II-637, II-644,  
 II-652  
 Lee, Sangjin II-777  
 Lee, Sang Wan III-248  
 Lee, Sang-Joong III-893  
 Lee, Sang-Wook I-122, I-130  
 Lee, Sang-Yun I-443  
 Lee, Seok-Joo I-590  
 Lee, SeongHoon I-110, I-1124, I-1139,  
 I-1147, I-1155  
 Lee, Seung Wook I-401  
 Lee, Seungjae III-556  
 Lee, Shaun H. I-1179, I-1206  
 Lee, Soon Woong I-78  
 Lee, Sungdoke I-260  
 Lee, Tsair-Fwu I-179, II-1  
 Lee, Tsang-Yean III-938  
 Lee, Yang-Weon I-598  
 Lee, Yong Kyu I-292, I-992  
 Lee, Young-Kyun III-541  
 Leem, Choon Seong I-1259  
 Lefkaditis, Dionisios I-1226  
 Lehtikunnas, Tuija I-720  
 Lemos, João M. II-1248  
 Leng, Jinsong II-472  
 Lenič, Mitja III-515  
 León, Carlos I-475  
 Levachkine, Sergei I-698  
 Levachkine, Serguei I-550  
 Lewis, Chris J. III-349  
 Lewkowicz, Myriam I-1131  
 Lhotská, Lenka II-409  
 Li, Francis F. III-1208  
 Li, Jiuyong I-212, I-976  
 Li, Ming I-21  
 Li, Peng III-507  
 Li, Xun III-90  
 Li, Yanhui I-647, I-655  
 Li, Yueli I-498, I-714, I-1029  
 Li, Zhonghua I-153  
 Liang, Liang I-1106  
 Liang, Yanchun I-21  
 Licchelli, Oriana III-606  
 Ligon, Gopinathan L. III-1192  
 Lim, Jounghoon I-559  
 Lim, Myo-Taeg III-225  
 Lin, Cheng II-561  
 Lin, Geng-Sian III-46  
 Lin, Kuang II-922  
 Lin, Lily III-930, III-956  
 Lin, Mu-Hua III-46  
 Lin, Zuoquan I-459  
 Liu, Baoyan I-145, I-498, I-714, I-1029  
 Liu, Hongbo II-500  
 Liu, Honghai I-639, I-825, I-1198  
 Liu, Ju II-28  
 Liu, Jun II-171  
 Liu, Qingshan II-47  
 Liu, Xianxing II-204  
 López-Peña, Fernando III-292

- Lops, Pasquale III-606  
 Lorenzo, Gianluca III-1092  
 Lorkiewicz, Wojciech III-1004  
 Lortal, Gaëlle I-1131  
 Lovrek, Ignac I-318  
 Lu, Hanqing II-47  
 Lu, Jianjiang I-647, I-655  
 Lu, Peng I-780  
 Lun, Xiangmin II-19  
 Luukka, Pasi III-383
- Ma, Songde II-47  
 Ma, Wanli III-1176, III-1184  
 Macaš, Martin II-409  
 Mackin, Kenneth J. III-820  
 Magalhães, Hugo II-1248  
 Mahalik, Nitaigour Premchand  
 I-122, I-130  
 Maisonnasse, Jérôme II-162  
 Mak, Raymond Yiu Wai II-652  
 Malski, Michał II-251  
 Mandić, Danilo P. III-1216, III-1232,  
 III-1248, III-1280  
 Mao, Yong I-171  
 Marcenaro, Lucio II-179  
 Martinez, Miguel I-698  
 Masuda, Tsuyoshi II-220  
 Mathur, Abhishek III-1176  
 Matijasevic, Stjepan I-284  
 Matsuda, Noriyuki II-620, II-628  
 Matsui, Nobuyuki III-699  
 Matsui, Tatsunori II-977  
 Matsumoto, Hideyuki II-561  
 Matsuno, Takuma III-181  
 Matsuura, Kyoichi II-1010  
 Matsuyama, Hisayoshi II-579  
 Matta, Nada I-687  
 Mattila, Jorma K. III-358  
 Mendonça, Teresa F. II-1248  
 Mera, Kazuya III-749  
 Miaoulis, Georgios I-425  
 Mikac, Branko I-318  
 Mille, Alain I-1171  
 Mineno, Hiroshi III-150, III-159,  
 III-166, III-189  
 Misue, Kazuo III-835, III-843  
 Mitani, Tsubasa II-1103  
 Mitsuishi, Takashi II-1027  
 Miura, Hirokazu II-620, II-628  
 Miura, Motoki II-1019
- Miyachi, Taizo II-603  
 Miyadera, Youzou II-1035  
 Miyaji, Masako III-725  
 Miyamoto, Sadaaki II-78  
 Mizuno, Tadanori III-150, III-159,  
 III-166, III-189  
 Mo, Eun Jong I-29  
 Molan, Gregor I-360  
 Molan, Marija I-360  
 Molina, Javier I-475  
 Montero-Orille, Carlos III-300  
 Moon, Daesung I-906  
 Moon, Il-Young I-467  
 Moreno, Antonin II-1256, III-758  
 Moreno, Francisco III-406  
 Moreno, Marco I-550, I-614, I-698  
 Moret-Bonillo, Vicente III-1136  
 Morihiro, Koichiro III-699  
 Morita, Kazuhiro II-303, II-317, II-325  
 Moriyama, Jun II-603  
 Mosqueira-Rey, Eduardo III-1136  
 Motoyama, Jun-ichi II-1162  
 Mouri, Katsushiro II-1045  
 Mouzakidis, Alexandros II-1152  
 Mukai, Naoto I-1059, II-1095  
 Munemori, Jun III-174, III-212  
 Murai, Takeshi II-393  
 Murakami, Jin'ichi III-715  
 Murase, Yosuke II-1053  
 Mure, Yuji II-1103  
 Musiał, Katarzyna II-417
- Na, Yun Ji III-1035, III-1050  
 Naganuma, Takefumi I-1021  
 Nagar, Atulya K. I-1051  
 Nagasawa, Isao II-1103  
 Nagasawa, Shin'ya III-980  
 Naito, Takeshi III-1272  
 Nakamatsu, Kazumi II-844, II-851,  
 II-858, II-871  
 Nakamura, Hiroshi II-1071  
 Nakamura, Shoichi II-1035  
 Nakano, Ryohei II-945, II-969  
 Nakano, Tomofumi II-1162  
 Nakano, Yukiko II-220  
 Nakao, Zensho II-55  
 Nakazono, Nagayoshi III-843  
 Nam, Mi Young I-368, I-532,  
 III-116, III-124  
 Naoe, Yukihisa III-189

- Nara, Yumiko III-64  
 Nehaniv, Chrystopher L. III-333  
 Neves, José I-1083  
 Nguyen, Ngoc Thanh II-267, III-805  
 Niimi, Ayahiko III-813  
 Niimura, Masaaki II-994  
 Nikiforidis, George I-515  
 Nishida, Toyoaki II-220  
 Nishikawa, Ikuko II-953  
 Nishimoto, Kazushi III-859  
 Nishimura, Haruhiko III-699  
 Nishimura, Shinichi III-174  
 Nitta, Tsuneo II-1010  
 Noda, Manabu II-1045  
 Nojima, Yusuke II-86  
 Nouno, Ikue II-953  
 Ntoutsis, Christos II-1152  
 Numao, Masayuki II-929  
 Nunes, Catarina S. II-1248  
 Nunohiro, Eiji III-820  
 Nürnberger, Andreas I-763
- O'Grady, Michael J. II-686, III-1201  
 O'Hare, Gregory M.P. II-686, III-1201  
 O'Riordan, Colm III-766  
 Odagiri, Kazuya II-379  
 Oeda, Shinichi III-742  
 Oehlmann, Ruediger III-57  
 Ogawa, Hisashi II-620  
 Oh, Chang-Heon II-829  
 Oh, Jae-Yong I-598  
 Oh, Tae-Kyoo II-812, III-900  
 Ohsawa, Yukio III-38  
 Ohshiro, Masanori III-820  
 Okamoto, Takeshi II-123  
 Okumura, Noriyuki I-506  
 Okuno, Masaaki II-387  
 Orłowski, Cezary I-671  
 Oyarzun, Joaquín I-679  
 Ozaki, Masahiro II-1170, II-1176  
 Ozaku, Hiromi Itoh III-22  
 Ozkarahan, Irem I-37
- Pacheco, Marco Aurélio C. III-307  
 Pahikkala, Tapio I-720  
 Palade, Vasile III-487  
 Paliouras, Georgios II-1152  
 Palmisano, Ignazio III-606  
 Pandzic, Igor S. II-220
- Pant, Millie I-951  
 Papageorgiou, Elpiniki I-515  
 Park, Byungkwan I-906  
 Park, Chang-Hyun III-241  
 Park, Choung-Hwan III-533  
 Park, Gil-Cheol II-694, III-1075  
 Park, Gwi-Tae I-590, III-255, III-670  
 Park, Hee-Un II-726  
 Park, Heejun I-1234  
 Park, Hyun-gun II-819  
 Park, Jeong-Hyun III-135  
 Park, Jin-Won I-906  
 Park, Jinsub III-1068  
 Park, Jong Hyuk II-777  
 Park, Jong Kang III-907  
 Park, Jong Sou I-935  
 Park, Jung-Il I-850  
 Park, KeeHyun I-276  
 Park, Kiheon II-545  
 Park, Namje I-924  
 Park, Soohong I-203  
 Park, Sun III-84  
 Patel, Meera III-57  
 Pei-Shu, Fan II-879  
 Peng, Lizhi III-398  
 Petridis, Kosmas III-633  
 Pham, Tuan D. I-524  
 Philips, Wilfried I-623  
 Phillips-Wren, Gloria II-515, II-531,  
 III-504  
 Pi, Daoying I-171  
 Pieczyńska, Agnieszka II-227, III-1012  
 Plemenos, Dimitri I-425  
 Polymenakos, Lazaros C. III-1224  
 Pontes, Beatriz II-1264, II-1272  
 Popek, Grzegorz III-997  
 Posada, Jorge I-679  
 Pousada-Carballo, J.M. III-263  
 Prado, João Carlos Almeida II-844  
 Prentzas, Jim I-968  
 Price, Don C. III-349  
 Prié, Yannick I-1171  
 Prieto, Abraham III-292  
 Prieto-Blanco, Xesus III-300  
 Prokopenko, Mikhail III-315, III-324  
 Pu, Geguang I-459
- Qian, Zuoqin I-1198  
 Qiao, Jianping II-28  
 Qin, Jun II-1087

- Qiu, Zongyan I-459  
 Qudeiri, Jaber Abu I-252  
 Quintero, Rolando I-550, I-614  
  
 Ratanamahatana, Chotirat Ann  
     III-733  
 Redavid, Domenico III-606  
 Regazzoni, Carlo S. II-179  
 Reignier, Patrick II-162  
 Ren, Xiang II-1191  
 Resta, Marina III-641  
 Rhee, Phill Kyu I-68, I-78, I-368,  
     I-532, I-569, I-582, III-116, III-124,  
     III-707  
 Ridgewell, Alexander III-1163  
 Ríos, Sebastián A. II-371  
 Rodríguez-Hernández, P.S. III-263  
 Rodríguez, Jesús Barrasa III-588  
 Roh, Seok-Beom III-219  
 Romero, Sixto III-406  
 Ross, Robert I-102  
 Rousselot, François I-1098  
 Ruiz, Francisco J. II-425  
 Ruiz, Roberto II-1272  
 Rutkowski, Tomasz M. III-1216,  
     III-1232  
  
 Saathoff, Carsten III-633  
 Sadi, Mohammed Golam I-874  
 Saito, Kazumi II-929, II-937,  
     II-945, II-969  
 Sáiz, José Manuel II-433  
 Sakakibara, Kazutoshi II-953  
 Sakamoto, Hirotaka II-953  
 Sakamoto, Junichi II-628  
 Sakurai, Kouichi II-737  
 Salakoski, Tapio I-720  
 Salanterä, Sanna I-720  
 Sánchez, Daniel II-355  
 Sánchez, David III-758  
 Sánchez-Fernández, Luis III-580  
 Sánchez-Solano, Santiago II-363  
 Sánchez, Omar III-406  
 Sanin, Cesar I-663  
 Sarker, M. Omar Faruque I-874  
 Saruwatari, Yasufumi II-281  
 Sasaki, Mizuho II-611  
 Sato, Eri III-725  
 Sato, Hideki II-1216  
 Sato, Yoshiharu II-94  
  
 Sato-Ilic, Mika II-102, II-110  
 Savvopoulos, A. I-960  
 Schetinin, Vitaly III-523  
 Schmidt, Rainer I-326, I-334  
 Schulz, Klaus U. III-614  
 Scott, D. Andrew III-349  
 Seising, Rudolf III-366, III-374  
 Semeraro, Giovanni III-606  
 Seno, Yasuhiro III-189  
 Seo, Duck Won I-542  
 Seo, Young Hwan I-1267  
 Settouti, Lotfi S. I-1171  
 Sharma, Dharmendra III-1150,  
     III-1163, III-1168, III-1176, III-1184,  
     III-1192  
 Shi, Ruihong I-780  
 Shi, Xiaohu I-21  
 Shih, Ching-Nan I-179  
 Shiizuka, Hisao III-988  
 Shim, Bo-Yeon II-803  
 Shim, Donghee I-110  
 Shimada, Yukiyasu II-553, II-579,  
     II-587  
 Shimizu, Toru III-189, II-1224  
 Shimodaira, Chie III-867  
 Shimodaira, Hiroshi III-867  
 Shin, Dong-Myung II-726  
 Shin, Ho-Jun II-803  
 Shin, Miyoung I-1043  
 Shin, Myong-Chul II-812, III-883,  
     III-900  
 Shin, Woochul III-90  
 Shinohara, Shuji II-1010  
 Shirai, Hirokazu II-1035  
 Shoji, Hiroko III-8, III-16  
 Sidhu, Amandeep S. I-728  
 Siemiński, Andrzej III-782  
 Silva, José Demisio S. da I-268  
 Sim, Kwee-Bo III-241  
 Simoff, Simeon I-228  
 Sinkovic, Vjekoslav I-236  
 Sioutis, Christos II-450, II-464  
 Sirois, Bill III-1264  
 Skourlas, Christos II-1152  
 Stanina, Marta III-797  
 Soak, Sang-Moon I-122, I-130  
 Sobecki, Janusz III-797  
 Soh, Wooyoung II-709, III-1068  
 Sohn, Hong-Gyoo III-533  
 Sohn, Sang-Wook I-817

- Sokhi, Dilbag I-1051  
 Solazzo, Gianluca III-1092, III-1102  
 Sommer, David III-1256, III-1264  
 Song, Hyunsoo I-916  
 Song, MoonBae I-1139  
 Song, Yeong-Sun III-533  
 Sorensen, Humphrey III-766  
 Sotiropoulos, D.N. I-960  
 Soto-Hidalgo, Jose M. II-355  
 Staab, Steffen III-633  
 Stathopoulou, Ioanna-Ourania II-1128  
 Sterpi, Dario II-442  
 Stiglic, Gregor II-1297  
 Stober, Sebastian I-763  
 Streichert, Felix III-647, III-655  
 Stylios, Chrysostomos I-515  
 Sucar, Enrique I-943  
 Sugano, Naotoshi III-948  
 Sugiki, Daigo II-63  
 Sugino, Yoshiki II-961  
 Suh, Il Hong I-559  
 Suh, Jae Won I-808, I-817  
 Suh, Jungwon III-480  
 Sujitjorn, Sarawut I-833  
 Sumitomo, Toru II-275  
 Sun, Yong I-171  
 Sun, Zhaohao I-1115  
 Suominen, Hanna I-720  
 Suzuki, Hideharu III-159, III-166  
 Suzuki, Nobuo II-296  
 Suzuki, Susumu II-393  
 Szczerbicka, Helena I-706  
 Szczerbicki, Edward I-663
- Tabakow, Iwan II-187  
 Tabata, Toshihiro II-737  
 Tadauchi, Masaharu II-379  
 Takahashi, Hiroshi II-310  
 Takahashi, Satoru II-289, II-310  
 Takahashi, Shin III-197  
 Takahashi, Masakazu II-289, II-310  
 Takai, Toshihiro II-401  
 Takata, Osamu II-1103  
 Takeda, Kazuhiro II-553, II-579, II-587  
 Taki, Hirokazu II-611, II-620, II-628  
 Tan, Hong-Zhou I-153  
 Tanahashi, Yusuke II-969  
 Tanaka, Jiro III-197, III-835, III-843  
 Tanaka, Kaoru III-851  
 Tanaka, Kiyoko III-159, III-166
- Tanaka, Toshihisa III-1248  
 Tanaka-Yamawaki, Mieko III-647, III-655  
 Tang, Ming Xi II-670  
 Taniguchi, Rin-ichiro II-212  
 Tarasenko, Kateryna II-220  
 Tatara, Kohei II-737  
 Tawfik, Hissam I-1051  
 Termenón, Maite I-679  
 Thatcher, Steve II-508  
 Tigan, Stefan III-1128  
 Timmermann, Norman III-633  
 Todirascu-Courtier, Amalia I-1131  
 Tokuhisa, Masato III-715  
 Tommasi, Maurizio De III-1083  
 Toro, Carlos I-679  
 Torres, Cláudio Rodrigo II-851  
 Torres, Germano L. II-851  
 Torres, Miguel I-550, I-614, I-698  
 Torres-Schumann, Eduardo III-614  
 Tran, Dat T. I-524, III-1176, III-1184  
 Trawiński, Bogdan III-774  
 Trutschel, Udo III-1264  
 Trzec, Krunoslav I-318  
 Tsang, Edward P.K. III-30  
 Tseng, Wei-Kuo II-910  
 Tsihrintzis, George A. I-376, I-384, I-960, II-1128, II-1289  
 Tsuchiya, Seiji I-1002  
 Tsuda, Kazuhiko II-281, II-296, II-310  
 Tsuge, Yoshifumi II-579  
 Tweedale, Jeffrey II-450, II-464, II-479, II-486, III-497
- Uchida, Seiichi II-212  
 Umeda, Masanobu II-1103  
 Unno, Shunsuke II-1071  
 Urlings, Pierre II-450  
 Ushiana, Taketoshi II-1111, II-1224, II-1232  
 Ushiro, Mika II-994
- Vales-Alonso, J. II-659  
 Valls, Aïda II-1256  
 Van der Weken, Dietrich I-623  
 Vansteenkiste, Ewout I-623  
 Vayanos, Phebe III-1216  
 Velásquez, Juan D. II-371, III-487  
 Vélez, Miguel A. III-406  
 Verastegui, Karina I-698



- Vialatte, Francois III-1232  
 Vidal, Juan Carlos III-623  
 Virvou, Maria I-376, I-435,  
     I-960, II-1289  
 Vorobieva, Olga I-334
- Wada, Masaaki III-813  
 Waldeck, Carsten I-753  
 Waligora, Tina I-326  
 Walters, Simon D. I-1179, I-1206  
 Wang, Hongmoon I-401  
 Wang, Hua I-976  
 Wang, Jian Xun II-670  
 Wang, Jin-Long II-888  
 Wang, Leuo-Hong III-1  
 Wang, Limin I-21  
 Wang, Peter III-315  
 Wang, Xiaodong II-11  
 Wang, Yupeng III-457  
 Wang, Zhengyou I-631  
 Wang, Zhijian II-1199  
 Wang, Zhong-xian I-52  
 Washizawa, Yoshikazu III-1248  
 Watabe, Hirokazu I-506, I-882, I-1002  
 Watada, Junzo III-922, III-964, III-972  
 Watanabe, Toyohide I-771, I-1037,  
     I-1059, II-1053, II-1062, II-1095,  
     II-1111, II-1119, II-1224,  
     II-1232, III-827  
 Watanabe, Yuji II-131  
 Weigel, Felix III-614  
 Wen, YuanLin II-879  
 Won, Chung-Yuen III-875  
 Won, Dongho I-924  
 Won, Hee-Sun I-443  
 Wong, Ka Yan III-269  
 Wong, S.T.C. I-171  
 Wu, Linyan I-1037  
 Wu, Menq-Jiun I-300  
 Wu, Shiqian I-631
- Xia, Zheng I-171  
 Xu, Baowen I-647, I-655  
 Xu, Hao I-714  
 Xu, Zhiping III-390  
 Xue, Peng III-441  
 Xue, Quan I-631
- Yaegashi, Rihito II-379  
 Yamada, Kunihiko III-150, III-189  
 Yamada, Takahiro II-393  
 Yamaguchi, Toru III-725  
 Yamakami, Toshihiko III-143  
 Yamamoto, Hidehiko I-252  
 Yamasaki, Kazuko III-820  
 Yamashita, Yoshiyuki II-595  
 Yamawaki, Shigenobu II-866  
 Yan, Wang II-47  
 Yang, Bingru I-145, I-498, I-714, I-1029  
 Yang, Byoungnak I-60  
 Yang, Chih Chieh I-1  
 Yang, Guosheng II-204  
 Yang, Hsiao-Fang III-46  
 Yang, Jueng-je I-52  
 Yang, Jung-Jin II-492  
 Yang, Yongqing II-1199  
 Yang, Yuxiang III-507  
 Yasuda, Hiroshi II-371  
 Yasuda, Takami II-1045, II-1079  
 Yeh, Chen-Huei II-917  
 Yildiz, Ali Bekir I-1075  
 Yip, Chi Lap III-269  
 Yokoi, Shigeki II-1045, II-1079  
 Yokoyama, Setsuo II-1035  
 Yoo, Dong-Wook III-875  
 Yoo, Sang Bong III-90  
 Yoo, Seong Joon I-753  
 Yoo, Weon-Hee I-451, I-898  
 Yoon, Chang-Dae III-900  
 Yoon, Seok Min III-914  
 Yoshida, Jun II-281  
 Yoshida, Koji III-189  
 Yoshino, Takashi III-181, III-205  
 You, Bum-Jae I-874  
 You, Ilsun II-785  
 You, Kang Soo I-542  
 You, Kwanho II-545  
 Youk, Sang Jo III-1042  
 Yu, Gang III-507  
 Yu, Jae-Sung III-875  
 Yuizono, Takaya III-174  
 Yukimura, Yoko III-205  
 Yüksel, Mehmet Emin II-331  
 Yun, Al-Chan I-817  
 Yun, Byoung-Ju I-1011  
 Yun, JaeMu I-890
- Zanni, Cecilia I-1098  
 Zatwarnicki, Krzysztof II-195  
 Zazula, Damjan III-515

Zeman, Astrid III-315, III-324  
Zeng, Weiming I-631  
Zhang, Changjiang II-11  
Zhang, Fan II-204  
Zhang, Guangde I-1198  
Zhang, Miao II-678  
Zhang, Naixiao II-1144  
Zhang, Shiyong III-390  
Zhang, Weishi II-500  
Zhang, Xinhong II-204  
Zhang, Yonggang III-1208

Zhao, Lei II-678  
Zhao, Shuo I-145, I-498  
Zharkov, Sergei III-523  
Zharkova, Valentina III-523  
Zhong, Yiping III-390  
Zhou, Xia I-171  
Zhou, Yi I-738  
Zhu, Chaopin III-1280  
Zhu, Yuelong I-1037, II-1191  
Zong, Ping II-1087

# ERRATUM

## Comparison of Squall Line Positioning Methods Using Radar Data

Ka Yan Wong and Chi Lap Yip

Dept. of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong  
{kywong, clyip}@cs.hku.hk

B. Gabrys, R.J. Howlett, and L.C. Jain (Eds.): KES 2006, Part III, LNAI 4253, pp. 269–276, 2006.  
© Springer-Verlag Berlin Heidelberg 2006

DOI 10.1007/11893011\_163

An error has been found in the above article.

In the original version of this paper, Fig. 3c was wrong. The correct version of figure is given below.

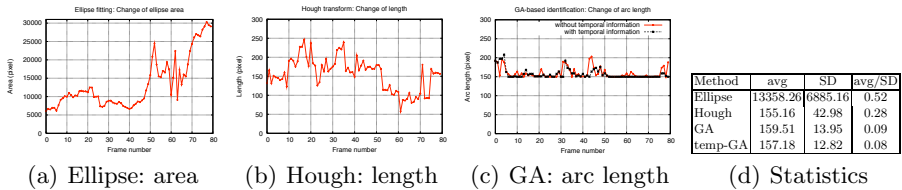


Fig. 3. Temporal stability