

James J. (Jong Hyuk) Park  
Joseph Kee-Yin Ng  
Hwa Young Jeong  
Borgy Waluyo *Editors*

# Multimedia and Ubiquitous Engineering

MUE 2013

Volume I

# Lecture Notes in Electrical Engineering

Volume 240

For further volumes:  
<http://www.springer.com/series/7818>

James J. (Jong Hyuk) Park  
Joseph Kee-Yin Ng · Hwa Young Jeong  
Borgy Waluyo  
Editors

# Multimedia and Ubiquitous Engineering

MUE 2013

 Springer

*Editors*

James J. (Jong Hyuk) Park  
Department of Computer Science  
Seoul University of Science  
and Technology (SeoulTech)  
Seoul  
Republic of South Korea

Hwa Young Jeong  
Humanitas College  
Kyung Hee University  
Seoul  
Republic of South Korea

Joseph Kee-Yin Ng  
Department of Computer Science  
Hong Kong Baptist University  
Kowloon Tong  
Hong Kong SAR

Borgy Waluyo  
School of Computer Science  
Monash University  
Clayton, VIC  
Australia

ISSN 1876-1100

ISSN 1876-1119 (electronic)

ISBN 978-94-007-6737-9

ISBN 978-94-007-6738-6 (eBook)

DOI 10.1007/978-94-007-6738-6

Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013936524

© Springer Science+Business Media Dordrecht(Outside the USA) 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law. The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))



# Message and Organization Committee

## Message from the MUE 2013 General Chairs

MUE 2013 is the FTRA 7th event of the series of international scientific conferences. This conference will take place in May 9–11, 2013, in Seoul Korea. The aim of the MUE 2013 is to provide an international forum for scientific research in the technologies and application of Multimedia and Ubiquitous Engineering. It is organized by the Korea Information Technology Convergence Society in cooperation with Korea Information Processing Society. MUE2013 is the next event in a series of highly successful international conferences on Multimedia and Ubiquitous Engineering, MUE-12 (Madrid, Spain, July 2012), MUE-11 (Loutraki, Greece, June 2011), MUE-10 (Cebu, Philippines, August 2010), MUE-09 (Qingdao, China, June 2009), MUE-08 (Busan, Korea, April 2008), and MUE-07 (Seoul, Korea, April 2007).

The papers included in the proceedings cover the following topics: *Multimedia Modeling and Processing, Ubiquitous and Pervasive Computing, Ubiquitous Networks and Mobile Communications, Intelligent Computing, Multimedia and Ubiquitous Computing Security, Multimedia and Ubiquitous Services, Multimedia Entertainment, IT and Multimedia Applications*. Accepted and presented papers highlight new trends and challenges of Multimedia and Ubiquitous Engineering. The presenters showed how new research could lead to novel and innovative applications. We hope you will find these results useful and inspiring for your future research.

We would like to express our sincere thanks to Steering Chairs: James J. (Jong Hyuk) Park (SeoulTech, Korea), Martin Sang-Soo Yeo (Mokwon University, Korea). Our special thanks go to the Program Chairs: Eunyoung Lee (Dongduk Women's University, Korea), Cho-Li Wang (University of Hong Kong, Hong Kong), Borgy Waluyo (Monash University, Australia), Al-Sakib Khan Pathan (IIUM, Malaysia), SangHyun Seo (University of Lyon 1, France), all Program Committee members and all the additional reviewers for their valuable efforts in the review process, which helped us to guarantee the highest quality of the selected papers for the conference.

We cordially thank all the authors for their valuable contributions and the other participants of this conference. The conference would not have been possible without their support. Thanks are also due to the many experts who contributed to making the event a success.

May 2013

Young-Sik Jeong  
Leonard Barolli  
Joseph Kee-Yin Ng  
C. S. Raghavendra  
MUE 2013 General Chairs

## Message from the MUE 2013 Program Chairs

Welcome to the FTRA 7th International Conference on Multimedia and Ubiquitous Engineering (MUE 2013), to be held in Seoul, Korea on May 9–11, 2013. MUE 2013 will be the most comprehensive conference focused on the various aspects of multimedia and ubiquitous engineering. MUE 2013 will provide an opportunity for academic and industry professionals to discuss recent progress in the area of multimedia and ubiquitous environment. In addition, the conference will publish high quality papers which are closely related to the various theories and practical applications in multimedia and ubiquitous engineering. Furthermore, we expect that the conference and its publications will be a trigger for further related research and technology improvements in these important subjects.

For MUE 2013, we received many paper submissions; after a rigorous peer review process, we accepted 62 articles with high quality for the MUE 2013 proceedings, published by Springer. All submitted papers have undergone blind reviews by at least two reviewers from the technical program committee, which consists of leading researchers around the globe. Without their hard work, achieving such a high-quality proceeding would not have been possible. We take this opportunity to thank them for their great support and cooperation. We would like to sincerely thank the following invited speaker who kindly accepted our invitations, and, in this way, helped to meet the objectives of the conference: Prof. Hui-Huang Hsu, Tamkang University, Taiwan. Finally, we would like to thank all of you for your participation in our conference, and also thank all the authors, reviewers, and organizing committee members. Thank you and enjoy the conference!

Eunyoung Lee, Korea  
Cho-Li Wang, Hong Kong  
Borgy Waluyo, Australia  
Al-Sakib Khan Pathan, Malaysia  
SangHyun Seo, France  
MUE 2013 Program Chairs

## Organization

- Honorary Chair: Makoto Takizawa, Seikei University, Japan
- Steering Chairs: James J. Park, SeoulTech, Korea  
Martin Sang-Soo Yeo, Mokwon University, Korea
- General Chairs: Young-Sik Jeong, Wonkwang University, Korea  
Leonard Barolli, Fukuoka Institute of Technology, Japan  
Joseph Kee-Yin Ng, Hong Kong Baptist University, Hong Kong  
C. S. Raghavendra, University of Southern California, USA
- Program Chairs: Eunyoung Lee, Dongduk Women's University, Korea  
Cho-Li Wang, University of Hong Kong, Hong Kong  
Borgy Waluyo, Monash University, Australia  
Al-Sakib Khan Pathan, IIUM, Malaysia  
SangHyun Seo, University of Lyon 1, France
- Workshop Chairs: Young-Gab Kim, Korea University, Korea  
Lei Ye, University of Wollongong, Australia  
Hiroaki Nishino, Oita University, Japan  
Neil Y. Yen, The University of Aizu, Japan
- Publication Chair: Hwa Young Jeong, Kyung Hee University, Korea
- International Advisory Committee: Seok Cheon Park, Gachon University, Korea  
Borko Furht, Florida Atlantic University, USA  
Thomas Plagemann, University of Oslo, Norway  
Roger Zimmermann, National University of Singapore, Singapore  
Han-Chieh Chao, National Ilan University, Taiwan  
Hai Jin, HUST, China  
Weijia Jia, City University of Hong Kong, Hong Kong  
Jianhua Ma, Hosei University, Japan  
Shu-Ching Chen, Florida International University, USA

	Hamid R. Arabnia, The University of Georgia, USA
	Stephan Olariu, Old Dominion University, USA
	Albert Zomaya, University of Sydney, Australia
	Bin Hu, Lanzhou University, China
	Yi Pan, Georgia State University USA
	Doo-soon Park, SoonChunHyang University, Korea
	Richard P. Brent, Australian National University, Australia
	Koji Nakano, University of Hiroshima, Japan
	J. Daniel Garcia, University Carlos III of Madrid, Spain
	Qun Jin, Waseda University, Japan
	Kyung-Hyune Rhee, Pukyong National University, Korea
Publicity Chairs:	Chengcui Zhang, The University of Alabama at Birmingham, USA
	Michele Ruta, Politecnico di Bari, Italy
	Bessam Abdulrazak, Sherbrooke University, Canada
	Junaid Chaudhry, Universiti Teknologi Malaysia, Malaysia
	Bong-Hwa Hong, Kyung Hee Cyber University, Korea
	Won-Joo Hwang, Inje University, Korea
Local Arrangement Chairs:	HyunsungKim, Kyungil University, Korea
	Eun-Jun Yoon, Kyungil University, Korea
Invited Speaker:	Hui-Huang Hsu, Tamkang University, Taiwan
Program Committee:	A Ra Khil, Soongsil University, Korea
	Afrand Agah, West Chester University of Pennsylvania, USA
	Akihiro Sugimoto, National Institute of Informatics, Japan
	Akimitsu Kanzaki, Osaka University, Japan
	Angel D. Sappa, Universitat Autònoma de Barcelona, Spain
	Bartosz Ziolkowski, AGH University of Science and Technology, Poland
	Bin Lu, West Chester University, USA

Brent Lagesse, BBN Technologies, USA  
 Ch. Z. Patrikakis, Technological Education  
 Institute of Piraeus, Greece  
 Chang-Sun Shin, Sunchon National Univer-  
 sity, Korea  
 Chantana Chantrapornchai, Silpakorn  
 University, Thailand  
 Chih Cheng Hung, Southern Polytechnic  
 State University, USA  
 Chulung Lee, Korea University, Korea  
 Dakshina Ranjan Kisku, Asansol Engineer-  
 ing College, India  
 Dalton Lin, National Taipei University,  
 Taiwan  
 Dariusz Frejlichowski, West Pomeranian  
 University of Technology, Poland  
 Debzani Deb, Winston-Salem State Univer-  
 sity, USA  
 Deqing Zou, Huazhong University of Sci-  
 ence and Technology, China  
 Ezendu Ariwa, London Metropolitan Uni-  
 versity, United Kingdom  
 Farid Meziane, University of Salford, UK  
 Florian Stegmaier, University of Passau,  
 Germany  
 Francisco Jose Monaco, University of Sao  
 Paulo, Brazil  
 Guillermo Camara Chavez, Universidade  
 Federal de Minas Gerais, Brazil  
 Hae-Young Lee, ETRI, Korea  
 Hai Jin, Huazhong University of Science and  
 Techn, China  
 Hangzai Luo, East China Normal University,  
 China  
 Harald Kosch, University of Passau, Ger-  
 many  
 Hari Om, Indian School of Mines University,  
 India  
 Helen Huang, The University of Queensland,  
 Australia  
 Hermann Hellwagner, Klagenfurt Univer-  
 sity, Austria  
 Hong Lu, Fudan University, China

Jeong-Joon Lee, Korea Polytechnic University, Korea  
Jin Kwak, Soonchunhyang University, Korea  
Jinye Peng, Northwest University, China  
Joel Rodrigue, University of Beira Interior, Portugal  
Jong-Kook Kim, Korea University, Korea  
Joyce El Haddad, Universite Paris-Dauphine, France  
Jungong Han, Civolution Technology, the Netherlands  
Jun-Won Ho, Seoul Women's University, Korea  
Kilhung Lee, Seoul National University of Science and Technology, Korea  
Klaus Schoffmann, Klagenfurt University, Austria  
Ko Eung Nam, Baekseok University, Korea  
Lidan Shou, Zhejiang University, China  
Lukas Ruf, CEO Consecom AG, Switzerland  
Marco Cremonini, University of Milan, Italy  
Maria Vargas-Vera, Universidad Adolfo Ibanez, Chile  
Mario Doeller, University of applied science, Germany  
Maytham Safar, Kuwait University, Kuwait  
Mehran Asadi, Lincoln University of Pennsylvania, USA  
Min Choi, Chungbuk National University, Korea  
Ming Li, California State University, USA  
Muhammad Younas, Oxford Brookes University, UK  
Namje Park, Jeju National University, Korea  
Neungsoo Park, Konkuk University, Korea  
Ning Zhou, University of North Carolina, USA  
Oliver Amft, TU Eindhoven, Netherlands  
Paisarn Muneesawang, Naresuan University, Thailand  
Pascal Lorenz, University of Haute Alsace  
Quanqing Xu, Quanqing Xu, Data Storage Institute, A\*STAR, Singapore  
Rachid Anane, Coventry University, UK

Rainer Unland, University of Duisburg-Essen, Germany  
 Rajkumar Kannan, Affiliation Bishop Heber College, India  
 Ralf Klamma, RWTH Aachen University, Germany  
 Ramanathan Subramanian, Advanced Digital Sciences Center, Singapore  
 Reinhard Klette, The University of Auckland, New Zealand  
 Rene Hansen, Aalborg University, Denmark  
 Sae-Hak Chun, Seoul National University of Science and Technology, Korea  
 Sagarmay Deb, University of Southern Queensland, Australia  
 Savvas Chatzichristofis, Democritus University of Thrace, Greece  
 Seung-Ho Lim, Hankuk University of Foreign Studies, Korea  
 Shingo Ichii, University of Tokyo, Japan  
 Sokratis Katsikas, University of Piraeus, Greece  
 SoonSeok Kim, Halla University, Korea  
 Teng Li, Baidu Inc., China  
 Thomas Grill, University of Salzburg, Austria  
 Tingxin Yan, University of Arkansas, USA  
 Toshihiro Yamauchi, Okayama University, Japan  
 Waleed Farag, Indiana University of Pennsylvania, USA  
 Wee Siong Ng, Institute for Infocomm Research, Singapore  
 Weifeng Chen, California University of Pennsylvania, USA  
 Weifeng Zhang, Nanjing University of Posts and Telecommunication, China  
 Wesley De Neve, Ghent University iMinds and KAIST,  
 Won Woo Ro, Yonsei University, Korea  
 Wookho Son, ETRI, Korea  
 Wei Wei, Xi'an University of Technology, China

Xubo Song, Oregon Health and Science University, USA

Yan Liu, The Hong Kong Polytechnic University, Hong Kong

Yijuan Lu, Texas State University, USA

Yingchun Yang, Zhejiang University, China

Yong-Yoon Cho, Suncheon University, Korea

Yo-Sung Ho, GIST, Korea

Young-Hee Kim, Korea Copyright Commission, Korea

Young-Ho Park, Sookmyung Women's University, Korea

Zheng-Jun Zha, National University of Singapore, Singapore

Zhu Li, Samsung Telecom America, USA

## **Message from ATACS-2013 Workshop Chair**

Welcome to the Advanced Technologies and Applications for Cloud Computing and Sensor Networks (ATACS-2013), which will be held from May 9 to 11, 2013 in Seoul, Korea.

The main objective of this workshop is to share information on new and innovative research related to advanced technologies and applications in the areas of cloud computing and sensor networks. Many advanced techniques and applications in these two areas have been developed in the past few years. Sensor networks are becoming increasingly large and produce vast amounts of raw sensing data, which cannot be easily processed, analyzed, or stored using conventional computing systems. Cloud computing is one promising technique of efficiently processing these sensing data to create useful services and applications. The convergence of cloud computing and sensor networks requires new and innovative infrastructure, middleware, designs, protocols, services, and applications. ATACS-2013 will bring together researchers and practitioners interested in both the technical and applied aspects of Advanced Techniques and Application for Cloud computing and Sensor networks. Furthermore, we expect that the ATACS-2013 and its publications will be a trigger for further related research and technology improvements in this important subject.

ATACS-2013 contains high quality research papers submitted by researchers from all over the world. Each submitted paper was peer-reviewed by reviewers who are experts in the subject area of the paper. Based on the review results, the Program Committee accepted eight papers.



I hope that you will enjoy the technical programs as well as the social activities during ATACS-2013. I would like to send our sincere appreciation to all of the Organizing and Program Committees who contributed directly to ATACS-2013. Finally, we special thank all the authors and participants for their contributions to make this workshop a grand success.

Joon-Min Gil  
Catholic University of Daegu  
ATACS-2013 Workshop Chair

## ATACS-2013 Organization

General Chair:	Joon-Min Gil, Catholic University of Daegu, Korea
Program Chair:	Jaehwa Chung, Korea National Open University, Korea
Publicity Chair:	Dae Won Lee, Seokyeong University, Korea
Program Committee:	Byeongchang Kim, Catholic University of Daegu, Korea
	Hansung Lee, Electronics and Telecommunications Research Institute (ETRI), Korea
	HeonChang Yu, Korea University, Korea
	Jeong-Hyon Hwang, State University of New York at Albany, USA
	JongHyuk Lee, Samsung Electronics, Korea
	Ki-Sik Kong, Namseoul University, Korea
	KwangHee Choi, LG Uplus, Korea
	Kwang Sik Chung, Korea National Open University, Korea
	Mi-Hye Kim, Catholic University of Daegu, Korea
	Shanmugasundaram Hariharan, TRP Engineering College (SRM Group), India
	Sung-Hwa Hong, Mokpo National Maritime University, Korea
	Sung Suk Kim, Seokyeong University, Korea
	Tae-Gyu Lee, Korea Institute of Industrial Technology (KITECH), Korea
	Tae-Young Byun, Catholic University of Daegu, Korea
	Ui-Sung Song, Busan National University of Education, Korea
	Yong-Hee Jeon, Catholic University of Daegu, Korea
	Yunhee Kang, Baekseok University, Korea
	Zhefu Shi, University of Missouri, USA

## Message from PSSI-2013 Workshop Chair

The organizing committee of the *FTRA International Workshop on Pervasive Services, Systems and Intelligence (PSSI 2013)* would like to welcome all of you to join the workshop as well as the FTRA MUE 2013. Advances in information and communications technology (ICT) have presented a dramatic growth in merging the boundaries between physical space and cyberspace, and go further to improve mankind's daily life. One typical instance is the use of smartphones. The modern smartphone is equipped with a variety of sensors that are used to collect activities, locations, and situations of its user continuously and provide immediate help accordingly. Some commercial products (e.g., smart house, etc.) also demonstrate the feasibility of comprehensive supports by deploying a rapidly growing number of sensors (or intelligent objects) into our living environments. These developments are collectively best characterized as ubiquitous service that promises to enhance awareness of the cyber, physical, and social contexts. As such, researchers (and companies as well) tend to provide tailored and precise solutions (e.g., services, supports, etc.) wherever and whenever human beings are active according to individuals' contexts. Making technology usable by and useful to, via the ubiquitous services and correlated techniques, humans in ways that were previously unimaginable has become a challenging issue to explore the picture of technology in the next era.

This workshop aims at providing a forum to discuss problems, studies, practices, and issues regarding the emerging trend of pervasive computing. Researchers are encouraged to share achievements, experiments, and ideas with international participants, and furthermore, look forward to map out the research directions and collaboration in the future.

With an amount of submissions (13 in exact), the organizing chairs decided to accept six of them based on the paper quality and the relevancy (acceptance rate at 46 %). These papers are from Canada, China, and Taiwan. Each paper was reviewed by at least three program committee members and discussed by the organizing chairs before acceptance.

We would like to thank three FTRA Workshop Chair, Young-Gab Kim from Korea University, Korea for the support and coordination. We thank all authors for submitting their works to the workshop. We also appreciate the program committee members for their efforts in reviewing the papers. Finally, we sincerely welcome all participants to join the discussion during the workshop.

James J. Park  
Neil Y. Yen  
Workshop Co-Chairs

# **FTRA International Workshop on Pervasive Services, Systems and Intelligence (PSSI-13)**

## **Workshop Organization**

Workshop Chairs:: James J. Park (Seoul National University of Science and Technology, Korea)  
Neil Y. Yen (The University of Aizu, Japan)

Program Committee:: Christopher Watson, Durham University, United Kingdom  
Chengjiu Yin, Kyushu University, Japan  
David Taniar, Monash University, Australia  
Jui-Hong Chen, Tamkang University, Taiwan  
Junbo Wang, the University of Aizu, Japan  
Lei Jing, the University of Aizu, Japan  
Marc Spaniol, Max-Planck-Institute for Informatic, Germany  
Martin M. Weng, Tamkang University, Taiwan  
Nigel Lin, Microsoft Research, United States  
Ralf Klamma, RWTH Aachen University, Germany  
Vitaly Klyuev, the University of Aizu, Japan  
Xaver Y. R. Chen, National Central University, Taiwan  
Wallapak Tavanapong, Iowa State University, United States  
Renato Ishii, Federal University of Mato Grosso do Sul, Brazil  
Nicoletta Sala, U. of Lugano, Switzerland and Università dell'Insubria Varese, Italy  
Yuanchun Shi, Tsinghua University, China  
Robert Simon, George Mason University, USA

# Contents

## Part I Multimedia Modeling and Processing

<b>Multiwedgelets in Image Denoising</b> . . . . .	3
Agnieszka Lisowska	
<b>A Novel Video Compression Method Based on Underdetermined Blind Source Separation</b> . . . . .	13
Jing Liu, Fei Qiao, Qi Wei and Huazhong Yang	
<b>Grid Service Matching Process Based on Ontology Semantic</b> . . . . .	21
Ganglei Zhang and Man Li	
<b>Enhancements on the Loss of Beacon Frames in LR-WPANs</b> . . . . .	27
Ji-Hoon Park and Byung-Seo Kim	
<b>Case Studies on Distribution Environmental Monitoring and Quality Measurement of Exporting Agricultural Products</b> . . . . .	35
Yoonsik Kwak, Jeongsam Lee, Sangmun Byun, Jeongbin Lem, Miae Choi, Jeongyong Lee and Seokil Song	
<b>Vision Based Approach for Driver Drowsiness Detection Based on 3DHead Orientation</b> . . . . .	43
Belhassen Akrouf and Walid Mahdi	
<b>Potentiality for Executing Hadoop Map Tasks on GPGPU via JNI</b> . . .	51
Bongen Gu, Dojin Choi and Yoonsik Kwak	
<b>An Adaptive Intelligent Recommendation Scheme for Smart Learning Contents Management Systems</b> . . . . .	57
Do-Eun Cho, Sang-Soo Yeo and Si Jung Kim	

**Part II Ubiquitous and Pervasive Computing**

**An Evolutionary Path-Based Analysis of Social Experience Design . . .** 69  
Toshihiko Yamakami

**Block IO Request Handling for DRAM-SSD in Linux Systems. . . . .** 77  
Kyungkoo Jun

**Implementation of the Closed Plant Factory System Based  
on Crop Growth Model. . . . .** 83  
Myeong-Bae Lee, Taehyung Kim, HongGeun Kim, Nam-Jin Bae,  
Miran Baek, Chang-Woo Park, Yong-Yun Cho and Chang-Sun Shin

**Part III Ubiquitous Networks and Mobile Communications**

**An Energy Efficient Layer for Event-Based Communications  
in Web-of-Things Frameworks. . . . .** 93  
G r me Bovet and Jean Hennebert

**A Secure Registration Scheme for Femtocell Embedded Networks . . .** 103  
Ikram Syed and Hoon Kim

**Part IV Intelligent Computing**

**Unsupervised Keyphrase Extraction Based Ranking Algorithm  
for Opinion Articles . . . . .** 113  
Heungmo Ryang and Unil Yun

**A Frequent Pattern Mining Technique for Ranking Webpages  
Based on Topics . . . . .** 121  
Gwangbum Pyun and Unil Yun

**Trimming Prototypes of Handwritten Digit Images with Subset  
Infinite Relational Model . . . . .** 129  
Tomonari Masada and Atsuhiko Takasu

**Ranking Book Reviews Based on User Influence. . . . .** 135  
Unil Yun and Heungmo Ryang

**Speaker Verification System Using LLR-Based Multiple  
Kernel Learning . . . . .** 143  
Yi-Hsiang Chao

**Edit Distance Comparison Confidence Measure for Speech Recognition** . . . . . 151  
 Dawid Skurzok and Bartosz Ziólko

**Weighted Pooling of Image Code with Saliency Map for Object Recognition** . . . . . 157  
 Dong-Hyun Kim, Kwanyong Lee and Hyeyoung Park

**Calibration of Urine Biomarkers for Ovarian Cancer Diagnosis** . . . . . 163  
 Yu-Seop Kim, Eun-Suk Yang, Kyoung-Min Nam, Chan-Young Park, Hye-Jung Song and Jong-Dae Kim

**An Iterative Algorithm for Selecting the Parameters in Kernel Methods** . . . . . 169  
 Tan Zhiying, She Kun and Song Xiaobo

**A Fast Self-Organizing Map Algorithm for Handwritten Digit Recognition** . . . . . 177  
 Yimu Wang, Alexander Peyls, Yun Pan, Luc Claesen and Xiaolang Yan

**Frequent Graph Pattern Mining with Length-Decreasing Support Constraints** . . . . . 185  
 Gangin Lee and Unil Yun

**An Improved Ranking Aggregation Method for Meta-Search Engine** . . . . . 193  
 Junliang Feng, Junzhong Gu and Zili Zhou

**Part V Multimedia and Ubiquitous Computing Security**

**Identity-Based Privacy Preservation Framework over u-Healthcare System** . . . . . 203  
 Kambombo Mtonga, Haomiao Yang, Eun-Jun Yoon and Hyunsung Kim

**A Webmail Reconstructing Method from Windows XP Memory Dumps** . . . . . 211  
 Fei Kong, Ming Xu, Yizhi Ren, Jian Xu, Haiping Zhang and Ning Zheng

**On Privacy Preserving Encrypted Data Stores** . . . . . 219  
 Tracey Raybourn, Jong Kwan Lee and Ray Kresman

**Mobile User Authentication Scheme Based on Minesweeper Game** . . . . . 227  
 Taejin Kim, Siwan Kim, Hyunyi Yi, Gunil Ma and Jeong Hyun Yi

**Design and Evaluation of a Diffusion Tracing Function for Classified Information Among Multiple Computers. . . . .** 235  
 Nobuto Otsubo, Shinichiro Uemura, Toshihiro Yamauchi and Hideo Taniguchi

**DroidTrack: Tracking Information Diffusion and Preventing Information Leakage on Android. . . . .** 243  
 Syunya Sakamoto, Kenji Okuda, Ryo Nakatsuka and Toshihiro Yamauchi

**Three Factor Authentication Protocol Based on Bilinear Pairing . . . .** 253  
 Thokozani Felix Vallent and Hyunsung Kim

**A LBP-Based Method for Detecting Copy-Move Forgery with Rotation . . . . .** 261  
 Ning Zheng, Yixing Wang and Ming Xu

**Attack on Recent Homomorphic Encryption Scheme over Integers. . .** 269  
 Haomiao Yang, Hyunsung Kim and Dianhua Tang

**A New Sensitive Data Aggregation Scheme for Protecting Data Integrity in Wireless Sensor Network. . . . .** 277  
 Min Yoon, Miyoung Jang, Hyoung-il Kim and Jae-woo Chang

**Reversible Image Watermarking Based on Neural Network and Parity Property . . . . .** 285  
 Rongrong Ni, H. D. Cheng, Yao Zhao, Zhitong Zhang and Rui Liu

**A Based on Single Image Authentication System in Aviation Security. . . . .** 293  
 Deok Gyu Lee and Jong Wook Han

**Part VI Multimedia and Ubiquitous Services**

**A Development of Android Based Debate-Learning System for Cultivating Divergent Thinking . . . . .** 305  
 SungWan Kim, EunGil Kim and JongHoon Kim

**Development of a Lever Learning Webapp for an HTML5-Based Cross-Platform . . . . .** 313  
 TaeHun Kim, ByeongSu Kim and JongHoon Kim

**Looking for Better Combination of Biomarker Selection and Classification Algorithm for Early Screening of Ovarian Cancer** . . . . . 321  
 Yu-Seop Kim, Jong-Dae Kim, Min-Ki Jang, Chan-Young Park and Hye-Jeong Song

**A Remote Control and Media Sharing System Based on DLNA/UPnP Technology for Smart Home.** . . . . . 329  
 Ti-Hsin Yu and Shou-Chih Lo

**A New Distributed Grid Structure for k-NN Query Processing Algorithm Based on Incremental Cell Expansion in LBSs.** . . . . . 337  
 Seungtae Hong, Hyunjo Lee and Jaewoo Chang

**A New Grid-Based Cloaking Scheme for Continuous Queries in Centralized LBS Systems** . . . . . 345  
 Hyeong-Il Kim, Mi-Young Jang, Min Yoon and Jae-Woo Chang

**New Database Mapping Schema for XML Document in Electronic Commerce** . . . . . 353  
 Eun-Young Kim and Se-Hak Chun

**A Study on the Location-Based Reservation Management Service Model Using a Smart Phone** . . . . . 359  
 Nam-Jin Bae, Seong Ryoung Park, Tae Hyung Kim, Myeong Bae Lee, Hong Gean Kim, Mi Ran Baek, Jang Woo Park, Chang-Sun Shin and Yong-Yun Cho

**A Real-time Object Detection System Using Selected Principal Components** . . . . . 367  
 Jong-Ho Kim, Byoung-Doo Kang, Sang-Ho Ahn, Heung-Shik Kim and Sang-Kyoon Kim

**Trajectory Calculation Based on Position and Speed for Effective Air Traffic Flow Management** . . . . . 377  
 Yong-Kyun Kim, Deok Gyu Lee and Jong Wook Han

**Part VII Multimedia Entertainment**

**Design and Implementation of a Geometric Origami Edutainment Application.** . . . . . 387  
 ByeongSu Kim, TaeHun Kim and JongHoon Kim



**Gamification Literacy: Emerging Needs for Identifying Bad Gamification** . . . . . 395  
Toshihiko Yamakami

**Automatic Fixing of Foot Skating of Human Motions from Depth Sensor** . . . . . 405  
Mankyu Sung

**Part VIII IT and Multimedia Applications**

**A Study on the Development and Application of Programming Language Education for Creativity Enhancement: Based on LOGO and Scratch** . . . . . 415  
YoungHoon Yang, DongLim Hyun, EunGil Kim, JongJin Kim and JongHoon Kim

**Design and Implementation of Learning Content Authoring Framework for Android-Based Three-Dimensional Shape**. . . . . 423  
EunGil Kim, DongLim Hyun and JongHoon Kim

**A Study on GUI Development of Memo Function for the E-Book: A Comparative Study Using iBooks**. . . . . 431  
Jeong Ah Kim and Jun Kyo Kim

**Relaxed Stability Technology Approach in Organization Management: Implications from Configured-Control Vehicle Technology** . . . . . 439  
Toshihiko Yamakami

**Mapping and Optimizing 2-D Scientific Applications on a Stream Processor**. . . . . 449  
Ying Zhang, Gen Li, Hongwei Zhou, Pingjing Lu, Caixia Sun and Qiang Dou

**Development of an Android Field Trip Support Application Using Augmented Reality and Google Maps**. . . . . 459  
DongLim Hyun, EunGil Kim and JongHoon Kim

**Implementation of Automotive Media Streaming Service Adapted to Vehicular Environment** . . . . . 467  
Sang Yub Lee, Sang Hyun Park and Hyo Sub Choi

**The Evaluation of the Transmission Power Consumption Laxity-Based (TPCLB) Algorithm . . . . .** 477  
 Tomoya Enokido, Ailixier Aikebaier and Makoto Takizawa

**The Methodology for Hardening SCADA Security Using Countermeasure Ordering . . . . .** 485  
 Sung-Hwan Kim, Min-Woo Park, Jung-Ho Eom and Tai-Myoung Chung

**Development and Application of STEAM Based Education Program Using Scratch: Focus on 6th Graders’ Science in Elementary School . . . . .** 493  
 JungCheol Oh, JiHwon Lee and JongHoon Kim

**Part IX Advanced Technologies and Applications for Cloud Computing and Sensor Networks**

**Performance Evaluation of Zigbee Sensor Network for Smart Grid AMI . . . . .** 505  
 Yong-Hee Jeon

**P2P-Based Home Monitoring System Architecture Using a Vacuum Robot with an IP Camera . . . . .** 511  
 KwangHee Choi, Ki-Sik Kong and Joon-Min Gil

**Design and Simulation of Access Router Discovery Process in Mobile Environments . . . . .** 521  
 DaeWon Lee, James J. Park and Joon-Min Gil

**Integrated SDN and Non-SDN Network Management Approaches for Future Internet Environment . . . . .** 529  
 Dongkyun Kim, Joon-Min Gil, Gicheol Wang and Seung-Hae Kim

**Analysis and Design of a Half Hypercube Interconnection Network . . . . .** 537  
 Jong-Seok Kim, Mi-Hye Kim and Hyeong-Ok Lee

**Aperiodic Event Communication Process for Wearable P2P Computing . . . . .** 545  
 Tae-Gyu Lee and Gi-Soo Chung

**Broadcasting and Embedding Algorithms for a Half Hypercube Interconnection Network . . . . .** 553  
 Mi-Hye Kim, Jong-Seok Kim and Hyeong-Ok Lee

**Obstacle Searching Method Using a Simultaneous Ultrasound Emission for Autonomous Wheelchairs** . . . . . 561  
 Byung-Seop Song and Chang-Geol Kim

**Part X Future Technology and its Application**

**A Study on Smart Traffic Analysis and Smart Device Speed Measurement Platform** . . . . . 569  
 Haejong Joo, Bonghwa Hong and Sangsoo Kim

**Analysis and Study on RFID Tag Failure Phenomenon.** . . . . . 575  
 Seongsoo Cho, Son Kwang Chul, Jong-Hyun Park and Bonghwa Hong

**Administration Management System Design for Smart Phone Applications in use of QR Code** . . . . . 585  
 So-Min Won, Mi-Hye Kim and Jin-Mook Kim

**Use of Genetic Algorithm for Robot-Posture** . . . . . 593  
 Dong W. Kim, Sung-Wook Park and Jong-Wook Park

**Use of Flexible Network Framework for Various Service Components of Network Based Robot** . . . . . 597  
 Dong W. Kim, Ho-Dong Lee, Sung-Wook Park and Jong-Wook Park

**China’s Shift in Culture Policy and Cultural Awareness.** . . . . . 601  
 KyooSeob Lim

**China’s Cultural Industry Policy** . . . . . 611  
 WonBong Lee and KyooSeob Lim

**Development of Mobile Games for Rehabilitation Training for the Hearing Impaired** . . . . . 621  
 Seongsoo Cho, Son Kwang Chul, Chung Hyeok Kim and Yunho Lee

**A Study to Prediction Modeling of the Number of Traffic Accidents** . . . . . 627  
 Young-Suk Chung, Jin-Mook Kim, Dong-Hyun Kim and Koo-Rock Park

**Part XI Pervasive Services, Systems and Intelligence**

**A Wiki-Based Assessment System Towards Social-Empowered Collaborative Learning Environment . . . . .** 633  
 Bruce C. Kao and Yung Hui Chen

**Universal User Pattern Discovery for Social Games: An Instance on Facebook. . . . .** 641  
 Martin M. Weng and Bruce C. Kao

**Ubiquitous Geography Learning Smartphone System for 1st Year Junior High Students in Taiwan . . . . .** 649  
 Wen-Chih Chang, Hsuan-Che Yang, Ming-Ren Jheng and Shih-Wei Wu

**Housing Learning Game Using Web-Based Map Service . . . . .** 657  
 Te-Hua Wang

**Digital Publication Converter: From SCORM to EPUB . . . . .** 665  
 Hsuan-pu Chang

**An Intelligent Recommender System for Real-Time Information Navigation . . . . .** 673  
 Victoria Hsu

**Part XII Advanced Mechanical and Industrial Engineering, and Control I**

**Modal Characteristics Analysis on Rotating Flexible Beam Considering the Effect from Rotation . . . . .** 683  
 Haibin Yin, Wei Xu, Jinli Xu and Fengyun Huang

**The Simulation Study on Harvested Power in Synchronized Switch Harvesting on Inductor . . . . .** 691  
 Jang Woo Park, Honggeun Kim, Chang-Sun Shin, Kyungryong Cho, Yong-Yun Cho and Kisuk Kim

**An Approach for a Self-Growing Agricultural Knowledge Cloud in Smart Agriculture . . . . .** 699  
 TaeHyung Kim, Nam-Jin Bae, Chang-Sun Shin, Jang Woo Park, DongGook Park and Yong-Yun Cho

**Determination of Water-Miscible Fluids Properties . . . . .** 707  
 Zajac Jozef, Cuma Matus and Hatala Michal

**Influence of Technological Factors of Die Casting on Mechanical Properties of Castings from Silumin. . . . .** 713  
 Stefan Gaspar and Jan Pasko

**Active Ranging Sensors Based on Structured Light Image for Mobile Robot. . . . .** 723  
 Jin Shin and Soo-Yeong Yi

**Improved Composite Order Bilinear Pairing on Graphics Hardware . . . . .** 731  
 Hao Xiong, Xiaoqi Yu, Yi-Jun He and Siu Ming Yiu

**Deployment and Management of Multimedia Contents Distribution Networks Using an Autonomous Agent Service . . . . .** 739  
 Kilhung Lee

**Part XIII Advanced Mechanical and Industrial Engineering, and Control II**

**Design Optimization of the Assembly Process Structure Based on Complexity Criterion . . . . .** 747  
 Vladimir Modrak, Slavomir Bednar and David Marton

**Kinematics Modelling for Omnidirectional Rolling Robot. . . . .** 755  
 Soo-Yeong Yi

**Design of Device Sociality Database for Zero-Configured Device Interaction . . . . .** 763  
 Jinyoung Moon, Dong-oh Kang and Changseok Bae

**Image Processing Based a Wireless Charging System with Two Mobile Robots . . . . .** 769  
 Jae-O Kim, Chan-Woo Moon and Hyun-Sik Ahn

**Design of a Reliable In-Vehicle Network Using ZigBee Communication . . . . .** 777  
 Sunny Ro, Kyung-Jung Lee and Hyun-Sik Ahn

**Wireless Positioning Techniques and Location-Based Services: A Literature Review . . . . .** 785  
 Pantea Keikhosrokiani, Norlia Mustafa, Nasriah Zakaria and Muhammad Imran Sarwar

**Part XIV Green and Human Information Technology**

**Performance Analysis of Digital Retrodirective Array Antenna System in Presence of Frequency Offset . . . . .** 801  
 Junyeong Bok and Heung-Gyoon Ryu

**A Novel Low Profile Multi-Band Antenna for LTE Handset . . . . .** 809  
 Bao Ngoc Nguyen, Dinh Uyen Nguyen, Tran Van Su, Binh Duong Nguyen and Mai Linh

**Digital Signature Schemes from Two Hard Problems . . . . .** 817  
 Binh V. Do, Minh H. Nguyen and Nikolay A. Moldovyan

**Performance Improvements Using Upgrading Precedences in MIL-STD-188-220 Standard . . . . .** 827  
 Sewon Han and Byung-Seo Kim

**Blind Beamforming Using the MCMA and SAG-MCMA Algorithm with MUSIC Algorithm. . . . .** 835  
 Yongguk Kim and Heung-Gyoon Ryu

**Performance Evaluation of EPON-Based Communication Network Architectures for Large-Scale Offshore Wind Power Farms . . . . .** 841  
 Mohamed A. Ahmed, Won-Hyuk Yang and Young-Chon Kim

**A User-Data Division Multiple Access Scheme . . . . .** 849  
 P. Niroopan, K. Bandara and Yeon-ho Chung

**On Channel Capacity of Two-Way Multiple-hop MIMO Relay System with Specific Access Control . . . . .** 857  
 Pham Thanh Hiep, Nguyen Huy Hoang and Ryuji Kohno

**Single-Feed Wideband Circularly Polarized Antenna for UHF RFID Reader . . . . .** 863  
 Pham HuuTo, B. D. Nguyen, Van-Su Tran, Tram Van and Kien T. Pham

**Experimental Evaluation of WBAN Antenna Performance for FCC Common Frequency Band with Human Body . . . . .** 871  
 Musleemin Noitubtim, Chairak Deepunya and Sathaporn Promwong

<b>Performance Evaluation of UWB-BAN with Friis’s Formula and CLEAN Algorithm . . . . .</b>	879
Krisada Koonchiang, Dissakan Arpasilp and Sathaporn Promwong	
<b>A Study of Algorithm Comparison Simulator for Energy Consumption Prediction in Indoor Space . . . . .</b>	887
Do-Hyeun Kim and Nan Chen	
<b>Energy Efficient Wireless Sensor Network Design and Simulation for Water Environment Monitoring . . . . .</b>	895
Nguyen Thi Hong Doanh and Nguyen Tuan Duc	
<b>An Energy Efficient Reliability Scheme for Event Driven Service in Wireless Sensor Actuator Networks . . . . .</b>	903
Seungcheon Kim	
<b>Efficient and Reliable GPS-Based Wireless Ad Hoc for Marine Search Rescue System . . . . .</b>	911
Ta Duc-Tuyen, Tran Duc-Tan and Do Duc Dung	
<b>Improved Relay Selection for MIMO-SDM Cooperative Communications . . . . .</b>	919
Duc Hiep Vu, Quoc Trinh Do, Xuan Nam Tran and Vo Nguyen Quoc Bao	
<b>Freshness Preserving Hierarchical Key Agreement Protocol Over Hierarchical MANETs . . . . .</b>	927
Hyunsung Kim	
<b>A Deployment of RFID for Manufacturing and Logistic . . . . .</b>	935
Patcharaporn Choeksuwan and Somsak Choomchuay	
<b>Real Time Video Implementation on FPGA . . . . .</b>	943
Pham Minh Luan Nguyen and Sang Bock Cho	
<b>Recovery Algorithm for Compressive Image Sensing with Adaptive Hard Thresholding . . . . .</b>	949
Viet Anh Nguyen and Byeungwoo Jeon	
<b>Estimation Value for Three Dimension Reconstruction . . . . .</b>	957
Tae-Eun Kim	
<b>Gesture Recognition Algorithm using Morphological Analysis. . . . .</b>	967
Tae-Eun Kim	

**Omnidirectional Object Recognition Based Mobile Robot Localization** . . . . . 975  
 Sungho Kim and In So Kweon

**Gender Classification Using Faces and Gaits** . . . . . 983  
 Hong Quan Dang, Intaek Kim and YoungSung Soh

**Implementation of Improved Census Transform Stereo Matching on a Multicore Processor** . . . . . 989  
 Jae Chang Kwak, Tae Ryong Park, Yong Seo Koo and Kwang Yeob Lee

**A Filter Selection Method in Hard Thresholding Recovery for Compressed Image Sensing** . . . . . 997  
 Phuung Minh Pham, Khanh Quoc Dinh and Byeungwoo Jeon

**Facial Expression Recognition Using Extended Local Binary Patterns of 3D Curvature** . . . . . 1005  
 Soon-Yong Chun, Chan-Su Lee and Sang-Heon Lee

**Overview of Three and Four-Dimensional GIS Data Models**. . . . . 1013  
 Tuan Anh Nguyen Gia, Phuoc Vinh Tran and Duy Huynh Khac

**Modeling and Simulation of an Intelligent Traffic Light System Using Multiagent Technology** . . . . . 1021  
 Tuyen T. T. Truong and Cuong H. Phan

**A Numerical Approach to Solve Point Kinetic Equations Using Taylor-Lie Series and the Adomian Decomposition Method** . . . . . 1031  
 Hag-Tae Kim, Ganduulga, Dong Pyo Hong and Kil To Chong

**Regional CRL Distribution Based on the LBS for Vehicular Networks** . . . . . 1039  
 HyunGon Kim, MinSoo Kim, SeokWon Jung and JaeHyun Seo

**Study of Reinforcement Learning Based Dynamic Traffic Control Mechanism** . . . . . 1047  
 Zheng Zhang, Seung Jun Baek, Duck Jin Lee and Kil To Chong

**Understanding and Extending AUTOSAR BSW for Custom Functionality Implementation** . . . . . 1057  
 Taeho Kim, Ji Chan Maeng, Hyunmin Yoon and Minsoo Ryu



**A Hybrid Intelligent Control Method in Application of Battery Management System . . . . .** 1065  
T. T. Ngoc Nguyen and Franklin Bien

**Interpretation and Modeling of Change Patterns of Concentration Based on EEG Signals. . . . .** 1073  
JungEun Lim, Soon-Yong Chun and BoHyeok Seo

**Design of Autonomic Nerve Measuring System Using Pulse Signal. . . . .** 1081  
Un-Ho Ji and Soon-Yong Chun

**Semiconductor Monitoring System for Etching Process . . . . .** 1091  
Sang-Chul Kim

**Enhancing the Robustness of Fault Isolation Estimator for Fault Diagnosis in Robotic Systems. . . . .** 1099  
Ngoc-Bach Hoang and Hee-Jun Kang

**Software-Based Fault Detection and Recovery for Cyber-Physical Systems . . . . .** 1107  
Jooyi Lee, Ji Chan Maeng, Byeonghun Song, Hyunmin Yoon, Taeho Kim, Won-Tae Kim and Minsoo Ryu

**Sample Adaptive Offset Parallelism in HEVC . . . . .** 1113  
Eun-kyung Ryu, Jung-hak Nam, Seon-oh Lee, Hyun-ho Jo and Dong-gyu Sim

**Comparison Between SVM and Back Propagation Neural Network in Building IDS . . . . .** 1121  
Nguyen Dai Hai and Nguyen Linh Giang

**Anomaly Detection with Multinomial Logistic Regression and Naïve Bayesian . . . . .** 1129  
Nguyen Dai Hai and Nguyen Linh Giang

**Implementation of Miniaturized Automotive Media Platform with Vehicle Data Processing. . . . .** 1137  
Sang Yub Lee, Sang Hyun Park, Duck Keun Park, Jae Kyu Lee and Hyo Sub Choi

**Design of Software-Based Receiver and Analyzer System for DVB-T2 Broadcast System. . . . .** 1147  
M. G. Kang, Y. J. Woo, K. T. Lee, I. K. Kim, J. S. Lee and J. S. Lee

**Age-Group Classification for Family Members Using Multi-Layered Bayesian Classifier with Gaussian Mixture Model . . .** 1153  
 Chuho Yi, Seungdo Jeong, Kyeong-Soo Han and Hankyu Lee

**Enhancing Utilization of Integer Functional Units for High-Throughput Floating Point Operations on Coarse-Grained Reconfigurable Architecture. . . . .** 1161  
 Manhwee Jo, Kyuseung Han and Kiyoung Choi

**An Improved Double Delta Correlator for BOC Signal Tracking in GNSS Receivers . . . . .** 1169  
 Pham-Viet Hung, Dao-Ngoc Chien and Nguyen-Van Khang

**Implementation of Automatic Failure Diagnosis for Wind Turbine Monitoring System Based on Neural Network . . .** 1181  
 Ming-Shou An, Sang-June Park, Jin-Sup Shin, Hye-Youn Lim and Dae-Seong Kang

**Development of Compact Microphone Array for Direction-of-Arrival Estimation . . . . .** 1189  
 Trình Quốc Võ and Udo Klein

**Design and Implementation of a SoPC System for Speech Recognition . . . . .** 1197  
 Tran Van Hoang, Nguyen Ly Thien Truong, Hoang Trang and Xuan-Tu Tran

**Erratum to: Design of a Reliable In-Vehicle Network Using ZigBee Communication . . . . .** E1  
 Sunny Ro, Kyung-Jung Lee and Hyun-Sik Ahn

**Index . . . . .** 1205

**Part I**  
**Multimedia Modeling and Processing**

# Multiwedgelets in Image Denoising

Agnieszka Lisowska

**Abstract** In this paper the definition of a multiwedgelet is introduced. The multiwedgelet is defined as a vector of wedgelets. In order to use a multiwedgelet in image approximation its visualization and computation methods are also proposed. The application of multiwedgelets in image denoising is presented, as well. As follows from the experiments performed multiwedgelets assure better denoising results than the other known state-of-the-art methods.

**Keywords** Multiwedgelets · Wedgelets · Multiresolution · Denoising

## 1 Introduction

Geometrical multiresolution methods of image approximation are widely used in these days. It follows from the multiscale nature of the world, especially of digital images. Such methods can better adapt to image singularities than the well known wavelets theory [1]. Many new, geometrical, representations have been proposed recently. They can be divided into two groups. The one is based on nonadaptive methods of computing, with the use of frames, like brushlets [2], ridgelets [3], curvelets [4], contourlets [5], shearlets [6]. In the second group the approximations are computed in an adaptive way. The majority of the representations are based on dictionaries, examples include wedgelets [7], beamlets [8], second order wedgelets [9, 10], platelets [11], surflets [12], smoothlets [13]. However, recently also the adaptive schemes based on basis have been proposed, like bandelets [14], grouplets [15], tetrolets [16]. More and more “X-lets” have been still defined.

---

A. Lisowska (✉)  
Institute of Computer Science, University of Silesia,  
ul. Bedzinska 39 41-200 Sosnowiec, Poland  
e-mail: alisow@ux2.math.us.edu.pl  
URL: [http://www.math.us.edu.pl/al/eng\\_index.html](http://www.math.us.edu.pl/al/eng_index.html)

Many theories, which are based on functions, are further extended on vectors of functions. In general, from the mathematical point of view, it is rather a simple task. However, the practical application of such theories is not easy, especially in the area of image processing. It follows mainly from the fact that an image is a two dimensional object and can be represented by a set of functions in a natural way. A set of vectors seems to be used rather in the representation of a set of images. So, the main question is not how to extend a theory on vectors but how to apply such an extended theory to image processing?

In this paper the answer to the above question is presented. In more details, firstly, the definition of multiwedgelet is proposed as a vector of wedgelets. Because the visualization of a multiwedgelet is not straightforward, the method of visualization is also proposed. It is used further in the image approximation. In order to justify the usefulness of the proposed approach the application of multiwedgelets to image denoising is presented. As follows from the experiments, the proposed method outperforms the known methods like wedgelets, second order wedgelets, curvelets and wavelets [17, 18].

## 2 Multiwedgelets

Let us define an image domain  $D = [0, 1] \times [0, 1]$ . Next, let us denote function  $h(x)$  defined within  $D$  as the “horizon”, that is any smooth function defined on the interval  $[0, 1]$ . In practical applications it is sufficiently to assume that the function  $h$  is of  $C^2$  class.

Further, consider the characteristic function

$$H(x, y) = \mathbf{1}\{y \leq h(x)\}, \quad 0 \leq x, y \leq 1. \quad (1)$$

Then function  $H$  is called a “horizon function” if  $h$  is a “horizon”. Function  $H$  models a black and white image with a horizon where the image is white above the horizon and black below.

Having an image domain  $D = [0, 1] \times [0, 1]$  one can, in some sense, discretize it on different levels of multiresolution. Consider the dyadic square  $D(j_1, j_2, i)$  as the two dimensional interval

$$D(j_1, j_2, i) = [j_1/2^i, (j_1 + 1)/2^i] \times [j_2/2^i, (j_2 + 1)/2^i], \quad (2)$$

where  $j_1, j_2 \in \{0, \dots, 2^i - 1\}$ ,  $i \in \mathbb{N}$ . Note that  $D(0, 0, 0)$  denotes the whole image domain  $D$ , that is the square  $[0, 1] \times [0, 1]$ . On the other hand  $D(j_1, j_2, I)$  for  $j_1, j_2 \in \{0, \dots, N\}$  denote appropriate pixels from  $N \times N$  grid, where  $N$  is dyadic (it means that  $N = 2^I$ ). From this moment on let us consider a domain of an image as such  $N \times N$  grid of pixels.

## 2.1 Basic Definitions

Having assumed that an image domain is the square  $[0, 1] \times [0, 1]$  and that it consists of  $N \times N$  pixels (or, more precisely, squares of size  $1/N$ ) one can note that on each border of any square  $D(j_1, j_2, i), j_1, j_2 \in \{0, \dots, 2^i - 1\}, i \in \{0, \dots, \log_2 N\}$  the vertices with distance equal to  $1/N$  can be denoted. Every two such vertices in any fixed square may be connected to form a straight line  $b$ —an edge (also called a *beamlet* after the work [8]).

Let us denote then  $B_{D(j_1, j_2, i)}$  as the set of all nondegenerated beamlets (that is no lying on the same side of a square border) within  $D(j_1, j_2, i)$  for any  $j_1, j_2 \in \{0, \dots, 2^i - 1\}, i \in \mathbb{N}$ . Consider then a vector of beamlets  $\mathbf{b}_{j_1, j_2, i}^M = [b_{j_1, j_2, i}^1, \dots, b_{j_1, j_2, i}^M]$ ,  $M \in \mathbb{N}$ . We call vector  $\mathbf{b}_{j_1, j_2, i}^M$  a *multibeamlet* if for all  $k \in \{1, \dots, M\} b_{j_1, j_2, i}^k \in B_{D(j_1, j_2, i)}$  for fixed  $j_1, j_2 \in \{0, \dots, 2^i - 1\}, i \in \mathbb{N}$ . In Fig. 1 some examples of multibeamlets are presented.

Let us consider a beamlet  $b$ . It splits any square  $D$  (we skip the subscripts denoting the location and the scale for a moment for better clarity) into two pieces. Let us consider one of the two pieces which is bounded by lines connecting in turn in clockwise direction, from the lower left corner, the first of the two edge vertices and the second one. Let us define then the indicator function of that piece

$$W(x, y) = \mathbf{1}\{y \leq b(x)\}, \quad (x, y) \in D. \quad (3)$$

Such a function we call a *wedgelet* defined by beamlet  $b$  [7].

Let us denote then  $W_{D(j_1, j_2, i)}$  as the set of all nondegenerated wedgelets within  $D(j_1, j_2, i)$  for any  $j_1, j_2 \in \{0, \dots, 2^i - 1\}, i \in \mathbb{N}$ . Consider then a vector of wedgelets  $\mathbf{W}_{j_1, j_2, i}^M = [W_{j_1, j_2, i}^1, \dots, W_{j_1, j_2, i}^M]$ ,  $M \in \mathbb{N}$ . We call vector  $\mathbf{W}_{j_1, j_2, i}^M$  a *multiwedgelet* if for all  $k \in \{1, \dots, M\} W_{j_1, j_2, i}^k \in W_{D(j_1, j_2, i)}$  for fixed  $j_1, j_2 \in \{0, \dots, 2^i - 1\}, i \in \mathbb{N}$ .

Consider the complete quadtree image partition. Each segment can be represented by two numbers  $i$  and  $j$  where  $j \in \{0, \dots, 4^i - 1\}$  and  $i \in \{0, \dots, \log_2 N\}$ . In other words, the pair of subscripts  $(j_1, j_2)$  from the above considerations can be replaced by the subscript  $j$ . Having defined a multiwedgelet and renumbering subscripts for better clarity one can define the dictionary of multiwedgelets as the following set

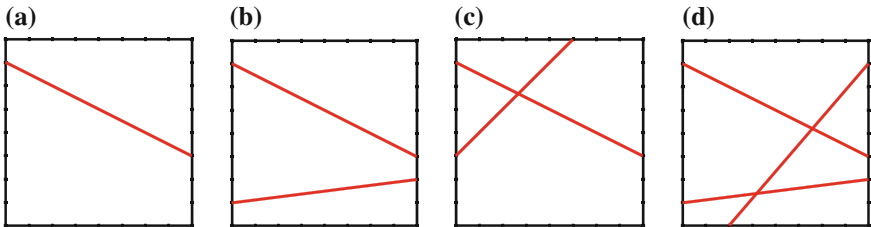


Fig. 1 Examples of multibeamlets: **a**  $M = 1$ , **b–c**  $M = 2$ , **d**  $M = 3$

$$W_M = \{\mathbf{W}_{ij}^M : i \in \{0, \dots, \log_2 N\}, j \in \{0, \dots, 4^i - 1\}\}, \quad (4)$$

where  $\mathbf{W}_{ij}^M = [W_{ij}^1, \dots, W_{ij}^M]$ .

Image approximation is performed in two steps. In the first step, for each quadtree partition segment, the best multiwedgelet has to be found (in the mean of the smallest Mean Square Error approximation), forming the complete quadtree with the optimal multiwedgelet parameters in each node. In the second step the bottom-up tree pruning algorithm has to be applied [7] in order to obtain the optimal image representation.

There is one drawback related to the image approximation by multiwedgelets. It is related to finding a reasonable method of a multiwedgelet visualization. Since the visualization of one wedgelet is quite simple, the vector of wedgelets has to be visualized in a tricky way. Below such a method is presented.

## 2.2 Multiwedgelet Visualization

Let us consider the example presented in Fig. 2 for  $M = 3$ . The multiwedgelet  $W$  is defined as  $\mathbf{W} = [W^1, W^2, W^3]$  where wedgelets  $W^i$  are based on appropriate beamlets  $b^i$  for  $i \in \{1, 2, 3\}$ . If the wedgelets are defined as

$$W^i = \begin{cases} h_1^i, y \leq b^i, \\ h_2^i, y > b^i, \end{cases} \quad \text{for } i \in \{1, \dots, M\},$$

the appropriate colors are defined as

$$c^a = \frac{1}{M} \sum_{k=1}^M h_u^k \quad \text{for } a \in \{1, \dots, \text{Number of Areas}\}, u \in \{1, 2\}. \quad (5)$$

In other words, the colors are the means of all wedgelets colors.

Let us note that such correlation between multiwedgelets coefficients and image colors causes that the image is defined as a mean of all wedgelets of the multiwedgelet. Indeed, for image segment  $F$  one obtains  $F = \frac{1}{M} \sum_{k=1}^M W^k$ .

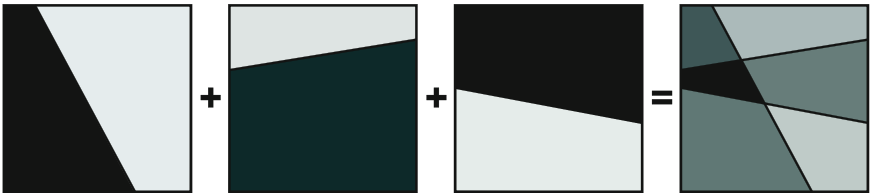


Fig. 2 The method of a multiwedgelet visualization ( $M = 3$ )

### 2.3 Multiwedgelet Computation

It can be defined plenty of methods of a multiwedgelet computation. In the paper the one is proposed. It is based on the fast wedgelet computation proposed in [19]. In order to compute multiwedgelet parameters one needs to proceed in the following way. Firstly, compute the wedgelet transform for the first wedgelet of a given multiwedgelet, then compute the wedgelet transform for the second wedgelet of the multiwedgelet for a slightly translated support (i.e. one pixel up and left), then do the same for all the rest wedgelets of multiwedgelet, each time changing slightly the support (by translating it in different directions). The support manipulation causes that the optimal wedgelets of multiwedgelet are different.

Let us note that the computational complexity of the proposed method is  $O(N^2 \log_2 N)$  for an image of size  $N \times N$  pixels. It follows from the fact that the method is based on the fast wedgelet transform [19] and it is performed  $M$  times. Since, usually,  $M = 3$  in practice it can be treated as a constant. The measured computation time of the multiwedgelet transform is as follows: for  $M = 1$  it equals 1.8 s, for  $M = 2$  it equals 3.2 s and for  $M = 3$  it equals 4.7 s for an image of size  $256 \times 256$  pixels. The computations were performed on Intel Core2 Duo 2 GHz processor.

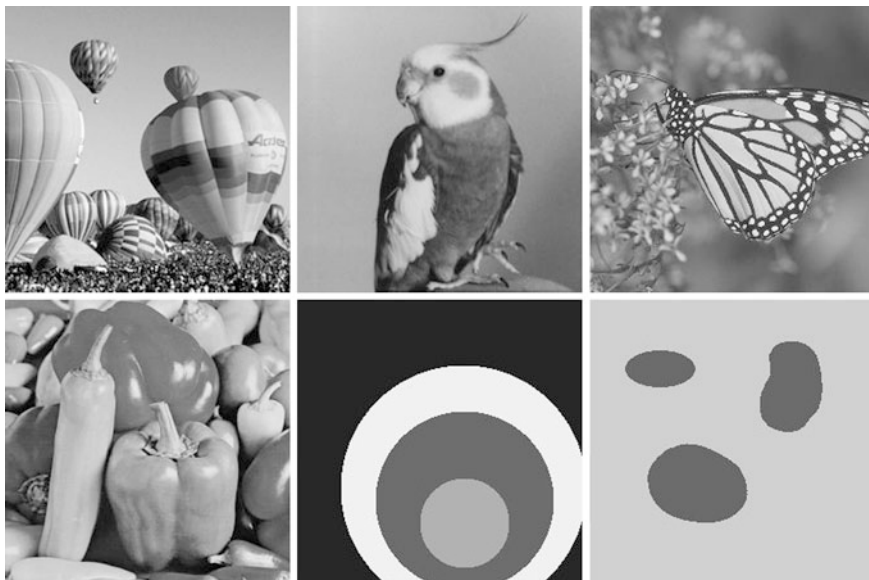
## 3 Experimental Results

In order to perform numerical computations the standard set of test images, presented in Fig. 3, was used. The images were additionally contaminated by Gaussian noise with zero mean and different values of variance with the help of Matlab Image Processing Toolbox. All computations were made with the help of the software written in C++ Builder 6 Environment.

In Table 1 the numerical results of image denoising for different methods and different values of noise variance are presented. The same set of images was used in the paper [18] from which it follows that the curvelets-based method of image denoising assures better results than denoising by wavelets. From Table 1 it follows that the method of image denoising by multiwedgelets (for  $M = 3$ ) outperforms the curvelets and wedgelets-based methods, it even outperforms the second order wedgelets-based method (wedgeletsII) [18]. Only in the case of images with strongly curvature geometry, like “Circles” and “Blobs”, second order wedgelets-based method assures better results of denoising than multiwedgelets. It is very natural since second order wedgelets were designed to best approximate images with curvature geometry.

In Fig. 4 two plots, arbitrarily chosen, of image denoising by multiwedgelets are presented. As one can see the use of multiwedgelets outperforms the use of wedgelets ( $M = 1$ ). The larger the value of  $M$  the better the result of image denoising. However, from the performed experiments follows that the choice of





**Fig. 3** The benchmark images, respectively: “Balloons”, “Bird”, “Monarch”, “Peppers”, “Circles”, “Blobs” [18]

$M = 3$  is the most flexible one. It gives satisfactory results and is not computationally expensive.

Sample results of image denoising are presented in Fig. 5 for images “Bird” and “Monarch” contaminated by zero-mean Gaussian noise with variance  $V = 0.022$ . The images were denoised by curvelets, second order wedgelets and multiwedgelets, respectively. As one can see the method based on multiwedgelets assured the best denoising results, both visually and in the mean of PSNR values (Table 1).

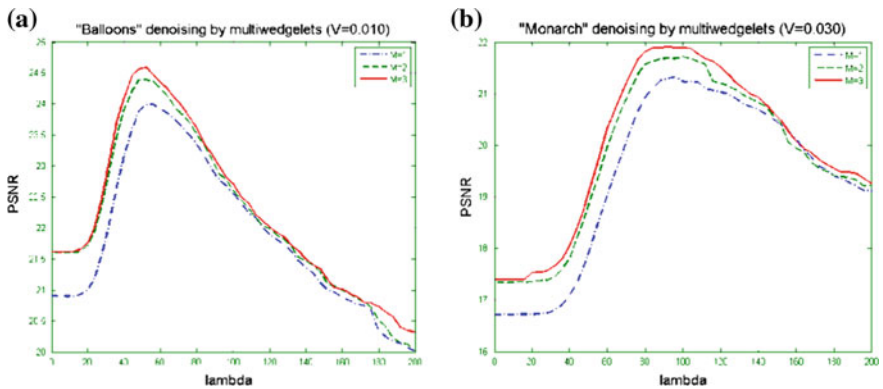
## 4 Conclusions

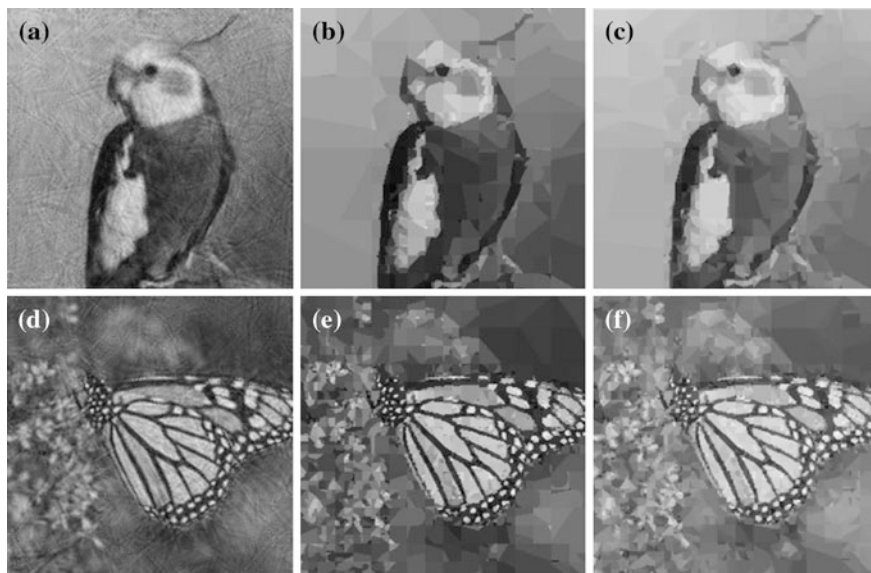
In this paper the new theory of multiwedgelets was presented. Instead of considering wedgelets, the vectors of these functions were used. In order to use them in image approximation the computation and visualization methods were also proposed.

In this paper also the application of multiwedgelets to image denoising was presented. In comparison to the other known state-of-the-art methods (like curvelets, wedgelets, second order wedgelets—directly, and wavelets—indirectly via comparison to the paper [18]) the method based on multiwedgelets assures the best denoising results. Probably, it is possible to apply multiwedgelets in other image

**Table 1** Numerical results of image denoising for methods based on curvelets, wedgelets, second order wedgelets and multiwedgelets for different values of noise variance (PSNR)

Image	Method	Noise variance							
		0.001	0.005	0.010	0.015	0.022	0.030	0.050	0.070
Balloons	Curvelets	24.97	24.22	23.87	23.28	21.54	20.04	17.46	15.89
	Wedgelets	30.50	26.10	24.03	23.17	22.29	21.72	20.60	19.94
	Wedg.II	30.40	25.92	24.00	23.12	22.26	21.71	20.67	19.97
	Multiwed.	29.23	26.14	24.59	23.72	22.91	22.33	21.24	20.45
Bird	Curvelets	20.34	20.95	26.94	25.87	23.40	21.02	18.09	16.19
	Wedgelets	34.24	30.24	28.76	28.05	27.35	26.82	25.71	25.21
	Wedg.II	34.07	30.24	28.76	28.02	27.29	26.79	25.66	25.09
	Multiwed.	34.55	31.05	29.70	28.95	27.99	27.50	26.47	25.72
Monarch	Curvelets	24.14	24.99	24.15	23.37	21.77	20.38	17.53	15.95
	Wedgelets	30.47	26.20	24.32	23.27	22.33	21.63	20.50	19.70
	Wedg.II	30.38	26.21	24.39	23.40	22.37	21.71	20.56	19.71
	Multiwed.	28.32	25.82	24.54	23.60	22.83	21.91	21.01	20.41
Peppers	Curvelets	22.52	25.04	23.91	24.41	22.89	20.85	17.65	15.95
	Wedgelets	31.71	27.44	25.82	24.89	24.10	23.41	22.43	21.75
	Wedg.II	31.56	27.31	25.81	24.79	24.04	23.37	22.36	21.68
	Multiwed.	31.63	27.81	26.58	25.64	24.83	24.21	22.99	22.32
Circles	Curvelets	25.56	22.35	23.90	21.49	20.41	18.71	16.61	15.32
	Wedgelets	41.97	35.32	31.84	29.95	28.21	26.84	24.56	23.18
	Wedg.II	43.19	36.60	32.60	30.51	28.59	26.93	24.58	23.17
	Multiwed.	33.16	32.22	30.67	29.58	28.31	26.96	24.66	23.06
Blobs	Curvelets	14.83	28.43	30.71	25.65	23.64	21.55	18.33	16.41
	Wedgelets	44.23	36.31	33.74	32.85	31.77	30.97	29.18	27.51
	Wedg.II	45.12	37.49	34.52	33.51	31.95	31.43	29.20	27.55
	Multiwed.	38.53	35.53	34.22	33.14	32.33	31.01	29.84	27.67

**Fig. 4** Denoising by multiwedgelets for different values of  $M$  for images contaminated by zero-mean Gaussian noise: **a** “Balloons”,  $V = 0.010$ , **b** “Monarch”,  $V = 0.030$ . Let us note that  $M = 1$  denotes wedgelets



**Fig. 5** Examples of image denoising: (*upper row*) “Bird”, (*lower row*) “Monarch”; by **a, d** curvelets, **b, e** second order wedgelets, **c, f** multiwedgelets. The images were contaminated by zero-mean Gaussian noise with variance  $V = 0.022$

processing tasks, like image compression or edge detection. It is the open problem for future research.

As follows from the performed experiments, the potential of multiwedgelets can be quite large. The methods proposed in this paper and the parameters fixed are both not optimal but multiwedgelets still outperform the state-of-the-art methods in such an image processing task like image denoising. There is still much to do in improving the multiwedgelets theory and in finding new applications.

## References

1. Mallat S (1999) A wavelet tour of signal processing. Academic Press, San Diego
2. Meyer FG, Coifman RR (1997) Brushlets: a tool for directional image analysis and image compression. *Appl Comput Harm Anal* 4:147–187
3. Candés E (1998) Ridgelets: theory and applications, Ph.D. thesis, Department of Statistics, Stanford University, Stanford
4. Candés E, Donoho D (1999) Curvelets—a surprisingly effective nonadaptive representation for objects with edges. In: Cohen A, Rabut C, Schumaker LL (ed) *Curves and surface fitting*, Vanderbilt University Press, Saint-Malo, pp 105–120
5. Do MN, Vetterli M (2003) Contourlets. In: Stoeckler J, Welland GV (ed) *Beyond wavelets*, Academic Press, San Diego, pp 83–105
6. Labate D, Lim W, Kutyniok G, Weiss G (2005) Sparse multidimensional representation using shearlets. *Proc SPIE* 5914:254–262

7. Donoho DL (1999) Wedgelets: nearly-minimax estimation of edges. *Ann Stat* 27:859–897
8. Donoho DL, Huo X (2000) Beamlet pyramids: a new form of multiresolution analysis, suited for extracting lines, curves and objects from very noisy image data. In: *Proceedings of SPIE*, vol 4119
9. Lisowska A (2000) Effective coding of images with the use of geometrical wavelets. In: *Proceedings of the decision support systems conference, Zakopane, Poland*
10. Lisowska A (2005) Geometrical wavelets and their generalizations in digital image coding and processing. Ph.D. thesis, University of Silesia, Poland
11. Willet RM, Nowak RD (2003) Platelets: a multiscale approach for recovering edges and surfaces in photon limited medical imaging. *IEEE Trans Med Imaging* 22:332–350
12. Chandrasekaran V, Wakin MB, Baron D, Baraniuk R (2004) Surflets: a sparse representation for multidimensional functions containing smooth discontinuities. In: *IEEE international symposium on information theory, Chicago*
13. Lisowska A (2011) Smoothlets—multiscale functions for adaptive representations of images. *IEEE Trans Image Process* 20(7):1777–1787
14. Pennec E, Mallat S (2005) Sparse geometric image representations with bandelets. *IEEE Trans Image Process* 14(4):423–438
15. Mallat S (2009) Geometrical grouplets. *Appl Comput Harm Anal* 26(2):161–180
16. Krommweh J (2009) Image approximation by adaptive tetrolet transform and International conference on sampling theory and applications,
17. Demare L, Friedrich F, Führ H, Szygowski T (2005) multiscale wedgelet denoising algorithms, *Proceedings of SPIE, San Diego, Wavelets XI, Vol. 5914, X1-12*
18. Lisowska A (2008) Image denoising with second order wedgelets. *Intern J Signal Imaging Syst Eng* 1(2):90–98
19. Lisowska A (2011) Moments-based fast wedgelet transform. *J Math Imaging Vis* 39(2):180–192. Springer

# A Novel Video Compression Method Based on Underdetermined Blind Source Separation

Jing Liu, Fei Qiao, Qi Wei and Huazhong Yang

**Abstract** If a piece of picture could contain a sequence of video frames, it is amazing. This paper develops a new video compression approach based on underdetermined blind source separation. Underdetermined blind source separation, which can be used to efficiently enhance the video compression ratio, is combined with various off-the-shelf codecs in this paper. Combining with MPEG-2, video compression ratio could be improved slightly more than 33 %. As for combing with H.264, twice compression ratio could be achieved with acceptable PSNR, according to different kinds of video sequences.

**Keywords** Underdetermined blind source separation • Sparse component analysis • Video surveillance system • Video compression

## 1 Introduction

Digital video is famous for its abundant information and its rigid demand for the bandwidth and process power as well. It leads to the emergence of multiple video coding standards, such as MPEG-2, H.264. However, new thoughts can be applied to compress video as well.

Blind Source Separation (BSS) provides a solution to recover original signals from mixed signals. It can be used in multiple fields, such as wireless communication to separate mixed radio signals, biomedicine to separate fetal electrocardiogram signals recorded by sensors, and typical “cocktail party” problem to separate mixed speech signals.

---

J. Liu · F. Qiao (✉) · Q. Wei · H. Yang  
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China  
e-mail: qiaofei@tsinghua.edu.cn

Independent Component Analysis (ICA) was widely accepted as a powerful solution of BSS since the past 20 years [1]. In 1999, A. Hyvarinen presented an improved ICA algorithm, called FastICA [2]. A detailed overview of many algorithms on BSS is made and their usages on image processing are presented as well [3]. However, few researchers focused on utilizing BSS into video processing.

In this paper, we apply Underdetermined BSS (UBSS, meaning the number of original signals is more than that of mixed signals) to compress video sequences. A new codec defined as Underdetermined Blind Source Separation based Video Compression (UBSSVC) is developed. As we explained later in detail, UBSSVC has good performance on video compression.

This paper is organized as follows. The next section briefly reviews BSS problem. In Sect. 3, detailed structure of UBSSVC is stated. And Sect. 4 shows simulation results. Finally, Sect. 5 summarizes the superior and deficiency of this video compression method. Also, future work is proposed in this section.

## 2 Blind Source Separation and Solution to UBSS

BSS was first established by J. Herault and C. Jutten in 1985 [1]. It can be described as following: multiple signals from separate sources  $s$  are somehow mixed into several other signals, defined as mixed signals  $x$ . Here  $n$  represents the number of source signals and  $m$  represents the number of mixed signals. The objective of BSS is to design an inverse system to get the estimation of source signals. The reason for the ‘‘Blind’’ here is neither the source signals nor the mixed process is known to the observer. The mixing model can be expressed as,

$$x = As; \quad A \in R^{m \times n}, s \in R^{n \times T} \quad (1)$$

where  $A$  is an  $m \times n$  mixing matrix. Both  $A$  and  $s$  are unknown, while  $x$  is known to observer.

Independent Component Analysis (ICA) is the main solution for overdetermined ( $n < m$ ) and standard ( $n = m$ ) case. However, it is not suitable for UBSS ( $n > m$ ). Other methods like Sparse Component Analysis (SCA) [4–7] and over-complete ICA [8, 9], are investigated for UBSS recent years.

In this work, SCA is adopted to solve the UBSS. SCA uses the sparsity of source signals to compensate information loss in the mixing process. So specific assumptions of mixing matrix  $A$  and source matrix  $s$  should be considered as follows [4]. Assumption 1: any  $m \times m$  square sub-matrix of mixing matrix  $A \in R^{m \times n}$  is nonsingular; assumption 2: there are at most  $m - 1$  nonzero elements of any column of matrix  $s$ . If the above assumptions are satisfied, the source matrix  $s$  can be recovered by SCA.

Let  $x_i$ ,  $i = 1, 2, \dots, m$  and  $s_i$ ,  $i = 1, 2, \dots, n$  represent mixed signals and source signals respectively; and  $a_j$ ,  $j = 1, \dots, n$  is the  $j$ th column of mixing matrix  $A$ . Therefore, the mixing process can also be described as following.

$$x(t) = (x_1(t) \quad x_2(t) \quad \cdots \quad x_m(t))^T = a_1 s_1(t) + a_2 s_2(t) + \cdots + a_n s_n(t) \quad (2)$$

Given the mixing matrix  $A$  satisfies the assumption 1, any  $m - 1$  columns of  $A$  span a  $m$ -dimensional linear hyperplane  $\mathcal{H}_q$ , which can be denoted as  $\mathcal{H}_q = \{h | h \in R^m, \lambda_{ik} \in R, h = \lambda_{i_1} a_{i_1} + \cdots + \lambda_{i_{m-1}} a_{i_{m-1}}\}$ , where  $q = 1, \dots, C_n^{m-1}$ . If source matrix  $s$  satisfies assumption 2, it is reasonable to suppose that at the  $t$  moment, all source signals except for  $s_{i_1}, s_{i_2}, \dots, s_{i_{m-1}}$  are zero, where  $\{i_1, i_2, \dots, i_{m-1}\} \subset \{1, 2, \dots, n\}$ . Consequently, at  $t$  moment, Eq. (2) can be rewritten as

$$(x_1(t) \quad x_2(t) \quad \cdots \quad x_m(t))^T = a_{i_1} s_{i_1}(t) + a_{i_2} s_{i_2}(t) + \cdots + a_{i_{m-1}} s_{i_{m-1}}(t) \quad (3)$$

From (3), it can be concluded that the  $t$ th column vector of observed signals matrix  $x$  is in one of  $C_n^{m-1}$  hyperplanes  $\mathcal{H}$ . Therefore, mixed frames can be recovered by the following algorithm.

- (a) Get the set  $\mathcal{H}$  of  $C_n^{m-1}$   $m$ -dimensional hyperplanes which are spanned by any  $m - 1$  columns of  $A$ ;
- (b)  $j$  repeat from 1 to  $m$ ,
  - (i) If  $x_j$ , which stands for the  $j$ th column of mixed signals matrix  $x$ , is in a hyperplane  $\mathcal{H}_q$ , then the following equation can be gotten

$$x_j = \sum_{v=1}^{m-1} \lambda_{i_v j} a_{i_v} \quad (4)$$

- (ii) Comparing Eqs. (3) and (4),  $s_i$ , the  $i$ th column of source signals matrix  $s$ , can be recovered: its components are  $\lambda_{i_v j}$  in the place  $i_v, v = 1, \dots, m - 1$ , and other components equal to zero.

### 3 Proposed UBSSVC Method

As explained above, for UBSS the number of mixed signals is less than that of source signals. Therefore, the mixing process of UBSS could be used to compress video sequences, and the separating process is used to decode the compressed video sequences.

#### 3.1 Mapping UBSS to Video Compression

Consider a video sequence with  $L$  frames,  $s_1, s_2, \dots, s_L$ , where  $s_i \in R^T$  is a T-pixel frame; we firstly divide the  $L$  video frames into  $b$  groups and in each group there are  $n$  frames. The encoder first chooses a matrix  $A \in R^{m \times n} (m < n)$  to mix  $n$  frames in each group. Thus, the compression ratio is  $n/m$ .

At the encoder side, unlike the traditional scenario of the UBSS issue, the mixing process is factitious in this proposed method. Thus, a specific mixing matrix  $A$ , known by both encoder and decoder, is chosen to mix raw video frames.

For standard BSS, there is only one restriction of mixing matrix  $A$ , that the columns of  $A$  should be mutually independent. However, in the proposed method, matrix  $A$  not only needs to satisfy the assumption 1, but also has to decrease the information loss in mixing process. Thus, in different mixed frames, the weight of different original frames should be varied. As each component of a row of  $A$  can be treated as the weight of every original frame in a mixed frame, the components of a row of  $A$  should be varied largely from each other. Experiments will be done to show  $A$ 's influence on the separation results in Sect. 4.

At the decoder side, the matrix  $A$  is known exactly, so the frames' order of recovered video sequence is not disturbed by mixing process and separating process, which is different from traditional BSS.

To ensure that the frames could satisfy the assumption 2, mixed frames are first transformed by a 2-D discrete Haar wavelet transform. And then SCA is used to recover the sparse high frequency components, while the recovered low frequency components are equal to multiply generalized inverse of mixing matrix  $A$  by mixed low frequency components.

### 3.2 Proposed UBSSVC Structure

The compression ratio for UBSS is only  $n/m$ . Therefore, to enhance the compression ratio more, we proposed UBSSVC framework that combines UBSS and conventional codec together, shown in Fig. 1.

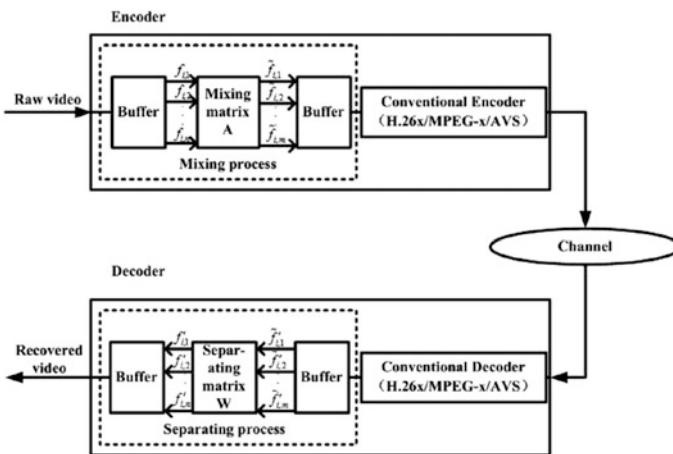


Fig. 1 Framework of UBSSVC



At the encoder side,  $n$  frames,  $f_{i,1}, f_{i,2}, \dots, f_{i,n}$ , are mixed into  $m$  frames,  $\tilde{f}_{i,1}, \tilde{f}_{i,2}, \dots, \tilde{f}_{i,m}$ . And then these mixed frames are encoded by traditional encoder such as MPEG-2, H.264. The buffer before mixing is used to buffer enough frames for being mixed. And the buffer after mixing is for storing mixed frames temporarily so that they can be encoded by conventional encoder one by one.

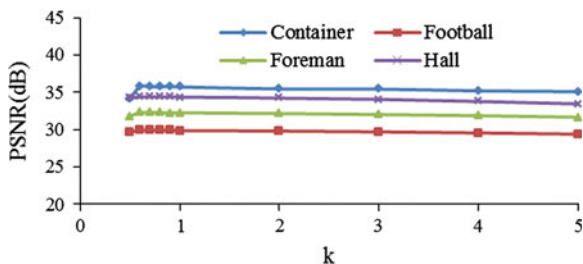
In the proposed decoder structure, received data is firstly decoded by traditional decoder; then the source recovery algorithm of underdetermined BSS is applied to recover original video sequence. In the separating process,  $m$  frames,  $\tilde{f}'_{i,1}, \tilde{f}'_{i,2}, \dots, \tilde{f}'_{i,m}$ , are separated to  $n$  frames,  $f'_{i,1}, f'_{i,2}, \dots, f'_{i,n}$ . The function of two buffers in decoder is similar to that of those two buffers in encoder.

## 4 Experiment Results

In order to validate this approach, multiple simulations are performed on four standard test video sequences: hall, container, foreman and football. The football sequence has the largest temporal variations, followed by foreman, and container ranks the third, while the hall sequence contains the most slowly scene variations. The first 40 frames of each sequence are used for test. Peak-Signal-to-Noise Ratio (PSNR) is used to evaluate the performance of recovery algorithm.

In the experiments, we just show an example of mixing 4 video frames into 3 frames, so the compression ratio of UBSS in the experiments is just 4/3. The mixing matrix  $A \in R^{3 \times 4}$ , shown in (5), is chosen to mix raw video sequence, where  $k \in Z, k \neq 0$ . The mixing process is performed as follows: continuous 4 frames are taken as source signals  $s$ , then  $A$  multiplies by  $s$  to calculate the mixed frames  $x$ . 30 mixed frames are generated after the mixing process. And then the above algorithm is applied to separate these mixed frames. Figure 2 shows the recovery PSNR on different video test sequences when  $k = 0.5-5$ . These plots show that the value of  $k$  has little influence on the separation PSNR. Although for some sequence, such as football, PSNR is a little low, it is still enough for monitor applications, which don't have very strict demands on high resolution.

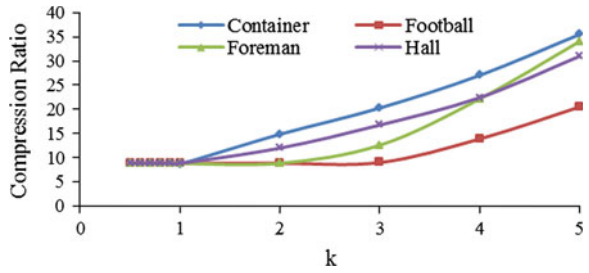
**Fig. 2** Separation PSNR related to different values of  $k$  on four videos



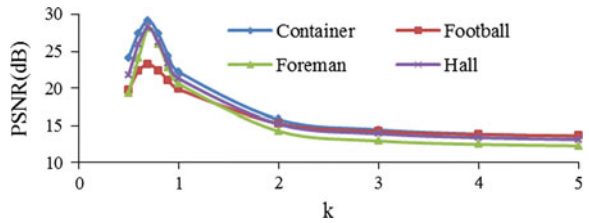
$$A = \frac{1}{k} \begin{pmatrix} 0.30 & 0.45 & 0.15 & 0.10 \\ 0.35 & 0.15 & 0.05 & 0.45 \\ 0.30 & 0.05 & 0.45 & 0.20 \end{pmatrix} \quad (5)$$

Experiments are done as well to show the  $k$ 's effect to the compression ratio and separation PSNR of UBSSVC+MPEG-2 which means that the conventional codec in Fig. 1 is MPEG-2, UBSSVC+H.264 which means that the conventional codec in Fig. 1 is H.264. Results are shown in Figs. 3, 4, 5 and 6. From the results, the  $k$  values indeed affect the UBSSVC+MPEG-2 and UBSSVC+H.264 compression ratio. That's because with the increment of  $k$ , most pixels values of the mixed frames approach to zero. Therefore, the compression ratios of MPEG-2 and H.264 for these mixed frames are much larger than that for the original frames. Meanwhile, it leads to a higher distortion. So the decoding PSNR decreases with the  $k$  increment when  $k > 0.7$ . For these four different test sequences, the largest PSNR and lowest compression ratio are almost gotten at the point  $k = 0.7$ . However, even the lowest compression ratio is larger than the corresponding

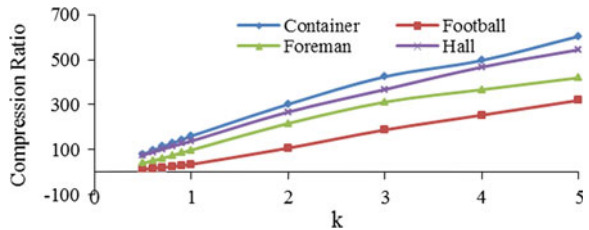
**Fig. 3** Compression ratio of UBSSVC+MPEG-2 related to different values of  $k$  on four video



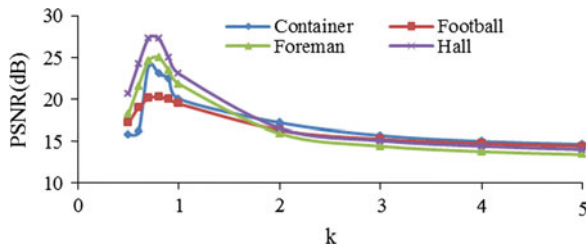
**Fig. 4** Decoding PSNR of UBSSVC+MPEG-2 related to different values of  $k$  on four videos



**Fig. 5** Compression ratio of UBSSVC+H.264 related to different values of  $k$  on four videos



**Fig. 6** Decoding PSNR of UBSSVC+H.264 related to different values of  $k$  on four videos



**Table 1** Compression results of test videos

	Hall (dB)	Container (dB)	Foreman (dB)	Football (dB)
MPEG-2	6.48	6.48	6.48	6.48
UBSSVC+MPEG-2 ( $k = 0.7$ )	8.84	8.84	8.84	8.84
H.264	76.74	92.28	62.04	12.15
UBSSVC+H.264 ( $k = 0.7$ )	102.41	161.06	74.25	25.38

**Table 2** PSNR results of test videos

	Hall (dB)	Container (dB)	Foreman (dB)	Football (dB)
MPEG-2	37.87	30.86	27.04	27.11
UBSSVC+MPEG-2 ( $k = 0.7$ )	28.23	29.18	28.15	23.23
H.264	36.19	35.43	35.13	31.7
UBSSVC+H.264 ( $k = 0.7$ )	27.26	24.16	25.01	20.29

compression ratio of MPEG-2 and H.264. Tables 1 and 2 show the comparison results. The PSNRs of UBSSVC+H.264 ( $k = 0.7$ ), and UBSSVC+MPEG-2 ( $k = 0.7$ ) is lower than those of H.264 and MPEG-2 respectively. Although the PSNR value is a little low, it is enough for some applications which don't have strict demands on high resolution, such as video surveillance system.

## 5 Conclusion

This paper initially develops the novel video compression approach UBSSVC. Furthermore, experiments are conducted to validate the efficiency of recovery algorithm, the influence of  $k$  values on separation PSNR and to measure the video compression ratio improvements of UBSSVC. The proposed method is suitable for video surveillance system perfectly. Firstly, it can achieve higher video compression ratio to decrease the bandwidth resource utilization. Secondly, the computation complexity of mixing process at encoder side is low, when improving the

video compression ratio. What's more, the mixing and separating process of UBSS has great potential in low-complexity video compression.

However, the presented new method still has more issues to be improved in our future work. Like the largest compression ratio the UBSS can achieve, and how to improve the compression ratio gained by the mixing process and enhance the separating results of video quality.

## References

1. Hyvarinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley, New York
2. Hyvarinen A (1999) Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans Neural Netw* 10:626–634
3. Cichocki A, Amari SI (2002) Adaptive blind signal and image processing: learning algorithms and applications. Wiley, Chichester
4. Georgiev P, Theis F, Cichocki A (2005) Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Trans Neural Netw* 16:992–996
5. Li YQ, Cichocki A, Amari SI, Shishkin S, Cao JT, Gu FJ (2004) Sparse representation and its applications in blind source separation. In: Thrun S, Saul K, Scholkopf B (eds) *Advances in neural information processing systems 16*, vol 16. MIT Press, Cambridge, pp 241–248
6. Ren M.-r, Wang P (2009) Underdetermined blind source separation based on sparse component in electronic computer technology, 2009 international conference on. pp 174–177
7. Zhenwei S, Huanwen T, Yiyuan T (2005) Blind source separation of more sources than mixtures using sparse mixture models. *Pattern Recogn Lett* 26:2491–2499
8. Lee TW, Lewicki MS, Girolami M, Sejnowski TJ (1999) Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Process Lett* 6:87–90
9. Waheed K, Salem FM (eds) (2003) Algebraic independent component analysis: An approach for separation of overcomplete speech mixtures. In: *Proceedings of the IEEE, international joint conference on neural networks 2003*, vols 1–4. New York, pp 775–780

# Grid Service Matching Process Based on Ontology Semantic

Ganglei Zhang and Man Li

**Abstract** Little prior communication between supplier and demander makes recognition of grid service ability helpful to make full use of some rich resources, such as software, hardware, information and others. Description and matching process about grid service based on ontology semantic were proposed, and its relative advantages comparing with the existing grid discovery mechanisms were also proved.

**Keywords** Grid computing · Ontology · Service matching

## 1 Introduction

Grid is a non centralized control across heterogeneous platform of collaborative working. It uses a standard, open and wide application protocol and seamless service quality. In addition, Grid technology allows sharing and collaboration in dynamic virtual Organizations. A new grid standard OGSA introduced Web Service technology, which has become industry standard, and presented the concept of grid service. It packages all software, hardware, network resources in a grid environment in the form of service, to provide a relatively uniform abstract interface. Grid resources are extremely rich, but if a specific user wants to find out adaptive service, he will face many problems. Some technologies such as UDDI, MDS2 mechanism in Globus Toolkit2 based on LDAP and Index Service based on service data matching. Their common shortcoming is little communication of

---

G. Zhang (✉) · M. Li  
Shandong Huayu Vocational and Technical College, DeZhou 253000, China  
e-mail: zhanganglei@163.com

M. Li  
e-mail: xdclm@126.com

service providers and demand side in advance [1]. Therefore, simple keywords matching can not provide enough flexibility and reasoning ability, and it is difficult to express the true ability of grid services. What's more, the retrieval result is not satisfied. This paper discussed the issues of description and retrieval of Grid Service on the basis of these results.

## 2 Grid Service and Its Ontology Description

### 2.1 *Ontology Theory and Description Language*

Ontology is derived from philosophy, intended to refer to the objective existence of nature. In computer science, an ontology is a group of shared concepts and deficit formal specifications. The type of concept and its bounds are defined definitely and ontology should be understood by machine. Sharing represents the concept must be public recognition, but not for private individuals or group. In addition, Concept represents ontology is the reflection of real world. Concepts such as class, object and inheritance are included in the development of ontology [2].

OWL Web ontology language is a universal language proposed by W3C organization after the standardization of DAML and OIL which is used to express Web Ontology. Compared with the original extensible markup language (XML) and resource description framework (RDF) technology, OWL is more convenient to express the semantic, and its ability of reading information is stronger. Its three sub-languages OWL Lite, OWL DL and OWL Full are respectively provided with different classification and binding capacity, supporting different levels of modeling needs.

### 2.2 *Grid Service and Globus*

Grid service is a Web service and provides a series of definitely defined interfaces to solve problems such as establishment of dynamic service, management of life circle, and notice. It can simplify some problems about large-scale network. Grid service used SOAP and WSDL technologies, and has some extension on the following aspects [3].

**Provide status service.** In Web Service, service never save the status information about users, while Grid Service can provide status service for users, and can generate a service instance and a mechanism of data inquiry.

**Provide instant service.** In Web Service, all the services are forever, while grid service not only provides forever services, but also provides instant services, that is, service instance and resource distribution only be carried out when the users

demand them. Resources and instances can be released after used by the customers to improve the availability.

### ***2.3 Ontology Description of Grid Service***

**OWL-S.** OWL-S is a kind of Web Service based on OWL, and it is derived from DAML-S. It provides a series of markup languages, and these languages can describe the features and abilities of Web Service exactly. According to the definition of OWL-S, a Web Service is described by presents, described by and supports, and their Range are respectively ServiceProfiles, ServiceModel and ServiceGrounding. They stated what the service can do, how to do and how to use service.

**Extension of OWL-S.** Grid service can be considered as an formal extension of Web Service. On one hand, OWL-S Service Profiles has its advantages on the description of ontology semantics. On the other hand, some extension is needed to describe the unique features of grid service. An extensible ontology class Grid-ServiceProfiles is introduced to provide a standard for the description of service ability of grid.

## **3 Semantic Matching Based on Ontology**

Semantic matching is to analyze the descriptions of two groups of semantics, and judge their conformation level. In factual application, requirement semantics description proposed by demand side is matched with description files from service publisher side to find out required service information. Ontology semantic information, the basis of matching, has strict concept definition and little ambiguity to realize strict matching. In addition, inheritance and intersection among concepts provide guarantee for the flexibility of matching. These are not exist in simple keywords matching. The matching includes main concept matching and service ability matching.

### ***3.1 Basic Concept Matching***

Grid service semantic matching is based on the concept model about its each property, while concept matching mainly depends on their positions in inheritance relationship. A matching function  $\text{concept\_match}(C1, C2)$  is defined, and its return results can be the following values.

**Exact.** It is an accurate matching, that is, C1 and C2 are the same concepts. In OWL, they point at the same URI node or the relative class through `<owl:sameClassOf>`.

**PlugIn.** C1 is the sub-concept of C2, that is  $C1 \subseteq C2$ , that is, C1 and C1 are combined by `<rdfs:subClassOf>`.

**Subsume.** In contrast with PlugIn, the meaning of Subsume is that C2 is the sub-concept of C1.

**Intersection.** There is an intersection between C1 and C2. It is defined by `<owl:intersectionOf>` in OWL.

**Fail.** In addition to the above four matching results, all the other results are Fail. It means the result is failed.

The five matching results are in descending order according to matching degree. It can be seen from definitions that Exact matching is the most accurate and the strictest matching, while the others provides alternatives with varying degrees when accurate matching can not be met. In addition, results of the middle three matching methods have transitivity. The more degrees of transition a matching results through, the lower the similarity degree is. In order to improve matching efficiency, the degrees can be limited by service requester.

### 3.2 Service Ability Matching

Overall evaluation of matching degree of service capability is seen as follows. Overall capacity matching should mainly consider category, input, output and data information. For the information of services provider, because it does not belong to semantic, it is only be matched when the demand side need.

For the the information of class service, if it is in the OWL framework, matches were nothing special; if it is in the external classification system,

There are two ways: one is mapping the existing classification system to its equivalent OWL ontology description, the other is to introduce external classification system with plugs. At present, the two programs are with high cost.

For the information of output, output description of demand side are removed one by one, and the corresponding matching results of its belonging concept are looked for in output of release side. If each item of output information can find a match, overall matching of output information is considered successful.

The process of input information matching is similar with output information matching. However, the judgment way is extracting each item of input information of release side to find matching results in input description of demand side. It is equivalent to swapping the positions of release side and demand side in output matching function. This is because whether output information is matched is relative to whether the service can be accepted. While matching degree of input information is related to whether the service can be normally executed.

Service data can be seen as a special kind of output information, so output information matching mode is adaptive to it.



Finally is the sorting of the whole service ability. It mainly refers to classification and service data whose return value is not Fail. Because of low importance of service data, only its matching results are returned as parameters. The sorting process followed by category information, input data, output data and service data.

## 4 Conclusions

This paper introduced the improved matching process and ontology semantic description of grid service ability. It is easy to be carried out under GT3 framework and has great flexibility and accuracy. However, it should be improved on specific matching information, the reliability of semantic information and the automation.

## References

1. Foster I, Kesselman C, Nick JM et al (2003) The physiology of the grid:an open grid services architecture for distributed systems integration. <http://www.globus.org>
2. <http://www.w3c.org/2002/ws> [EB/OL] (2004)
3. UDDI: Universal Description, Discovery and Integration (2004) <http://www.uddi.org>

# Enhancements on the Loss of Beacon Frames in LR-WPANs

Ji-Hoon Park and Byung-Seo Kim

**Abstract** In beacon-enabled LR-WPANs, the high reliability of beacon frame transmission is required because all transmissions are controlled by the information in the beacon frame. However, the process for the beacon-loss scenario is not carefully considered in the standard. The method proposed in this paper allows a node not receiving a beacon frame to keep transmit its pending frames only within the minimum period of CAP based on the previously received beacon frame while the standard prevents the node from sending any pending frame during a whole superframe. The method is extensively simulated and proven the enhancements on the performances.

**Keywords** LR-WPAN · Beacon · Sensor · IEEE802.15.4

## 1 Introduction

Applications using such IEEE802.15.4 standards-based Low Rate-Wireless Personal Area Networks (LR-WPANs) have been increasing in broad areas including public safety, home entertainment system, home automation systems, ubiquitous building systems, traffic information systems, and so on. IEEE802.15.4 standard [1] defines two types of LR-WPANs: beacon-enabled and nonbeacon-enabled networks. Any transmission of any node in the beacon-enabled LR-WPANs is controlled by the information in the beacon frames transmitted by the central PAN

---

J.-H. Park (✉)  
Nongshim Data System, XXX, Korea  
e-mail: topsicrit@daum.net

B.-S. Kim  
Department of Computer and Information Communications Engineering, Hongik  
University, Hongik, Korea  
e-mail: jsnbs@hongik.ac.kr

coordinator. Therefore, the high reliability of beacon transmission is essential for beacon-enabled LR-WPANs.

However, there are some factors to cause the loss of beacons such as collisions between beacons and interferences from other devices. The focus of this paper is the beacon loss due to the latter. The loss of beacon due to interference occurs because many communication networks like LR-WPANs and Wireless Local Area Networks (WLANs) and even microwaves uses same frequency bands of 2.4 GHz which is called Industrial Scientific Medical (ISM) band [2]. As a consequence, LR-WPANs experience severe interferences from other devices. The performance degradations of LR-WPAN due to interferences are reported by many experiments and studies as shown in [3–9]. As the electric power grid systems recently utilize LR-WPANs and WLANs, the interference issues in the power grid system is reported as shown in [9]. Especially, as the number of deployed WLANs rapidly increases, the impacts on LR-WPANs of interferences from WLANs are actively researched in [4–8] and it is shown that LR-WPANs coexisting with WLANs experience 10–100 % degradations on the performances depending on the distances between LR-WPANs and WLANs, locations, the channels used by LR-WPANs, and the traffic loads of WLANs. There are many studies to avoid the interference. To resolve the problem, the most of methods switch the channels to non-interference channel.

While all aforementioned method proposed methods to avoid a beacon loss, no aforementioned studies mentions the process when a device fails to receive a beacon. Even though the many solutions have been proposed, the beacon can still be lost because of the channel characteristics like noise, fading, Doppler effects and so on. Based on IEEE802.15.4 standard, devices failed to receive a beacons have to hold their pending transmissions during a superframe associated with the beacon which cause the performance degradations. Therefore, we need to a better method to improve the network performances when the beacon is lost.

In this paper, a method is proposed to improve the performances of beacon-enabled LR-WPANs by allowing nodes to transmit its pending frames during a Contention Access Period (CAP).

In Sect. 2, IEEE802.15.4 standard-based LR-WPANs and the process when the beacon is lost are introduced. In Sect. 3, the proposed protocol is described. After evaluating the performances of the proposed method with extensive simulations, finally conclusions are made.

## 2 Preliminary Researches

### 2.1 IEEE802.15.4 Standard

In beacon-enabled LR-WPANs in IEEE802.15.4 standard, the time is subdivided into consecutive superframes. The structure of the superframe is shown in Fig. 1. The standard optionally allows the superframe to be divided in two parts: active

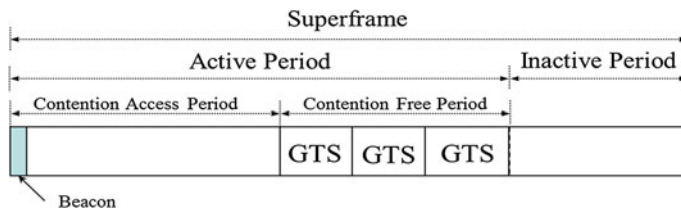


Fig. 1 Superframe structure for beacon-enabled LR-WPANs

and inactive periods. Actual data packets between devices are transmitted during active period while any packet transmission is prohibited during the inactive period for saving the power. Therefore, all devices with pending packets have to hold their transmissions until the next active period. The active period is composed of 16 slots and each slot is 960us based on [1]. The superframe is started with the transmission of beacon from a PAN coordinator. The beacon is used to synchronize with participating devices, to identify the WPAN, and to inform the participating devices the structure of the superframes. After beacon transmission, contention access period (CAP) is followed. The CAP adopts the contention-based data transmissions like carrier sense multiple access with collision avoidance (CSMA/CA). After CAP, Contention Free Period (CFP) is follows. CFP is composed of multiple Guaranteed Time Slots (GTSs). The lengths of GTSs are varied unlike time slots in conventional Time Division Multiple Access (TDMA)-based system. The maximum number of GTSs in CFP is 7 and a GTS can occupy more than one slot. GTSs in CFP are allocated by the PAN coordinator when devices requests. The information on the GTS allocation is included in the beacon. During GTS, only designated device transmits its packet without contention and collision.

## 2.2 Beacon Loss in IEEE802.15.4 Standard

The process when a device fails to receive a beacon is not clearly described in the standard and any literatures except for the case in which GTSs are allocated in the superframe. When a device's GTSs are allocated in a superframe and it loses a beacon, the device is not allowed to transmit its packet during its GTS. Since a beacon contains the information on superframe structure like period of CAP, the allocation of GTS, and so on, and the superframe structure can vary in every superframe, if a device fails to receive a beacon, it can be assumed that it needs better to hold its transmissions during the superframe to prevent from collisions with other scheduled transmissions. This assumption is clear for the cases that the network parameters like the number of nodes, traffic loads, etc. are frequently fluctuated.

Furthermore, based on the IEEE802.15.4 standard [1], if a device does not receive beacons an  $aMaxLostBeacons$  times, it declares synchronization loss and starts orphan channel scan after discarding all buffered packets in MAC layer. The orphan channel scan scans the channels in a specified set of logical channels to search a PAN coordinator to re-associate with. When starting the orphan channel scan, the device sends an orphan notification command, and waits a PAN coordinator realignment command from a PNC within a  $macResponseWaitTime$  symbols. This process is repeated for the channels in the set of logical channels.

Overall, the losses of beacons cause holding devices' transmissions as well as the synchronization loss, and as consequences it degrades the network performances.

### 3 Proposed Method

In this paper, we propose an enhancement on IEEE802.15.4 standard-based and beacon-enabled LR-WPAN for the case that the beacons are not received by participating devices. As mentioned in Sect. 2.2, the loss of beacon causes two issues: holding transmissions and re-association. This paper focuses to the first issue.

The proposed method in this paper is to allow devices failed to receive a beacon (hereinafter the device is called 'failed-device') to transmit their pending frames during the minimum period of CAP. The proposed method is focused at the scenario that the lengths of superframe size and active period are fixed which traffic is not much fluctuated like sensor networks.

As described in Sect. 2.1, the active period is composed of CAP and CFP. If the beacon is not received, any transmission in CFP by the failed-device causes a problem because each GTS in CFP is assigned to a specific device and the assigned device transmits its own data without any collision. If a failed-device transmits its data in CFP because it does not know the current structure of the superframe, it causes collisions with transmission that is supposed to be collision-free. In order to prevent this case, any device failed to receive beacon is allowed to transmit its data only in CAP while the device discard all pending frames defined IEEE802.15.4 standard. However, CAP can be varied due to varying the length of CFP. Therefore, the period that a failed-device can be allowed to transmit is defined as follows:

$$T_{BeaconLoss} = aNumSuperframeSlots - MaxNumofSymbol_{CFP} \quad (1)$$

where  $aNumSuperframeSlots$  is the number of slots in active period defined in IEEE802.15.4 standard and  $MaxNumofSymbol_{CFP}$  is a maximum number of symbols that can be assigned for CFP.

## 4 Performance Evaluations

In this section the enhancements by using the proposed method is evaluated in terms of throughput. The proposed method is compared with IEEE802.15.4-base LR-WPANs. For the simplicity in the mathematical analysis, the saturated traffic model and no CFP is considered. The synchronization loss is not considered because both of the proposed and IEEE802.15.4-based LR-WPANs have same effects on the loss of synchronizations. The throughput of the proposed method is

$$Thr_p = \frac{D(1 - PER_D)(1 - PER_B) + D'(1 - PER_D)PER_B}{T}, \quad (2)$$

where  $D$  and  $D'$  represent the amounts of data transmitted during active periods when a beacon is successfully and unsuccessfully received, respectively, and  $T$  means the duration of two superframes. The reason of two superframe durations is that one superframe with successfully receiving a beacon and with the miss of beacon. In addition,  $PER_D$  and  $PER_B$  represent packet error rates of data and beacons, respectively. The throughput of the IEEE802.15.4-base WPANs is

$$Thr_{IEEE} = \frac{D(1 - PER_D)(1 - PER_B)}{T}. \quad (3)$$

Because the amount of transmitted data is proportional to active period and  $T_{BeaconLoss}$  in (3), the ratio of  $D$  and  $D'$  is the ratio of active and  $T_{BeaconLoss}$  periods in a superframe. When the ratio of  $D'$  to  $D$  is  $\gamma$ , the enhancements obtained by using the proposed method is

$$E = \frac{Thr_p - Thr_{IEEE}}{Thr_{IEEE}} = \gamma \frac{PER_B}{1 - PER_B}. \quad (4)$$

Based on the Eq. (6), the performance enhancement in the throughput depends on the  $PER_B$  and the ratio of  $D'$  to  $D$ .

Figure 2 shows the enhancements in throughput performances as a function of data packet error rates and the lengths of beacons.  $\gamma$  is set to 7/15 because active period is composed of 16 slots, a beacon uses one slot, and the maximum slots for CFP is recommended 7 in [1]. As shown in Eq. (6) the performance enhancements are depending on the packet error rate of beacons. As shown in Fig. 2, the performance enhancements increase as the error rate of beacons increase. The proposed method allows devices to keep transmitting their pending frames during the minimum required times while the conventional method does not. Therefore, as the number of missed beacons increase, the conventional method loses opportunities for transmitting devices' pending frames, so that the proposed method shows the better performances. The enhancements are from 2.5 % with 5 %  $PER_D$  to 31 % with 60 %  $PER_D$ . As the measurement studies for interference issue with WLANs are shown in [3–9], the  $PER_D$  is varied from 10 % to 100 %. Therefore,

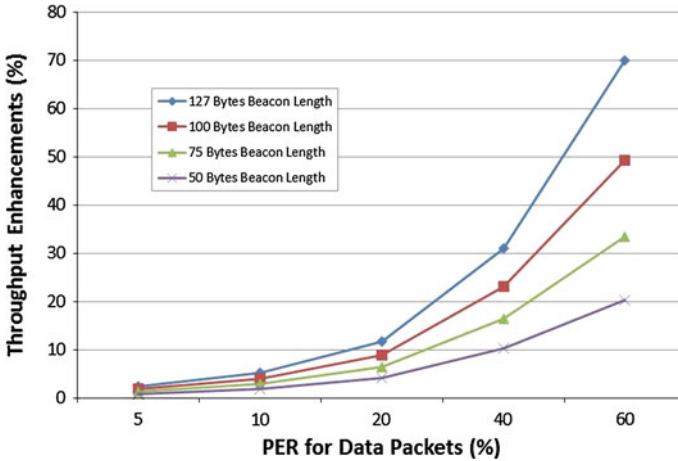


Fig. 2 Throughput enhancements as function of PER for data packets and the lengths of beacons

even though the standard [1] requires 10 %  $PER_D$ , analyzing performances over temporal high  $PER_D$  scenarios is valuable. Even at 10 %  $PER_D$ , 6 % improvements is achieved.

## 5 Conclusions

The reliability in the beacon transmissions is very critical on the performance of Beacon-enabled LR-WPANs because the loss of beacon causes for devices to hold their transmissions during the superframe. Unlike specification in the standard, the method proposed in the paper allows devices to transmit its pending packet only during the minimum period of CAP that is guaranteed in the superframe. Therefore, the proposed method improves the network performances.

**Acknowledgments** This research is supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (2012-0003609) and in part by the International Science and Business Belt Program through the Ministry of Education, Science and Technology (2012K001556).

## References

1. IEEE Std 802.15.4, Part 15.4: wireless medium access control (MAC) and Physical Layer (PHY) specification for Low Rate
2. Lau S-Y, Lin T-H, Huang T-Y, Ng I-H, Huang P (2009) A measurement study of zigbee-based indoor localization systems under RF interference. In: Proceedings of the 4th ACM

- international workshop on experimental evaluation and characterization (WINTECH'09), pp 35–42, Beijing 20–25 Sept 2009
3. Howitt I, Gutierrez JA (2003) IEEE 802.15.4 Low rate -wireless personal area network coexistence issues. In: Proceedings of the IEEE WCNC'03, vol 3, pp 1481–1486
  4. Sikora A, Groza, VF (2005) Coexistence of IEEE802.15.4 with other Systems in the 2.4 GHz-ISM-Band. In: Proceedings of the IEEE instrumentation and measurement technology conference (IMTC'05), vol 3, pp 1786–1791, Canada, 17–19 May 2005
  5. Shin S et al, Packet error rate analysis of IEEE 802.15.4 under IEEE 802.11b interference. In: Proceedings of the WWIC'05, pp 279–288
  6. Yoon DG, Shin SY, Kwon WH, Park HS (2006) Packet error rate analysis of IEEE 802.11b under IEEE 802.15.4 interference. In: Proceedings of the IEEE 63rd vehicular technology conference, pp 1186–1190, Australia, 7–10 May 2006
  7. Petrova M, Gutierrez JA (2006) IEEE 802.15.4 Low rate—wireless personal area network coexistence issues. In: Proceedings of the IEEE WCNC'06, USA
  8. Yuan W, Wang X, Linnartz J-PMG (2007) A coexistence model of IEEE 802.15.4 and IEEE 802.11b/g. In: 14th IEEE symposium on communications and vehicular technology in the Benelux, 15 Nov. 2007
  9. Shin SY, Park HS, Kwon WH (2007) Mutual interference analysis of IEEE 802.15.4 and IEEE 802.11b. *Comput Netw* 51(12):3338–3353, 22 August 2007
  10. Stanciulescu G, Farhangi H, Palizban A et al (2012) Communication technologies for BCIT Smart Microgrid. 2012 IEEE PES innovative smart grid technologies (ISGT), 16–20 Jan 2012
  11. The network simulator NS-2, Web site <http://www.isi.edu/nsnam/ns>



# Case Studies on Distribution Environmental Monitoring and Quality Measurement of Exporting Agricultural Products

Yoonsik Kwak, Jeongsam Lee, Sangmun Byun, Jeongbin Lem,  
Miae Choi, Jeongyong Lee and Seokil Song

**Abstract** In this paper, we present monitoring the distribution environmental factors for exporting agricultural products in real time based on sensor networks and packaging technologies, and how the distribution environmental factors would influence the quality of agricultural products has been studied by measuring the actual quality of agricultural products when this distribution process has been completed. For this, sensor nodes and communication system, optimized to a monitoring process for the distribution environmental factors of agricultural products, have been designed and implemented. With the paprika exported to overseas, information on temperature/humidity/path (distribution environmental factors) has been monitored in real time. The possibility of utilization of sensor networks based distribution environmental factors monitoring technology could be verified through such case studies.

---

Y. Kwak (✉) · S. Song

Department of Computer Engineering, Korea National University of Transportation,  
Chungju, South Korea  
e-mail: yskwak@ut.ac.kr

S. Song

e-mail: sisong@ut.ac.kr

J. Lee · S. Byun · J. Lem

Marketing Policy Division, Ministry for Food, Agriculture, Forestry and Fisheries,  
Gwacheon, South Korea  
e-mail: gnothi@hanmail.net

M. Choi

Postharvest Research Scientist, National Institute of Horticultural & Herbal Science,  
RDA, Suwon, South Korea  
e-mail: choma818@korea.kr

J. Lee

Agribusiness Development Team, FACT, Suwon, South Korea  
e-mail: dfy0928@daum.net

**Keywords** Monitoring · Distribution environmental factors · Agricultural products

## 1 Introduction

Due to the rapid globalization of the world economy and an aggravated competition between countries, to secure competitiveness and differentiate itself in the agricultural and fishery industry field, various efforts have been made. To secure competitiveness in the agricultural and fishery industry fields is being more and more important to the both of nations and related producers because of WTO and FTA systems. Thus, in an effort to secure competitiveness for each country and its producers, the agricultural industry has been promoted as a nation's growing potential industry. Subsequently, each country has invested heavily in agricultural technology. On the other hand, the significance is being emphasized even more as the agricultural and fishery products are rendered to resources or utilized as a scale of nation's competitiveness.

Consequently, for the past couple of years, information and communication technologies in agriculture industry have been received heavy attention as research issues. One of the research issues is sensor network based monitoring system for distribution and cultivation of agricultural products. In this paper, we develop a wireless sensor network (WSN) based monitoring system for agricultural products distribution. The monitoring system gathers physical environment during agricultural products distribution such as temperature and humidity which are main factors to affect the freshness of agricultural products. Subsequently, we analyze the monitoring results how the physical environment affect the commercial value of agricultural products with various packaging techniques [1–5].

We deploy our developed system in a container box that where paprika boxes packaged with some techniques are loaded. Among the exporting agricultural products in Korea, paprika has been the one of the biggest export quantities. In the year 2000, the amount of exported paprika was 2,207 ton, but the amount has increased by 733 % to 16,168 ton in the year 2010. However, in case of the exporting quantity of paprika, the needs of pioneering a new market is on the rise, as the domestic producers and export related industries are greatly influenced by change of the Japanese market as more than 99 % of it is biased to Japan.

Also, although various recent attempts to venture Australian or the U.S. markets have been done, the distribution period and freshness maintenance problems are still at the forefront as a prerequisite. In particular, in terms of the distribution period problem, long period transportation period is required as the transportation period by shipment takes 20 days to the U.S., 25 days to Canada and 23 days to Australian. Thus, long period transportation is impossible since the shelf life of paprika is only 2 weeks.

In order to solve such problems, efforts are continues to secure quality through freshness maintenance, during the long term distribution period, utilizing the research and developed management technology (packaging) after the harvest. Additionally, as efforts to increase export quantity and to construct a stable production supply system, export nations of diversified are being made consistently, and also, there are efforts to secure price in the overseas market (high quality, high price) [6, 7].

In this paper, determination process of the harvest time, chlorine dioxide processing and MA packaging technology have been applied as the packaging technology, and the distribution environmental factors of the distribution process from harvest to sales have been monitored in real time using the sensor network technologies. In other word, a temperature and humidity sensor was installed inside the container during the full-period distribution process of paprika, and change of the environmental factors (temperature, humidity) is measured on the whole of the distribution period. The information on temperature and humidity is measured with 30 min interval, and at each measurement, the location information is measured and stored at the same time. By doing this, what kind of temperature and humidity at which location was measured could be monitored in real-time? Additionally based on the obtained data for distribution environmental factors, through analyzing relationship of the changes for quality and temperature/humidity changes after the distribution, figure out how the changes of temperature/humidity during the distribution process had influenced the paprika quality. Quality management strategies of exporting agricultural products were proposed based on this.

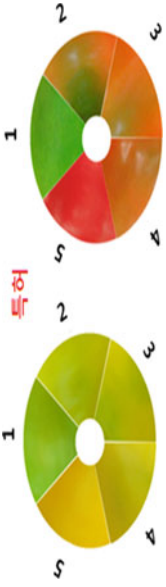




## 2 Packaging and Sensor Networks Technology

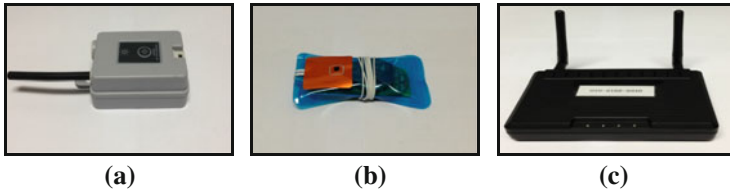
Table 1 shows the post-harvest management and packaging technologies used in paprika transportation process. Two packaging technologies have been applied to our experiments. The first technology is to sterilize the top and surface parts of paprika with chlorine dioxide for 30 min with 0.1 ppm. The second technique is to maintain the internal humidity of the box to an adequate level using the internal packaging materials such as MA packaging. The last one applies the both technologies together. As shown in Table 1, for the agricultural products used in the process, the harvest time of paprika are controlled using color chart.

For the sensor network technologies, temperature and humidity sensors and the communication hub that transmits the data collected by the sensors to remote server is used for real time monitoring the distribution environmental factors during the export.

400 MHz frequency bandwidth is used for the sensor nodes, and through the super low power technology, they are designed to sufficiently operate during the transportation period to Australia. Additionally, in case of the sensor nodes, they are designed and realized classified into fixed and box feeding types. The fixing

**Table 1** Packaging technologies

Applied technologies	Contents	Remark
Determine harvest period (color chart)	Adjust harvest period using a color chart Yellow: For export 3, for domestic demand 4–5 Red: For export 4, for domestic demand 5	
Chlorine dioxide processing	Sterilize paprika's top and surface parts Processing condition: 0.1 ppm, 30 min	 
MA packaging	Maintain adequate humidity within the box using the internal packaging materials External packaging materials: 4 air drains Internal packaging materials: 0.03 mm PE film, Vent ratio: 1.5– 2.0 %	 



**Fig. 1** Sensor nodes and communication hub. **a, b** Node. **c** Communication hub

type is designed to use by fixing into storage and the box feeding type is designed using soft plastic bags to minimize the impact to the quality of agricultural products such as paprika. Fig. 1 shows the sensor nodes (fixing type, box feeding type) and the communication hub used in the experiment.

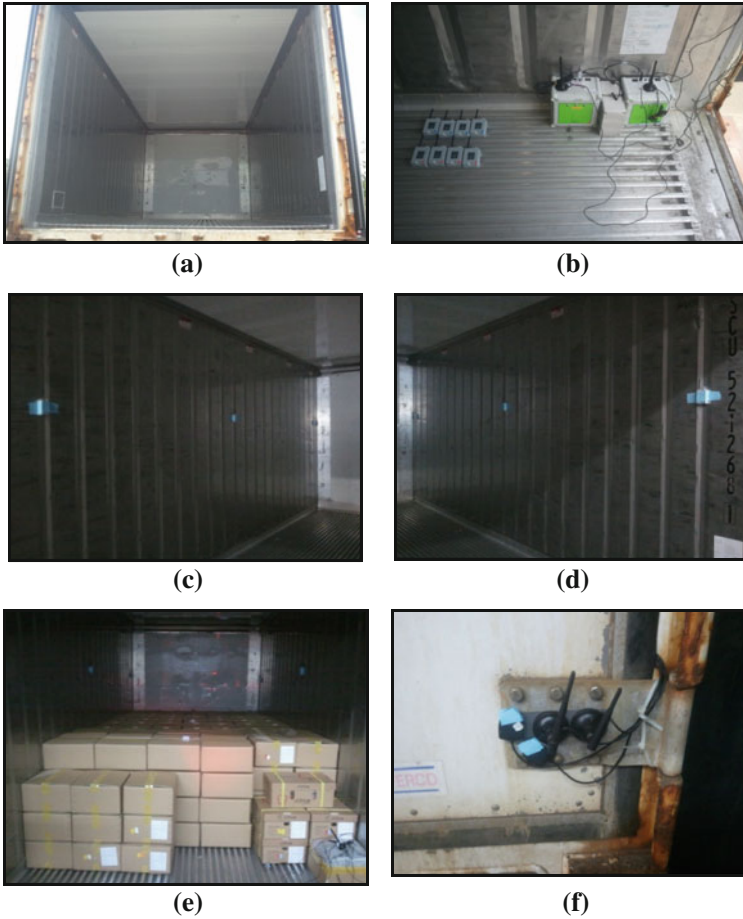
In order to transmit temperature and humidity data to the remote server in real time, the communication hub is equipped with a WCDMA module. However, transmission in open waters is not possible and can transmit the data to the server through automatic roaming when the ship anchors in the port. Also, it is designed to transmit environmental factors simultaneously through a GPS equipped in the communication hub to figure out the location where temperature and humidity data is collected. Through this, it is possible to analyze how the external environments affect to internal environments of the containers.

### 3 Case Studies

In this paper, the distribution environmental factors (temperature, humidity, location data) have been monitored in real time for the paprika, agricultural products exported to overseas, as the exported object from its domestic origin to Australia passing Busan Port. Figure 2 shows the sensor nodes and the communication hub that are deployed in the container box. Figure 2a shows the container box and Fig. 2b shows the fixing type sensor nodes and the communication hub. The fixing type sensor nodes to acquire the temperature and humidity data inside the container are shown in Fig. 2c and d, they have been installed at the middle height from the container floor. Figure 2e shows the loaded paprika boxes and Figure 2f shows the antenna installed at the outside of the container box to enable data transmission of the communication hub through the mobile telephone network.

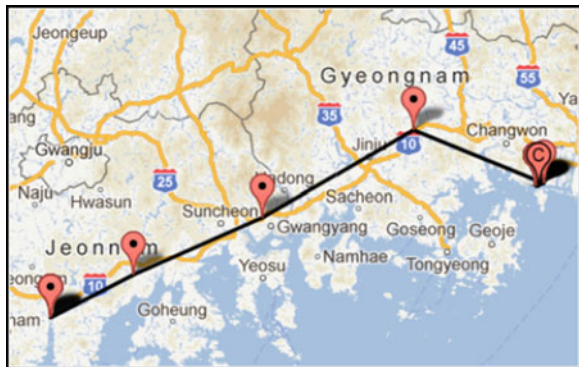
Figure 3 illustrates the mapping of the map and the data of the GPS mounted on the communication hub during the real-time monitoring on the distribution environmental factors. The map shows the transfer route of the container box by a transportation means.

Figure 4 shows the graph of the temperature and humidity data collected by the sensor nodes when selecting pre-installed sensor nodes. The figure shows which temperature and humidity has been measured at each location by interlocking with the graph. In the figure, the red at the right side is the humidity, and the blue at the



**Fig. 2** Deployed sensor networks system. **a** Container. **b** Sensor nodes and hub. **c, d** Deployed sensor nodes. **e** Paprika box. **f** Installed antenna

**Fig. 3** Information of GPS



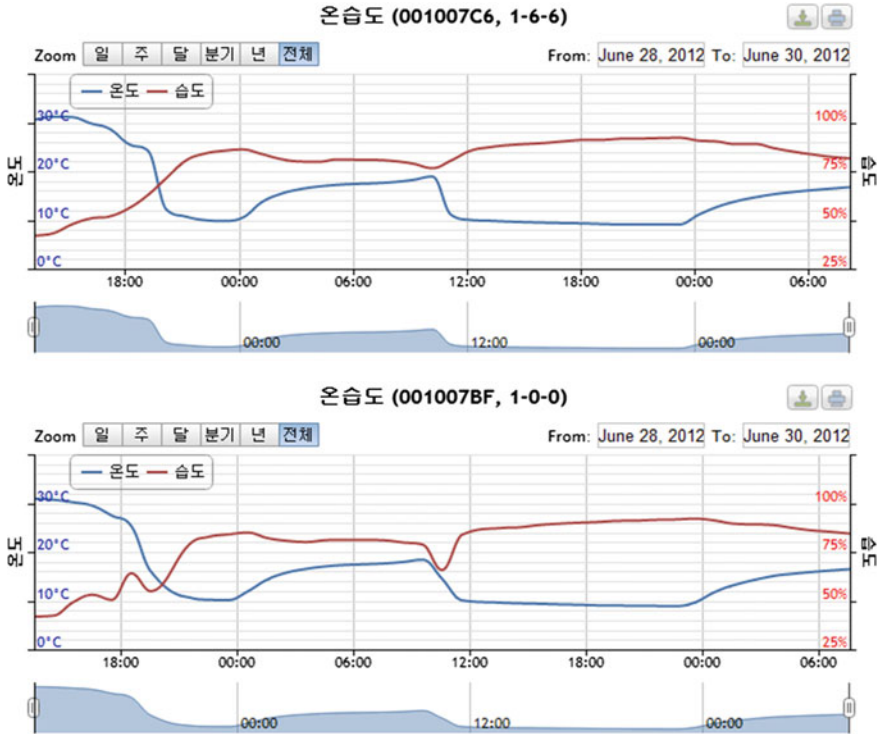


Fig. 4 Information on temperature and humidity

left side is the temperature. As can be known from the picture, it is the result obtained by real-time monitoring of the temperature change during transportation started from the initial loading of room's temperature 30°, and it can be known that it is maintained at the temperature/humidity pre-set by the user as time passed starting from the initial humidity of 30 %. Also, a drastic change of temperature and humidity at a certain period of time is observed and this implies that a random event (container power cut, container door problem etc.) occurs during transportation.

## 4 Conclusion

From the case studies, taking a real-time action was possible by monitoring various environmental problems likely to occur during the transportation process through applying packaging and sensor network technologies as the advancement technology for agricultural and fisheries products, and securing the commercial value was also possible through quality management.

Through real-time monitoring the distribution environmental factors for high value-added agricultural products, the conclusion that sales routes of paprika as well as agricultural products which require long time transportation could be ventured if the quality and freshness of agricultural products could be adequately maintained and estimated. Furthermore, it is determined to diversify exporting countries, enable stable production and construct of supply systems, and thus can promote profits of farmers and stabilization of market price.

The obtained information could be utilized as various reference data and such a technology is planned to be applied diversely to prosperous exporting agricultural products in the future.

**Acknowledgments** This research was supported by Technology Development Program for 'Bio- Industry Technology Development', Ministry for Food, Agriculture, Forestry and Fisheries, Republic of Korea.

## References

1. Hwang J, Shin C, Yoe H (2010) Study on an agricultural environment monitoring server system using wireless sensor networks. *Sensors* 10(12):11198–11211
2. Ruiz-Garcia L, Lunadei L, Barreiro P, Robla JI (2009) A review of wireless sensor technologies and applications in agriculture and food industry: state of the art and current trends. *Sensors* 9(6):4728–4750
3. Akyildiz IF, Su W (2002) A survey on sensor networks. *IEEE Commun Mag* 40(8):102–114
4. Culler D, Estrin D, Srivastava M (2004) Overview of sensor networks. *Computer* 37(8):41–49
5. Yoneki E, Bacon J (2005) A survey of wireless sensor network technologies: research trends and middleware's role. Technical Report, University of Cambridge
6. Kwak YS (2010) Design and implementation of sensor node hardware platform based on sensor network environments. *J Korea Navig Inst* 14(2):227–232
7. Kwak YS (2011) Design and implementation of the control system of automatic spry based on sensor network environments. *J Korea Navig Inst* 15(1):91–96



# Vision Based Approach for Driver Drowsiness Detection Based on 3D Head Orientation

Belhassen Akrouf and Walid Mahdi

**Abstract** The increasing number of accidents is attributed to several factors, among which is the lack of concentration caused by fatigue. The driver drowsiness state can be detected with several ways. Among these methods, we can quote those which analyze the driver eyes or head by video or studying the EEG signal. We present, in this paper an approach which makes it possible to determine the orientation of the driver head to capture the drowsiness state. This approach is based on the estimation of head rotation angles in the three directions yaw, pitch and roll by exploiting only three points face features.

**Keywords** Driver drowsiness detection • 3D head orientation • Perspective Projection • Haar features • Harris detector

## 1 Introduction

In literature, many systems based on video analysis have proposed for drowsiness detecting [1]. Special attention is given to the measures related to the speed of eye closure. Indeed, the analysis of the size of the iris that changes its surface according to its state in the video allows the determination of the eye closure [2]. Other work is based on detecting the distance between the upper and the lower eyelids in order to locate eye blinks. This distance decreases if the eyes are closed

---

B. Akrouf (✉) · W. Mahdi  
Laboratory MIRACL, Institute of Computer Science and Multimedia of Sfax,  
Sfax University, Sfax, Tunisia  
e-mail: akrouf\_belhassen@yahoo.fr

W. Mahdi  
e-mail: walid.mahdi@isimsf.rnu.tn

and increases when they are open [3]. These so-called single-variable approaches can prevent the driver in case of prolonged eye closure, of its reduced alertness. The second type of approach is called multi-variable [4, 5]. In this context, the maximum speed reached by the eyelid when the eye is closed (velocity) and the amplitude of blinking calculated from the beginning of blink until the maximum blinking are two indications that have been studied by Murray [6]. Takuhiro [7] uses an infrared camera and suggests five levels of vigilance namely non-drowsy, slightly drowsy, sleepy, rather sleepy, very sleepy and asleep. Picot [4] presents a synthesis of different sizes as the duration to 50 %, the PERCLOS 80 %, the frequency of blinking and the velocity amplitude ratio. These variables are calculated every second on a sliding window of the length of 20 s. Some multi-variable approaches require technical cooperation between the hardware and the driver. Moreover, these methods need the use of wide range of parameters, which calls for more data for learning. Other studies estimate orientation of the head driver [8] to detect drowsiness state. These researches are based on the face shape and calculate the local descriptors such as eyes, mouth and nose to estimate head angles rotation. In this paper, we present an approach called geometric, based on the nose tip and mouth corners to determine the angles (Yaw, Pitch and Roll).

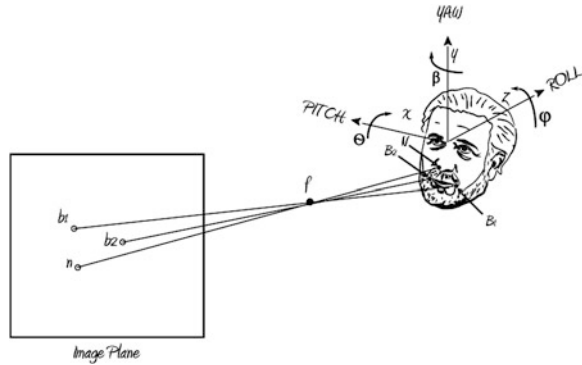
## 2 Proposed Approach for 3D Head Pose Estimation

Our approach requires primordial stages in order to pose estimation. The first step allows detecting driver nose and mouth with Haar features method [9]. Interest points of the face are located by Harris detector [10] to define mouth corners. The center of the box which encompasses the nose is the tip. We can calculate 3D rotation angles for head driver, from these three points.

### 2.1 Proposed Perspective Model

We present, in this section, a perspective model to estimate the driver head pose. We suppose that the subject is installed in front of fixed and calibrated RGB camera.  $f$  is the focal point (Fig. 1). We suppose that  $\Theta$ ,  $\beta$  and  $\varphi$  are the head angle rotations for X, Y and Z axes respectively. We consider that the image plane is parallel to X-Y axis of our subject. Let  $B_1$  and  $B_2$  the two corners points of the mouth and  $N$  the tip nose in 3D space. Projections of these last points in the image plane are  $b_1$ ,  $b_2$  and  $n$  respectively.

**Fig. 1** Geometric system coordinates for 3D head pose and its projection in the image plane



### 2.2 Roll ( $\varphi$ ) Estimation

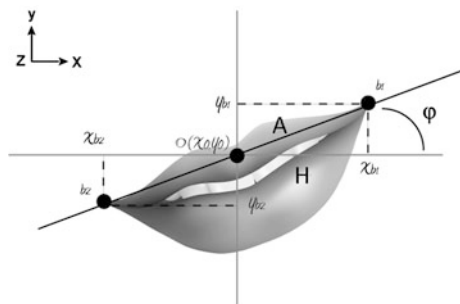
Let  $b_1$  and  $b_2$  the two corners points of the mouth,  $O$  with  $(X_0, Y_0)$  coordinates is the center of the segment  $[b_1, b_2]$ . Let  $A$  the distance between  $O$  and  $b_1$ ,  $H$  the distance between  $O$  and the projection of point  $b_1$  on the horizontal axis which passes by the point  $O$  (Fig. 2). The rotation angle  $\varphi$  of head Roll on  $Z$  axis is calculated as follows

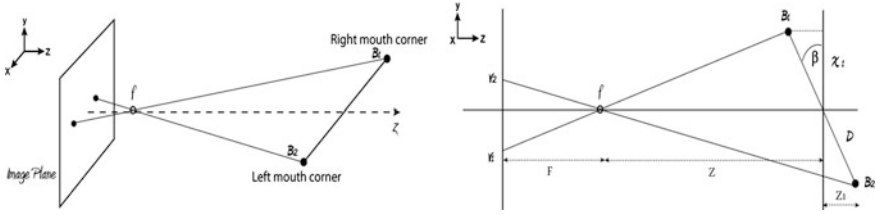
$$\varphi = \arccos\left(\frac{A}{H}\right) \tag{1}$$

with

$$A = \sqrt{(X_{b_1} - X_o)^2 + (Y_{b_1} - Y_o)^2} \text{ and } H = |X_{b_1} - X_o| \tag{2}$$

**Fig. 2** Calculate angle  $\varphi$  starting from the two points of mouth corners





**Fig. 3** Representation of mouth corners in 3D space (*top*), their projections in 2D image plane (*bottom*)

### 2.3 Yaw ( $\beta$ ) Estimation

For the estimation of the angle  $\beta$ , we propose a new method based on the perspective projection. Let us pose  $D$  is the distance between the center of the mouth and one of two mouth corners in 3D space (Fig. 3). With  $F$  is the focal distance from RGB camera.  $v_1$  and  $v_2$  are the two points mouth corners projections in the image plane respectively given by Eqs. 3 and 4.

$$v_1 = \frac{FX_1}{Z - Z_1} = \frac{F \times D \times \cos(\beta)}{Z - D \times \sin(\beta)} \quad (3)$$

and

$$v_2 = \frac{FX_1}{Z + Z_1} = \frac{F \times D \times \cos(\beta)}{Z + D \times \sin(\beta)} \quad (4)$$

We can calculate the distance  $Z$  according to  $v_1$  and  $v_2$

$$Z = \frac{D \times (F \times \cos(\beta) + v_1 \times \sin(\beta))}{v_1}, \quad Z = \frac{D \times (F \times \cos(\beta) - v_2 \times \sin(\beta))}{v_2} \quad (5)$$

Insofar  $Z$  is not known, Eq. 6 is advantageously replaced by Eq. 6.

$$\frac{D \times v_2 \times (F \times \cos(\beta) + v_1 \times \sin(\beta))}{v_1 \times v_2} + \frac{D \times v_1 \times (v_2 \times \sin(\beta) - F \times \cos(\beta))}{v_1 \times v_2} = 0 \quad (6)$$

When  $D$  is factoring, we obtain

$$\frac{D \times (S + C)}{v_1 \times v_2} = 0 \quad (7)$$

With  $S$  is represented as

$$S = v_2 \times (F \times \cos(\beta) + v_1 \times \sin(\beta)) \quad (8)$$

While  $C$  represents the following equation

$$C = v_1 \times (v_2 \times \sin(\beta) - F \times \cos(\beta)) \tag{9}$$

The distance  $D$  between the center of the mouth and one of its corners is strongly different from zero. On the other hand, the denominator of Eq. 7 is also different from zero, we can conclude that

$$S + C = 0 \tag{10}$$

By development of Eq. 10 gives the value of the angle  $\beta$  is calculated finally according to Eq. 11

$$\beta = \arctg\left(\frac{F \times (v_2 - v_1)}{-2v_2v_1}\right) \tag{11}$$

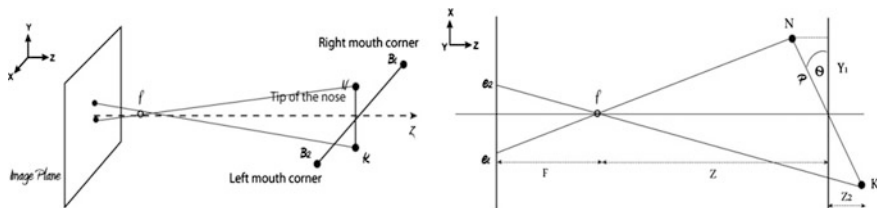
Equation 11 shows that we can calculate the yaw head driver without the influence of the distance between the driver and camera. This equation depends only on the classical camera calibration to determine the focal distance calculated only once.

### 2.4 Pitch ( $\Theta$ ) Estimation

The estimation of angle  $\Theta$ , depend on the tip of the nose, presented with point  $N$  and one of the mouth corners, let us take in our case the point  $B_1$  in 3D space. We suppose  $K$  the image of point  $N$  relative to  $(B_1B_2)$  axis. Let the distance  $P = ([NK]/2)$ .

The points  $e_1$  and  $e_2$  are the projection of the points  $N$  and  $K$  in the image plane respectively (Fig. 4). We obtain the following equation

$$e_1 = \frac{FY_1}{Z - Z_2} = \frac{F \times P \times \cos(\theta)}{Z - P \times \sin(\theta)} \quad \text{and} \quad e_2 = \frac{FY_1}{Z + Z_2} = \frac{F \times P \times \cos(\theta)}{Z + P \times \sin(\theta)} \tag{12}$$



**Fig. 4** Representation of the noise tip and its projection relative to  $(B_1B_2)$  axis in 3D coordinates (top) and their 2D projection in the image plane (bottom)

The distance  $Z$  is determined as follows

$$Z = \frac{P \times (F \times \cos(\theta) + e_1 \times \sin(\theta))}{e_1} \text{ and } Z = \frac{P \times (F \times \cos(\theta) - e_2 \times \sin(\theta))}{e_2} \quad (13)$$

Equation 13 give

$$\frac{P \times (Q + W)}{e_1 \times e_2} = 0 \quad (14)$$

then

$$Q = e_2 \times (F \times \cos(\theta) + e_1 \times \sin(\theta)) \quad (15)$$

and

$$W = e_1 \times (e_2 \times \sin(\theta) - F \times \cos(\theta)) \quad (16)$$

Since the distance  $P \neq 0$  and the product  $(e_1 \times e_2) \neq 0$ . It is possible to conclude

$$Q + W = 0 \quad (17)$$

After the development of Eq. 17 the angle  $\Theta$  is estimated according to the following equation

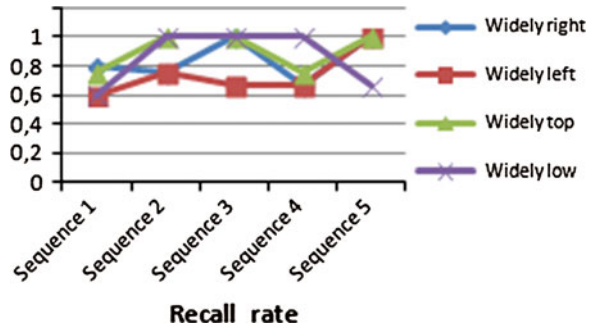
$$\theta = \arctg\left(\frac{F \times (e_2 - e_1)}{-2e_2e_1}\right) \quad (18)$$

### 3 Experimental Study

We describe in this section the experimental study. We used an RBG camera with  $640 \times 480$  of resolution and 30 fps. In order to evaluate our approach of head orientation, we tested our algorithm on five subjects.

We limit the degree of rotation angles value between  $[-35^\circ, 35^\circ]$  for all axes. Our work is not planned for the exact estimate of the 3D pose of installed subject. Indeed, we can tolerate in the error rates in order to know the face orientation. If rotation exceed  $[-20^\circ, 20^\circ]$  in the pitch and the yaw and  $[-15^\circ, 15^\circ]$  in the roll, we alert the driver of his drowsiness state. We divide these states in four categories: widely right, widely left, widely top and widely bottom. We note that there are states which are not detected. Figure 5 shows an average rate of recall which is equal to 84.2, 73.4, 90 and 85.2 % for four classes widely right, left, top and bottom respectively.

**Fig. 5** Curve recall rates for different head orientation



### 3.1 Conclusion and Future Work

This paper presents a new approach for 3D driver pose estimation. This method makes it possible to determine the states of drowsiness if the driver directs his head in four positions: widely right, left, top and bottom, compared to the vision angle. The choice of the two mouth corners and the nose tip is improved by their visibility in the majority of rotation angles. The average recall of orientation classes is near 85 %. Our method proves a success under various light conditions. On the other hand, it presents limits if one of the feature points is not detected correctly. These errors are explained by the noises or blurs effects in the recorded videos. We propose in our next work to resolve this kind of problem.

### References

1. Garcia I, Bronte S, Bergasa LM et al (2012) Vision-based drowsiness detector for real driving conditions. In: IEEE intelligent vehicles symposium, Spain
2. Horng W, Chen C, Chang Y (2004) Driver fatigue detection based on eye tracking and dynamic template matching. In: Proceeding of the IEEE international conference on networking sensing and control, New York, pp 7–12
3. Hongbiao M, Zehong Y, Yixu S, Peifa J (2008) A fast method for monitoring driver fatigue using monocular camera. In: Proceedings of the 11th joint conference on information sciences, China
4. Picot A, Caplier A, Charbonnier S (2009) Comparison between EOG and high frame rate camera for drowsiness detection. In: Proceedings of the IEEE workshop on applications of computer vision, USA
5. Akrouf B, Mahdi W, Ben hamadou A (2013) Drowsiness detection based on video analysis approach. In: Proceedings of the 8th international conference on computer vision theory and applications (VISAPP), Spain
6. Murray J, Andrew T, Robert C (2005) A new method for monitoring the drowsiness of drivers. In: Proceedings of the international conference on fatigue management in transportation operations, USA
7. Takuhiro O, Fumiya N, Takashi K (2008) Driver drowsiness detection focused on eyelid behavior. In: Proceedings of the 34th congress on science and technology of Thailand, Thailand

8. Lee JJSJ, Jung HG, Park KR, Kim J (2011) Vision-based method for detecting driver drowsiness and distraction in driver monitoring system. In: Proceedings of the optical engineering
9. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the computer vision and pattern recognition, USA
10. Harris C, Stephens M (1988) A combined corner and edge detector. In: Proceedings of the 4th Alvey vision conference



# Potentiality for Executing Hadoop Map Tasks on GPGPU via JNI

Bongen Gu, Dojin Choi and Yoonsik Kwak

**Abstract** Hadoop has good features for storing data, task distribution, and locality-aware scheduler. These features make Hadoop suitable to handle Big data. And GPGPU has the powerful computation performance comparable to super-computer. Hadoop tasks running on GPGPU will enhance the throughput and performance dramatically. However the interaction way between Hadoop and GPGPU is required. In this paper, we use JNI to interact between them, and write the experimental Hadoop program with JNI. From the experimental results, we show the potentiality GPGPU-enabled Hadoop via JNI.

**Keywords:** Hadoop · GPGPU · Map Task · Map/Reduce · CUDA · JNI · Cluster

## 1 Introduction

Hadoop [1] is suitable to handle Big Data. The reason of using Hadoop is that it has good features for handling Big Data as following: MapReduce programming model, HDFS, data locality-aware scheduler for multiple nodes on cluster. MapReduce is a programming model to simply express tasks which are concurrently executed for handling data, and developed by Google, Inc [2, 3]. HDFS (Hadoop Distributed File System) is a distributed file system on Hadoop cluster. It

---

B. Gu · D. Choi · Y. Kwak (✉)

Department of Computer Engineering, Korea National University of Transportation,  
ChungJu-Si 380-702, Chungbuk-Do, South Korea

e-mail: yskwak@ut.ac.kr

B. Gu

e-mail: bggoo@ut.ac.kr

D. Choi

e-mail: mycdj91@gmail.com

partitions files off, and stores each partition on multiple nodes redundantly for fault-tolerant data accessing service. Hadoop scheduler takes into account data locality to efficiently assign tasks to nodes [4], and reassigns abnormally terminated or delayed tasks to other nodes to prevent them from delaying job completion due to the abnormal execution state task. So Hadoop programmer can make his/her code without consideration of data storage, task/data assignment and migration, etc.

GPGPU is used to enhance computing throughput and performance [5]. Graphics Processing Units (GPU) is designed to process a huge number of graphics objects such as points, polygons, etc. To get enough graphics performance, GPU has many processing elements which can operate in parallel manner. These processing elements on recent GPU became to have additional functions suitable to perform general operations, and can be used to perform computation executed by CPU [6]. General Purpose computing on Graphics Processing Units (GPGPU) is using GPU to perform general computation handled by CPU. To efficiently handle a huge number of data in parallel manner, GPU has many processing elements with large number of register.

To make use of the computing power of GPU while MapReduce tasks are executed, there are many researches. Mars [7] is GPGPU-based MapReduce framework. Mars partitions data stored in local disk, and assigns it to threads executed on GPGPUs in parallel manner. This framework can enhance the computing throughput and performance. However it cannot handle data whose size is larger than the available capacity of local disk. And it cannot be operated on multiple nodes. DisMaRC [8] is GPGPU-based MapReduce framework for multiple nodes. However it cannot handle data whose size is larger than the available disk capacity of a node because it does not support distributed file system. And it does not have any mechanism to resolve fault state generated by GPGPU nodes due to hardware failure and data transferring problem, etc.

We think that Hadoop is suitable to resolve the problems previously described. To the best of our knowledge, the meaningful result about GPGPU-enabled Hadoop has not been reported yet. In this paper, we show the potentiality that GPGPU can be used by Hadoop framework. Our approach to show potentiality for executing Hadoop Map tasks on GPGPU is using Java Native Interface (JNI). JNI is the interaction mechanism between Java and other programming language like C. Java is the basic language for Hadoop MapReduce program. However Java is not the suitable language for programming GPGPU until now. Therefore it is necessary to use JNI for interaction between Hadoop MapReduce and GPGPU code.

This paper is organized as follows: [Section 2](#) describes interaction between Hadoop Mapper and GPGPU code, [Sect. 3](#) describes implementations of our approach. [Section 4](#) concludes and describes further studies.

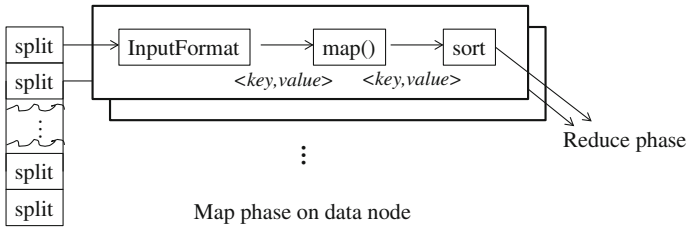


Fig. 1 Simplified data flow of Hadoop map phase

## 2 Interaction Between Mapper and GPGPU Code via JNI

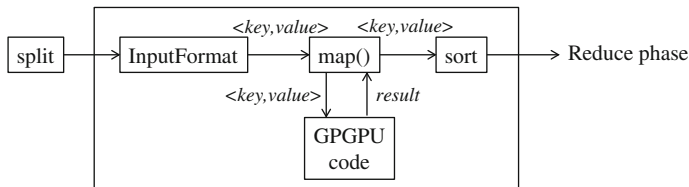
In HDFS, data file is partitioned into blocks whose default size is 64 MB, and blocks are distributed among nodes. One Map task is scheduled for each block by Hadoop scheduler. To maximize parallelism for handling big data, Hadoop assigns data blocks to all available Map tasks on cluster.

Figure 1 shows the data flow for Map task in Hadoop. Block assigned to a task is called ‘split’. Hadoop schedules a Map task executed on data node for handling a split in HDFS. Map task loads split and partitions it into  $\langle key,value \rangle$ -pair records. Each record fetched from split is passed to  $map()$  method to handle it.  $Map()$  method makes  $\langle key,value \rangle$  pair as a result. And then, this pair is passed to sort step to sort all  $\langle key,value \rangle$  pairs generated by Map task, and all sorted  $\langle key,value \rangle$  pairs are passed to reduce phase.

Our approach in this paper to use GPGPU for GPGPU-enabled Hadoop is changing the function of  $map()$  in Map phase. The function of  $map()$  in original Hadoop handles data expressed as  $\langle key,value \rangle$  pair. However  $map()$  in our approach transfers data to GPGPU, initiates GPGPU code, fetches results from GPGPU, and converts the result into  $\langle key,value \rangle$  format. Figure 2 shows the simplified data flow of GPGPU-enabled Hadoop Map phase.

Normally Hadoop Map/Reduce task is written in Java. Hadoop also has mechanisms, such as stream and pipeline, to write Map task with other language. But Java is standard language for programming Hadoop task. However C/C++ is currently standard language for GPGPU, and the famous GPGPU programming framework such as CUDA and OpenCL is based on C/C++. Therefore we use JNI for interaction between  $map()$  and GPGPU code. The JNI enables Java code running in JVM to call, and to be called by, native program written in other languages like C. JNI is normally used to call a hardware and operating system function written in a native language from Java application.

The procedure for executing Hadoop Map task on GPGPU is as following: At first,  $map()$  is called by Hadoop framework with  $\langle key,value \rangle$  record.  $Map()$  transfers record to GPGPU via a system bus in host. The record transferred by  $map()$  is stored in a memory for GPGPU. NVIDIA calls this memory as *global memory*. And then  $map()$  initiates GPGPU code, and waits until the execution of GPGPU code is complete. GPGPU code processes data stored in its memory, and



**Fig. 2** Simplified data flow of GPGPU-enabled Hadoop map phase

stores the result in the memory. Of course, the GPGPU code is previously prepared by programmer to process the record. When GPGPU code is done, *map()* fetches the result from the memory of GPGPU, and make it *<key,value>* record for the following step such as sort, merge, etc.

Using JNI to execute Hadoop task on GPGPU has the advantages as following: In Hadoop cluster consisted of multiple GPGPU-enabled node, the computation throughput and performance are dramatically enhanced without the consideration about how to store data, how to distribute tasks, and how to recover the faulty task. The feature of the throughput and performance enhancement is due to GPGPU. And the feature of programmability without the consideration about data store, task distribution, and fault tolerant task processing is due to Hadoop. So Hadoop application developers only focus on his/her algorithm to handle record, and get the maximized throughput and performance via GPGPU computing power at the same time.

### 3 Implementation of Hadoop Map Task on GPGPU via JNI

To show the potentiality for executing Hadoop Map task on GPGPU via JNI, we configure the small Hadoop cluster, and implement very simple experimental Hadoop program. The configuration of Hadoop cluster for our approach in this paper is as Table 1. The Hadoop cluster consists of three nodes, and two GPGPU add-on boards are installed in two data nodes.

**Table 1** Configuration of Hadoop cluster to show the potentiality of our approach

The number of nodes: 3	Name node	1 (GPGPU is not installed)	
		CPU	Intel Xeon
		Memory	4G
	Data node	2 (GPGPU is installed in all data node)	
	CPU	Intel Core2Duo/AMD Phenom II x6	
	Memory	2 GB/8GM	
OS	CentOS 6.3		
GPGPU	Nvidia GeForce GTX 670		
Hadoop	1.0.1		

To write the simple experimental Hadoop program executed on GPGPU, we use CUDA developed by Nvidia. The Compute Unified Device Architecture (CUDA) is a parallel computing platform and programming model. It enables dramatic increases in computing performance by using the computing power of GPU. The experimental program simply adds two numbers in each record. To do this, we create about 110 MB data file which consists of about seventeen million records. The 110 MB data file is partitioned into two splits by HDFS, and assigned two Map tasks.

The execution time of this experimental program is about 110 min even though GPGPUs installed in nodes are up-to-date devices, and each node has the good performance shown in other experiments. It is very long execution time in our cluster configuration. The reason of very long execution time is as follows: As shown in Fig. 2, *map()* is called once for each record in split. And each split averagely has about 8.5 million records. So *map()* is called about 8.5 million times, and for each call repeatedly executes the processing steps: transferring record to GPGPU, initiating GPGPU code, fetching the result. This repetition is very big overhead.

However the result due to the overhead cannot obscure our approach to use JNI for executing Hadoop task on GPGPU. Using JNI enables GPGPU to execute Hadoop Map tasks. And this shows the potentiality for executing Hadoop task on GPGPU via JNI. We think that our result is the first step for implementing GPGPU-enabled Hadoop.

## 4 Conclusion

Hadoop is known to be suitable platform for handling Big data because it has good features such as simple MapReduce programming model, distributed file system, and data locality-aware scheduling policy, etc. Therefore many researchers study on Hadoop and application fields. And recently many researchers are also interested in GPGPU because it has a powerful computation power comparable to supercomputer.

If Hadoop application can use the computation power of GPGPU, the throughput and performance will be dramatically enhanced. To the best of our knowledge, the GPGPU-enabled Hadoop is not reported yet. Some researchers reported GPGPU-enabled MapReduce frameworks. But they didn't target to Hadoop. To realize the GPGPU-enabled Hadoop, the interaction mechanism between Hadoop tasks and GPGPU code is required.

In this paper, we used the JNI to interact between Hadoop tasks and GPGPU code via JNI. And we experimentally implemented Hadoop Map tasks running on GPGPU. The execution time is very long due to the previous described reasons. However we showed the potentiality for executing Hadoop tasks on GPGPU via JNI.

In the future, we will revise Hadoop framework. In the current version of Hadoop, *map()* is called for each record in split. This strategy is good for normal Hadoop cluster. But this makes very big overhead in Hadoop cluster with GPGPU. So the additional *map()* calling strategy is required that *map()* is called for one split or splits in Hadoop cluster with GPGPU.

**Acknowledgments** This research was supported by a grant from the Academic Research Program of Chungju National University in 2010. And this research was partially supported by Technology Development Program for ‘Bio-Industry Technology Development’, Ministry for Food, Agriculture, Forestry and Fisheries, Republic of Korea.

## References

1. Tom W (2011) Hadoop: The definitive guide. O’relilly: 1–13
2. Dean J, Ghemawat J (2004) MapReduce: Simplified data processing on large cluster. In: ‘04: Sixth symposium on operating system design and implements (OSDI ’04), San Francisco, pp 137–150
3. Jorda P, David C, Yolanda B, Jordi T, Eduard A, Malgorzata S (2010) Performance-Driven task co-scheduling for MapReduce environments. In: IEEE network operations and management symposium (NOMS), pp 373–380
4. Matei Z, Dhruba B, Joydeep SS, Khaled E, Scott S, Ion S (2010) Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. In: Proceedings of the 5th European conference on computer systems (EuroSys’ 10), New York, pp 265–278
5. GPGPU <http://en.wikipedia.org/wiki/GPGPU>
6. Cayrel PL, Gerhard H, Michael S (2011) GPU implementation of the Keccak Hash function family. IJSA 5:123–132
7. He B, Fang W, Govindaraiu N, Luo Q, Yang T (2008) Mars: a MapReduce framework on graphics processors. In: PACT ’08: Proceedings of the 17th international conference on Parallel architectures and compilation techniques, New York, pp 260–269
8. Mooley A, Murthy K, Singh H (2008) DisMaRC: A distributed map reduce framework on CUDA. TechRep, The University of Texas, Austin, pp 65–66

# An Adaptive Intelligent Recommendation Scheme for Smart Learning Contents Management Systems

Do-Eun Cho, Sang-Soo Yeo and Si Jung Kim

**Abstract** This study aims to provide personalized contents recommendation services depending on a learner's learning stage and learning level in the learning management system using open courses. The intelligent recommendation system proposed in this study selects similar neighboring groups by performing user-based collaborative filtering process and recommends phased learning contents by using prior knowledge information between contents and considering the relevance and levels of learning contents. The proposed learning contents recommendation is applied flexibly according to a user's learning situation and situation-specific contents recommendation link is created by performing the intelligent learning process of recommendation system. This service allows a variety of industrial classification learners using open course to effectively choose more accurate curriculum.

**Keywords** Personalization · Recommendation system · Collaborative filtering · E-learning · Learner's preference

---

D.-E. Cho

Innovation Center for Engineering Education, Mokwon University, Daejeon, Korea  
e-mail: decho@mokwon.ac.kr

S.-S. Yeo

Division of Computer Engineering, Mokwon University, Daejeon, Korea  
e-mail: sangsooyeo@gmail.com

S. J. Kim (✉)

Center for Teaching and Learning, Hannam University, Daejeon, Korea  
e-mail: sjkim6183@gmail.com

# 1 Introduction

Recently entering the era of lifelong learning, E-learning, which is the new education paradigm, provides a variety of learning contents. The advantages of E-learning are to promote learner-centered education and enable customized education for individual learners. Therefore, a variety of lifelong learning programs or learning contents using it are provided. Generally, most of E-learning is conducted based on web and learning is carried out by choosing materials posted by instructors and materials posted by learners. In these circumstances, it is very hard for learners themselves to select necessary matters from a variety of learning contents and determine the learning process. Therefore, personalization strategy is needed in order for learners to obtain academic efficiency and learning effect [1–3]. For this personalization strategy, the recommendation system identifying learners' learning objectives and automatically filtering differentiated information by individual is required [4].

As representative recommendation research classification, there is a Content-based recommendation classification method by considering a user's previous preferences first and recommending first items. The content-based classification method is the method using the fact that the preferences of the past are highly likely to choose the future. In addition, there are other methods such as Demographic-based recommendation method recommending items by referring to the use form of learners showing similar patterns with using demographic information, Rule-based recommendation method which is the recommendation method according to several rules with existing data and Collaborative filtering recommendation method using approach value of groups with similar contents access data [5]. Recently, these recommendation methods are used variously for movies, music, video and other services.

This paper attempts to present contents recommendation services for providing personalized contents depending on the learning step and learning level of a learner in the learning management system using open courses. The method proposed in this paper is to gather individual learning information first based on learning information performed by learners in the learning management system. And then, it recommends learning contents deemed to be best suited for learners by using prior knowledge information between contents and considering the relevance and levels of learning contents. This paper is organized as follows. First, [Sect. 2](#) learns about the existing recommendation method of the recommendation system and E-learning system and [Sect. 3](#) describes the proposed intelligent recommendation system and service model. [Section 4](#) makes conclusions.



## **2 Related Work**

### ***2.1 Recommendation System***

#### **2.1.1 Content-Based Recommendation Schemes**

Content-based recommendation is based on information retrieval and recommendations are made by comparing the user profile and the contents to improve performance. Information about the user's tastes, preferences, need is included in the user profile. Profile information can be obtained in the explicit way by asking questions to the user or in the implicit way by observing the user's behaviors. The content-based recommendation has the characteristics that the user's attention on specific areas can be reflected and recommendation is available when new areas of interest occur.

However, it is difficult to independently use it in multimedia information such as music, photos, pictures which are hard to define the characteristics of contents and the user's potential interests cannot be indicated by solely relying on the user profile. Also, in order to improve the accuracy of recommendations, it is important to accurately extract the characteristics of contents well reflecting users' intention. Therefore, sufficient prior information about a user such as contents preferred by a user in the past, feedback etc. is required.

#### **2.1.2 Collaborative Filtering Recommendation Schemes**

Collaborative filtering recommendation methods used in the recommendation system are classified into user-based collaborative filtering method and item-based collaborative filtering method [6]. The user-based collaborative filtering method is the method to recommend contents that a particular user may prefer based on contents evaluated by other users with similar preferences by measuring the similarity between users. The techniques to select neighbors with similar preferences and any particular user based on the association between users include clustering, best N-neighbor, Bayesian network etc. The item-based collaborative filtering method is the method to recommend by predicting which items a specific user prefers by measuring the similarity, that is, similarity between existing items that a user entered preferences and items to be recommended. If using the collaborative filtering method, when sufficient preference information of users showing a similar tendency, contents can be recommended actively to those who accessed to the system for the first time. Eventually, in case of the collaborative recommendation method, if the number of users who have the similar preferences with them is less, the selection probability of the recommended list is lowered. Also, the disadvantage is that the evaluation on specific contents is not made, the system cannot be applied.

## ***2.2 E-Learning Recommendation System***

The methods of E-learning recommendation system currently ongoing include the recommendation system using contents-based collaborative filtering, recommendation system using user based collaborative filtering and automatic recommendation system using hybrid filtering [7, 8]. The recommendation system using contents-based collaborative filtering is the method used in currently active online education sites and is the method to recommend the courses of similar themes based on the courses that learners have taken in the past. Like this, if other courses similar to those that a learner has taken are uploaded newly, recommended courses can be offered easily. However, in case of a new learner, recommendation is impossible because any information does not exist indicating in which course he/she is interested. First, the recommendation system using user-based collaborative filtering calculates similarity with neighbor learners with the same idea by using the Pearson correlation coefficient based on what a learner evaluated after listening to a course. At this time, courses are recommended by extracting the list that the learner did not take from the courses taken by neighbor learners with high similarity value. Like this, if recommended by similar neighbor learners, the recommendation of unnecessary courses is reduced so the reliability of learners can be improved. However, its disadvantage is that if a learner's learning activities are not active, it is not easy to configure neighbor learners and therefore, correct recommendation is difficult. Hybrid recommendation system creates the learner profile by using the log of learners and identifies neighbor learners with similar interests through collaborative filtering. And it recommends a new list by text-mining courses and applying content-based filtering method to created contents profile. By mixing the list recommended through these two filtration methods, it finally recommends top N lists to learners.

## **3 Intelligent Learning Content Recommendation System Design and Service Model**

Intelligent learning recommendation system proposed in this study performs user-based collaborative filtering process and selects similar neighboring groups and then recommends phased contents according to levels by considering prior knowledge information between contents. Learning content recommendation is applied flexibly depending on the learner's learning situation and creates context-sensitive contents recommendation links by performing intelligent learning process of the recommendation system.

### 3.1 System Structure for Intelligent Learning Content Recommendation

In the recommendation system proposed in this study, learners are classified into learners taking learning and pre-learners who already took learning. The manager enters the contents profile for contents registered in the system.

The contents profile includes learning difficulty and prior knowledge information. Learning contents are saved in knowledge save location in the form of complex knowledge considering basic information as well as information occurring during learning etc. Through information gathering handler, information entered extracts contents use frequency and contents preference information of similar learners. By using values generated in information gathering handler, the candidate content recommendation engine provides the recommendation list by group selected by learners. A candidate content recommendation list is finally recommended to learners through weighting and ranking. The following Fig. 1 shows the overall configuration of proposed intelligent learning-based content recommendation system.

### 3.2 Information Creation and DB Configuration

When using the system for the first time, a learner must enter learner’s personal profile information such as his/her interest parts, log-in information etc. An manager performs the contents registration process for users who set content

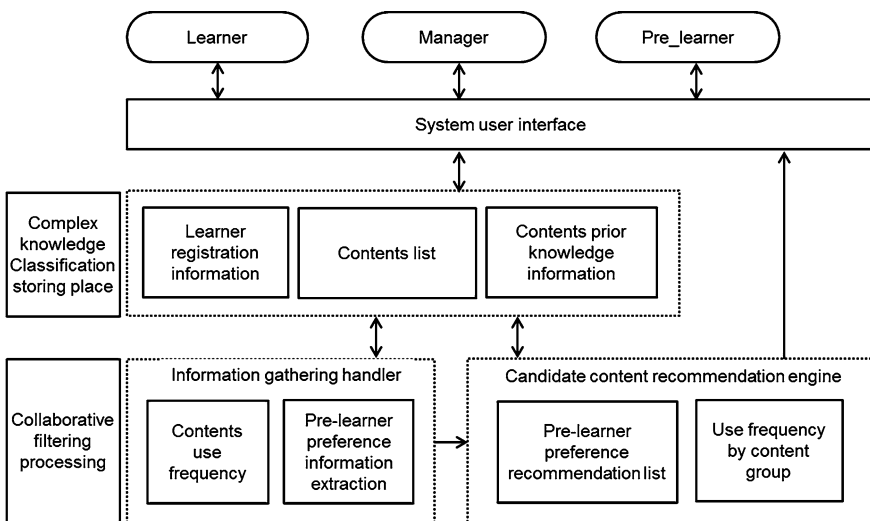
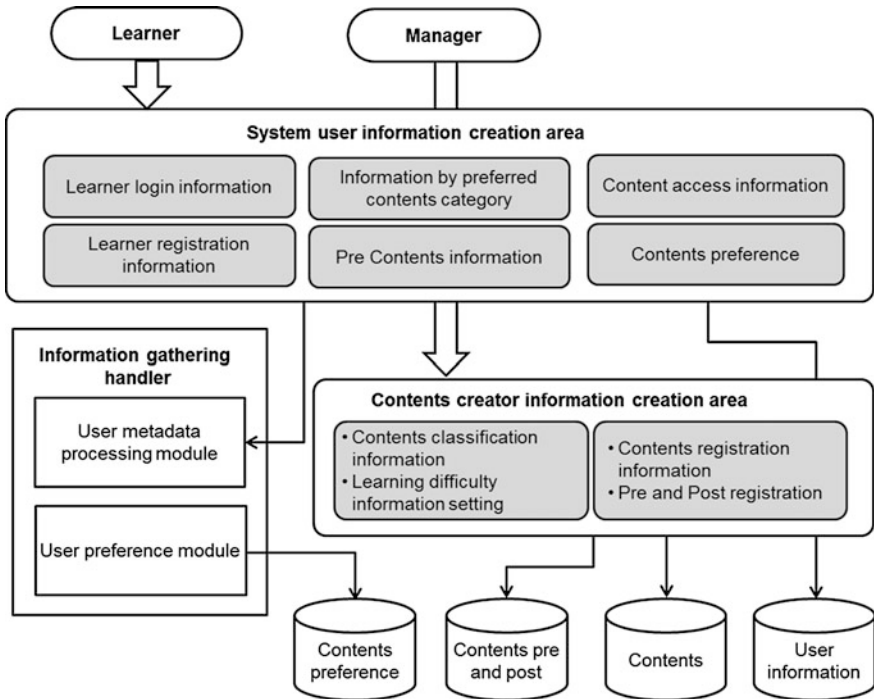


Fig. 1 Recommendation system configuration

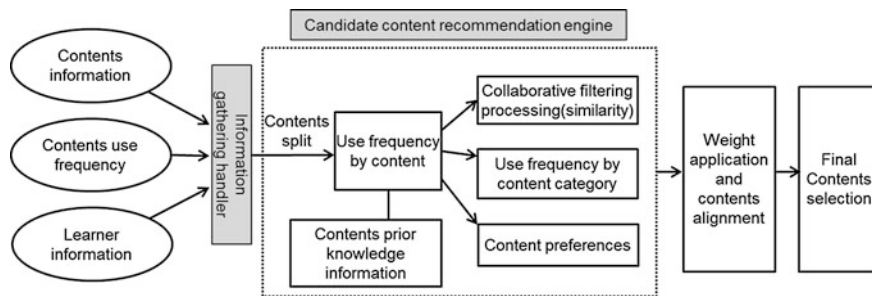


**Fig. 2** Recommendation system DB configuration

provision. In this process, the manager classifies the contents into by category and each content is saved in content DB by setting the difficulty rating (high, medium, low). Also, the manager creates pre/post course links of all contents by specifying pre-course for learning contents by the course. This information is saved in pre/post DB. And learners who completed learning and are classified into pre-learners perform the process of entering course information on each learning content and information on content satisfaction after learning. This information is used for creating a recommendation list corresponding to learning content category of next learners. The following Fig. 2 is the configuration on information creation of each DB and configuration items of the proposed system. Pre contents information receives values entered when registering learning contents based on association degree between learning topics and learning contents of learners.

### 3.3 Candidate Content Recommendation Engine

Contents attribute data on initial data are collected through information gathering handler and analyzed and the value of created metadata is passed to the candidate content recommendation engine. Based on prior knowledge information of



**Fig. 3** Configuration of candidate content recommendation engine

learning level information and learning contents of a learner, the candidate content recommendation engine derives best n-neighborhood and derives recommendation lists by calculating learners and similarity. Also, by applying weight according to contents registration time, it recommends recently registered contents first. Figure 3 shows candidate contents recommendation method, candidate contents recommendation process and necessary elements.

The similarity of learners is found by calculating Pearson correlation coefficient [9]. And by using the evaluation value and similarity within Best n-neighborhood, the evaluation value of the contents that learners did not learn is predicted. At this time, by applying the mean values and similarity of the learning results of each learner as weight, evaluation predictive value for items of the learners should be calculated [10].

The calculated candidate contents list performs the functions of applying weight to select contents best meeting learners' preferences and selecting final recommendation contents through ranking. For weight,  $w$ , which is the weight value according to contents registration time, is applied. For weight, the method of subtracting 0.1 depending on year is used. Recently registered contents will have 1 and older contents 0.1. As Top-N technique, it creates and provides top N recommendation lists and learners study at least one learning content.

### 3.4 Content Recommendation Service Model

In order to gain access to the system, each learner basically performs the registration process creating his/her registration information. The system provides services by classifying accessors into learners and contents managers. A learner enters basic information at the initial access and searches lists of 1st contents category of learning process that he/she wants. After checking the results, he/she searches contents lists with corresponding difficulty and receives the results. And then, to search for recommendation lists, the system determines whether there are students who take the contents and if so, creates recommendation lists by using

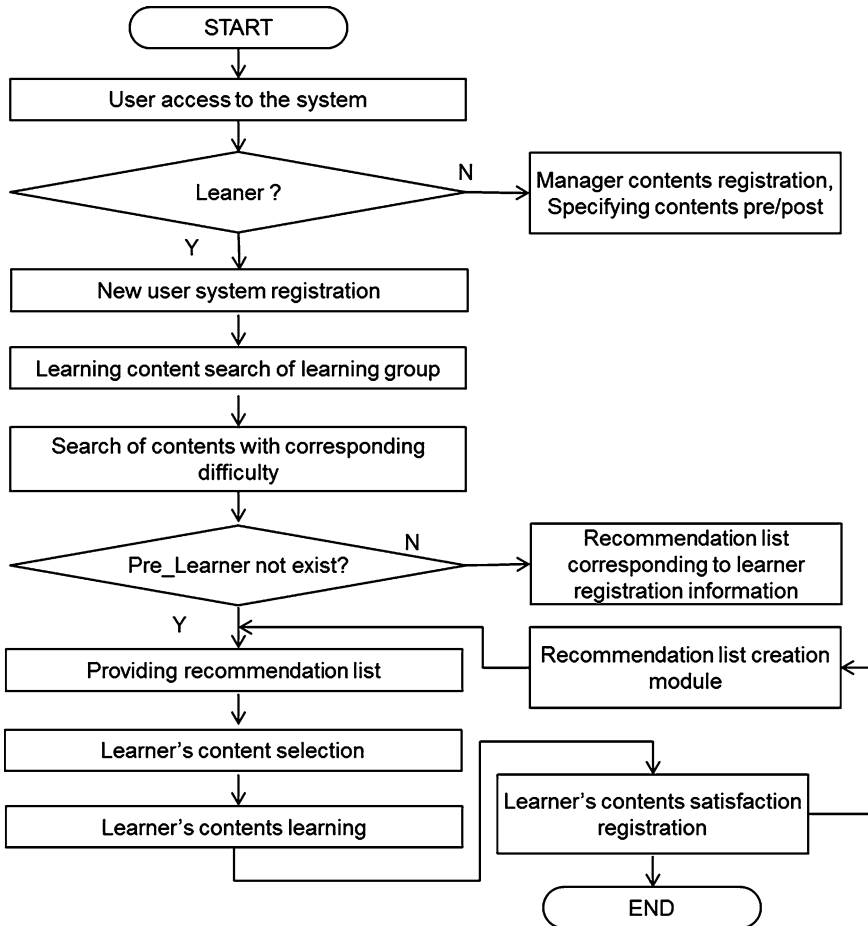


Fig. 4 Proposed system flowchart

input data for pre-learners and if there are no pre-learners, it provides recommendation lists based on contents registration information and learner basic registration information. Figure 4 shows the flow of the system service provision.

## 4 Conclusion

This study aims to provide personalized contents recommendation services depending on a learner’s learning stage and learning level in the learning management system using open courses. Intelligent learning recommendation system proposed in this study performs user-based collaborative filtering process and selects similar neighboring groups and then recommends phased contents

according to levels by considering prior knowledge information between contents. Also, by applying weight according to contents registration time, it recommends latest contents first. Also, even when a learner accesses for the first time, the contents are recommended by each interest part category by using user profile information provided at the beginning of registration. In the system, learners receive next learning contents by using prior knowledge information set by the manager and receive recommendation list top information on the corresponding contents category. Currently, a variety of contents are provided in many learning management systems but a lot of content information is required in order for learners to select contents appropriate for them. Also, even if a lot of information is provided, it is not easy for learners themselves to select contents appropriate for learning progress. The existing various recommendation systems use the method of recommending based on how many students chose different contents or simply based on learners' marks.

This study provides recommendation lists appropriate for the personal environment by using a variety of complex knowledge such as prior knowledge information of learning contents, selection frequency of pre-learners and learner preferences without learners' contents selection information in the open course learning management system of various learning areas commonly utilized in the lifelong learning environment. System implementation of contents recommendation service and performance and use assessment of models proposed as future research project will be carried out.

**Acknowledgments** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0014394).

## References

1. Cho D-E, Kim S-J, Kwak Y (2011) A study of personalized contents recommendation method based on user preference learning. *J Korean Inst Inf Technol* 9(9):229–235
2. Schafer et al (1999) Recommender system in E-Commerce. In: Proceedings of the ACM E-Commerce 1999 conference
3. Arazy O, Kumar N, Shapira B (2010) A theory-driven design framework for social recommender systems. *J Assoc Inf Syst* 11(9):455–490
4. Liang T, Lai H, Ku Y (2007) Personalized content recommendation and user satisfaction: theoretical synthesis and empirical findings. *J Manag Inf Syst* 23(3):45–70
5. Burke R (2002) Hybrid recommender systems: survey and experiments. *User Model User Adapt Interact* 12(4):331–370
6. Konstan J, Miller B, Maltz D, Herlocker J, Gordon K, Riedl J (1997) GroupLens: applying collaborative filtering to usenet news. *Commun ACM* 40(3):77–87
7. Kang Y-J, Sun C-Y, Park K-S (2010) A Study of IPTV-VOD program recommendation system using hybrid filtering. *J Institute Electron Eng Korea* 47(4):9–19
8. Inay Ha, Song G-S, Kim H-N et al (2009) Collaborative recommendation of online video lectures in e-learning system. *J Korean Soc Comput Inf* 9(14):87–94

9. Herlocker JJ, Konstan A, Borchers R et al (1999) An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22th ACM SIGIR conference on research and development in information retrieval, pp 230–237
10. Sarwar B, Karypis G, Konstan J et al (2000) Item-based collaborative filtering recommendation algorithm. WWW10, pp 285–295



**Part II**  
**Ubiquitous and Pervasive Computing**

# An Evolutionary Path-Based Analysis of Social Experience Design

Toshihiko Yamakami

**Abstract** Service engineering is quickly moving forward to social services. Social service engineering is one of the most promising arenas of service engineering in the 2010s. The term social experience represents the analogy of user experience in a social service context. The author proposes an evolutionary path model of social experience design in order to highlight the design principles of social experience design.

## 1 Introduction

Social interaction is difficult to design, manage and measure. These difficulties prevent social service engineering from being analyzed in a scientific manner.

The term user experience gained visibility in the 1990s as the computing power enabled human-centric affective user interface design with rich-media capabilities. This transition from user interface to user experience provides the basic insight for this research.

The concept of social experience design was proposed in a previous paper by the author. In this paper, the author examines the social experience using the transition paths and changes invoked by each transition.

The concept of social experience design was coined in order to provide an umbrella concept to guide social service designs. In this paper, the author extends the concept of social experience design using a transition view model from user interface to social experience.

---

T. Yamakami (✉)

ACCESS, Software Solution, 1-10-2 Nakase, Mihama-ku, 261-0023 Chiba-shi, JAPAN  
<http://www.access-company.com/>

## 2 Backgrounds

The aim of this research is to identify the unique characteristics of, and guidelines for social experience design.

The term User Experience Design was coined by Don Norman while he was Vice President of the Advanced Technology Group at Apple Computer in the 1990s. The term User Experience has been impacting user interaction design for two decades with the departure from computer–human interface toward high-level interaction design. He also discussed emotional design and mentioned that emotion is a necessary part of life, affecting how we feel, how we behave and think. He mentioned that usability and pleasure should go hand in hand.

Grudin presented eight challenges for groupware from social dynamics [1]. Social aspects of information technology research focused organizational ones.

The originality of this paper lies in the examination of key factors of social experience design using evolutionary-path-based analysis.

## 3 Definition and Method

### 3.1 Definition

The definitions of user interface, user experience, social interface, and social experience, are depicted in Table 1. In these definitions, the social interface is similar to the multi-user interface in this paper. The definition of social experience is coined by the author. Examples of each term are depicted in Table 2.

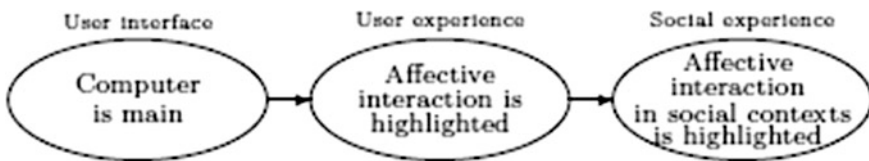
In the early stages, computers were precious assets and their ability to deal with human interactions was limited. This leads to a design where human beings were limited to following the computer-side restrictions. The drop in computer hardware prices and the increase of computing power brought an increase in the human-

**Table 1** Definitions

Term	Description
User interface	Design of human–machine interaction where interaction between humans and machines takes place. It aims at effective operation and control of the machine with usable feedback from the machine
User experience	Design of how a person feels about using a product, system or service. It highlights valuable aspects of human–computer interaction and product ownership
Social interface	User–computer interface that deals with human–human interactions. User–computer interface that deals with Multi-user interactions. (This is the definition used in this paper. Social interface may represent human-like computer interface in other contexts)
Social experience	Design of the way a person feels about other humans through computer-user interface

**Table 2** Examples of each term

Term	Examples
User interface	Artifacts to provide interfaces to a computer, network, or system. Artifacts that consist of each modality computer–human interaction. Menus, icons, command sequences, command parameters, and so on
User experience	Total experience aspect that governs multiple aspects of user interface. Holistic aspect of space-dimensional and time-dimensional integration of multiple components of user interface. For example, creating architecture or interaction models that affect the user’s perception of computer, device or system. It deals with the improvement of perception of total systems in order to satisfy both technical needs and business needs
Social interface	User interface that deals with multi-user factors. User interface to deal with roles, role-taking, conflict-resolution, collective culture, social awareness, and so on
Social experience	Total experience aspects that deal with different social roles with a single user interface



**Fig. 1** Shift of key concepts of design through transitions

interaction capabilities of computers. This provided a challenge to the legacy concept of user interface with the implication that humans should have to follow computer ways. The word “user experience” was coined to provide the best experience for users in terms of human–computer interaction.

Improved network capabilities brought opportunities for multi-user interactions. Multi-user interactions were encumbered with conflicts between human and computers, as well as conflicts among humans. Early multi-user interfaces needed to address control arbitration and other exclusive control matters.

Further advances of the Internet brought the new infrastructure of world-scale real-time human interactions. With this transition, we have to re-focus on the importance of user experience in social contexts. The shifts in key concepts of design through transitions are shown in Fig. 1.

### 3.2 Method

The research method is as follows:

- identify transition paths towards social experience,
- for each path, the characteristics of transition are examined,
- using the transition semantics, the key aspects of social experience are parsed.

### 4 Evolutionary Path-Based Analysis

An evolutionary path model of social experience design is depicted in Fig. 2.

The analysis of Path 1 is depicted in Table 3.

From this analysis, the transition from user interface to social interface takes place where the entity to be designed accepts simultaneous operations from multiple users. Then, the transition from social interface to social experience takes place where social emotion is invoked or where social relationship is built up over a span of time.

From this consideration, the transition to the final social experience takes place where a time-dimensional long-term approach is taken or where social semantics such as social emotion is taken in interaction models.

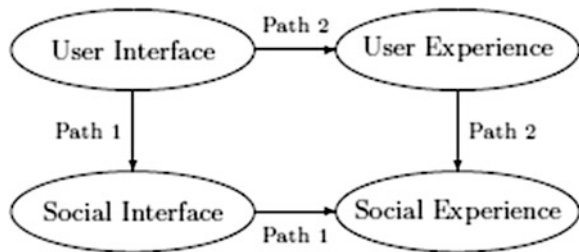
The important factors in social experience in path 1 are depicted in Table 4.

A typical example of this transition is the structure of knowledge-sharing as depicted in Fig. 3.

There are two layers in the knowledge sharing structure. One is an inner core layer, where core members exchange their expert knowledge. In this layer, information sharing is bidirectional. Experts actively engage in sharing the knowledge of other experts.

The other is an outer follower layer, where follower members connect to a core member. A follower member actively makes use of the knowledge of an expert. In this layer, information mainly flows from an expert to followers.

**Fig. 2** Design path model of social experience



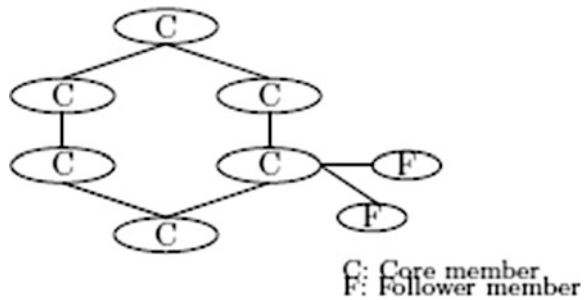
**Table 3** Path 1 analysis

Path	Changes
From user interface to social interface	Architecture shifts toward for multi-user interface, interpersonal interaction, and mutual exclusion. Interaction model deals with multi-use interference, synchronization, arbitration, role-taking and conflict resolution
From social interface to social experience	Design deals with socially-leveraged experience. It deals with satisfaction of role distribution. Simultaneous satisfaction with different roles. It aims at building collective experience with social causes. It covers collective cultural factors in satisfaction. It deals with collective satisfaction with a variety of skills and experience

**Table 4** Important social experience factors in path 1

Factor	Description
Time dimensional factors	Long-term relationships. High-level social roles
Fits with business goals	Satisfaction with overall experience. Maintaining high level satisfaction in the social contexts
Integrating social interface into positive social experience	Creating positive social experience with social rewards with extending the social interface

**Fig. 3** Knowledge sharing structure



Generally, knowledge sharing system do not pay attention to the values of information depending upon which layer a user belongs to. Social experience design deals with this social structure which stands for a relatively long-term.

The path 2 analysis is depicted in Table 5.

In this path, the transition from user interface to user experience takes place where high level requirements of user satisfaction or business goals are created.

Then, the transition from user experience to social experience takes place where the user satisfaction or business goals are tightly bound to multi-user interaction.

The important factors in social experience in path 2 are depicted in Table 6.

A typical example of this takes place in the beginner-veteran collaboration in a social game as depicted in Fig. 4.

In a mobile social game, veterans have knowledge, experience, have accumulated game points, and have paid premium items. It is difficult for a beginner to match these veterans. Mobile social game design has to deal with generating

**Table 5** Path 2 analysis

Path	Changes
From user interface to user experience	Architecture and interaction models to deal with higher level satisfaction and affective aspects rather than individual aspects of computer-human interaction
From user experience to social experience	Experience shifts from computer-human interaction to interpersonal interaction. Satisfaction has origins in social interactions such as support, gifts, thanks, greetings, and so on. It deals with collective experience of achievement and shared excitement. It also deals with bidirectional interactions, such as reciprocity and mutual education

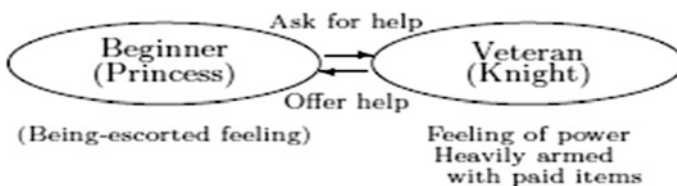
**Table 6** Important social experience factors in path 2

Factor	Description
Satisfaction portfolio	Different satisfaction for different types of users in a shared context
Emotion engineering	Engineering of social-context-based positive emotions. Creating rewarding social experiences (group achievement, collaboration, reciprocal support, being acknowledged as a member, greetings, and so on)
Improving user experience in social contexts	Upgrading user experience into socially-positive experience
Short-term time management	Human beings are asynchronous except in cases of strict time-keeping and real-time meeting. This asynchronicity makes people gradually accept asynchronous triggers

satisfaction among users with multiple skill levels. When we examine the relationship between helping and being helped in battles in mobile social games, it becomes clear that each battle brings different kinds of satisfaction and affective factors. A beginner receives support from a veteran. One feels like a princess, receiving protection and services from a guardian. A veteran helps a beginner, with the feeling of knight. One also exercises skills and premium items, as feeling like a “knight in shining armor.” The same game scene serves as different affective aspects depending on user experience and skills. It is a typical example of social experience design.

The term experience is broad. In order to further examine the best practices of social experience design, it is necessary to parse multiple layers of social experience design. The experience can be used in the total perception of long-term use of social systems, or in the individual socially-leveraged emotion. The detailed analysis of this spectrum remains for further research.

Considering the above analysis, the author presents the design components of social experience design, as depicted in Table 7.



**Fig. 4** Beginner-veteran collaboration in a social game

**Table 7** Components of social experience design

Time-scale	Component	Description
Long-term	Total experience	Total design of each interface with integrated delivery of social experience
	Culture	Design fits with social interaction with culture of corresponding group
Mid-term	Design of mid-term social experience	Creating social experience over a span of time. Experience utilizing reciprocity principle
	Awareness experience	Creating socially-leveraged awareness, we-feeling
Short-term	Emotion design	Creating positive social experience, greetings, thanks, gifts, social acknowledgement, feeling as a member, social achievement, and so on
	Shared emotion	Success of group action, social achievement

## 5 Discussion

### 5.1 Advantages of the Proposed Approach

Social experience design is a natural extension of user experience design in the domain of social service engineering. Social service engineering increases in importance according to the increasing stay time of people in social services. The natural extension from single-user to multi-user is not easy because it involves multi-faceted challenges that take a long time to identify and resolve.

There are three approaches that identify the unique characteristics of social experience design as depicted in Table 8.

The author takes the transition analysis approach. This approach focuses on what takes place at the time of transition. The author proposes an evolutionary path model where two evolutionary paths toward social experience design are identified.

**Table 8** Approaches that identify the uniqueness of social experience design

Approach	Description
Model-based approach	Highlighting the high level aspects of social factors, roles, social task, role-taking, conflicts, coordination, collective culture, and so on
Bottom-up approach	Collect examples of social experience in the socially-connected system examples
Transition analysis	Observe transitions from single user experience to multi-user experience from the design perspective to collect the distinguishing factors that separate single-user experience and social experience



## 5.2 *Limitations*

This research is a qualitative study. The quantitative measures for identifying multiple aspects of social experience design discussed in this paper remain for further study.

User acceptance of social experience design in the real world environment is beyond the scope of this paper. Quantitative analysis of performance and user satisfaction of social experience design requires future research. The concrete design methodology of social experience design is beyond the scope of this paper.

## 6 **Conclusion**

Social service engineering is increasing in importance as the Internet changes its primary role from information access to social interaction. Facebook reached one billion active users this year. This demonstrates that the social aspect of human lives is penetrating into the virtual world.

As Donald Norman conceived the concept of user experience, the increased capabilities of social interaction in the virtual world can lead to a new concept “social experience” coined by the author. This is analogous to how increased power from dealing with rich-media user interfaces led to the concept of user experience with the departure from the concept of the user interface.

Social experience design is different from user experience design with emphasis on satisfaction in social interactions. Social experience design aims at creating different types of user satisfaction depending upon each user’s expectation.

There are multiple approaches to social experience design. One is to focus on the social aspects of user experience, such as roles, role-taking, collective culture, conflicts, shared values, and so on. Another is to collect socially-connected design examples of user experience design. Another is to highlight the uniqueness of the social experience design in comparison with other approaches in design.

The author takes an approach that deals with the analysis of the evolutionary paths toward social experience design. The author examines two paths from user interface to social experience. With the examination of these two paths, the author clarifies the unique characteristics of social experience design.

## References

1. Grudin J (1994) Groupware and social dynamics: eight challenges for developers. *CACM* 37(1):92–105

# Block IO Request Handling for DRAM-SSD in Linux Systems

Kyungkoo Jun

**Abstract** This paper proposes a method to improve the performance of DRAM-SSD in Linux systems by modifying the block device driver. Currently, it processes requests in a segment-by-segment way. However it involves overheads because it needs to perform overlapped works repetitively when finishing one segment and starting next one. It prevents DRAM-SSD from running in full speed. The proposed method reduces the overheads by grouping multiple segments into one request, removing unnecessary duplicated steps. But, the grouping also involves overhead. Thus we propose to determine adaptively whether to do grouping or not according to the number of segments contained in requests. From the evaluation results, the throughput of the proposed method improved compared with the segment-by-segment way.

**Keywords** SSD · DISK IO · Throughput · Block device

## 1 Introduction

As cloud-based storage services are growing, demands for high performance storage are increasing. However, I/O performance of hard disks is still relatively slower than processors because hard disks depend on mechanical operations. Recently, Solid State Drives (SSD) [1, 2] are widely employed as high performance storages. Because SSD does not have mechanical operations, it is superior in reliability and I/O performance.

SSDs are categorized depending on the type of memory; flash-SSD and DRAM-SSD. Flash-SSD is already widely used in diverse computing devices due

---

K. Jun (✉)

Department of Embedded Systems Engineering, University of Incheon, Incheon, Korea  
e-mail: kjun@incheon.ac.kr

to high speed read and low power characteristic. DRAM-SSD inherits all the advantages of flash-SSD. In addition, it is faster in read/write than flash-SSD and more reliable, thus more suitable for handling massive data.

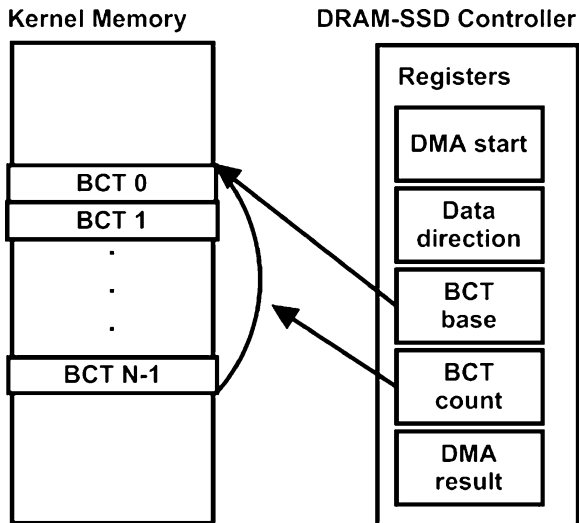
The DRAM-SSD that we consider in this paper is configured as follows. It consists of DRAM modules, DRAM controller, and PCIe [3] controller. The DRAM-controller is FPGA-based and it controls read/write to the DRAM modules, which is performed via DMA.

Figure 1 shows a set of registers that controls DMA data transfer between DRAM-SSD and a host to which DRAM-SSD is installed. DMA start register signals the beginning of transfer, direction register determines whether transfer is read or write, and DMA result register is a flag indicating whether transfer succeeds or fails. The region of DMA transfer is specified by using BCT base address and BCT counter register. One BCT defines one memory area for which DMA transfer performs and usually a set of BCTs are defined for one read or write operation. The BCT base address is the start address from which BCTs are stored consecutively. In detail, a BCT specifies the start address of memory, the start block address of DRAM-SSD, and a length. The addresses are 64-bit long.

Figure 2 shows the procedure of DMA transfer in Linux. Firstly, a set of BCTs are configured according to given read or write request. Then the direction register and the BCT counter register are set. On writing to the DMA start register, DMA transfer starts. And, an interrupt signals the end of the transfer and whether it is completed or not can be found by reading the DMA result register.

Considering such DMA transfer procedure, it should be noted that the number of DMA transfer has more influence on I/O performance than the transfer size. Due to high speed of DRAM, most of time is wasted in configuring BCTs and waiting for the completion interrupt. Therefore, it is obvious to minimize the number of

**Fig. 1** Registers that control DRAM-SSD



transfers while increasing per-transfer size in order to maximize I/O performance of DRAM-SSD.

However, in reality, Linux operating system which is widely adopted for storage systems takes the opposite action; decreasing per-transfer size while increasing the number of transfers. It is because Linux block layer is designed in a hard-disk oriented way. In the sense of spinning time of hard disks, reduced per-transfer size is more desirable for Linux; it divides one IO requests into multiple segments, resulting in increased number of transfers and decreased size. However, such behavior is not adequate for DRAM-SSD.

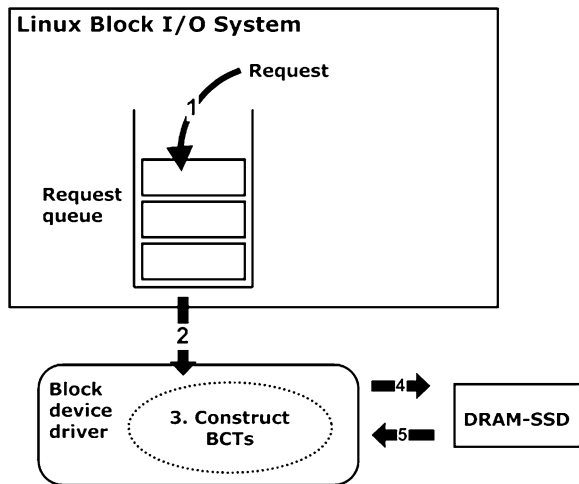
Regarding IO performance of SSD, scheduling without queue [4] for reducing scheduling overhead and SSD-oriented IO scheduler [5, 6, 7] are proposed. However, these methods are designed for flash-SSD.

In this paper, we propose a method that adaptively determines transfer size according to the whole IO size in order to optimize the number of transfers. This paper is organized as follows. Section 2 proposes an adaptive method with the explanation about the operation of block device drivers. Section 3 compares the performance of the proposed method with existing methods and Sect. 4 concludes this paper.

## 2 Adaptive Block Handling considering sizes

When Linux kernel performs the read or write on DRAM-SSD, a request queue and a block device driver are used as shown in Fig. 2. The block device driver processes the requests by fetching them from the request queue in sequence. Requests consist of a set of segments and each segment can specify maximum 4 KB transfer. The number of segment is different depending between the requests.

Fig. 2 Procedure to handle IO requests from Linux kernel



The block device driver processes a request segment by segment. Read/write on DRAM-SSD is also handled in this way. Given a request, the block device driver configures a BCT only for a first segment and begins DMA transfer, and then proceeds to a next segment after the completion interrupt. If it is successful, the device driver releases DMA mapping information regarding the transfer. As a result, each segment processing repeats the step 3, 4, and 5. It involves overheads to wait for the interrupt and release the mapping information. Such overheads increase linearly as the transfer size of a request increases because of the limitation of the maximum 4 KB segment.

The overheads can be easily reduced by performing only one DMA transfer for all of the segments, namely request-by-request. However, it requires hardware support. The controllers that we used in this paper can support up to 1024 DMA transfers in sequence at once. Also, it has another type of overheads to save DMA mapping information separately for the segments because the mapping should be freed after transfer completion. The segment-by-segment processing is free from this overhead.

Another way to reduce the overheads is to perform only one DMA transfer for multiple requests, but it is impractical because of the need to modify kernel. Current kernel, in some cases, is not allowed to proceed to next requests until previous requests complete. Another reason is that some requests cannot be combined together, for example, a read and a write.

This paper proposes a method to alternate between the segment-by-segment way and the request-by-request way depending on the number of contained segments. If the number of segments is less than a threshold  $N$ , the segment-by-segment is preferable because the overhead of the request-by-request way is larger. On the contrary, if the number is more than  $N$ , the request-by-request way is adopted. The request-by-request way requires that the BCTs for each segment be configured in advance which is different from the segment-by-segment. Note that the DMA mapping information should be saved separately to be freed after completion.

### 3 Performance Evaluation

The performance of the proposed method is evaluated and compared with the case when only the segment-by-segment is used and also with the case of the request-by-request. For the evaluation, we used the system running the Linux kernel version of 2.6.31. A benchmark program IOMeter [8] is used to generate four types of loads; sequential read/write, random read/write. Also the transfer size can vary ranging from 512 byte up to 256 KB. As a performance metric, throughput (MB/s) is measured.

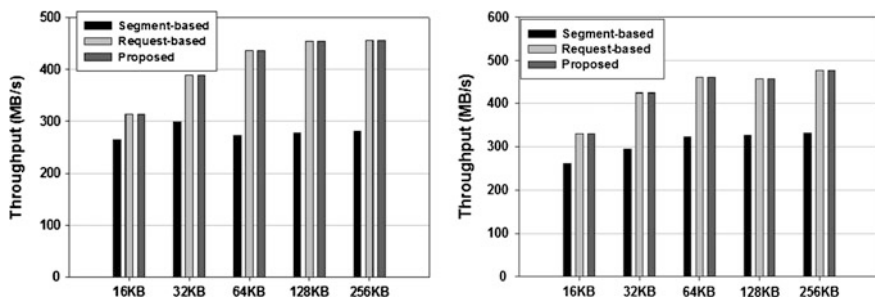
We firstly perform a set of experiment to determine an optimal  $N$ . We measure the throughput as we increase the request size step by step from 512 byte. When  $N = 2$ , its throughput is superior to other cases. Particularly when the request size

is 8 KB, the throughput gap between  $N = 2$  and  $N = 3, 4, 5$  is the largest. Since 8 KB is a multiple of the maximum segment size of 4 KB, it is one of the perfect conditions for the request-by-request processing. The results of 16 KB can be explained in the same way. However, the case of 32 KB shows similar throughput for all  $N$  because it is large enough to use the request-by-request for any  $N$ . Not only for random ready, but also for the other work loads, similar performance was observed. Since  $N = 2$  shows the best performance, the following experiments set  $N = 2$ .

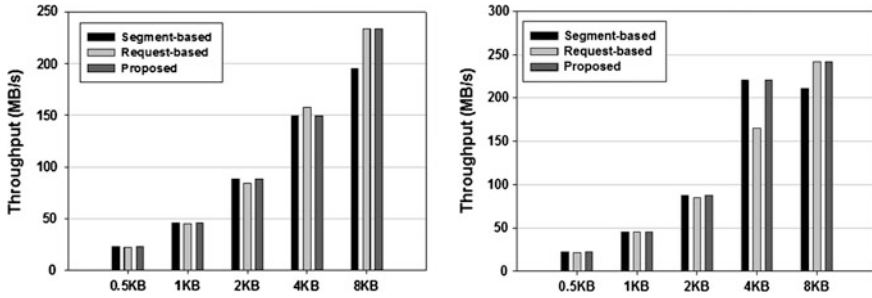
Figure 3 shows the throughput of the sequential read and the sequential write when the request sizes are large such as 16 KB or larger. The throughput of the segment-based processing is lower than the request-based and the proposed method in all the cases. As the request sizes increases, the number of the included segments in a request also increases. Therefore the segment-based processing incurs more overheads than in the case of smaller sizes of requests. And the increased overheads lower the throughput. On the other hand, the request-based processing shows the similar performance as that of the proposed method. It is because it operates in the same way as the proposed method when the request size is bigger than 4 KB.

Figure 4 shows the throughput of the sequential read and the sequential write when the request sizes are small such as less than 16 KB. Different from the results of Fig. 3, the throughput of the segment-based processing increases as the request sizes increases. It is because its overhead does not increase as the number of the included segments in a request does not increase. However, the throughput when the request size is 8 KB shows differences. It can be explained by the effect of the overheads.

Generally hard disks show different performance between random access and sequential access. However, DRAM-SSD is not affected by access pattern. Therefore, the performance of random access is similar to Figs. 3 and 4. We do not present the results in this paper because of the limitation of space.



**Fig. 3** Throughput of sequential read (*left*) and sequential write (*right*) according to large request sizes



**Fig. 4** Throughput of sequential read (*left*) and sequential write (*right*) according to small request sizes

## 4 Conclusions

This paper proposed a method to improve the throughput of DRAM-SSD by modifying the request handling procedure of Linux block device driver. It adaptively decides whether to use the request-by-request handling or the segment-by-segment according to the number of the contained segments in a request size. The number of the segments increases as the request sizes increases. If more than one segment is included, the request-based handling is more advantageous than the segment-based way. It is because of the overheads concerning the processing of the segments in sequence. On the other hand, if the number of the segments is less than two, the segment-based way is better. The request-based handling has its own overhead. Depending on the number of the segments, our method chooses a proper handling method. We evaluated the performance of our proposed method and observed that it is effective in improving the throughput.

## References

1. Takeuchi K (2013) Flash signal processing and NAND/ReRAM SSD. In: Inside solid state drives, vol 37. Springer Series in Advanced Microelectronics, pp 357–374
2. Zambelli C, Olivo P (2013) SSD reliability. In: Inside solid state drives, vol 37. Springer Series in Advanced Microelectronics, pp 203–231
3. PCI Express Base Specification (2010) PCI SIG
4. Seppanen E, OKeefe M, Jilja D (2010) High performance solid state storage under Linux. In: The 26th IEEE symposium on MSST, pp 1–12
5. Hui S, Rui Z, Jin C, Lei L, Fei W, Sheng X (2011) Analysis of the file system and block IO scheduler for SSD in performance and energy consumption. In: 2011 IEEE Asia Pacific services computing conference, pp 48–55
6. Zhang X, Davis K, Jiang S (2012) iTransformer: using SSD to improve disk scheduling for high-performance I/O. In: 2012 IEEE parallel and distributed processing symposium, pp 715–726
7. Kang S, Park H, Yoo C (2011) Performance enhancement of I/O scheduler for solid state devices. In: 2011 IEEE ICCE, pp 31–32
8. <http://www.iometer.org>

# Implementation of the Closed Plant Factory System Based on Crop Growth Model

**Myeong-Bae Lee, Taehyung Kim, HongGeun Kim, Nam-Jin Bae, Miran Baek, Chang-Woo Park, Yong-Yun Cho and Chang-Sun Shin**

**Abstract** The paper proposed the Closed Plant Factory System (CPFS) applied the crop growth model. The CPFS monitors climate data in a closed building or room and the actuator's status for control devices, and provides optimized operations for controlling growth environments. The CPFS monitors environmental data, plant growth data and the control devices' status data. This system can analyse the optimal growth environment and the correct control environment. We implemented the system and applied it to a testbed, also confirmed that the CPFS operated real-time monitoring service and controlling service correctly.

**Keywords** Vertical farm · USN · Growth monitoring · Plant factory

---

M.-B. Lee · T. Kim · H. Kim · N.-J. Bae · M. Baek · C.-W. Park · Y.-Y. Cho  
C.-S. Shin (✉)

Department of Information and Communication Engineering, Suncheon National University,  
Suncheon, South Korea  
e-mail: csshin@suncheon.ac.kr

M.-B. Lee  
e-mail: lmb@suncheon.ac.kr

T. Kim  
e-mail: taehyung@suncheon.ac.kr

H. Kim  
e-mail: khg\_david@suncheon.ac.kr

N.-J. Bae  
e-mail: bakkepo@suncheon.ac.kr

M. Baek  
e-mail: tm904@suncheon.ac.kr

C.-W. Park  
e-mail: jwpark@suncheon.ac.kr

Y.-Y. Cho  
e-mail: yycho@suncheon.ac.kr



## 1 Introduction

Recently, environmental pollution and climate change cause concern for many in terms of a future food production system. Additionally, consumers' change in demand (easier access to cleaner and organic foods) raised the need for local farms. CPFS is created to address such demands and needs [1].

A CPFS is a new farming format that maximizes the production by optimizing light, temperature, humidity, nutrients and moisture, etc. in a controlled environment. It enables the highly optimized environmental control and more accurate estimation of production through active monitoring than existing greenhouses. In short, it is an agricultural IT technology combined with BT technology that seeks to identify the most optimized growth points in areas such as light source technology such as LED, automated manufacturing process, USN and integrated control, etc.

A CPFS artificially controls the growth environment enabling planned farming during anytime of the year, while eliminating external elements such as climate, pollution or geographic limitation, etc. Therefore, a development of an optimized crop growth model is an essential field of study to maximize the benefits of the CPFS. The crop optical growth model will be a basis for the development of research in the field of standardization of crop growth process and quality, and automation control framework for the plant factory [2, 3].

Generally, plant factories can be categorized as plant production factories and vertical plant production factories. The most important task of such facilities is how to monitor and control the growth environment. In particular, to develop a crop growth model requires a continuous monitoring service and accumulated data crop growth cycle.

In this paper, we construct a testbed for the CPFS. Through this, we are able to research a monitoring about a variety of environment elements in the CPFS. [Section 2](#) explains the actualization of the suggested system and results of its performance in [Sect. 3](#); and then draws conclusions and provides additional topics for future studies in [Sect. 4](#).

## 2 Design of Closed Plant Factory System

### 2.1 Structure of CPFS

A CPFS can be categorized into; a physical level that consists of controlling devices that sense and adjust the factory environment accordingly and an application that processes gathered data and makes necessary adjustments to the system.

The physical level transmits all obtained data that sensors receive around the facility to the server's middleware which identifies any abnormalities in the data,

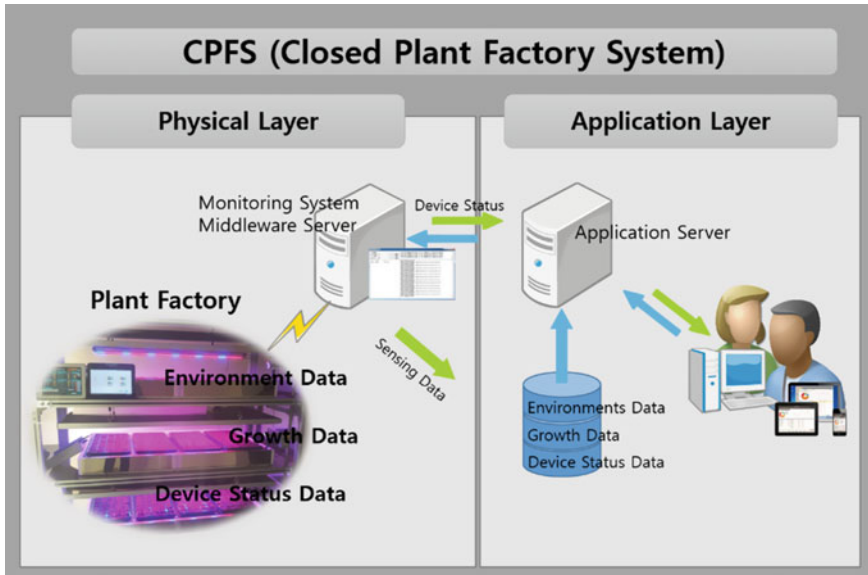


Fig. 1 System structure

then stores them in the database. The application server provides the factory necessary monitoring services so users can have real-time monitoring capabilities through PCs or smart terminals. Figure 1 describes the monitoring system of a CPFS.

## 2.2 Components of a CPFS

Unlike a typical glasshouse, a CPFS is rarely influenced by external elements therefore eliminating a concern for weather related issues. On the other hand, it must satisfy all necessary conditions that crops require in a controlled environment; therefore a continuous monitoring of internal environment is critical. Such monitoring must involve factory’s environment data, a crop’s growth data and conditions of controlling devices [4–6]. Table 1 describes each of the elements.

Figure 2 identifies all elements that form the monitoring system. The physical level consists of a climate sensor, an integrated sensor node, a manual controller

Table 1 Monitoring elements

Division	Elements
Environment info	Temperature., humidity, illumination, CO <sub>2</sub>
Growth info	Leaf temp., nutrient solution EC, PH
Control device info	Irrigation pumps, LED, heater, fan, humidifier, CO <sub>2</sub> generator

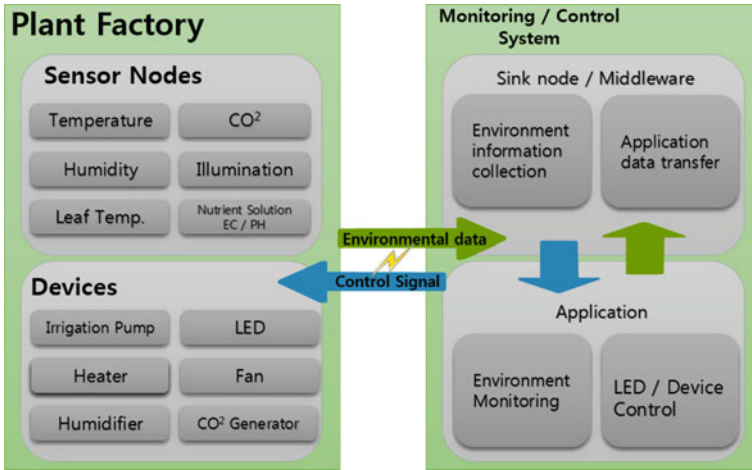


Fig. 2 System integration

that can override equipment's that are stationed within a factory (water pumps, lighting, CO<sub>2</sub> generator, ventilator, humidifier, etc.).

The application level consists of an application server that monitors the factory environment and monitoring software for users.

Figure 3 provides pictures of an integrated controller and sensor node that were used at the physical level. Sensor nodes are equipped with a climate sensor and they are stationed around the factory to measure different conditions, they then transmit the data wirelessly. The middleware receives data from the sink node through USB communication and stores them to the database. Then, the stored data become available to users through the application server.

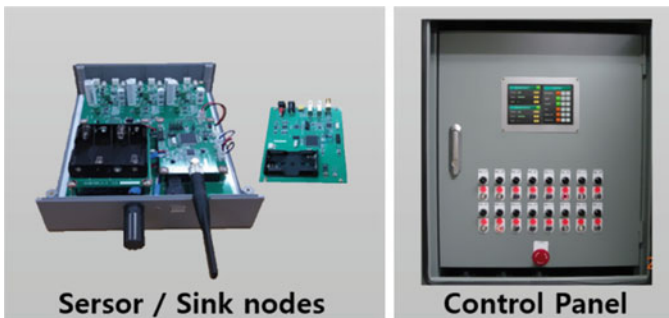


Fig. 3 Sensor node and control panel



**Fig. 4** CPFS testbed

### **3 Realization of CPFS**

#### ***3.1 Test Settings***

In order to confirm the functionality of the suggested system, this study constructed a test bed as described in Fig. 4. A vertical plant production factory that is installed at the Rural Development Administration served as a role model and we modified to fit the purpose of this study. The study also chose leaf lettuce as a test crop because it has a relatively short growth span and less impacted by its environment [7].

The testbed has two parts that are laid side by side and each part consists of four layers. The first part utilized florescent lighting and it was used to germinate seeds, while the second part used red/blue LED light to grow the germinated seeds. Each light source was to be exchanged to another if future tests required doing so. Any data that were obtained from the test bed became available users through a monitoring program, so users can easily analyze data to optimize the factory environment for the crop.

#### ***3.2 Plant Factor's Monitoring Program***

The factory's supporting software which was used to monitor its environment was developed in three different types; a middleware, an application server and a monitoring software. The middleware not only stores data that sensors gather but also transmits each device's control signals to the integrated controller.

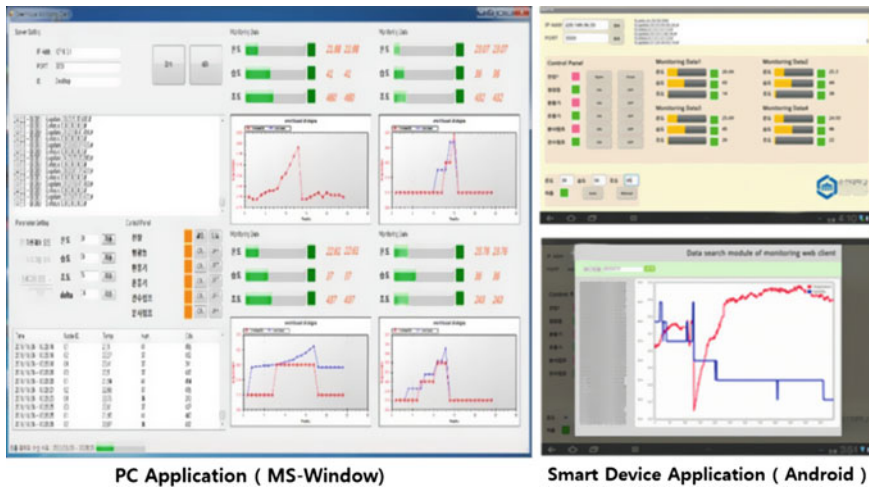


Fig. 5 Monitoring applications

The application server transmits data that the middleware gathers and stores to the client program, in addition to transmitting back to the middleware its decision on appropriateness of control signals.

Data from the factory are sent to users through an application server and users can monitor such data through their PCs or Smartphone applications. Figure 5 shows the monitoring software for PC and Android devices that was specifically developed for this study.

Each type of software shows the real-time state of the factory environment and controlling devices, and users can utilize such information to actively monitor and adjust the environment accordingly to ensure an optimized condition for the crops.

## 4 Conclusion

This study designed a monitoring system of a CPFS, constructed a test environment while enabling a continuous monitoring and analyzing of the plant growth. This study is to serve as a basis for future studies. Data that were retrieved in this research shall be further reviewed and analyzed in order to identify the most optimized conditions for crops, and to develop an automated controlling system that utilizes the most efficient algorithm. Through this, we will define the optical control set point. In addition, we will design an optimal control algorithm and develop an optimal control system throughout the crop growth cycle.

**Acknowledgments** This work was supported by the Industrial Strategic technology development program, 10040125, Development of the Integrated Environment Control S/W Platform for Constructing an Urbanized Vertical Farm Funded by the Ministry of Knowledge Economy (MKE, Korea).

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0014742).

## References

1. Despommier D (2012) Advantages of the vertical farm. *Sustainable environmental design in architecture. Springer optimization and its application* 2012. pp 259–275
2. Gim BG, Lee WJ, Heo SY (2010) Construction of a testbed for ubiquitous plant factory monitoring system using artificial lighting. *Korea Institute of Information Technology*, pp 272–275
3. Lee EJ, Lee KL, Kim HS, Kang BS (2010) Development of agriculture environment monitoring system using integrated sensor module. *Korea Contents Soc* 10(2):63–71
4. Yiming Z (2007) A design of green house monitoring & control system based on ZigBee wireless sensor network. In: *Proceedings of wireless communications, networking and mobile computing*. pp 2563–2567
5. Song Y, Ma J, Zhang X, Feng Y (2012) Design of wireless sensor network-based greenhouse environment monitoring and automatic control system. *J Network* 7(5):838–844
6. Cha MK, Lee SH, Cho YY (2012) Selection of leaf vegetables and set-up of planting density and light intensity in the plant factory. *J Asian Agric Biotechnol* 28(1):17–23
7. Park DH, Park CY, Cho SE, Park JW (2010) Greenhouse environment monitoring and automatic control system based on dew condensation prevention. In: *Proceedings of EMC 2010: embedded and multimedia computing 2010*, pp 1–5

**Part III**  
**Ubiquitous Networks and Mobile**  
**Communications**

# An Energy Efficient Layer for Event-Based Communications in Web-of-Things Frameworks

G r me Bovet and Jean Hennebert

**Abstract** Leveraging on the Web-of-Things (WoT) allows standardizing the access of things from an application level point of view. The protocols of the Web and especially HTTP are offering new ways to build mashups of things consisting of sensors and actuators. Two communication protocols are now emerging in the WoT domain for event-based data exchange, namely WebSockets and RESTful APIs. In this work, we motivate and demonstrate the use of a hybrid layer able to choose dynamically the most energy efficient protocol.

**Keywords** Web-of-things • RESTful services • WebSockets

## 1 Introduction

In the last few years, a vision of inter-connected sensors and actuators attached to physical objects has emerged, leading to the concept of Internet-of-Things (IoT) [14]. This idiom includes the concept of Wireless Sensor Networks (WSN) and goes beyond with all kind of physical objects able to communicate. The field of building automation is a potential target for IoT approaches where numerous communicating sensors and actuators are in use [2]. In such smart-buildings, new communicating objects are also appearing, for example to provide the user with feedback on the energy consumption [7]. The IoT has since then been extended from the IP usage towards the inclusion of well-known Web patterns to ease the

---

G. Bovet (✉)

LTCI, Telecom ParisTech, Paris, France

e-mail: gerome.bove@telecom-paristech.fr

J. Hennebert

ICT Institute, University of Applied Sciences of Western Switzerland, Fribourg, Switzerland

e-mail: jean.hennebert@hefr.ch



the integration and communication with things at the application level, leading to the concept of Web-of-Things (WoT) [10]. One of the main problems of the IoT is certainly in the management of the energy consumption of this multitude of communicating nodes. Although new low-power standards like 6LoWPAN, IEEE802.15.4 and RPL are being established at the network layer, the WoT framework is actually not energy aware at the application level. We believe that the protocol and data structure used for communicating with things at the highest layers could contribute to a significant reduction of the energy consumption.

In this paper, we show the feasibility of using an additional layer at the application level able to select the most suitable communication method in order to reduce the energy consumption of things connected to the Internet through Wi-Fi. We rely on the *Web-of-Things* paradigm proposing to use WebSockets or RESTful APIs for event-based data exchange. Instead of forcing application developers choosing a communication method, they can rely on an hybrid layer dynamically selecting which method is less energy consuming depending on how much and how frequently data should be sent. This represents a meaningful advantage letting developers focus on other tasks than thinking about costs. Sections 2 and 3 summarize related work and the principles of event-based WoT communications. In Sect. 4, we present our proposal for improving the event-based communication. In Sect. 5, we present the experimental measurements and their analysis. Section 6 provides details on the implementation of our hybrid layer and energy consumption measurements. Section 7 concludes our paper and provides insights on further research.

## 2 Related Work

The Cooltown project [13] is one of the early projects considering people, places and things as Web resources, using HTTP GET and POST requests for manipulating things. The recent progresses in embedded devices are now enabling the integration of Web servers on things. The tendency is clearly shown with, for example, the WebPlug WoT framework where sensors and actuators used to build so-called mashups [16]. An important step towards a standardization of the communication at the application level for web services was the introduction of the SOAP protocol. However, SOAP is not optimized in terms of energy consumption due to the large overhead of XML and of the protocol itself [8]. Much lighter, RESTful APIs provided a clear answer to this problem, with an increased adoption for many IS, especially in the domain of IoT and WoT [1, 11]. Recently, persistent TCP connections called WebSockets have been proposed for the communication between things [17]. Preliminary comparisons between HTTP and WebSockets in terms of energy consumption have been reported in [3]. This previous research shown differences between these protocols in terms of energy consumption, with complex variations as a function of the payload and frequency of the communication. Motivated by this previous work, the research presented in

this paper focuses on the analysis of the optimal choice between RESTful APIs and persistent TCP connections targeting energy efficiency. More specifically, we open the question if rules may be implemented on things for choosing automatically the most efficient way of communicating.

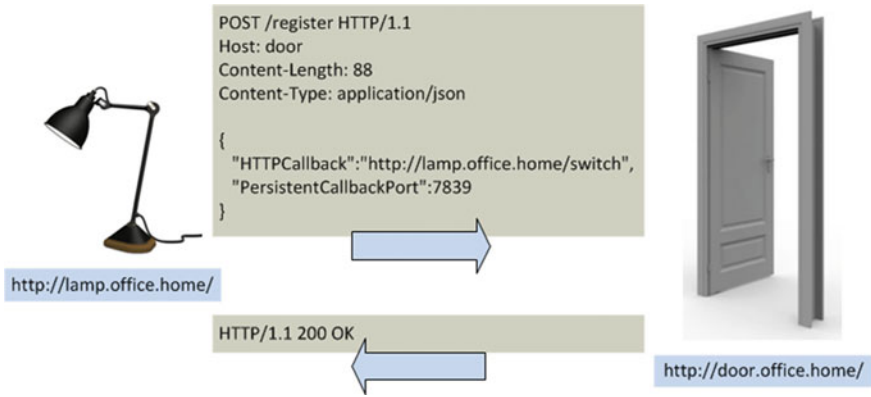
### 3 WoT Event-Based Communications

Sensor and actuator data can vary in quantity and frequency according to the context of use. For example a power outlet will continuously notify about the electricity consumption when a device is plugged in, while on the other side a presence sensor will only signal a change of state. This kind of behavior is leading to so-called event-based communications. The WoT proposes two fundamentally different approaches for managing event-based communication: HTTP callbacks and persistent TCP connections [9]. Both approaches are detailed below.

**Registration.** The first step for event-based system is the registration of the consumer at the producer. Using things with REST, we can simply expand the API with a service dedicated to registration [9]. A thing interested in being notified by change of states of another object will announce itself by providing the required callback information. For example, a lamp actuator will register a door contact sensor to be notified when someone enters or leaves the room. The lamp sends a HTTP POST request to <http://door.office.home/register>. This request can be of two types: (1) REST service—containing a JSON message indicating a REST service as callback, (2) WebSocket—containing the HTTP upgrade header field for switching to WebSocket, keeping open the connection.

**HTTP requests.** The WoT relies on REST for exposing things as resources to the Web [5]. Unlike SOAP, REST uses HTTP as application protocol for interacting with things and not only as transport protocol. The advantages of REST over SOAP are in having less overhead, and being resource oriented, which fits naturally with physical objects. With WoT, every object is embedding a built-in Web server exposing an API for interacting with its sensing, actuating and configuration capabilities. Self-descriptives URLs are used through common HTTP requests, like GET, PUT, POST and DELETE. For example, reading a sensor value is done using the GET verb and actuating using POST. For event-based communications, POST is actually the only necessary operation. A “consumer” object typically provides a REST service to be notified of changes in another object. The service URL is provided as callback at the registration on the producer. This is a significant aspect of our approach as we can link sensors with actuators.

**Persistent TCP connections.** The second way of managing event-based communications proposed in the Web-of-Things framework is using persistent TCP connections also known as WebSockets [4]. This kind of communication is mostly used in push scenarios where data has to be sent from a server to a client not running a Web server, as for example Web browsers. The channel is kept open on both sides as long as possible.



**Fig. 1** Example of the registration process

## 4 Proposal for Energy Efficient Communications

The main idea of our proposal is to let the producer decide the most energy efficient way to communicate, either through REST HTTP or through WeSockets. As it will be shown later in [Sect. 5](#), either mode become optimal as a function of the frequency and payload of the messages exchanged. To enable dynamic switching between modes, we explain here the modifications that are requested. It concerns mainly the registration process and persistent TCP connections concept explained above. In our vision, both modes are supported and therefore, the registration JSON message has to include the available callbacks for REST and persistent TCP connection. The producer will further select automatically which method is best suited for exchanging data from an energy efficiency point of view. This is illustrated in [Fig. 1](#) with an example involving a lamp and a door.

For persistent TCP connections, our proposal slightly differs from what is currently done with WoT. Indeed, WoT approaches suppose that the consumer initiates the persistent connection, keeping the channel open while the producer sends its data. In our approach, the producer has to select between HTTP or TCP and therefore initiates the connection. If the connection is lost due to network faults, the producer will retry to open a connection on the same port, unless the consumer registers with another one.

## 5 Experimental Measurements and Analysis

HTTP requests and persistent TCP connections have different impact on energy consumption. This is especially true for objects connected to the Internet with a Wi-Fi transceiver. We show here how each method can influence the energy consumption of things.

## 5.1 Test Environment

We used the openPICUS FLYPORT programmable Wi-Fi module. This tiny module ( $35 \times 48 \times 11$  mm), is Wi-Fi IEEE 802.11 certified and embeds a full TCP/IP stack, able to connect to IEEE 802.11b/g/n networks. It supports 1 or 2 Mbit/s rates as well as security protocols such as WEP, WPA-PSK and WPA2-PSK. The FLYPORT can be powered either at 5 V or at 3.3 V and drains 128 mA current at 3.3 V when connected to Wi-Fi. An IDE is available for developing application in C [15]. We set up an isolated test environment composed of a FLYPORT module acting as the producer, an access point and a PC acting as the consumer. The wireless network set up is 802.11 g, no encryption and long preamble. We also used a Hameg HM8115-2 for measuring the energy consumption of the FLYPORT [12]. Having a dedicated test bench ensures that no other device will be disturbing the proper running of the experiment as it would be in a public network.

## 5.2 Power Consumption Measurements

We describe here our measurement campaign for both TCP and HTTP. During each test of 30 s, the producer sent packets with a fixed payload size at a specific interval. For measuring precisely the consequence of each method on the power consumption, the FLYPORT was only running a minimal program sending events. The values of payload and interval are chosen to match the behavior of some specific devices and therefore to perform more realistic measurements. We made the payload size in bytes vary from 1 to 400 and the intervals between packets in milliseconds from 50 to 800, which correspond to certain devices one can find in smart buildings. The combination of the payload sizes and intervals gives us a campaign of 30 measurements as illustrated in Fig. 2.

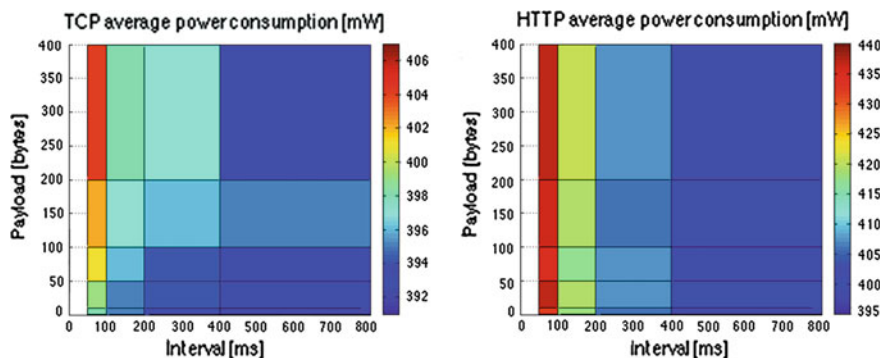


Fig. 2 Results of the average power consumption measurements for TCP and HTTP

From Fig. 2, we see that TCP is overall less energy consuming than HTTP. TCP appears to be on average 4 % less consuming than HTTP with a maximal gain of 9.5 %. The quantity of transmitted packets is indeed lower for TCP than HTTP. With TCP, once the connection is established, only one packet is necessary to send the JSON message. HTTP is more complex as a connection has to be established every time a JSON message must be sent. An HTTP connection includes the potential TCP window negotiation, the HTTP header, the HTTP response, and finally the connection closing. All this overhead causes an increase in consumption. The measurements also show that the amount of payload data plays a less important role in the power consumption. This is especially true for HTTP consumption. On the other hand, a factor influencing the consumption is clearly the sending interval. The main observation is that both modes are overlapping in terms of efficiency, with TCP is becoming less optimal than HTTP in some conditions.

### 5.3 Consumption Approximation for TCP and HTTP

**HTTP.** With TCP, the variable is the necessary time needed to send data, including all underlying protocols. With Wi-Fi (802.11 g) frame composition is taken from [19]. The energy consumption for one packet of data can be computed with the following function:  $E(\text{payload}) = \{\text{PLCP preamble} + (\text{MAC header} + \text{IP header} + \text{TCP header} + \text{payload}) * \text{ByteRate}\} * \text{TransmitPower}$  with IPheader, TCPheader and ByteRate known from [6, 18] and TransmitPower previously measured. When comparing the theoretical values of the approximation to the measurements of the FLYPORT in Fig. 2, it comes out this function is accurate enough with an average error of 0.86 %.

**TCP.** As explained earlier, the HTTP case is more complicated as for TCP. Instead of using a theoretical model, we opted for a parametric model where the parameters are fit to the observations. We converged to an exponential function, approximated as in  $P(\text{interval}) = a * \exp(b * \text{interval}) + c * \exp(d * \text{interval})$ . Through a numerical fitting algorithm, we computed the parameters a, b, c and d for every case of payload (1, 10, 50, 100, 200 and 400), ending up with 6 functions for the different payload sizes. The computed parameters allow the functions to be quite precise with an average error of 0.05 %.

## 6 Implementation and Evaluation of the Hybrid Layer

We had first to develop a **REST server** library in C for the FLYPORT. The services are registered by indicating a URL scheme corresponding to the Web service, and providing a pointer to a callback function that will be called when the server receives a request for this particular service. We then implemented the

hybrid layer in charge of dynamically choosing the appropriate method between TCP and HTTP when sending events to registered consumers. We first implemented a history structure for recording the past events sent to consumers. An instance of this structure is created for every consumer registered. This allows computing the energy consumed to send the previous events. According to this result, the layer then switches to the most efficient method and this every time a new event must be sent. For computing the TCP mode energy consumption, we implemented the function described in Sect. 5.3. The implementation includes a rule for intervals higher than 10 s to consider the keep-alive packets (specific to the FLYPORT as it may differ on other modules). The final value is computed as follows: *energy of each packet sent in history + energy at idle between the shipments + energy of keep-alive packets*. For HTTP, we implemented the function as in Sect. 5.4. Using the history, we know the interval and the average payload. Those values are then used as parameters for our approximation function. Linear interpolations are used in the case of payload different as our reference values (1, 10, 50, 100, 200 or 400). The obtained power value is then converted in energy by knowing the time duration of the history.

Table 1 shows the energy measurements of our hybrid layer where some relevant saves were achieved. For comparison purposes, we had to rerun the campaign for each TCP and HTTP modes as our REST server running on the module is also consuming some energy. The column *Gain* shows the percentage of energy saved relative to the highest value between TCP and HTTP. The column *Loss* shows the percentage of energy lost relative to the lowest value between TCP and HTTP. The negative values in the *Gain* can be explained by the consumption due to the hybrid layer. Nevertheless, our hybrid layer clearly shows its usefulness

**Table 1** Power consumption comparison between TCP, HTTP and the hybrid layer

Payload (bytes)	Interval (ms)	TCP (mW)	HTTP (mW)	Hybrid (mW)	Gain (%)	Loss (%)
1	50	406	429	407	5.41	0.25
1	100	404	420	406	3.45	0.49
1	200	402	409	403	1.49	0.25
10	50	405	430	405	6.17	0.00
10	100	405	422	405	4.20	0.00
10	200	403	411	404	1.73	0.25
50	50	408	432	409	5.62	0.24
50	100	404	422	404	4.46	0.00
50	200	402	409	403	1.49	0.25
100	50	407	430	408	5.39	0.25
100	100	403	422	404	4.46	0.25
100	200	402	411	402	2.24	0.00
200	50	411	431	414	4.11	0.72
200	100	406	423	406	4.19	0.00
400	50	415	429	418	2.63	0.72
400	100	410	423	412	2.67	0.49

allowing saving 6.2 % of energy in the best case and 2.1 % on average. The hybrid layer also chooses the best method for higher intervals above 10 s as it selects HTTP, which is theoretically the best one for higher intervals.

## 7 Discussion and Conclusion

Our measurements showed that TCP and HTTP are not equivalent in terms of energy, even if their purpose is the same. By offering a hybrid layer, we expect to globally reduce the energy consumption and lengthen battery life of Web-of-Things. Although our hybrid layer allows energy savings for sensors sending at a fixed interval, the behavior remains open for varying intervals. The number of records saved in the history will play a role on how the layer will respond to changes of interval. Another unresolved issue concerns the rate of symbols sent over Wi-Fi. The approximation function for TCP requires knowing at which rate the module sends its data. Due to changes in the surrounding environment, traffic congestions and other reasons, this rate may be changing. In our case, we forced a rate of 2 Mb/s in our test infrastructure.

In this paper, we explored a new way on how to reduce the energy consumption of things working inside the WoT framework. Instead of giving the responsibility of choice between TCP and HTTP for event notifications to developers, we introduce an hybrid layer doing the job for them. Our results show that energy savings can be achieved by selecting the most appropriate transport protocol. Further to this, we believe that our approach simplifies callbacks between things. Future work includes addressing the varying interval of events and finding the best history size to conciliate reaction time and filtering of outlier intervals. While the measured energy savings are relatively limited, we believe our hybrid layer has further potentials, for example if used as caching method of events by considering time penalties to limit the radio's use.

## References

1. Aijaz F, Chaudhary M, Walke B (2009) Performance comparison of a SOAP and REST mobile web server. In: Proceeding of the 3rd international conference on open-source systems and technologies, Lahore, Pakistan
2. Bovet G, Hennebert H (2012) The web-of-things conquering smart buildings. *Bulletin* 10s:15–19
3. Bovet G, Hennebert H (2012) Communicating with things: an energy consumption analysis. In: Proceeding of the 10th International conference on pervasive computing, Newcastle, UK
4. Fette I, Melnikov A (2011) The WebSocket protocol. RFC
5. Fielding R, Taylor R (2002) Principled design of the modern Web architecture. *ACM Trans Internet Technol* 2:115–150
6. Gast M (2005) 802.11 wireless networks: the definitive guide, 2nd ed. O'Reilly Media

7. Gisler C, Barchi G, Bovet G, Mugellini H, Hennebert J (2012) Demonstration of a monitoring lamp to visualize the energy consumption in houses. In: Proceedings of the 10th international conference on pervasive computing, Newcastle, UK
8. Groba C, Clarke S (2010) Web services on embedded systems: a performance study. In: Proceeding of the 8th IEEE international conference on pervasive computing and communications, Mannheim, Germany
9. Guinard D (2011) A web of things application architecture: integrating the real-world into the web. ETHZ, Zurich
10. Guinard D, Trifa V, Mattern F, Wilde E (2011) From the internet of things to the web of things: resource oriented architecture and best practices In: Uckelmann D, Harrison M, Michahelles F (eds) Architecting the internet of things. Springer, Heidelberg, p 97
11. Hamad H, Saad M, Abed R (2010) Performance evaluation of RESTful web services. *Comput Eng* 2:72–78
12. Hameg (2012) HM8115-2 power meter description. <http://www.hameg.com/0.147.0.html>
13. Kindberg T et al (2002) People, places, things: web presence for the real world. *Mobile Netw Appl* 7:365–376
14. Mattern F, Floerkemeier C (2010) From the internet of computers to the internet of things. In: Sachs K, Petrov I, Guerrero P (eds) From active data management to event-based systems and more. Springer, Heidelberg, p 242
15. OpenPicus (2012) FLYPORT datasheet. [http://space.openpicus.com/u/ftp/datasheet/flyport\\_wifi\\_datasheet\\_rev8.pdf](http://space.openpicus.com/u/ftp/datasheet/flyport_wifi_datasheet_rev8.pdf)
16. Ostermaier B, Schlup F, Römer K (2010) WebPlug: a framework for the web of things. In: Proceedings of the first IEEE international workshop on the web of things (WOT2010), Mannheim, Germany
17. Priyantha N, Kansal A, Goraczko M et al (2008) Tiny web services: design and implementation of interoperable and evolvable sensor networks. In: Proceeding of the 6th ACM conference on embedded network sensor systems, Raleigh, USA
18. Stevens R (1993) TCP/IP illustrated: the protocols. Addison-Wesley Longman Publishing Co, Boston
19. Vassis D, Rouskas A, Maglogiannis I (2005) The IEEE 802.11 g standard for high data rate WLANs. *IEEE Netw J* 9:21–26



# A Secure Registration Scheme for Femtocell Embedded Networks

Ikram Syed and Hoon Kim

**Abstract** Recently, femtocell received a signification interest to improve the indoor coverage and provide better voice and data services. Lots of work has been done to improve the femtocell security, but still there are some issues which need to be addressed. Our contribution to the femtocell security is to protect secure zone (femtocell coverage area within macrocell) from unauthorized (non-CSG) users. In this paper, we propose a secure registration scheme for femtocell embedded network. In this scheme, only Closed Subscriber Group (CSG) users are allowed to access both the femtocell and macrocell services within the secure zone. By prioritizing the femtocell over macrocell within the secure zone, every user will try to camp on femtocell and invoke location registration to the femtocell as the user enters to the femtocell coverage area. If the user is within the allowed users list, the femtocell will allow the user otherwise femtocell will send a reject message to the user and also send the user information to the core network.

**Keywords** Femtocell · Macrocell · Closed subscriber group · Location area update · Secure zone

## 1 Introduction

Femtocell is small base station, connected with the service provider network through broadband (DSL, cable modem), it typically designs for home use and office use, they are short range, low cost and low power base stations that provides

---

I. Syed · H. Kim (✉)  
University of Incheon, Incheon, Korea  
e-mail: hoon@incheon.ac.kr

I. Syed  
e-mail: ikram@incheon.ac.kr

better coverage, better indoor voice and data services [1, 2]. Recent research shows that more than 50 % of voice calls and more than 70 % of data traffic are generated indoors [3]. Femtocell improves indoor coverage and capacity of the cellular providers with very low cost compare to the traditional macrocell base station [4]. The femtocell operates in licensed spectrum and may use the same or different frequency from the macrocell [5].

Recently, femtocells received a significant interest in the telecommunications industry. According to the ABI Research, 5.3 million femtocells will be deployed by end of 2012 [6]. Many of the major issues in femtocell have been studied, especially in security, lots of work has been done on femtocell security [7–9], but still there are some issues which needs to be addressed, especially on embedded networks security. In this paper, we focus on the protection of specific coverage area of femtocell within the macrocell coverage area, we called it secure zone. Our contribution is to protect the secure zone from unauthorized (non-CSG) users. Only authorized users are allowed to access both the femtocell and macrocell services within the secure zone.

There are some restrictions in the femtocell network for unauthorized users, when femtocell is working in CSG mode, but there is no restriction in macrocell for unauthorized (non-CSG) users of accessing the macrocell services. When the non-CSG users are in femtocell coverage area, they can access the macrocell services if the macrocell service is available. Our main contribution is to protect the femtocell coverage within the macrocell coverage area from non-CSG users. The secure zone would be used for security purpose in security agencies and military organizations.

The remainder of this paper is organized as follows. In [Sect. 2](#) Access Control methods in femtocell, Location Area Update and cell selection and reselection criterion are discussed, [Sect. 3](#) presented the proposed scheme for the secure zone and conclusion are presented in [Sect. 4](#).

## 2 Background

### 2.1 Access Control Methods in Femtocell

There are basically three types of access control in femtocell, namely Closed Access, Open Access and Hybrid Access.

1. *Closed Access*: In closed access mode, the femtocell doesn't want to share their resources with other users due to limited resources or security reasons. Only authorized users are allowed to access the CSG cell [10]. In 3GPP the closed access is known as the CSG cell.
2. *Open Access*: In open Access, all users are allowed to access the femtocell. There is no restriction on any user [10].

3. *Hybrid Access*: In Hybrid access mode, both CSG users and non-CSG users are allowed to camp on femtocell but there are some exceptions for non-CSG users. In this mode CSG users are given more priority over non-CSG users [10].

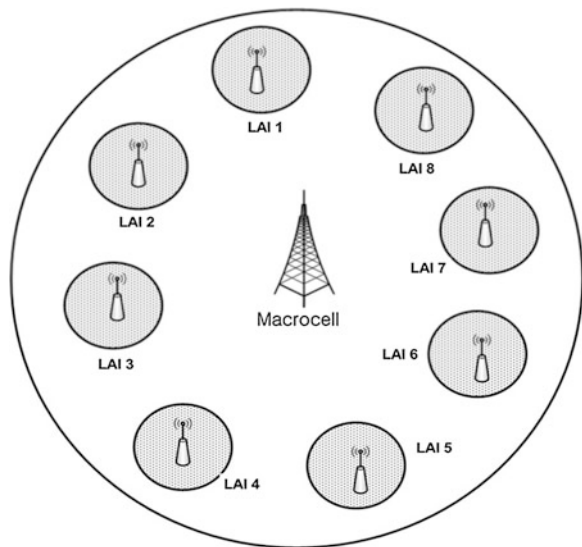
### 2.2 Location Area Update

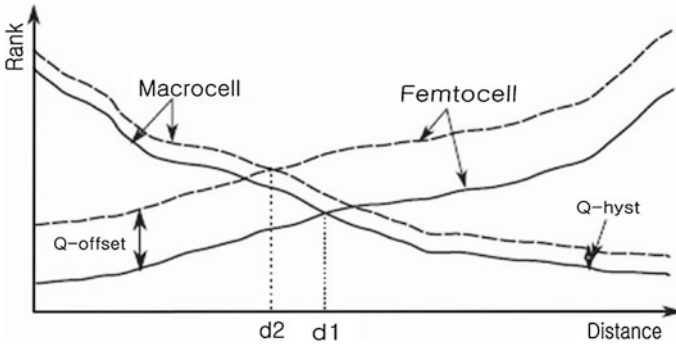
The Location Area Update (LAU) normally performs when a user turn on the power or enter to a new location area different from the old location area of the user. The access control need to be invoked, when a user move from one LAI to another LAI. In macrocell network, access control is normally invoked during the LAU [10]. Each femtocell has assigned a femtocell specific LAI different from Macrocell LAI [10]. Figure 1 shows the deployment of femtocells within the macrocell coverage area. In this approach, each femtocell has assigned different and unique LAI from the other femtocell and also from the macrocell, whenever a user enters to the femtocell coverage area and try to camp on femtocell. If a user is not allowed to a specific femtocell and try to camp on it, the user will received negative response in location update procedures.

### 2.3 Cell Selection and Reselection

Cell selection and reselection are still more complex problem in femtocell network, during the cell selection and reselection, the users need to carry out cell measurement parameters for intra-frequency, inter-frequency and inter-RAT

**Fig. 1** Femtocells deployment





**Fig. 2** Prioritizing femtocell over the macrocell in cell reselection [10]

neighbor cells, and rank the cells based on the policy used for specific cell [10]. Searching criterion (S-criterion) invokes the users to do monitoring and measurements for the intra-frequency, inter-frequency and inter-RAT, for the corresponding cell [10]. A lower S-criterion with the macrocell can help the users to start searching for the femtocell as soon as it reached to the femtocell coverage area.

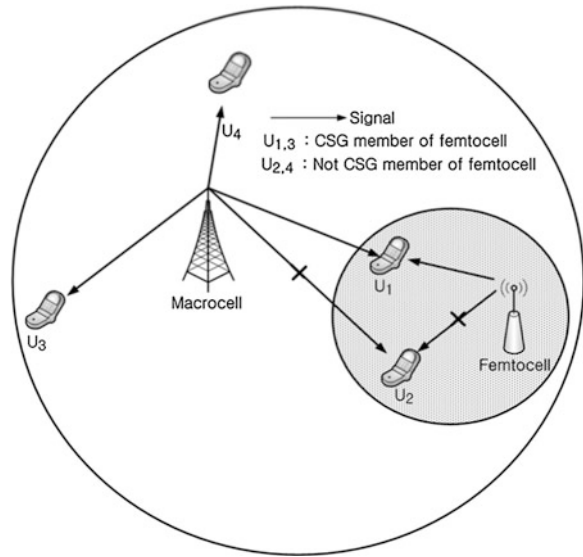
Further we also consider cell Ranking Criterion (R-Criterion) in which the Hysteresis values for the serving cell (Q-hyst) and offset values for neighboring cells (Q-offset) jointly affect the cell ranking. Figure 2 shows that by setting the Qoffset values negative and low Q-hyst in the macrocell neighbor cell list (NCL), the cell searching point  $d1$  moves toward  $d2$ . It prioritizes the femtocell over the macrocell within the macrocell NCL. The main advantage of these methods is that all users in femtocell coverage area will prioritize the femtocell over the macrocell and will try to camp on femtocell. The non-CSG users will receive a negative response from the femtocell.

### 3 Proposed Scheme

Our proposed scheme focused on the secure zone protection from non-CSG users. We proposed a secure registration scheme for femtocell embedded network. It is assumed that the femtocell will operate in CSG mode in the secure zone. The authorized user information is stored in CSG list. The CSG members list is located in the core network (CN) [11]. The CN entities such as Mobile Switching Centre (MSC)/Visitor Location Registration (VLR), Serving GPRS Support Node (SGSN), Mobility Management Entity (MME) and Home Subscriber Server (HSS) [10].

In normal case non-CSG users can access the macrocell service if the macrocell coverage is available. Due to Security reasons, we want to block non-CSG users of accessing both the femtocell and macrocell services within the secure zone. Figure 3 shows the basic scenarios of the secure zone within the macrocell coverage. U1, U3 are CSG members while U2, U4 are non-CSG members. U1 and U3

**Fig. 3** Secure zone in macrocell coverage area



can access both the femtocell and macrocell services. U2 and U4 can access the macrocell service out of secure zone, but within secure zone U2 and U4 can't access the femtocell and macrocell services. In the Cell selection and reselection, femtocell will be prioritized over macrocell in Searching Criterion by setting the lower S-criterion within the macrocell NCL. Every user will start searching femtocell as soon as they reached to the femtocell coverage area. We also prioritized the femtocell over macrocell in Ranking Criterion by setting lower value of Q-offset for femtocell within macrocell, by doing this, every user will try to camp on femtocell as the user enter femtocell coverage area and initiates the initial Non Access Stratum (NAS) procedure by establishing the RRC Connection with the femtocell.

The user capabilities are reported to the Femtocell as a part of the RRC connection establishment procedure [12]. The RRC connection message includes user identity (IMSI or TMSI) and establishment cause etc. by sending the RRC connection procedure to femtocell, the femtocell will check the user capabilities. If no context ID exists for the user, the femtocell will initiate user registration request to the Femtocell gateway (FGW). The FGW will check the user capabilities provided in RRC Connection message. If the user is a CSG user, the FGW may accept the user registration and allocate a context ID for the user [12]. If the user is non-CSG user the FGW will send reject message to the femtocell and will also send the user information to the CN to block the user within the specific femtocell coverage area.

The user information includes IMSI or TMSI and the serving femtocell location area identity. Figure 4 shows the flow chart of the proposed scheme, and the signal flow diagram of the proposed scheme is shown in Fig. 5.

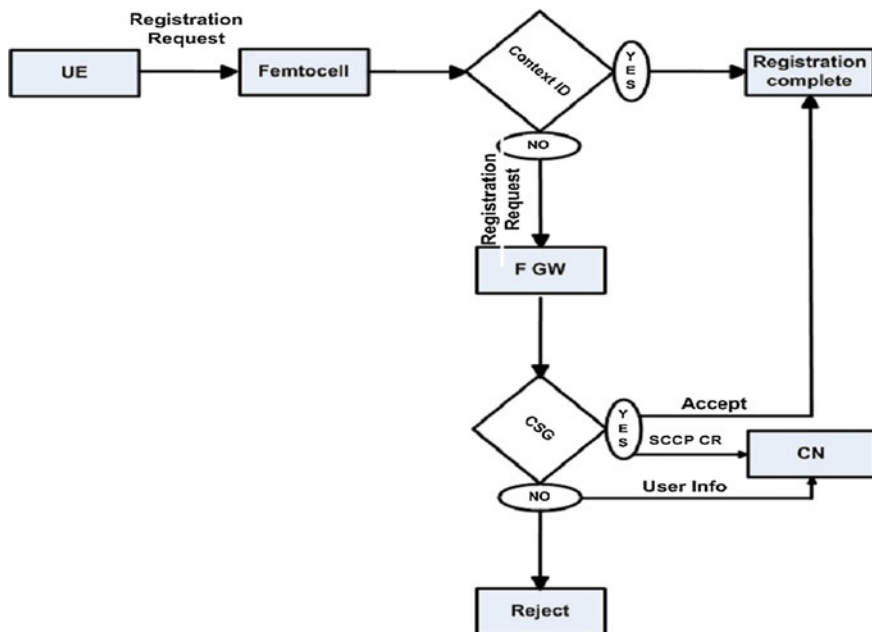


Fig. 4 Flow chart of the proposed scheme

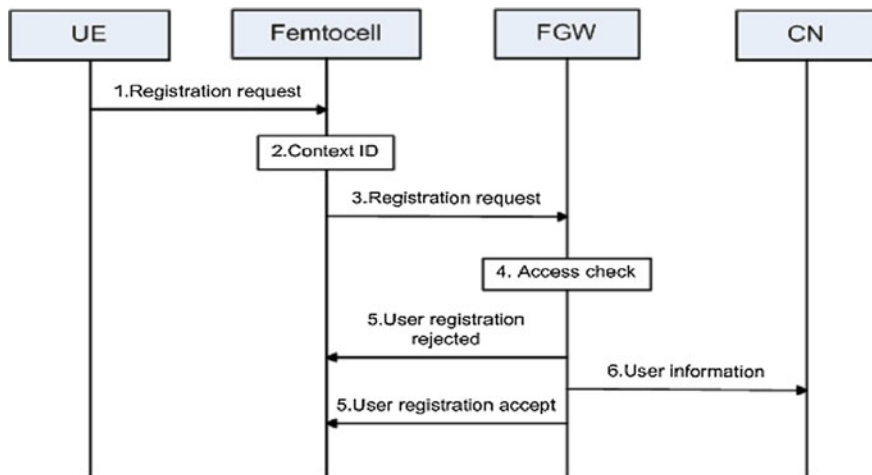


Fig. 5 Signal flow chart of user registration process

After receiving a rejected response from the femtocell, the user will try to camp on macrocell, but femtocell already sent the user information to the macrocell to block this user within the serving femtocell coverage area. The femtocell sends this information to the macrocell in the user information messenger. When the user

try to camp on macrocell, the macrocell will first check the user capabilities (the user location area and user membership), If the user is CSG user then the macrocell will allow the user to access his services, but if the user is not in the CSG member list then check the LA of the user. If the user is in the secure zone, the macrocell will block the user.

## 4 Conclusions

This paper focused on the protection of secure zone, to make the secure zone secure from non-CSG users. We proposed a new user registration scheme for both CSG and non-CSG users, by prioritizing the femtocell over macrocell within the secure zone. By doing this, every user will try to camp on femtocell and invoke location registration to the femtocell as soon it enter the secure zone. The femtocell will reject the non-CSG users and send the user information to the CN.

## References

1. Namgeol O, Han S, Kim H (2010) System capacity and coverage analysis of femtocell networks. In: WCNC, Sydney, pp 1–5
2. Chandrasekhar V, Andrews J, Gatherer A (2008) Femtocell networks: a survey. *IEEE Commun Mag* 46(9):59–67
3. Sang YJ, Hwang HG, Kim KS (2009) A self-organized femtocell for IEEE 802.16e system. In: GLOBECOM, pp 1–5
4. Chowdhury MZ, Trung BM, Jang YM (2011) Neighbor cell list optimization for femtocell-to-femtocell handover in dense femtocell networks. In: ICUFN, pp 241–245
5. Cao F, Fan Z (2010) The tradeoff between energy efficiency and system performance of femtocell deployment. ISWCS, UK, pp 315–319
6. Femtocell shipments flatline due to slow inventory burn rate. ABI research, Research report (2012)
7. Han C-K, Choi H-K, Kim I-H (2009) Building femtocell more secure with improved proxy signature. In: GLOBECOM, pp 1–6
8. Vanek T, Rohlik M (2011) Perspective security procedures for femtocell backbone. In: ICUMT, pp 1–4
9. Brassil J, Manadhata PK, (2012) Securing a femtocell-based location service. *Mobile and wireless networking*, pp 30–35
10. Zhang J, de la Roche G (2009) Femtocells: technologies and deployment
11. 3GPP TS 25.304 (2009) User equipment (UE) procedures in idle mode and procedures for cell reselection in connected mode. 3GPP-TSG RAN, 8.5.0
12. 3GPP TS 25.467 (2010) UTRAN architecture for 3G Home Node B (HNB). 3GPP-TSG RAN, 9.3.0

**Part IV**  
**Intelligent Computing**



# Unsupervised Keyphrase Extraction Based Ranking Algorithm for Opinion Articles

Heungmo Ryang and Unil Yun

**Abstract** Keyphrase extraction is to select the most representative phrases within a given text. While supervised methods require a large amount of training data, unsupervised methods can perform without prior knowledge such as language. In this paper, we propose a ranking algorithm based on unsupervised keyphrase extraction and develop a framework for retrieving opinion articles. Since the proposed algorithm uses an unsupervised method, it can be employed to multi-language systems. Moreover, our proposed ranking algorithm measures the importance in three aspects, the amount of information within articles, representativeness of sentences, and frequency of words. Our framework shows better performance than previous algorithms in terms of precision and NDCG.

**Keywords** Opinion article · Ranking algorithm · Unsupervised keyphrase extraction

## 1 Introduction

With the expanse of the e-commerce, people are using the Internet to check opinion article of products written by other people before buying them. The rapid growth of online stores not only leads to explosive increasing of opinion information but also presents new challenges to Information Retrieval (IR) field. As the result of the expanding information, it becomes difficult to find helpful opinion articles. Thus, appropriate ranking algorithms are required for searching and retrieving meaningful opinion articles. Ranking algorithms used in the IR measure

---

H. Ryang · U. Yun (✉)

Department of Computer Science, Chungbuk National University, Chungbuk, South Korea  
e-mail: yunei@chungbuk.ac.kr

H. Ryang

e-mail: riangs@chungbuk.ac.kr

the importance of targets such as web pages and words in documents, and many algorithms [1, 3, 5] have been proposed. Although these traditional ranking algorithms have played an important role, it is hard to retrieve relevant opinion articles since the algorithms do not consider characteristics of them. To address this issue, opinion ranking algorithms [2, 6, 7] have been proposed, and they can be divided into two approaches: (1) sentimental and semantic analyze opinions using dictionaries which contain opinion words; (2) measure the importance of opinion articles based on additional information. The former approaches [2, 7] need the prior data such as dictionary, and thus these approaches can be called supervised methods. Because they demand sentimental and semantic words list, it is difficult for multi-language system to apply the methods. In view of this, the latter approaches [6] can be easily employed to the multi-language system due to ability of computing rankings without any prior knowledge. Thus, these approaches can be called unsupervised methods. However, they do not consider the importance of sentences. In this paper, therefore, motivated by the above, we propose a ranking algorithm, called ROU (Ranking Opinion articles based on Unsupervised keyphrase extraction), for reflecting the importance of sentences to rankings. The proposed algorithm computes ranking scores in three aspects, the amount of information within articles, representativeness of sentences, and frequency of words. The remainder of this paper is organized as follows. In Sect. 2, we introduce the related work. In Sect. 3, we describe the proposed algorithm and framework for retrieving opinion articles in detail. In Sect. 4, we show and analyze experimental results for performance evaluation. Finally, conclusions are given in Sect. 5.

## 2 Related Work

Various ranking algorithms were proposed for IR. TF-IDF [1] is one of the most popular and important algorithms. In this algorithm, terms are given more weights when they appear frequently in a single document (Term Frequency) or they are included smaller set of documents in the corpus (Inverse Document Frequency). Although TF-IDF can measure the importance of words, it cannot know how important sentences or documents are. Nevertheless, it still can be used as effective measurement to words. Meanwhile, ranking algorithms [3, 5] for web pages were also proposed, such as PageRank [5] adopted by Google (<http://www.google.com>), and they have played significant role. Recently, with the rapid growth of the e-commerce, the amount of opinion information is increasing explosively and appropriate ranking algorithms are required since the general purpose algorithms cannot reflect the characteristics of opinion articles. To address this issue, opinion ranking algorithms [2, 6, 7] were proposed. There are two types of work, sentimental analysis based and additional information based ranking algorithms. The former methods [3, 5] identify nouns, adjectives, and inversion words using dictionary which contains positive, negative, and neutral words and calculate ranking

based on the found words. It means that they need prior knowledge for certain language, and thus they can be called supervised methods. Therefore, it is difficult for multi-language systems to apply the algorithms. In contrast, the latter measures the importance of opinion articles without any prior knowledge, and thus they can be easily adopted by multi-language systems. RLRank [6] is one of the latter algorithms and employs weight of words and the amount of information. However, they do not reflect the importance of sentences. For these reasons, this study aims to develop framework for opinion article retrieval and reflect importance of sentences. On the other hand, TextRank [4] is one of algorithms for unsupervised keyphrase extraction. It calculates ranking score of sentences using both similarity and other graph based ranking algorithms such as PageRank and HITS [3]. Since TextRank is an unsupervised method, it can be performed regardless of language. Thus, our proposed ranking algorithm applies TextRank for measuring the importance of sentences based on representativeness.

### 3 Unsupervised Keyphrase Extraction Based Algorithm

The proposed and developed framework in this study is a keyword based retrieval system for opinion articles, and it searches important opinion articles. In addition, the framework consists of three steps, data preprocessing, ranking, and indexing. In the first step, data of the collected dataset is preprocessed for applying the proposed ranking algorithm; at the same time, information for the algorithm such as title is extracted. In the second step, ranking scores are computed using the preprocessed data. Especially, the ranking algorithm measures the importance of keywords in three aspects. Let *key* be a certain keyword. The first aspect is about how useful an opinion article includes *key*, and the proposed algorithm uses the amount of information as the measurement. The second aspect is how representative sentences containing *key* are. The last aspect is about how important *key* is in both the opinion article and dataset.

Figure 1a is a diagram of the three aspects. In the last step, index is constructed by creating inverted index files including index information such as ranking scores. Figure 1b shows the system architecture of the framework. It consists of not only the three modules described in the above but also searching module for providing search results. The searching module first accepts queries and extracts keywords from the queries. Then, it finds inverted index files containing each extracted keyword, and calculates ranking score by employing the information in the files. Finally, the module provides results by sorting in ranking score descending order.

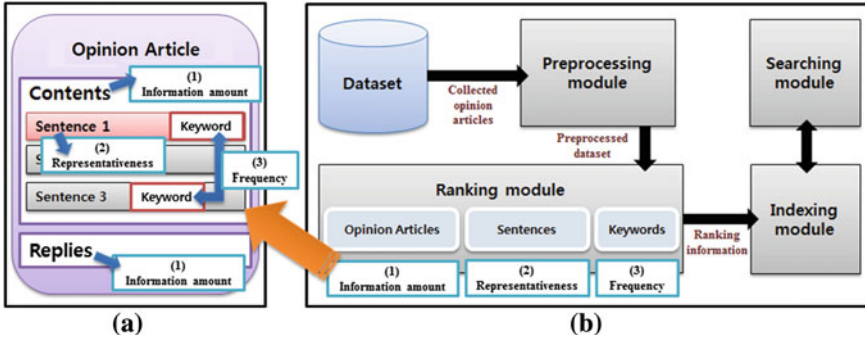


Fig. 1 The proposed framework

### 3.1 Measuring the Amount of Information in Opinion Articles

The proposed ranking algorithm, ROU first analyzes the usefulness of each opinion article in the dataset by measuring the amount of information with two assumptions. The first assumption is that if opinion articles include more contents, they contain more information related to products. It can be measured using the length of contents. The second assumption is that reply involves information about not only products but also opinion articles. Thus, for this purpose, preprocessing module extracts information of reply and the number of words. Ranking module computes the importance of opinion articles using the extracted information through following equation.

$$Importance(OA) = L(OA) + R(OA) \times (1 + RO(OA)/R(OA)) \quad (1)$$

In the equation,  $OA$  is a certain opinion article,  $L(OA)$  is the length of contents in  $OA$ ,  $R(OA)$  is the number of reply contained by  $OA$ , and  $RO(OA)$  is the number of reply written by other users.

### 3.2 Measuring Representativeness of Sentences

After calculating of  $Importance(OA)$ , ranking module measures the importance of each sentence in  $OA$  based on representativeness, and TextRank [4] is used for this purpose. Most of global online stores such as Amazon (<http://www.amazon.com>) provide multi-language systems. Thus, ROU algorithm applies the unsupervised method so as to be performed without prior knowledge of languages. In this stage, ranking module first divides  $OA$  into sentences (i.e. keyphrases) and computes similarity between the divided sentences. The similarity is measured by a function which analyzes their content overlap [4]. Then, by using the computed similarity,

TextRank score, that is representativeness of the sentences is obtained based on PageRank [5]. In TextRank, the more representative sentences, the more scores are assigned. The obtained scores are normalized as [0, 1] for calculating the importance of keywords in *OA*, and the details are described in the following section.

### 3.3 Ranking Technique Based on Unsupervised Keyphrase Extraction

Once both *Importance(OA)* and the normalized TextRank score of sentences in *OA* are computed, ranking module calculates final ranking scores of keywords based on the results obtained through previous steps. First, TF-IDF [1] of each keyword is computed to measure how the each keyword is important in both *OA* and dataset. After that, the average importance of sentences containing the each keyword is calculated for reflecting representativeness of the sentences to ranking. Following equation is used for this purpose.

$$\text{Representative}(key) = tf \cdot Idf \times \left( 1 + \sum N\text{TextRank}(\text{sentence}, key) / S(key) \right) \quad (2)$$

In the equation, *tf·Idf* is TF-IDF value of a certain keyword, *key*, *sentence* is a sentence which has *key*, *NTextRank(sentence, key)* is a normalized TextRank score

```

Construct_Index(dataset, ratio)
1. For each opinion article OA in dataset
2.   Extract information of title, contents, and reply from OA
3.   Analyze the extracted information and Save the analyzed data into files PF
4. For each preprocessed data POA in PF
5.   L ← the length of contents
6.   R ← the number of reply
7.   RO ← the number of reply written by other users
8.   Importance ← L + R × (1 + RO / R) /* Eq. (1) */
9.   Divide contents of POA into sentences
10.  Scores[][] ← ϕ
11.  For each sentence S in the sentences
12.    Calculate TextRank of S and Normalize TextRank as [0, 1]
13.    TR ← the normalized TextRank score
14.    Extract keywords from S
15.    For each keyword key of the extracted keywords
16.      Scores[key][score] ← Scores[key][score] + TR
17.      Scores[key][count] ← Scores[key][count] + 1
18.  For each Scores[key] in Scores
19.    Calculate TF-IDF of key
20.    tf·Idf ← the calculated TF-IDF value
21.    Representative ← tf·Idf × (1 + Scores[key][score] / Scores[key][count]) /* Eq. (2) */
22.    Ranking ← (Importance × ratio) + (Representative × (1 - ratio)) /* Eq. (3) */
23.    Save information of Ranking and OA into inverted index file for key

```

Fig. 2 Algorithm for constructing index with the ranking algorithm

of all *sentence* including *key*, and  $S(key)$  is the number of all *sentence*. Then, final ranking score of keywords is calculated through following equation based on Eqs. (1) and (2).

$$Ranking(key) = (Importance(OA) \times r) + (Representative(key) \times (1-r)) \quad (3)$$

In the equation,  $r$  is the adoption rate of two ranking factors. After this process of ranking module, indexing module creates inverted index files with respect to each keyword with information such as the ranking scores for construction of index. The constructed index is used by searching module to provide service of keyword based opinion article retrieval. Figure 2 shows our ranking algorithm for constructing index with ROU algorithm of the framework.

## 4 Performance Evaluation

In performance evaluation, all experiments were performed on 3.3 GHz Intel processor with 8 GB main memory, and run with Microsoft Windows 7 operating system. Algorithms are implemented in C++ language. In addition, about 58,000 opinion articles have been collected from Amazon. To evaluate performance of algorithms, we compare our ROU algorithm with RLRank [6] and TF-IDF [1]. Common settings for performance evaluation are as follows. First, top-50 searching results are extracted in respect to sampled keywords, which are related to product names selected from categories in the Amazon. Second, the number of relevance articles is counted from the searched results. In the experiments, relevance article is an article containing contents closely related to a given keyword. In addition, it is an article having no less than the average number of evaluations such as helpful in regard to the all articles in the collected dataset, and the average number is 5.414 in our collected dataset. Third, performances of algorithms are measured according to precision and NDCG.

We first perform precision test of ROU with RLRank and TF-IDF. For given retrieved results, precision is defined as the percentage of retrieved relevant articles to the results. In the experiment, five sampled keywords are used. The left figure of Fig. 3 shows the results of precision evaluation. From the figure, we can observe that our ROU algorithm outperforms RLRank and TF-IDF in the sampled keywords.

Next, we evaluate performance of compared algorithms in terms of Normalized Discounted Cumulate Gain (NDCG). NDCG is used to measure performance of IR systems using graded relevance and ranking order since users usually refer to the top results as important, not the bottom results. In the right figure of Fig. 3, the proposed algorithm, ROU shows better NDCG results than the previous algorithms in the sampled keywords. Although previous algorithms show better performance with respect to a keyword “jewelry”, all performances of compared algorithms is almost the same. It means that the more retrieved relevant articles appear in the top of the results.

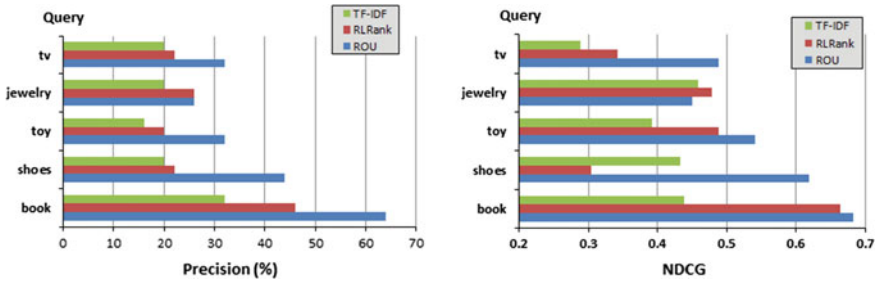


Fig. 3 Precision and NDCG evaluations

## 5 Conclusions

In this paper, we proposed unsupervised keyphrase extraction based ranking algorithm, ROU which measures the importance in three aspects, the amount of information in articles, representativeness of sentences, and frequency of words. To reflect the representativeness to ranking, ROU calculates TextRank scores of sentences. Moreover, we conducted precision and NDCG experiments for performance evaluation. The experimental results showed that our ranking algorithm, ROU outperformed previous ranking algorithms. In addition, our framework can provide service of keyword based opinion article retrieval.

**Acknowledgements** This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

## References

1. Aizawa AN (2003) An information-theoretic perspective of Tf-idf measures. *J Info Process Manage* 39(1):45–65
2. Eirinaki M, Pital S, Singh J (2012) Feature-based opinion mining and ranking. *J Comp Syst Sci* 78(4):1175–1184
3. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. *J ACM* 46(5):604–632
4. Mihalcea R, Tarau P (2004) TextRank: bringing order into text. In: *Proceedings of EMLNP 2004*, Barcelona, pp 404–411
5. Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the Web. Technical report, Stanford InfoLab
6. Yun U, Ryang H, Pyun G, Lee G (2012) Efficient opinion article retrieval system. *Lecture Note in Computer Science*. In: *Proceedings of ICHIT 2012*, Daejeon. pp 566–573
7. Zhang L, Liu B, Lim SH, O’Brien-Strain E (2010) Extracting and ranking product features in opinion documents. In: *Proceedings of COLING 2010*, Beijing, pp 1462–1470

# A Frequent Pattern Mining Technique for Ranking Webpages Based on Topics

Gwangbum Pyun and Unil Yun

**Abstract** In this paper, we propose a frequent pattern mining technique for ranking webpages based on topics. This technique shows search results according to selected topics in order to give users exact and meaningful information, where we use an indexer with the frequent pattern mining technique to comprehend webpages' topics. After mining frequent patterns related to topics (i.e. frequent topics) in collected webpages, the indexer compares new webpages with the generated patterns and calculates degree of topic proximity to rank the new ones, where we also propose a special tree structure, named RP-tree, to compare the new webpages to the frequent patterns. Since our technique reflects topic proximity scores to ranking scores, it can preferentially show webpages which users want.

**Keywords** Frequent pattern mining · Ranking · RP-tree · Topic search

## 1 Introduction

Information retrieval algorithms find and show users webpages related to their needs. Previous information retrieval algorithms generally use a method that finds webpages through keywords inputted by users, and the found results contain a number of webpages with unrelated topics. A topic search is a method which provides webpages related to specific topics selected by users. To conduct the topic search, topic analysis of webpages is needed. FIIR [1], one of topic analysis

---

G. Pyun · U. Yun (✉)

Department of Computer Science, Chungbuk National University,  
Chungbuk, Republic of Korea  
e-mail: yunei@chungbuk.ac.kr

G. Pyun

e-mail: pyunb@chungbuk.ac.kr



methods, uses a word dictionary associated with a specific topic and assigns weights to webpages according to the number of food-related words. However, this method has a limitation that does not consider word combinations since it analyzes topics in terms of the number of related words. In this paper, we propose a frequent pattern mining technique for precisely analyzing webpages' topics and reasonably assigned weights. The technique mines frequent word patterns from topic-related webpages and calculates topic scores for new webpages through the mining results. Then, the topic score is applied as a weight for ranking, and our algorithm computes exact ranking scores by using the weight. As related works, DTM [2] preferentially shows webpages for frequently accessed topics by considering queries inputted by users so far. In LDA [3], a method for finding hidden topics through webpage's domain information was proposed. LDA can discover the hidden information by calculating relations between users' queries and webpages. As a food information retrieval algorithm, FIIR [1] computes the number of words matched with its food dictionary and sets weights to webpages. In contrast to FIIR which only uses word quantity information derived from the food dictionary. FP-growth [4], one of the frequent pattern mining methods, finds meaningful patterns.

## 2 Frequent Pattern Mining Technique for Analyzing Topics of Webpages

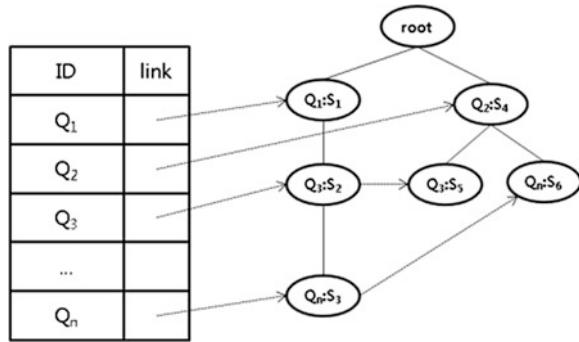
### 2.1 *Extracting Patterns of Webpages Related to Topics*

To analyze topics of webpages, appropriate analysis criteria are required, and we use patterns of webpages associated with topics, where the patterns are composed of special combinations of words in webpages. Therefore, patterns, which occur frequently among patterns derived from a number of topic-related webpages, become important data representing topics. We can determine whether new webpages are associated with a current topic or not, according to the topic score. To find frequent patterns for any topic, an indexer searches them through the frequent pattern mining technique, and for this purpose, we propose a preprocessing method regarding webpage's words.

**Definition 1** (*Preprocessor*) A preprocessor generates unique IDs reflected to words consisting of webpages.

Assuming that any webpage,  $P$  has  $n$  words,  $i$ ,  $P$  is denoted as  $P = \{i_1, \dots, i_n\}$ , where IDs are uniquely assigned for each  $i$ . For  $k < n$ ,  $r < n$ ,  $k \neq r$ , any ID,  $Q$  is expressed as  $Q = k$ , iff  $Q \neq r$ , and vice versa. After the preprocessor converts words of any webpage as IDs, the webpage is denoted as a set of IDs, and the set from one webpage is considered as one transaction. Our indexer extracts frequent patterns by pattern mining approach after constructing a database from the webpages. However, general pattern mining approach generates too many frequent

Fig. 1 RP-tree structure



patterns. To solve the disadvantage, we propose an efficient comparison method using a new tree structure, named RP-tree.

**Definition 2** RP-tree (Result Pattern–tree) is a tree structure for storing frequent patterns.

RP-tree’s basic structure is shown in Fig. 1, where the tree has a link table and a tree, and the link table has an ID list composing RP-tree. Note that sorting order of the list is the same as that of frequent pattern mining (i.e. support descending order). Each ID in the link table has a pointer, and the pointer is connected to all of the nodes with the same ID. RP-tree’s pattern storage procedure is as follows. We first create a link table according to sort order of frequent patterns, and then insert frequent patterns’ IDs into a tree, starting from a root. When IDs are inserted in the tree, we assign 0 if a new node is generated, and we only change the support for not middle nodes but the end node after any transaction is inserted. Thereby, the node where the last ID of the frequent pattern is entered has support information for the pattern. Thus, RP-tree represents the characteristics as shown in Lemma 1.

**Lemma 1** Support of all nodes in RP-tree does not interfere in other supports of frequent patterns.

*Proof* Given a frequent pattern,  $F = \{i_1, i_2, \dots, i_k\}$  with a length,  $k$ , and its support,  $S$ , the only node with  $i_k$  has  $S$  in RP-tree. Let us assume that a sub-pattern of  $F$ ,  $F' = \{i_1, i_2, \dots, i_{k-1}\}$  and a super pattern of  $F$ ,  $F'' = \{i_1, i_2, \dots, i_{k+1}\}$  are inserted into the tree after  $F$ . Then, the nodes with  $i_{k-1}$  and  $i_{k+1}$  are set as their corresponding supports. However, these two patterns do not have any effect in the support of  $F$ . The reason is that RP-tree only updates the last node’s support when any transaction is inserted. Thus, since  $F'$  is a sub-pattern of  $F$ , there is no change for  $F$ . In the case of  $F''$ ,  $F''$  has an effect on the node with  $i_k$ , since  $F''$  is a super pattern of  $F$  and has all items in  $F$ . However, any problem is not caused as in the cases of  $F'$  since RP-tree only updates a support for the last node with  $i_{k+1}$ , not changing the node with  $i_k$ . Namely, the information for  $F$  is preserved. As a result, RP-tree does not cause any interference among all of the patterns.  $\square$

## 2.2 Analyzing Topics of Webpages by RP-Tree

If a web robot collects new webpages, our indexer compares the new webpages with the frequent patterns in RP-tree and computes topic scores, where our algorithm converts words of the new ones to ID forms to calculate topic scores. Thereafter, the generated IDs are sorted according to RP-tree's sort order. Then, the algorithm selects IDs from the bottom one by one and finds nodes with the current ID as using the link table containing ID and link information.

**Definition 3** (*Matching List*) A matching list stores a set of paths matched with new webpages as traversing RP-tree and supports corresponding to the paths.

When the algorithm visits nodes linked to the selected ID, IDs and supports for the nodes are stored into the matching list. After that, a pointer moves to the parent of the current node, and its ID is added in the matching list if the parent has ID of the new webpage. These operations are iterated by the root. For all of the IDs from the link table, the above steps are performed. Then, the finally generated matching list has match information between the new webpages and RP-tree's frequent patterns. If any webpage is related to the current topic, the patterns in the matching list have long lengths and high supports. The topic score,  $W$  is computed as

$$W = \text{maximal length} * (1 + \text{sum of maximal patterns' supports})$$

where *maximal length* means a length value of the pattern with the longest length in the matching list and *sum of maximal patterns' supports* are to add all of the patterns' supports with *maximal length*. We add 1 to the sum to increase an effect by *maximal length*. In the equation, *sum of maximal patterns' supports* has a value between 0 and 1, and we assign higher scores to frequent patterns with 2 or more lengths rather than 1 length. Topic scores and words' supports are saved into an inverted file, and they are used as important factors when we consider ranking scores. As an example, the left list in Fig. 2 represents that any new webpage is converted to its IDs. Then, we first select the last item in the ID list, G, and search the tree by using the link table as shown in the figure. After searching the tree from each node with G to the root, {A, E, G} and {G} are included in the matching list, where the corresponding *maximal length* and *sum of maximal patterns' supports* are 3 and 0.042 respectively. Thus, its topic score,  $W$  is calculated as  $W = 3 * (1 + 0.042) = 3.126$ . Figure 3 shows an algorithm for analyzing webpages' topics and generating an inverted file. The algorithm mines frequent patterns related to the current topic as using the FP-growth method [4], and then RP-tree is generated though the mined patterns. If new webpages are collected, the algorithm removes a few words unrelated to the topic in the webpages and converts them as IDs. After comparing the ID list to the RP-tree, the score of corresponding topic is computed, and finally, an inverted file for P is generated.

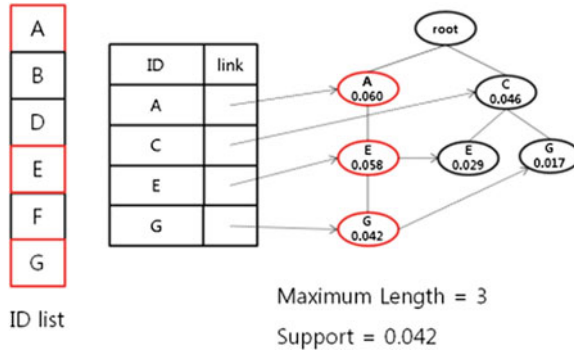


Fig. 2 A webpage pattern and its RP-tree

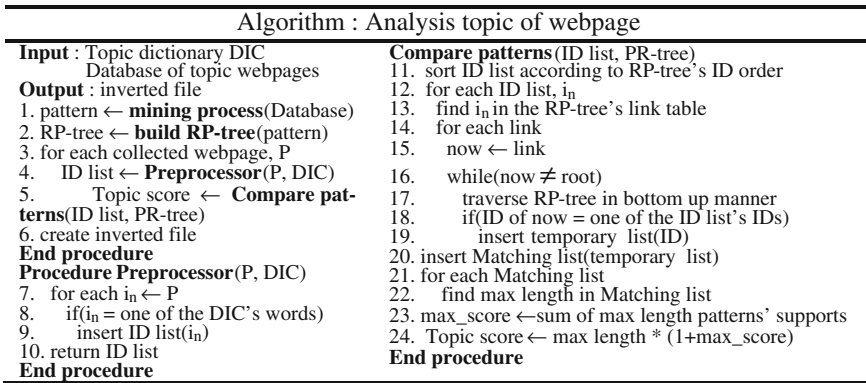


Fig. 3 Algorithm for topic analysis

### 2.3 Ranking Technique by Frequent Pattern Mining

In general information retrieval, a ranking score is computed by multiplying a frequency by idf [5]. However, the above calculation is not suitable for the topic search in this paper. Therefore, we calculate ranking scores as reflecting topic scores computed through frequent patterns, and the proposed ranking score, RS is as follows.

$$RS = tf * idf * \ln(W + e)$$

Here, tf, idf, and W are a relative frequency reflecting how many a searched word is included in a webpage, an inverted document frequency related to a searched word, and a topic score comparing frequent patterns for topics with the current webpage respectively. W is utilized as a weight condition changing ranking scores of webpages, and we need to adjust this value to obtain better

scores. If  $W$  is too low, a role as a weight condition for topics is lost. In contrast, if it is too high, webpages simply related to topics are presented regardless of user-inputted keywords. To solve these problems, we use the natural logarithm, where this is a logarithm which has a natural constant,  $e$  as a base and is used as a normal distribution value in statistics. Accordingly, we adjust  $W$  by using the natural logarithm to compute reasonable ranking scores which both contain as many queries as possible and consider topics. However, the ranking score using the adjusted  $W$  has a limitation, if  $0 \leq W < e$ . Topic scores of webpages unrelated to topics become 0 while those highly related to topics are more than 0. However, if  $W < e$ , the ranking score is lower than the score when  $W$  is not included. Therefore, our ranking technique computes the scores as  $tf * idf$ , if  $W = 0$ , while it reflects  $W$  to  $RS$  after adding  $e$ , if  $W > 0$ .

### 3 Performance Evaluation

In this section, we compare our FPR (Frequent Pattern mining for Ranking webpages by topics) with the other topic search algorithm, FIIR [1], in terms of precision, recall, and NDCG. Webpages used in these experiments were gathered from [www.washingtonpost.com](http://www.washingtonpost.com) between 01/01/2011 and 12/31/2011, where the number of webpages is 32,248. In addition, “A Dictionary of food [6]” was utilized for the evaluation. FPR first performs frequent pattern mining process to construct RP-tree, and then calculates ranking scores based on the food dictionary and the comparison technique by RP-tree. Queries used in the experiments are {Travel, Friday, Beach, Young, Water, Chief, Car, Train Apple, Meat, Food, Corn}. X-axis in Figs. 4 and 5 denotes the above 12 queries in sequence. To evaluate precision [7], we measure ratios of food-related webpages in the top-30 webpages gained from the two algorithms. In Fig. 4, FPR shows outstanding precision compared to FIIR [1] in all of the cases except for the 9th query, “Apple”. In the recall test [8], given webpages including both the current query and food-related information, we calculate ratios for how many food-related webpages exist in the top-30 webpages found by the algorithms. Figure 5

Fig. 4 Precision test

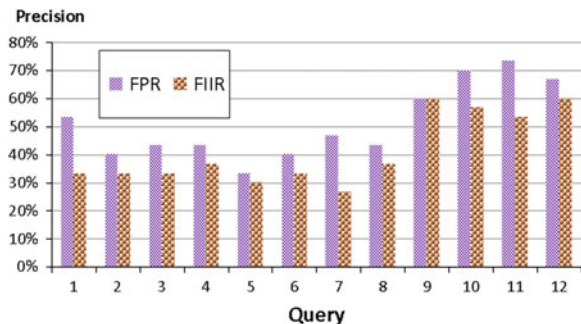


Fig. 5 Recall test

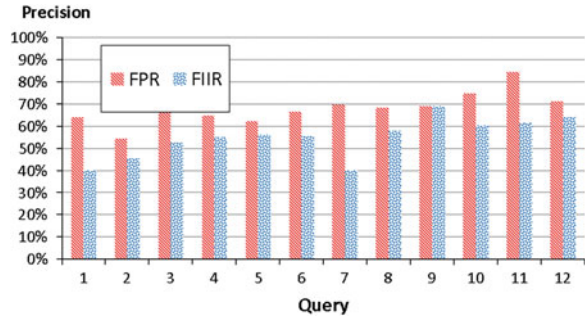
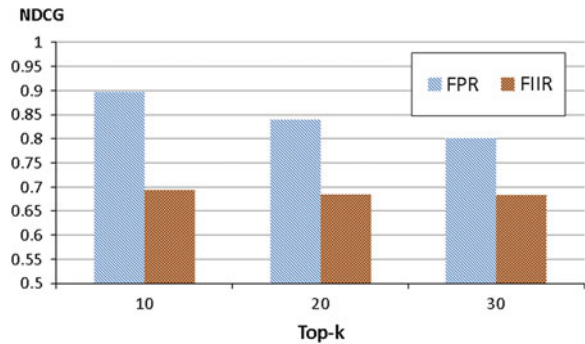


Fig. 6 NDCG test



represents recall results, where FPR guarantees higher recall ratios than those of FIIR in most cases. NDCG [8] has a high value if any webpage with food information is located in upper ranks while it has a low value if this is in lower ranks. A food-related score,  $rel_k$  is set as 0 when there is no information related to food, 1 when there is food-related information in the other topics, 2 when the topic is only contained to food, or 3 when the topic belongs to food and there are many food-related contents. Figure 6 shows average values of NDCG for the 12 queries, where top-k is set as 10–30. In this result, FPR also guarantees outstanding NDCG performance compared to FIIR in every case.

## 4 Conclusion

In this paper, we proposed a frequent pattern mining technique and the corresponding algorithm which can rank webpages based on topics. To analyze topics exactly, we conducted frequent pattern mining operations regarding topic-related webpages, and thereafter we calculated topic scores by comparing webpages with the mined frequent patterns and applied the topic scores to ranking scores. In the various experiments, it was observed that our algorithm outperforms the previous topic-based algorithm in terms of precision, recall, and NDCG. Through the

proposed techniques and algorithm, we expect that they will contribute to improving the level of both information retrieval and frequent pattern mining fields.

**Acknowledgment** This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

## References

1. Pyun G, Yun U (2011) Efficient food retrieval techniques considering relative frequencies of food related words. *Lect note Comput Sci* 368–375
2. Chen KY, Wang HM, Chen B (2012) Spoken document retrieval leveraging unsupervised and supervised topic modeling techniques. *IEICE Trans* 1195–1205
3. Andrzejewski D, Buttler D (2011) Latent topic feedback for information retrieval. *Knowl Discovery Data Min* 600–608
4. Han J, Pei J, Yin Y, Mao R (2004) Mining frequent patterns without candidate generation: a frequent pattern tree approach. *DMKD* 8(1):53–87
5. Donald M (2008) Generalized inverse document frequency. Paper presented at conference on information and knowledge management, pp 399–408
6. Charles S (2005) *A dictionary of food: international food and cooking terms from A to Z*, 2nd edn. A&C Black Publishers Ltd
7. Kato MP, Ohshima H, Tanaka K (2012) Content-based retrieval for heterogeneous domains: domain adaptation by relative aggregation points. Paper presented at ACM SIGIR conference on Research and development in information retrieval, pp 811–820
8. Croft WB, Metzler D, Strohman T (2010) *Search engines: information retrieval in practice*. Addison-Wesley, Boston

# Trimming Prototypes of Handwritten Digit Images with Subset Infinite Relational Model

Tomonari Masada and Atsuhiko Takasu

**Abstract** We propose a new probabilistic model for constructing efficient prototypes of handwritten digit images. We assume that all digit images are of the same size and obtain one color histogram for each pixel by counting the number of occurrences of each color over multiple images. For example, when we conduct the counting over the images of digit “5”, we obtain a set of histograms as a *prototype* of digit “5”. After normalizing each histogram to a probability distribution, we can classify an unknown digit image by multiplying probabilities of the colors appearing at each pixel of the unknown image. We regard this method as the baseline and compare it with a method using our probabilistic model called Multinomialized Subset Infinite Relational Model (MSIRM), which gives a prototype, where color histograms are clustered column- and row-wise. The number of clusters is adjusted flexibly with Chinese restaurant process. Further, MSIRM can detect *irrelevant* columns and rows. An experiment, comparing our method with the baseline and also with a method using Dirichlet process mixture, revealed that MSIRM could neatly detect irrelevant columns and rows at peripheral part of digit images. That is, MSIRM could “trim” irrelevant part. By utilizing this trimming, we could speed up classification of unknown images.

**Keywords** Bayesian nonparametrics · Prototype · Classification

---

T. Masada (✉)

Nagasaki University, 1-14 Bunkyo-machi, Nagasaki-shi, Nagasaki 852-8521, Japan  
e-mail: masada@nagasaki-u.ac.jp

A. Takasu

National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan  
e-mail: takasu@nii.ac.jp



## 1 Introduction

This paper considers image classification. While there are a vast variety of methods, we focus on *prototype*-based methods. We construct a prototype for each image category and classify an unknown image to the category whose prototype is the most similar. In this paper, we describe prototypes with probability distributions and classify an unknown image to the category whose prototype gives the largest probability.

We assume that all images are of the same size, say  $N_1$  by  $N_2$  pixels, because we consider *handwritten digit images* in this paper. We can obtain a set of color histograms by counting the number of occurrences of each color at each pixel. Consequently, we obtain  $N_1N_2$  color histograms, each at a different pixel. When we conduct this counting over the images of the same category, e.g. the images of handwritten digit “5”, the resulting set of histograms gives a color configuration specific to the category and can be regarded as a *prototype* of the category. Formally, we describe each prototype with parameters  $\{g_{ijw}^h\}$ , where  $g_{ijw}^h$  is a probability that the  $w$  th color appears at the 2D pixel location  $(i, j)$  for  $w = 1, \dots, W$ ,  $i = 1, \dots, N_1$ ,  $j = 1, \dots, N_2$ , where  $W$  is the number of different colors. The superscript  $h$  is a category index. By using the parameters, we can calculate the log probability of an unknown image as  $\sum_{i,j,w} n_{ijw} \ln g_{ijw}^h$ , where  $n_{ijw}$  is 1 if the image has the  $w$  th color at the pixel  $(i, j)$  and 0 otherwise. Based on the obtained probabilities, a category for the image can be determined by  $\arg \max_h \sum_{i,j,w} n_{ijw} \ln g_{ijw}^h$ . We regard this method as the baseline method and would like to improve it in terms of *efficiency*.

Any prototype the baseline gives has  $N_1N_2W$  parameters, whose number can be reduced by *clustering* histograms. We propose a new probabilistic model called Multinomialized Subset Infinite Relational Model (MSIRM) for clustering. MSIRM clusters histograms column- and row-wise. Denote the numbers of column and row clusters as  $K_1$  and  $K_2$ , respectively. In MSIRM, each pixel is assigned to a pair of column and row clusters, and the colors of the pixels assigned to the same pair of column and row clusters are assumed to be drawn from the same distribution. Consequently, we can reduce the complexity of prototypes from  $N_1N_2W$  to  $K_1K_2W$ . MSIRM determines  $K_1$  and  $K_2$  flexibly with Chinese restaurant process (CRP) [4]. MSIRM has an important feature: it detects columns and rows *irrelevant* for constructing prototypes. This feature can speed up classification, because we can reduce execution time of classification by skipping irrelevant columns and rows. The skipping technically means giving probability one to all pixels in irrelevant columns and rows. We will show that the skipping lead to only a small degradation in classification accuracy.

The rest of the paper is organized as follows. [Section 2](#) gives preceding proposals important for us. [Section 3](#) provides details of MSIRM. [Section 4](#) presents the results of our comparison experiment. [Section 5](#) concludes the paper with discussions.

## 2 IRM and SIRM

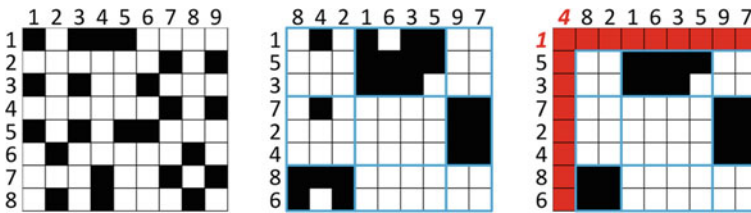
We propose MSIRM as an extension of Subset Infinite Relational Model (SIRM) [1], which is, in turn, an improvement of Infinite Relational Model (IRM) [2].

Assume that we have  $N_1$  entities of type  $T_A$  and  $N_2$  entities of type  $T_B$ , and that a binary relation  $R$  is defined over a domain  $T_A \times T_B$ . The relation can be represented by an  $N_1$  by  $N_2$  binary matrix as the left most panel of Fig. 1 shows. IRM discovers a column- and row-wise clustering of entities so that the matrix gives a relatively clean block structure when sorted according to the clustering as the center panel of Fig. 1 shows. IRM can determine the numbers of column and row clusters flexibly with CPR. Further, IRM associates each block, enclosed by thick lines in the center panel of Fig. 1, with a binomial distribution determining the probability that the pairs of entities in the block fall under the relation  $R$ . While IRM can be extended to the cases where there are more than two types of entity, we do not consider such cases.

When constructing a prototype from the images of the same category, we can consider a relation between pixel columns and rows by taking an IRM-like approach, because contiguous pixels are likely to give similar color distributions. However, IRM is vulnerable to noisy data. Therefore, a mechanism for detecting *irrelevant* columns and rows is introduced by SIRM, where clustering is conducted only on a *subset* of columns and rows, i.e., on columns and rows other than irrelevant ones. In SIRM, we flip a coin at each column and row to determine whether relevant or not. Irrelevant pixels are bundled into the same group, as the right most panel of Fig. 1 depicts with red cells. However, both IRM and SIRM can only handle binary data. Therefore, we “multinomialize” SIRM for handling multiple colors and obtain our model, MSIRM.

## 3 MSIRM

We below describe how observed data are generated by MSIRM.



**Fig. 1** Clustering of a binary relation (*left*) by IRM (*center*) and by SIRM (*right*). Each cluster is enclosed by *thick lines*. *Red cells* in the *right* most panel correspond to irrelevant pixels

1. Draw the parameter  $\lambda_1$  of a binomial  $\text{Bi}(\lambda_1)$  from a Beta prior  $\text{Be}(a_1, b_1)$ . Then, for each column, draw a 0/1 value from  $\text{Bi}(\lambda_1)$ . Let  $r_{1i}$  denote the value for the  $i$ th column, which is irrelevant if  $r_{1i} = 0$  and is relevant otherwise.
2. Draw the parameter  $\lambda_2$  of a binomial  $\text{Bi}(\lambda_2)$  from a Beta prior  $\text{Be}(a_2, b_2)$ . Then, for each row, draw a 0/1 value from  $\text{Bi}(\lambda_2)$ . Let  $r_{2j}$  denote the value for the  $j$ th row, which is irrelevant if  $r_{2j} = 0$  and is relevant otherwise.
3. Draw the parameters  $\phi_{kl1}, \dots, \phi_{klW}$  of a multinomial distribution  $\text{Mul}(\phi_{kl})$  from a Dirichlet prior  $\text{Dir}(\beta)$ .  $\text{Mul}(\phi_{kl})$  is the multinomial for generating colors of the pixels belonging to the  $k$ th column and, at the same time, to the  $l$ th row clusters.
4. Draw the parameters  $\psi_1, \dots, \psi_W$  of a multinomial  $\text{Mul}(\psi)$  from a Dirichlet prior  $\text{Dir}(\gamma)$ .  $\text{Mul}(\psi)$  is the multinomial for generating colors of the irrelevant pixels.
5. For each relevant column, draw a cluster ID based on a Chinese restaurant process  $\text{CRP}(\alpha_1)$ . We introduce a latent variable  $z_{1i}$ , which is equal to  $k$  if the  $i$ th column is relevant and belongs to the  $k$ th column cluster.
6. For each relevant row, draw a cluster ID based on a Chinese restaurant process  $\text{CRP}(\alpha_2)$ . We introduce a latent variable  $z_{2j}$ , which is equal to  $l$  if the  $j$ th row is relevant and belongs to the  $l$ th column cluster.
7. For each pixel  $(i, j)$ , draw a color from the multinomial  $\text{Mul}(\psi)$  if  $r_{1i}r_{2j} = 0$  (i.e., the pixel is irrelevant) and from the multinomial  $\text{Mul}(\phi_{z_{1i}z_{2j}})$  otherwise.

We adopt Gibbs sampling technique for inferring posterior distribution of MSIRM. For each column, we update the relevant/irrelevant coin flip  $r_{1i}$  and the cluster assignment  $z_{1i}$  based on the following posterior probability:

$$p(z_{1i}, r_{1i} | \mathbf{X}, \mathbf{Z}^{\setminus i}, \mathbf{R}^{\setminus i}) \propto p(\mathbf{X}^{+i} | z_{1i}, r_{1i}, \mathbf{X}^{\setminus i}, \mathbf{Z}^{\setminus i}, \mathbf{R}^{\setminus i}) p(z_{1i}, r_{1i} | \mathbf{Z}^{\setminus i}, \mathbf{R}^{\setminus i}),$$

where the details of the two terms on the right hand side are omitted due to space limitation. The derivation is almost the same with that for SIRM [1]. We also conduct a similar update for each row. Additionally, we update the hyperparameters of the Dirichlet and the Beta priors by Minka's method.<sup>1</sup>

## 4 Experiment

Our experiment compared MSIRM with the baseline and also with a method using Dirichlet process mixture of multinomial (DP-multinomial) [3] for clustering histograms. Our target was MNIST data set,<sup>2</sup> consisting of 60,000 training and

<sup>1</sup> <http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/>

<sup>2</sup> <http://yann.lecun.com/exdb/mnist/>

10,000 test images. All images are  $28 \times 28$  pixels in size, i.e.,  $N_1 = N_2 = 28$ . We quantized 8 bit gray-scaled colors into 4 bit colors uniformly, i.e.,  $W = 16$ .

Let  $n_{ijw}^h$  be the number of times the  $w$ th color appears at the pixel  $(i, j)$  in the training images of category  $h$ , where  $h$  ranges over ten digit categories. The baseline method calculates a probability  $g_{ijw}^h$  that the  $w$ th color appears at the pixel  $(i, j)$  in the images of category  $h$  as  $g_{ijw}^h = (n_{ijw}^h + \eta_w) / \sum_w (n_{ijw}^h + \eta_w)$ , where a Dirichlet smoothing with the parameters  $\eta_1, \dots, \eta_W$  is applied. We determine the category of an unknown image by  $\arg \max_h \sum_{i,j} \hat{n}_{ijw} \ln g_{ijw}^h$ , where  $\hat{n}_{ijw}$  is 1 if the  $w$ th color appears at the pixel  $(i, j)$  in the unknown image and is 0 otherwise. For the DP-multinomial method, we set  $g_{ijw}^h$  to a posterior probability estimated by a Gibbs sampling [3] and determine the category of an unknown image in the same manner with the baseline method. For MSIRM, we determine the category of an unknown image as follows:

$$\arg \max_h \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \hat{n}_{ijw} \left\{ r_{1i} r_{2j} \ln \frac{n_{z_{1i}z_{2j}w}^h + \beta_w^h}{\sum_w (n_{z_{1i}z_{2j}w}^h + \beta_w^h)} + (1 - r_{1i} r_{2j}) \frac{q_w^h + \gamma_w^h}{\sum_w (q_w^h + \gamma_w^h)} \right\},$$

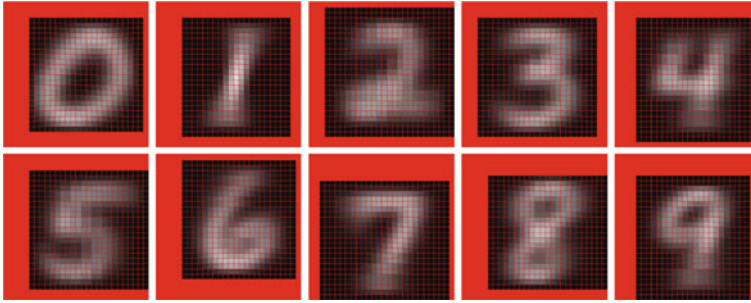
where  $n_{klw}^h$  is the number of times the  $w$ th color appears at the pixels belonging to the  $k$ th column and the  $l$ th row clusters in the training images of category  $h$ .  $q_w^h$  is the number of times the  $w$ th color appears at the irrelevant pixels of the training images of category  $h$ .  $\beta_w^h$  and  $\gamma_w^h$  are the hyperparameters for category  $h$ .

We give results of the experiment. Let  $K$  be the number of clusters given by DP-multinomial. With respect to the complexity of prototypes, DP-multinomial was superior to MSIRM. While  $K_1 \approx 20$  and  $K_2 \approx 20$  for MSIRM,  $K \approx 85$  for DP-multinomial with respect to all digits. That is,  $K < K_1 K_2$  for all digits.

Classification accuracies of the baseline, DP-multinomial, and MSIRM were 0.840, 0.839 and 0.837 respectively. While the accuracies were far from the best reported at the Web site of the data set, they were good enough for a meaningful comparison. MSIRM and DP-multinomial gave almost the same accuracies with that of the baseline. Therefore, it can be said that we could reduce the complexity of prototypes by clustering histograms without harming clustering accuracy.

While DP-multinomial and MSIRM showed no significant difference in terms of accuracy, MSIRM could neatly detect irrelevant columns and rows at peripheral part of images. Figure 2 visualizes the prototypes obtained by MSIRM. We mixed grayscale colors linearly by multiplying their probabilities at each pixel and obtained the visualization in Fig. 2. The red-colored area corresponds to irrelevant columns and rows. Astonishingly, MSIRM precisely detected the area irrelevant for digit identification.

By utilizing this ‘‘trimming’’ effect, we could speed up classification. We skipped irrelevant columns and rows when we calculated the log probabilities of unknown images. Technically, this means that we assigned probability one to all irrelevant pixels. For the prototypes in Fig. 2, we could skip 32.2 % of the  $28^2 \times 10 = 7,840$  pixels. This led to 32.2 % speeding up of classification.



**Fig. 2** Irrelevant portion (*red-colored pixels*) trimmed out by MSIRM from each prototype

The achieved accuracy was 0.819, which shows a small degradation from 0.837. Therefore, it can be said that MSIRM could speed up classification with only a small degradation in accuracy.

## 5 Conclusion

This paper proposes a prototype-based image classification using MSIRM. We could neatly detect irrelevant pixels and could speed up classification by skipping the irrelevant pixels. This led to only a small degradation in accuracy. While DP-multinomial was comparable with MSIRM in accuracy and was superior to MSIRM in reduction of the complexity of prototypes, it has no mechanism for detecting irrelevant pixels.

We have a future plan to extend MSIRM for realizing a clustering of training images by introducing an additional axis aside from the column and the row axes.

## References

1. Ishiguro K, Ueda N, Sawada H (2012) Subset infinite relational models. In: Proceedings of AISTATS 2012, JMLR W&CP 22, pp 547–555
2. Kemp C, Tenenbaum JB, Griffiths TL, Yamada T, Ueda N (2006) Learning systems of concepts with an infinite relational model. In: Proceedings of AAAI'06. p 381–388
3. Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. *J Comput Graph Stat* 9(2):249–265
4. Pitman J (2002) Combinatorial stochastic processes. Notes for Saint Flour Summer School

# Ranking Book Reviews Based on User Influence

Unil Yun and Heungmo Ryang

**Abstract** As the number of people who buy books online increases, it is becoming a way of life for people. Recently, algorithms to retrieve book reviews have been proposed for searching meaningful information since ranking algorithms for general purpose are not suitable. Although the previous algorithms consider features of book review, they calculate ranking scores without reflecting user influence. In this paper, thus, we propose a novel algorithm for ranking book reviews based on user influence. To apply user influence, the proposed algorithm uses recommendations by other users. For performance evaluation, we perform precision and recall tests. The experimental results show that the proposed algorithm outperforms previous algorithms for searching book reviews.

**Keywords** Book review · Information retrieval · User influence

## 1 Introduction

Since World Wide Web has emerged, the number of documents on the Internet has been increasing exponentially. In face of the overwhelming information volumes, people focus on solving information overload rather than its shortage, and it is a difficult task to find meaningful information on the Internet. Thus, the importance of searching relevant and useful documents has risen. IR (Information Retrieval) systems look for the information by measuring its importance. If given a query of user, general search engines evaluate the relation between the query and indexed

---

U. Yun (✉) · H. Ryang

Department of Computer Science, Chungbuk National University, Chungcheongbuk-do, South Korea

e-mail: yunei@chungbuk.ac.kr

H. Ryang

e-mail: riangs@chungbuk.ac.kr

documents on the Internet, and it lists results based on the relation in descending order of ranking scores. This process is called ranking documents, and algorithms used for such retrieval are called ranking algorithms. Various ranking algorithms [1, 2, 4–7] have been proposed, and some of them are based on references or quotations between documents through hyperlinks [1, 2, 5]. One of the most famous algorithms is PageRank [5] adopted by Google (<http://www.google.com>) and it becomes a fundamental algorithm of IR. Although PageRank has played an important role in IR area, it is not suitable for the system finding meaningful book reviews since there are few references or quotations between book reviews. To address this issue, algorithms, LengthRank and ReplyRank [6], were proposed, and the algorithms reflect features of book review. Nevertheless, ranking scores are calculated without reflecting influence of users in the algorithms. That is, only the features of book review are considered in the ranking scores. For this reason, we propose a novel algorithm for ranking book reviews based on user influence. Major contributions of this paper are summarized as follows: (1) a novel algorithm, called IRRank (Influence based Review Rank), is proposed and (2) various experiments are conducted for performance evaluation of the proposed algorithm.

The remainder of this paper is organized as follows. In Sect. 2, we introduce the influential related work. The proposed ranking algorithm is described in Sect. 3. In Sect. 4, we show performance of the proposed algorithm through various experimental evaluations. Finally our contributions are summarized in Sect. 5.

## 2 Related Work

In the IR field, extensive studies [1–3, 7] have been proposed to search and retrieve relevant information, and ranking algorithms are divided into content-based and inbound link-based methods. Content-based ranking algorithms determine the relevant degrees between user queries and documents through inside information such as the frequency and distance. TF-IDF [9] is a popular weighting scheme and a numerical statistics which is reflected how important words are to documents in the collection, and it is calculated by multiplying  $tf$  by  $idf$ , where  $tf$  refers to the frequency of words in documents and  $idf$  is a measure as general importance of the words. In this paper, we employ TF-IDF for measuring the importance of words in book reviews. On the other hand, inbound link-based ranking algorithms improve the quality of search results based on external information. They measure the importance of documents based on concept of casting votes such as inbound hyperlinks, as votes from some documents are regarded as having a greater ranking score. RakeRank [5] estimates the ranking score of documents using hyperlinks between the documents. Although PageRank has become the most famous method after it was adopted by Google, it is not suitable for the systems searching meaningful book reviews. The reason is that there are few references or quotations between book reviews by hyperlinks. To address this issue, LengthRank and ReplyRank [6] were proposed for evaluating the importance of book reviews,

and the algorithms consider features of book review regarding the length of contents and the number of reply, respectively. In the algorithms, book reviews which have more length or contain a larger number of replies receive higher ranking. However, they cannot consider user influence, and thus this study aims to evaluate the relevance of book reviews by reflecting user influence.

### 3 Ranking Algorithm Based on User Influence

The framework of the proposed method consists of three steps. In the first step, influence of each user is analyzed. In the second step, features of book review are extracted, and then morphemic analysis is performed. In the last step, ranking scores are calculated using information with the analyzed influence and extracted features. In the following subsections, we first describe and define user influence. Then, we illustrate the proposed ranking algorithm in detail.

#### 3.1 User Influence

The proposed ranking algorithm computes ranking scores by reflecting user influence as well as features of book review. The user influence of a certain review used in the proposed algorithm indicates that the average number of recommendations in reviews written by a user. On the other hand, reply means additional information as well as more interesting in the reviews. That is, people can gain more information about books from reviews having more reply. Especially, reply written by users with high influence can be more useful. The user influence is defined as follows.

$$Influence(user) = \sum Rec(rv) / Cnt(user) \quad (1)$$

In the equation, *user* is a certain user, *rv* is a review written by *user*, *Rec(rv)* is the number of recommendations of *rv* such as *helpful* evaluation in Amazon (<http://www.amazon.com>) which is a well-known online book store, and *Cnt(user)* is the total number of book reviews written by *user*. For example, if there is a certain user who wrote *n* reviews and the total number of recommendations of the reviews is *r*, influence of the user,  $Influence(user) = r/n$ . Figure 1 is an example of user influence.

#### 3.2 Ranking Book Reviews

For calculating ranking score, it is needed to compute influence of each user. Thus, in the first step of the framework, the influence is calculated with Eq. (1). First, each user who wrote book review or reply is extracted from the collected dataset.



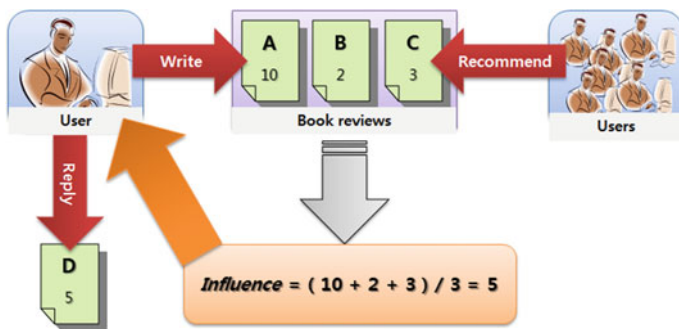


Fig. 1 Example of user influence

Then, the number of reviews written by each user is counted; at the same time, the total number of recommendations with respect to the reviews is computed. After that, values of user influence are calculated. The importance of book reviews can be represented by two aspects. The first is influence of users who wrote reviews or reply. This aspect indicates how an influential user wrote the reviews and how many influential other users participate in the reviews through reply. The other aspect is how much information is contained in the reviews. This aspect can be measured easily by checking the number of replies. On the other hand, keywords involved in each review have to be analyzed and extracted for keyword-based searching. To extract the keyword information, morphemic analysis is performed. The analyzed information is saved to indexed files for employing in the stage of calculating ranking scores and searching book reviews. In addition, each indexed file is created with respect to each keyword except for stop words. In the last step, ranking scores are obtained using the extracted and analyzed information in the indexed files. In this stage, the importance of book reviews is first computed based on user influence. Since the proposed framework is a keyword-based retrieval system, the relevance of keywords is also calculated. Basically, TF-IDF [9] is used for this purpose, and thus we also apply TF-IDF in the proposed ranking algorithm. TF-IDF represents multiplication of the term frequency and the inverse document frequency which is a measure of the general importance of the term as following formula.

$$tf \cdot Idf = \log_2 N - \log_2 d_k + 1 \quad (2)$$

After that, the relevance of keywords is adjusted using the importance of book reviews containing the keywords. To compute the importance of a review  $rv$ , influence of users who are involved in the review and the number of replies are used through the following equation.

$$Irv(rv) = \sum Influence(u_{rep}) + Influence(u_{rv}) + Reply(rv) \quad (3)$$

In the equation,  $Irv(rv)$  is the importance of a review  $rv$ ,  $u_{rv}$  and  $u_{rep}$  refer to a writer of  $rv$  and each user who wrote reply in regard to  $rv$  except for  $u_{rv}$ , respectively, and  $Reply(rv)$  is the number of replies.  $Reply(rv)$  is used as the factor for reflecting the amount of additional information in  $rv$ . Figure 2 shows an algorithm for ranking book reviews based on user influence.

First, we calculate TF-IDF of each keyword in the indexed files (lines 3–10). That is, the importance of keywords is calculated first. After that, the relevance of each review is computed using Eq. (3) (lines 11–20), and then ranking scores of each keyword is obtained by employing both TF-IDF and  $Irv(rv)$  as the following Eq. (4) (line 21).

$$IRRank(keyword, rv) = tf \cdot Idf(keyword) + (tf \cdot Idf(keyword) \times Irv(rv)) \quad (4)$$

## 4 Performance Evaluation

In this section, we provide performance evaluation of IRRank with LengthRank, ReplyRank [6], and Google. Especially, we perform the test with Google by using search operator to limit target dataset as book reviews in online bookstores. We have collected about 114,409 reviews from GoodReads (<http://www.goodreads.com>) which is a collecting book review site and Amazon. Algorithms are written in Microsoft Visual C++ 2010. In addition, they run with the Windows 7 operating system on an Intel Pentium Quad-core 3.2 GHz CPU with 8 Giga bytes main memory.

---

```

Function IRRank( dataset, indexedFiles )
1. IRL ←  $\phi$  /* a list of the importance of each review in dataset */
2. N ← the total number of reviews in dataset
3. For each indexed file idx in indexedFiles
4.   Let kw be a keyword of idx and info be information stored in idx
5.   Count the number of reviews contained kw using info
6.   Set  $d_k$  as the counted value
7.    $idf \leftarrow \log_2 N - \log_2 d_k + 1$  /* Equation (2) */
8.   For each review rv in idx
9.     inf ← 0
10.    Let tf be the frequency of kw in rv
11.     $TF-IDF \leftarrow tf \times idf$ 
12.    Check a value of the importance of rv is stored in the IRL
13.    If  $Irv(rv)$  is not stored in the IRL then
14.       $inf \leftarrow Influence(u_{rv})$  /* Equation (1) */
15.      For each user  $u_{rep}$  in rv
16.        If  $u_{rep}$  is  $u_{rv}$  then continue
17.        Increase inf by  $Influence(u_{rep})$ 
18.      Add inf to IRL
19.       $Irv(rv) \leftarrow inf + Reply(rv)$  /* Equation (3) */
20.    Else Set  $Irv(rv)$  as the value stored in IRL
21.     $IRRank \leftarrow TF-IDF + (TF-IDF \times Irv(rv))$  /* Equation (4) */

```

---

Fig. 2 IRRank algorithm

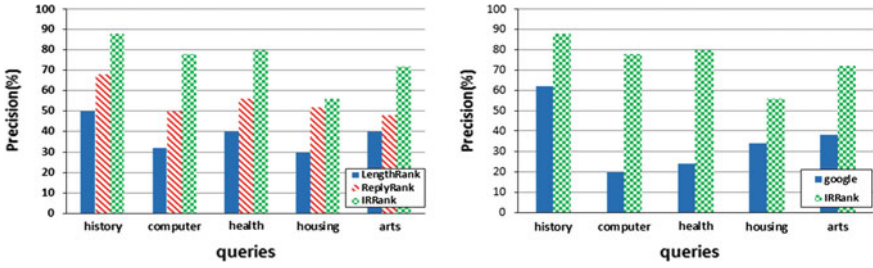


Fig. 3 Precision evaluation

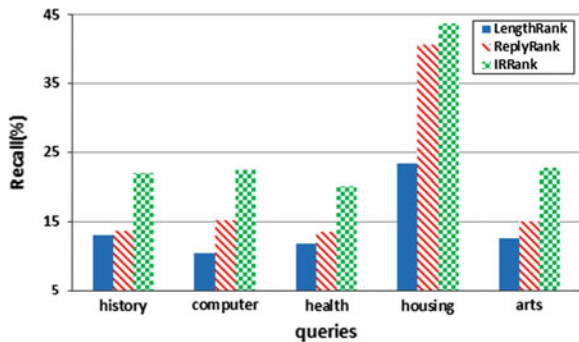
### 4.1 Precision Evaluation

We first compare performance of IRRank with LengthRank, ReplyRank, and Google by evaluating precision. Precision refers to the rate of retrieved relevant reviews by the searched  $K$  reviews. We used five keywords selected from categories in the Amazon and  $K$  is set to 50. Figure 3 is the precision results of IRRank with LengthRank and ReplyRank which use the same dataset and Google, respectively. In this paper, we define the important review as a review having no less than the average value of evaluations such as *helpful* in the collected dataset. From Fig. 3, we can observe that IRRank outperforms other algorithms in the sampled keywords. Especially, the reason in regard to result compared with Google is that ranking technique adopted by Google measures the importance based on references or quotations by hyperlinks while there are little references or quotations between book reviews.

### 4.2 Recall Evaluation

Second, we evaluate performance through recall which is the fraction of relevant reviews to the retrieved reviews with the five queries which are used in the previous experiment. Note that answer dataset is required for performing

Fig. 4 Recall evaluation



evaluation of recall. Thus, we only perform recall experiment with LengthRank and ReplyRank so that they can use the same answer dataset. Figure 4 shows results of recall evaluation, and we can know that IRRank shows better performance than the previous algorithms in the sampled five keywords.

## 5 Conclusions

In this paper, we proposed a ranking algorithm, IRRank based on user influence for evaluating the importance of book reviews. Moreover, we conducted precision and recall experiments for performance evaluation. The experimental results showed that the proposed ranking algorithm, IRRank outperformed previous ranking algorithms in terms of both precision and recall. We expect that our research will take effects to not only searching book reviews but also retrieving area based on user influence.

**Acknowledgments** This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

## References

1. Aktas MS, Nacar MA, Menczer F (2004) Using hyperlink features to personalize web search. *WebKDD*. Seattle, pp 104–115
2. Alyguliev RM (2007) Analysis of hyperlinks and the ant algorithm for calculating the ranks of web pages. *Autom Control Comput Sci* 41(1):44–53
3. Chen MY, Chu HC, Chen YM (2010) Developing a semantic-enable information retrieval mechanism. *Expert Syst Appl* 37(1):322–340
4. Kritikopoulos A, Sideri M, Varlamis I (2009) BLOGRANK: ranking weblogs based on connectivity and similarity features. *CoRR*
5. Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the Web. Technical report, Stanford InfoLab
6. Ryang H, Yun U (2011) Effective ranking techniques for book review retrieval based on the structural feature. *Lecture note in computer science, ICHIT*, pp 360–367
7. Tayebi MA, Hashemi S, Mohades A (2007) B2Rank: an algorithm for ranking blogs based on behavioral features. *Web intelligence*. Silicon Valley, pp 104–107
8. Weng J, Lin EP, Jiang J, He Q (2010) TwitterRank: Finding topic-sensitive influential Twitterers. *WSDM*. New York, pp 261–270
9. Xia T, Chai Y (2011) An improvement to TF-IDF: term distribution based term weight algorithm. *J Softw* 6(3):413–420

# Speaker Verification System Using LLR-Based Multiple Kernel Learning

Yi-Hsiang Chao

**Abstract** Support Vector Machine (SVM) has been shown powerful in pattern recognition problems. SVM-based speaker verification has also been developed to use the concept of sequence kernel that is able to deal with variable-length patterns such as speech. In this paper, we propose a new kernel function, named the Log-Likelihood Ratio (LLR)-based composite sequence kernel. This kernel not only can be jointly optimized with the SVM training via the Multiple Kernel Learning (MKL) algorithm, but also can calculate the speech utterances in the kernel function intuitively by embedding an LLR in the sequence kernel. Our experimental results show that the proposed method outperforms the conventional speaker verification approaches.

**Keywords** Log-Likelihood ratio · Speaker verification · Support vector machine · Multiple kernel learning · Sequence kernel

## 1 Introduction

The task of speaker verification problem is to determine whether or not an input speech utterance  $U$  was spoken by the target speaker. In essence, speaker verification is a hypothesis test problem that is generally formulated as a Log-Likelihood Ratio (LLR) [1] measure. Various LLR measures have been designed [1–4]. One popular LLR approach is the GMM-UBM system [1], which is expressed as

$$L_{\text{UBM}}(U) = \log p(U|\lambda) - \log p(U|\Omega), \quad (1)$$

---

Y.-H. Chao (✉)

Department of Applied Geomatics, Chien Hsin University of Science and Technology,  
Taoyuan, Taiwan  
e-mail: yschao@uch.edu.tw

where  $\lambda$  is a target speaker Gaussian Mixture Model (GMM) [1] trained using speech from the claimed speaker, and  $\Omega$  is a Universal Background Model (UBM) [1] trained using all the speech data from a large number of background speakers. Instead of using a single model UBM, an alternative approach is to train a set of background models  $\{\lambda_1, \lambda_2, \dots, \lambda_N\}$  using speech from several representative speakers, called a cohort [2], which simulates potential impostors. This leads to several LLR measures [3], such as

$$L_{\text{Max}}(U) = \log p(U|\lambda) - \max_{1 \leq i \leq N} \log p(U|\lambda_i), \quad (2)$$

$$L_{\text{Ari}}(U) = \log p(U|\lambda) - \log \left( \sum_{i=1}^N p(U|\lambda_i) / N \right), \quad (3)$$

$$L_{\text{Geo}}(U) = \log p(U|\lambda) - \left( \sum_{i=1}^N \log p(U|\lambda_i) \right) / N, \quad (4)$$

and a well-known score normalization method called T-norm [4]:

$$L_{\text{Tnorm}}(U) = L_{\text{Geo}}(U) / \sigma_U, \quad (5)$$

where  $\sigma_U$  is the standard deviation of  $N$  scores,  $\log p(U|\lambda_i)$ ,  $i = 1, 2, \dots, N$ .

In recent years, Support Vector Machine (SVM)-based speaker verification methods [5–8] have been proposed and successfully found to outperform traditional LLR-based approaches. Such SVM methods use the concept of sequence kernels [5–8] that can deal with variable-length input patterns such as speech. Bengio [5] proposed an SVM-based decision function:

$$L_{\text{Bengio}}(U) = a_1 \log p(U|\lambda) - a_2 \log p(U|\Omega) + b, \quad (6)$$

where  $a_1$ ,  $a_2$ , and  $b$  are adjustable parameters estimated using SVM. An extended version of Eq. (6) using the Fisher kernel and the LR score-space kernel for SVM was investigated in [6]. The supervector kernel [7] is another kind of sequence kernel for SVM that is formed by concatenating the parameters of a GMM or Maximum Likelihood Linear Regression (MLLR) [8] matrices. Chao [3] proposed using SVM to directly fuse multiple LLR measures into a unified classifier with an LLR-based input vector. All the above-mentioned methods have the same point that must convert a variable-length utterance into a fixed-dimension vector before a kernel function is computed. Since the fixed-dimension vector is formed independent of the kernel computation, this process is not optimal in terms of overall design.

In this paper, we propose a new kernel function, named the LLR-based composite sequence kernel, which attempts to compute the kernel function without needing to represent utterances into fixed-dimension vectors in advance. This kernel not only can be jointly optimized with the SVM training via the Multiple Kernel Learning (MKL) [9] algorithm, but also can calculate the speech utterances in the kernel function intuitively by embedding an LLR in the sequence kernel.

## 2 Kernel-Based Discriminant Framework

In essence, there is no theoretical evidence to indicate what sort of LLR measures defined in Eqs. (1)–(5) is absolutely superior to the others. An intuitive way [3] to improve the conventional LLR-based speaker verification methods would be to fuse multiple LLR measures into a unified framework by virtue of the complementary information that each LLR can contribute. Given  $M$  different LLR measures,  $L_m(U)$ ,  $m = 1, 2, \dots, M$ , a fusion-based LLR measure [3] can be defined as

$$L_{\text{Fusion}}(U) = \mathbf{w}^T \Phi(U) + b, \quad (7)$$

where  $b$  is a bias,  $\mathbf{w} = [w_1 \ w_2 \ \dots \ w_M]^T$  and  $\Phi(U) = [L_1(U) \ L_2(U) \ \dots \ L_M(U)]^T$  are the  $M \times 1$  weight vector and LLR-based vector, respectively. The implicit idea of  $\Phi(U)$  is that a variable-length input utterance  $U$  can be represented by a fixed-dimension characteristic vector via a nonlinear mapping function  $\Phi(\cdot)$ . Equation (7) forms a nonlinear discriminant classifier, which can be implemented by using the kernel-based discriminant technique, namely the Support Vector Machine (SVM) [10]. The goal of SVM is to find a separating hyperplane that maximizes the margin between classes. Following [10],  $\mathbf{w}$  in Eq. (7) can be expressed as  $\mathbf{w} = \sum_{j=1}^J y_j \alpha_j \Phi(U_j)$ , which yields an SVM-based measure:

$$L_{\text{SVM}}(U) = \sum_{j=1}^J y_j \alpha_j k(U_j, U) + b, \quad (8)$$

where each training utterance  $U_j$ ,  $j = 1, 2, \dots, J$ , is labeled by either  $y_j = 1$  (the positive sample) or  $y_j = -1$  (the negative sample), and  $k(U_j, U) = \Phi(U_j)^T \Phi(U)$  is the kernel function [10] represented by an inner product of two vectors  $\Phi(U_j)$  and  $\Phi(U)$ . The coefficients  $\alpha_j$  and  $b$  can be solved by using the quadratic programming techniques [10].

### 2.1 LLR-Based Multiple Kernel Learning

The effectiveness of SVM depends crucially on how the kernel function  $k(\cdot)$  is designed. A kernel function must be symmetric, positive definite, and conform to Mercer's condition [10]. There are a number of kernel functions [10] used in different applications. For example, the sequence kernel [6] can take variable-length speech utterances as inputs. In this paper, we rewrite the kernel function in Eq. (8) as

$$k(U_j, U) = [L_1(U_j) \ \dots \ L_M(U_j)] [L_1(U) \ \dots \ L_M(U)]^T = \sum_{m=1}^M k_m(U_j, U). \quad (9)$$

Complying with the closure property of Mercer kernels [10], Eq. (9) becomes a composite kernel represented by the sum of  $M$  LLR-base sequence kernels [11] defined by

$$k_m(U_j, U) = L_m(U_j) \cdot L_m(U), \quad (10)$$

where  $m = 1, 2, \dots, M$ . Since the design of Eq. (9) does not involve any optimization process with respect to the combination of  $M$  LLR-base sequence kernels, we further redefine Eq. (9) as a new form, named the LLR-base composite sequence kernel, in accordance with the closure property of Mercer kernels [10]:

$$k_{\text{com}}(U_j, U) = \sum_{m=1}^M \beta_m k_m(U_j, U), \quad (11)$$

where  $\beta_m$  is the weight of the  $m$ -th kernel function  $k_m(\cdot)$  subject to  $\sum_{m=1}^M \beta_m = 1$  and  $\beta_m \geq 0, \forall m$ . This combination scheme quantifies the unequal nature of  $M$  LLR-base sequence kernel functions by a set of weights  $\{\beta_1, \beta_2, \dots, \beta_M\}$ . To obtain a reliable set of weights, we apply the MKL [9] algorithm. Since the optimization process is related to the speaker verification accuracy, this new composite kernel defined in Eq. (11) is expected to be more effective and robust than the original composite kernel defined in Eq. (9).

The optimal weights  $\beta_m$  can be jointly trained with the coefficients  $\alpha_j$  of the SVM in Eq. (8) via the MKL algorithm [9]. Optimization of the coefficients  $\alpha_j$  and the weights  $\beta_m$  can be performed alternately. First we update the coefficients  $\alpha_j$  while fixing the weights  $\beta_m$ , and then we update the weights  $\beta_m$  while fixing the coefficients  $\alpha_j$ . These two steps can be repeated until convergence. In this work, the MKL algorithm is implemented via the SimpleMKL toolbox developed by Rakotomamonjy et al. [9].

## 3 Experiments

### 3.1 Experimental Setup

Our speaker verification experiments were conducted on the speech data extracted from the extended M2VTS database (XM2VTSDB) [12]. In accordance with “Configuration II” described in Table 1 [12], the database was divided into three subsets: “Training”, “Evaluation”, and “Test”. In our experiments, we used “Training” to build each target speaker GMM and background models, and “Evaluation” to estimate the coefficients  $\alpha_j$  in Eq. (8) and the weights  $\beta_m$  in Eq. (11). The performance of speaker verification was then evaluated on the “Test” subset.

As shown in Table 1, a total of 293 speakers in the database were divided into 199 clients (target speakers), 25 “evaluation impostors”, and 69 “test impostors”. Each speaker participated in 4 recording sessions at approximately one-month intervals, and each recording session consisted of 2 shots. In a shot, every speaker was prompted to utter 3 sentences “0 1 2 3 4 5 6 7 8 9”, “5 0 6 9 2 8 1 3 7 4”, and “Joe took father’s green shoe bench out”. Each utterance, sampled at 32 kHz, was converted into a stream of 24-order feature vectors, each consisting of 12 mel-



**Table 1** Configuration of the speech database

Session	Shot	199 clients	25 impostors	69 impostors
1	1	Training	Evaluation	Test
	2			
2	1			
	2			
3	1	Evaluation		
	2			
4	1	Test		
	2			

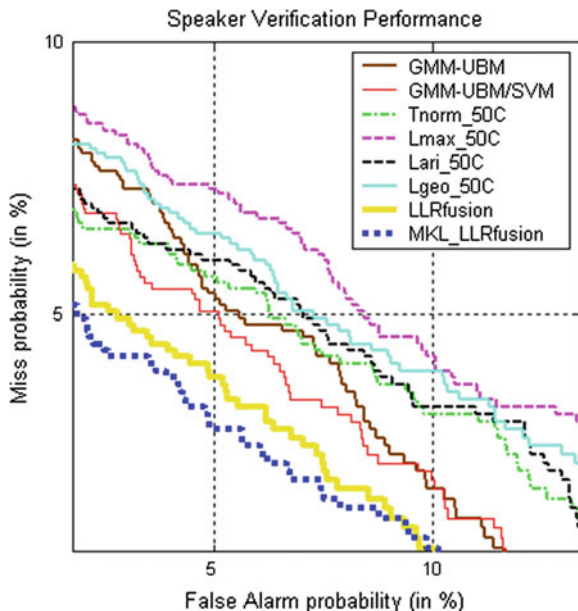
frequency cepstral coefficients (MFCCs) [13] and their first time derivatives, by a 32-ms Hamming-windowed frame with 10-ms shifts.

We used 12 ( $2 \times 2 \times 3$ ) utterances/client from sessions 1 and 2 to train the client model, represented by a GMM with 64 mixture components. For each client, the other 198 clients’ utterances from sessions 1 and 2 were used to generate the UBM, represented by a GMM with 256 mixture components; 50 closest speakers were chosen from these 198 clients as a cohort. Then, we used 6 utterances/client from session 3, along with 24 ( $4 \times 2 \times 3$ ) utterances/evaluation-impostor, which yielded 1,194 ( $6 \times 199$ ) client examples and 119,400 ( $24 \times 25 \times 199$ ) impostor examples, to estimate  $\alpha_j$  and  $\beta_m$ . However, recognizing the fact that the kernel method can be intractable when a huge amount of training examples involves, we downsized the number of impostor examples from 119,400 to 2,250 using a uniform random selection method. In the performance evaluation, we tested 6 utterances/client in session 4 and 24 utterances/test-impostor, which produced 1,194 ( $6 \times 199$ ) client trials and 329,544 ( $24 \times 69 \times 199$ ) impostor trials.

### 3.2 Experimental Results

We implemented two SVM systems,  $L_{\text{Fusion}}(U)$  in Eq. (7) (“LLRfusion”) and  $k_{\text{com}}(U_j, U)$  in Eq. (11) (“MKL\_LLRFusion”), both of which are fused by five LLR-based sequence kernel functions defined in Eqs. (1)–(5). For the purpose of performance comparison, we used six baseline systems,  $L_{\text{UBM}}(U)$  in Eq. (1) (“GMM-UBM”),  $L_{\text{Bengio}}(U)$  in Eq. (6) (“GMM-UBM/SVM”),  $L_{\text{Max}}(U)$  in Eq. (2) (“Lmax\_50C”),  $L_{\text{Ari}}(U)$  in Eq. (3) (“Lari\_50C”),  $L_{\text{Geo}}(U)$  in Eq. (4) (“Lgeo\_50C”), and  $L_{\text{Tnorm}}(U)$  in Eq. (5) (“Tnorm\_50C”), where 50C represents 50 closest cohort models were used. Figure 1 shows the results of speaker verification evaluated on the “Test” subset in terms of DET curves [14]. We can observe that the curve “MKL\_LLRFusion” not only outperforms six baseline systems, but also performs better than the curve “LLRfusion”. Further analysis of the results via the minimum Half Total Error Rate (HTER) [14] showed that a 5.76 % relative improvement was achieved by “MKL\_LLRFusion” (the minimum HTER = 3.93 %), compared to 4.17 % of “LLRfusion”.

**Fig. 1** DET curves for “Test”



## 4 Conclusion

In this paper, we have presented a new kernel function, named the Log-Likelihood Ratio (LLR)-based composite sequence kernel, for SVM-based speaker verification. This kernel function not only can be jointly optimized with the SVM training via the Multiple Kernel Learning (MKL) algorithm, but also can calculate the speech utterances in the kernel function intuitively by embedding an LLR in the sequence kernel. Our experimental results have shown that the proposed system outperforms the conventional speaker verification approaches.

**Acknowledgments** This work was funded by the National Science Council, Taiwan, under Grant: NSC101-2221-E-231-026.

## References

1. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. *Digit Signal Proc* 10:19–41
2. Rosenberg AE, DeLong J, Lee CH, Juang BH, Soong FK (1992) The use of Cohort Normalized scores for speaker verification. *Proc, ICSLP*
3. Chao YH, Tsai WH, Wang HM, Chang RC (2006) A kernel-based discrimination framework for solving hypothesis testing problems with application to speaker verification. *Proceedings of the ICPR*
4. Auckenthaler R, Carey M, Lloyd-Thomas H (2000) Score normalization for text-independent speaker verification system. *Digit Signal Proc.* 10:42–54

5. Bengio S, Mariéthoz J (2001) Learning the decision function for speaker verification. Proceedings of the ICASSP
6. Wan V, Renals S (2005) Speaker verification using sequence discriminant support vector machines. IEEE Trans Speech Audio Proc 13:203–210
7. Campbell WM, Sturim DE, Reynolds DA (2006) Support vector machine using GMM supervectors for speaker verification. IEEE Signal Proc Lett 13
8. Karam ZN, Campbell WM (2008) A multi-class MLLR Kernel for SVM speaker recognition. Proceedings of the ICASSP
9. Rakotomamonjy A, Bach F.R, Canu S, Grandvalet Y (2008) SimpleMKL. J. Mach Learn Res 9:2491–2521
10. Herbrich R (2002) Learning Kernel classifiers: theory and algorithms, MIT Press
11. Chao YH, Tsai WH, Wang HM (2010) Speaker verification using support vector machine with LLR-based sequence kernels. Proceedings of the ISCSLP
12. Luettin J, Maître G (1998) Evaluation protocol for the extended M2VTS database (XM2VTSDB). IDIAP-COM 98-05, IDIAP
13. Huang X, Acero A, Hon HW (2001) Spoken language processing. Prentics Hall
14. Bengio S, Mariéthoz J (2004) The expected performance curve: a new assessment measure for person authentication. Proceedings ODYSSEY

# Edit Distance Comparison Confidence Measure for Speech Recognition

Dawid Skurzok and Bartosz Ziółko

**Abstract** A new possible confidence measure for automatic speech recognition is presented along with results of tests where they were applied. A classical method based on comparing the strongest hypotheses with an average of a few next hypotheses was used as a ground truth. Details of our own method based on comparison of edit distances are depicted with results of tests. It was found useful for spoken dialogue system as a module asking to repeat a phrase or declaring that it was not recognised. The method was designed for Polish language, which is morphologically rich.

**Keywords** Speech recognition decisions · Polish

## 1 Introduction

Research on automatic speech recognition (ASR) started several decades ago. Most of the progress in the field was done for English. It has resulted in many successful designs, however, ASR systems are always below the level of human speech recognition capability, even for English. In case of less popular languages, like Polish (with around 60 million speakers), the situation is much worse. There is no large vocabulary ASR (LVR) commercial software for continuous Polish. Polish speech contains high frequency phones (fricatives and plosives) and the language is highly inflected and non-positional.

---

D. Skurzok (✉) · B. Ziółko

Department of Electronics, AGH University of Science and Technology,  
Al. Mickiewicza 30, 30-059 Kraków, Poland

e-mail: skurzok@agh.edu.pl

URL: www.dsp.agh.edu.pl

B. Ziółko

e-mail: bziolko@agh.edu.pl

URL: www.dsp.agh.edu.pl

It is crucial in a spoken dialogue system to not only provide a hypothesis of what was spoken but also to evaluate how likely it is. A simple probability is not always a good measure because its value depends on too many conditions. In case of dialogue systems, additional measure evaluating if the recognition is creditable or not is very useful. A relation to other, non-first hypothesis can provide it. It allows to repeat a question by a spoken dialogue system or choose a default answer for an unknown utterance. The purpose of Confidence Measures (CMs) is to estimate the quality of a result. In speech recognition, confidence measures are applied in various manners.

Existing types and applications of CMs were well summarised [1–3]. CMs can help to decide to keep or reject a hypothesis in keyword spotting applications. They can be also useful in detecting out-of-vocabulary words to not confuse them with some similar vocabulary words. Moreover, for acoustic adaptation, CM can help to select the reliable phonemes, words or even sentences, namely those with a high confidence score. They can be also used for the unsupervised training of acoustic models or to lead a dialogue in an automatic call-centre or information point in order to require a confirmation only for words with a low confidence score. Recently, applying Bayes based CM for reinforced learning was also tested [4]. A CM based on comparison of phonetic substrings was also described [5]. CMs were also applied in a new third-party error detection system [6]. CMs are even more important in speaker recognition. A method based on expected log-likelihood ratio was recently tested in speaker verification [7]. CMs can be classified [2] according to the criteria which they are based on: hypothesis density, likelihood ratio, semantic, language syntax analysis, acoustic stability, duration, lattice-based posterior probability.

## 2 Literature Review

Results and views on CMs for speech recognition found in the latest papers were analysed while we worked on our method. In some scenarios it is very important to compute CMs without waiting for the end of the audio stream [2]. The frame-synchronous ones can be computed as soon as a frame is processed by the recognition system and are based on a likelihood ratio. They are based on the same computation pattern: a likelihood ratio between the word for which we want to evaluate the confidence and the competing words found within the word graph. A relaxation rate to have a more flexible selection of competing words was introduced.

Introducing a relaxation rate to select competing words implies managing multiple occurrences of the same word with close beginning and ending times. The situation can be solved in two ways. A summation method adds up the likelihood of every occurrence of the current word and adds up the likelihood of every occurrence of the competing words. A maximisation method keeps only the occurrence with the maximal acoustic score.

The frame-synchronous measures were implemented in three ways regarding a context: unigram, bigram and trigram. The trigram one gave the best results on a test corpus.

The local measures estimate a local posterior probability in the vicinity of the word to analyse. They can use data slightly posterior to the current word. However, this data is limited to the local neighbourhood of this word and the confidence estimation does not need the recognition of the whole sentence. Local measures gave better results on a test set.

Two  $n$ -gram CMs based evaluations were also recently tested [8] 7-gram based on part-of-speech (POS) tags and 4-gram based on words. The latter was not succesful in detecting wrong recognitions. Applying POS tags in a CM was efficient, probably because it enables analysis on a larger time scale (7-gram instead of 4-gram).

A new CM based on phonetic distance was described [9]. It uses distances between subword units and density comparison (called anti-model by authors). The method employs separate phonetic similarity knowledge for vowels and consonants, resulting in more reliable performance. Phonetic similarities between a particular subword model and the remaining models are identified using training data

$$P(X^{(i)}|\lambda_{i,1}) \geq P(X^{(i)}|\lambda_{i,2}) \geq \dots \geq P(X^{(i)}|\lambda_{i,M}) \quad (1)$$

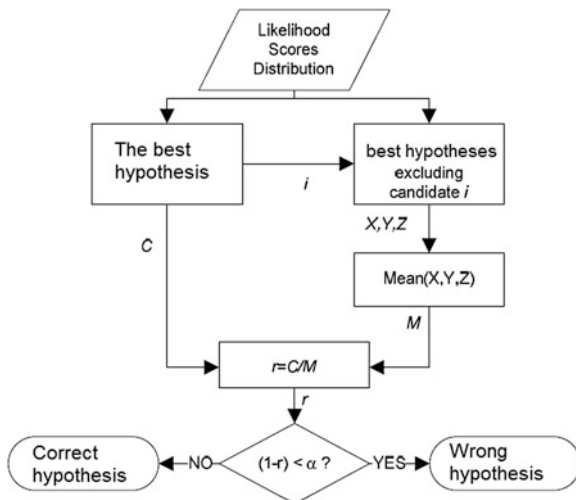
where  $X^{(i)}$  is a collection of training data labeled as model  $\lambda_i$  and  $\lambda_{i,m}$  indicates the  $m$ th similar model among  $M$  subword models compared to the pivotal model  $\lambda_i$ .

Applying of conditional random fields was recently tested [10] for confidence estimation. They allow comparison of features from several sources, namely lattice arc posterior ratio, lattice-based acoustic stability and Levenshtein alignment feature.

### 3 1-to-3 Comparison

The most widely known CM is based on hypothesis density. It compares the strongest hypothesis with an average of the following  $n$  weaker ones (Fig. 1). In our experiments  $n = 3$  was empirically found useful and it is a common value for this parameter in other systems as well. Our evaluations were done for sentence error rate. In the first evaluations it worked very well but later on, we found out, that its usefulness is limited in real dialogue applications because it had similar ratio for sentences allowed by a dictionary as for the ones which were not allowed. It was confirmed in later statistical tests with larger dictionaries. This is why we searched for a method based on edit distance comparison and earlier on phonetic substrings [5].

**Fig. 1** Algorithm of a standard method of CM by analysis of hypotheses density



## 4 Edit Distance Comparison

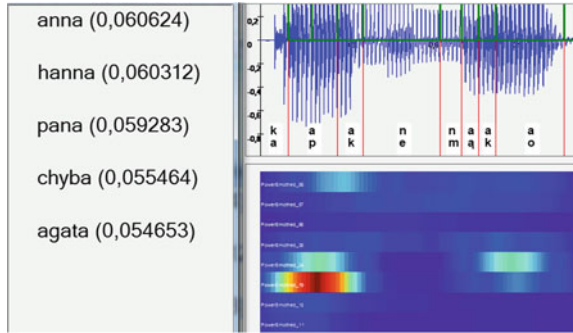
Edit distance comparison CM was designed and implemented for scenarios where there are several utterances very similar to each other. Such situation is especially common in morphologically rich languages like Polish [11], Czech [12] or Finnish [13]. In this type of scenarios classical CMs frequently fail to help detect wrong recognitions. Our new approach operates by measuring Levenshtein distance [14] in phonetic domain between the strongest hypothesis and the following ones. In this method, the mean of edit distances between the first hypothesis and  $m$  following ones is taken as the confidence value. We found that  $m = 6$  gives the best results (Fig. 2).

Considering only the mean of distances as the confidence indicator, gives worse results than simple 1-to-3 probability comparison, although both methods can be connected to improve final results. Both methods returns numbers from different range and with different variance. We suggest a following formula as a hybrid approach

$$C = C_{1to3} + \alpha \bar{D}^\beta, \quad (2)$$

where  $C$  is a final confidence,  $C_{1to3}$  is a confidence calculated using previous method and  $\bar{D}$  is a mean of edit distances between the strongest and  $m$  following hypothesis. Coefficients  $\alpha$  and  $\beta$  are used to scale distance confident and were chosen through optimization. We found that  $\alpha = 0.8$  and  $\beta = -2$  give the best results.

As it can be concluded, the suggested edit distance comparison method is quite a new approach, which does not fall directly into any of the CM types presented above and listed in literature [2] (Table 1).



**Fig. 2** A screenshot of the developer version of our ASR SARMATA system presents an example of how the described edit distance CM can be applied. The *left* part shows the ranking of top 5 hypotheses and the right one, the time and frequency representation of the analysed audio file. The first three hypothesis have small edit distance between them and the recognition is actually correct

**Table 1** Example of calculation of edit distance CM

	Hypothesis	Likelihood	Distance
1	<i>/anna/</i>	0.120	0
2	<i>/xanna/</i>	0.095	1
3	<i>/panna/</i>	0.080	1
4	<i>/pana/</i>	0.065	2

For this case, let us assume that  $m = 3$ . The 1-to-3 confidence is  $C_1 \text{ to } 3 = 0.12 / ((0.095 + 0.08 + 0.065) / 3) = 0.12 / 0.08 = 1.5$ . The hybrid confidence (2) is  $C = 1.5 + 0.8 \cdot ((1 + 1 + 2) / 3)^{-2} = 1.5 + 0.8 \cdot 1.33^{-2} = 1.5 + 0.45 = 1.95$

## 5 Tests and Results

The standard 1-to-3 method was compared with the edit distance method in a sequence of experiments on as test corpus based on CORPORA [15]. The recordings consists of 4435 audio files, each with one word spoken by various male speakers. The audio files have sampling rate 16 kHz and 16-bit rate. No language model was used in the tests. Some of the words in test corpora were recorded as isolated word, while others were extracted from longer sentences. All tests were made using SARMATA ASR system [11]. The dictionary has 9177 words. 1492 of total 4435 words were recognised correctly (Table 2).

**Table 2** Result for different methods ED is an abbreviation of edit distance confidence

	1-to-3	ED	1-to-3 + ED
Precision	0.71	0.38	0.72
Recall	0.65	0.77	0.65
Accuracy	0.79	0.50	0.80
F-measure	0.68	0.51	0.70



## 6 Conclusions

The suggested CM method based on edit distance enhanced the classical 1-to-3 method in an experiment motivated by real applications and end-user tests. The method was designed for morphologically rich languages, like Polish, as it gives better scores if the strongest hypotheses are phonetically similar. The presented method gives 2 % improvement in F-measure and 1 % improvement in accuracy.

**Acknowledgments** The project was funded by the National Science Centre allocated on the basis of a decision DEC-2011/03/D/ST6/00914.

## References

1. Guo G, Huang C, Jiang H, Wang RH (2004) A comparative study on various confidence measures in large vocabulary speech recognition. Proceedings of international symposium on Chinese spoken language, pp 9–12
2. Razik J, Mella O, Fohr D, Haton J (2011) Frame-synchronous and local confidence measures for automatic speech recognition. *Int J Pattern Recognit Artif Intell* 25:157–182
3. Wessel F, Schluter R, Macherey K, Ney H (2001) Confidence measures for large vocabulary continuous speech recognition. *IEEE Trans Speech Audio Proc* 9(3):288–298
4. Molina C, Yoma N, Huenupan F, Garreton C, Wuth J (2010) Maximum entropy-based reinforcement learning using a condense measure in speech recognition for telephone speech. *IEEE Trans Audio, Speech Lang Proc* 18(5):1041–1052
5. Ziółko B, Jadczyk T, Skurzok D, Ziółko M (2012) Confidence measure by substring comparison for automatic speech recognition. ICALIP, Shanghai
6. Zhou L, Shi Y, Sears A (2010) Third-party error detection support mechanisms for dictation speech recognition. *Interact Comput* 22:375–388
7. Vogt R, Sridharan S, Mason M (2010) Making confident speaker verification decisions with minimal speech. *IEEE Trans Audio Speech Lang Process* 18(6):1182–1192
8. Huet S, Gravier G, Sebillot P (2010) Morpho-syntactic post-processing of n-best lists for improved French automatic speech recognition. *Comput Speech Lang* 24:663–684
9. Kim W, Hansen J (2010) Phonetic distance based condense measure. *IEEE Signal Process Lett* 17(2):121–124
10. Seigel M, Woodland P (2011) Combining information sources for confidence estimation with crf models. Proceedings of InterSpeech
11. Ziółko M, Gałka J, Ziółko B, Jadczyk T, Skurzok D, Mąsior M (2011) Automatic speech recognition system dedicated for Polish. Proceedings of Interspeech, Florence
12. Nouza J, Zdansky J, David P, Cervá P, Kolorenc J, Nejedlova D (2005) Fully automated system for Czech spoken broadcast transcription with very large (300 k+) lexicon. Proceedings of InterSpeech, pp 1681–1684
13. Hirsimäki T, Pytkkonen J, Kurimo M (2009) Importance of high-order n-gram models in morph-based speech recognition. *IEEE Trans Audio Speech Lang Process* 17(4):724–732
14. Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Phys Doklady* 10:707–710
15. Grochowski S (1998) First database for spoken Polish. Proceedings of international conference on language resources and evaluation, Grenada, pp 1059–1062

# Weighted Pooling of Image Code with Saliency Map for Object Recognition

Dong-Hyun Kim, Kwanyong Lee and Hyeyoung Park

**Abstract** Recently, codebook-based object recognition methods have achieved the state-of-the-art performances for many public object databases. Based on the codebook-based object recognition method, we propose a novel method which uses the saliency information in the stage of pooling code vectors. By controlling each code response using the saliency value that represents the visual importance of each local area in an image, the proposed method can effectively reduce the adverse influence of low visual saliency regions, such as the background. On the basis of experiments on the public Flower102 database and Caltech object database, we confirm that the proposed method can improve the conventional codebook-based methods.

**Keywords** Object recognition · Visual saliency map · Codebook-based recognition · Code pooling

## 1 Introduction

Subsequent to the development of the bag-of-features (BoF) method [1] and spatial pyramid matching (SPM) method [2] for object recognition, many studies have been conducted on these types of codebook-based object recognition methods. Some of the studies focused on finding good coding schemes [3], while others

---

D.-H. Kim  
Infraware, Seoul, Korea

K. Lee  
Korea Open National University, Seoul, Korea

H. Park (✉)  
School of Computer Science and Engineering, Kyungpook National University,  
Daegu, Korea  
e-mail: hypark@knu.ac.kr

were devoted to pooling of code vectors [4]. However, most of the studies commonly treat all the code vectors from the main object and from the background image with the same importance, and therefore, the codes from the background can have an adverse influence on the recognition performance. To resolve this problem, we propose the use of saliency information, which is often calculated for detecting the reason of interest (ROI), in order to alleviate the effect of the code vectors of the background in the pooling stage. We also propose a generalized pooling method with saliency weight based on the concept of  $\alpha$ -mean [5], which is a generalized version of mean operation.

## 2 Overall Structure of Proposed System

Figure 1 shows the overall process of the proposed object recognition method involving the use of a saliency map; the method is based on a codebook and SPM. We represent an input image  $I$  by using the set of scale invariant feature transform (SIFT) descriptors  $\{\mathbf{x}_m\}_{m=1\dots M}$ , and we then apply the locality-constrained linear coding (LLC) method [3] to obtain a code vector  $\mathbf{c}_m$  for each descriptor  $\mathbf{x}_m$ . Once the set of code vectors  $\{\mathbf{c}_m\}_{m=1\dots M}$  is obtained, we perform pooling with SPM to get the histogram features of the sub-regions structured in three levels. In the pooling stage, we also use the additional weight value  $w_m$  of each  $\mathbf{c}_m$ , which is obtained by using a saliency map  $\mathcal{S}(x,y)$ . Finally, all the histogram features from the sub-regions are concatenated to obtain a single feature vector for the given image. The feature vector is fed to a linear support vector machine (SVM) classifier to get the recognition result.

In the overall process, the novelty of the present work lies in two steps. First, we calculate the weight value of each code vector by using a saliency map. Second, we propose a generalized pooling method involving the use of the weight values, which we call  $\alpha$ -pooling. These two processes are explained in detail in the next section.

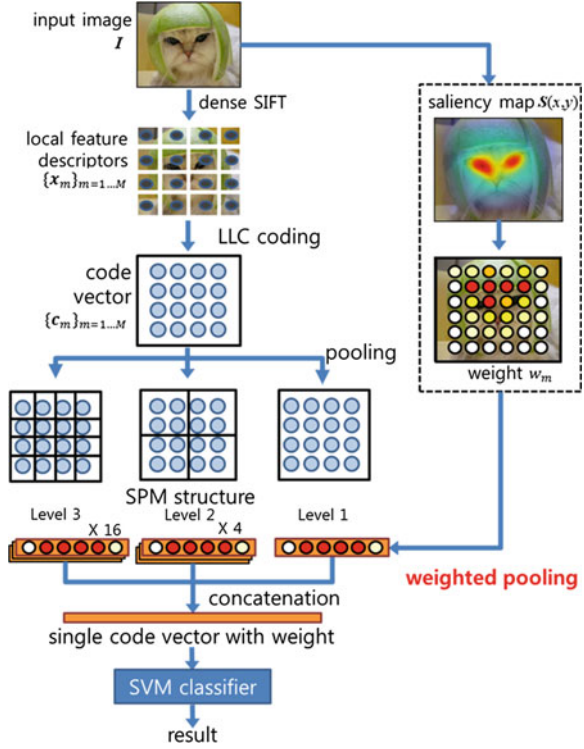
## 3 Pooling of Image Code with Saliency Weight

When we have a saliency map  $\mathcal{S}(x,y)$  for a given image  $I(x,y)$ , we first calculate the weight value  $w_m$  for each code vector  $\mathbf{c}_m$  that has been obtained by the LLC method. When a code vector is obtained for a feature descriptor  $\mathbf{x}_m$  that has been extracted from a local image patch  $\mathbf{i}_m$ , the corresponding weight value  $w_m$  is calculated by simply taking the average of the saliencies in the local image patch, which can be written as

$$w_m = \frac{1}{|\mathbf{i}_m|} \sum_{(x,y) \in \mathbf{i}_m} \mathcal{S}(x,y), \quad (1)$$

where  $|\mathbf{i}_m|$  denotes the number of pixels in  $\mathbf{i}_m$ .

**Fig. 1** Overall structure of the proposed method, which combines a codebook-based object recognition method and a saliency map



Once the weight values and code vectors are obtained, we conduct pooling to obtain the histogram features for the given image. In the SPM method, we need to calculate a histogram feature for each sub-region defined in the SPM structure. For example, when we use three-level SPM with grids of dimensions  $1 \times 1$ ,  $2 \times 2$ , and  $4 \times 4$ , we can have a set of 21 sub-regions,  $\{R_{1,1}^{1 \times 1}, R_{1,1}^{2 \times 2}, \dots, R_{2,2}^{2 \times 2}, R_{1,1}^{4 \times 4}, \dots, R_{4,4}^{4 \times 4}\}$ . When a sub-region  $R$  is composed of  $N$  code vectors and  $N$  weight values, represented as  $\{(c_j, w_j)\}_{j=1 \dots N}$ , the corresponding histogram feature  $\mathbf{h}$  is obtained by  $\alpha$ -pooling, which is defined as

$$\mathbf{h}_\alpha = f_\alpha^{-1} \left( \sum_{j=1}^N f_\alpha(w_j c_j) \right), \tag{2}$$

$$f_\alpha(\mathbf{u}) = \begin{cases} \mathbf{u}^{\frac{1-\alpha}{2}} & \alpha \neq 1 \\ \log \mathbf{u} & \alpha = 1 \end{cases}. \tag{3}$$

This  $\alpha$ -pooling is based on a stochastic integration method that is a generalization of various types of mean operations, which is called  $\alpha$ -mean [5]. When  $\alpha = -1$  and  $w_j = 1 (j = 1 \dots N)$  are satisfied, the Eq. (2) becomes equivalent to the well-known sum pooling. Further, when  $\alpha = -\infty$  and  $w_j = 1 (j = 1 \dots N)$  are satisfied, it gives the formula for the conventional max pooling. By taking arbitrary real

values of  $\alpha$  and  $w_j$ , we can obtain various types of weighted pooling methods. In the experiments, we will show the dependence of the recognition performance on the value of  $\alpha$  as well as the weight.

## 4 Experimental Comparisons

In order to confirm the effect of the saliency weight, we conducted computational experiments with three benchmark datasets: Caltech101 [6], Caltech256 [6], and Flower102 [7]. In each experiment, we extracted a dense set of multi-scale SIFT (PHOW) descriptors with a grid size of  $16 \times 16$ ,  $24 \times 24$ , and  $32 \times 32$  at every 6-pixel steps. In applying LLC, we set its parameter  $K$  to be 5. To obtain the saliency value, we adopted the saliency map proposed in [8] and [9].

Figure 2 shows the effect of the weight value and the dependence of the performance of the proposed method on the value of  $\alpha$ . From the figure, we can confirm that the use of weights improves the performance for all the databases, regardless of the value of  $\alpha$ . We can also see a consistent tendency: a smaller value of  $\alpha$  gives better performance. This observation implies that max pooling is superior to sum pooling.

In Table 1, we showed the best accuracies of the proposed method among the results for different values of  $\alpha$  shown in Fig. 2. We also showed the results of state-of-the-art methods reported in the literatures. In the case of Flower102, we listed two representative results reported in the original works [7] that has built the database: one was obtained by using only SIFT descriptor and the other was obtained by using the combination of four different descriptors. Though the proposed method cannot achieve the best accuracy obtained by using four descriptors, we can say that our results is promising in the sense that it is superior to the original work under the same condition of single SIFT descriptor. In the case of Caltech databases, we compared the proposed method with a number of recent codebook-based methods. As shown in Table 1, the proposed method gives the best result on Catech101, and the second best results on Catech256. Concerning

**Table 1** Experimental results on three benchmark datasets

Method	Flower 102	Caltech 101	Caltech 256
	(20 train)	(15 train)	(30 train)
Weighted $\alpha$ -pooling	0.633( $\alpha = -5$ )	0.691( $\alpha = -\infty$ )	0.403( $\alpha = -\infty$ )
$\alpha$ -pooling	0.579( $\alpha = -5$ )	0.663( $\alpha = -\infty$ )	0.392( $\alpha = -\infty$ )
Nilsback [7] (SIFT)	0.551	–	–
Nilsback [7] (4 descriptors)	0.728	–	–
Yang [2] (sparse code + SIFT)	–	0.670	0.340
Wang [3] (LLC + HOG)	–	0.654	0.412
McCann [10] (SPM variant)	–	0.686	0.395

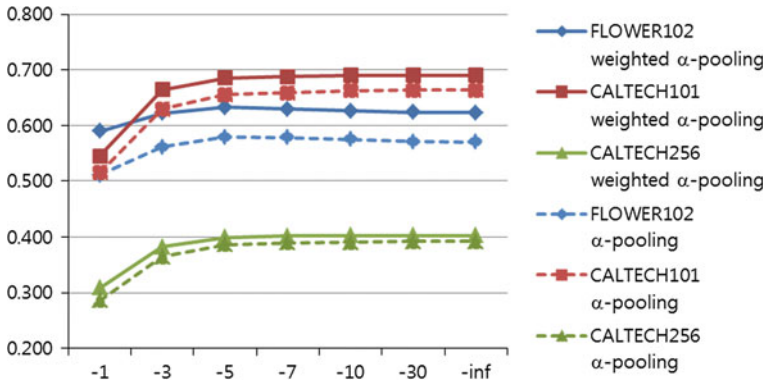


Fig. 2 Dependence of the performance of the proposed method on the value of  $\alpha$  and weights

the method suggested in [3], which uses LLC and max pooling, we need to note that it is a special case of the proposed  $\alpha$ -pooling method with  $\alpha = -\infty$  and  $w_j = 1 (j = 1 \dots N)$ . Thus, the difference of accuracy between  $\alpha$ -pooling and Wang [3] shown in Table 1 is mainly due to the use of different local descriptors (SIFT vs. HOG).

## 5 Conclusions

On the basis of the LLC and SPM methods, which are state-of-the-art methods of object recognition, we propose a novel weighted  $\alpha$ -pooling method in which saliency information and the concept of  $\alpha$ -mean are used. By using weighted pooling, we can expect to achieve an efficient representation of a given object image by excluding the background information. By using  $\alpha$ -mean, we can have a generalized pooling formula, which can cover sum pooling and max pooling. Experiments with benchmark data show the positive effect of the weight value: it leads to the proposed method showing performance comparable to state-of-the-art methods. The proposed method may be improved further by using more sophisticated local feature descriptors and saliency maps.

**Acknowledgments** This research was partially supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2012- H0301-12-2004) supervised by the NIPA(National IT Industry Promotion Agency); and by the Converging Research Center Program funded by the Ministry of Education, Science and Technology (2012K001342).

## References

1. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. Proceedings of ICCV'03, vol. 2. Los Alamitos, USA, p 1470
2. Yang J, Yu K, Gong Y, Huang T (2009) Linear spatial pyramid matching using sparse coding for image classification. Proceedings of CVPR'09, Miami, USA, pp 1794–1801
3. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality-constrained linear coding for image classification. Proceedings of CVPR'10, San Francisco, USA, pp 3360–3367
4. Boureau Y-L, Roux N, Bach F, Ponce J, LeCun Y (2011) Ask the locals: multi-way local pooling for image recognition. ICCV'11, Barcelona, Spain
5. Amari S (2007) Integration of stochastic models by minimizing  $\alpha$ -divergence. Neural Comput 19(10):2780–2796
6. [http://www.vision.caltech.edu/Image\\_Datasets](http://www.vision.caltech.edu/Image_Datasets). Accessed 20 July 2012
7. Nilsback M-E, Zisserman A (2008) Automated flower classification over a large number of classes. Proceedings of ICVGIP'08, Bhubaneswar, India, pp 722–729
8. Harel J, Koch C, Perona P (2006) Graph-based visual saliency. Proceedings of NIPS'06, Vancouver, Canada
9. Cheng M-M, Zhang G-X, Mitra NJ, Huang X, Hu S-M (2011) Global contrast based salient region detection. Proceedings of CVPR'11, Colorado Springs, USA, pp 409–416
10. McCann S, Lowe DG (2012) Local naïve Bayes nearest neighbor for image classification. Proceedings of CVPR'12, Providence, USA, pp 3650–3656

# Calibration of Urine Biomarkers for Ovarian Cancer Diagnosis

Yu-Seop Kim, Eun-Suk Yang, Kyoung-Min Nam, Chan-Young Park,  
Hye-Jung Song and Jong-Dae Kim

**Abstract** For the ovarian cancer diagnosis with biomarkers in urine samples, various calibration functions are selected and investigated to compensate the variability of their concentrations. The 15 biomarkers tested in this paper were extracted and measured for the urine samples of 178 patients. Three types of functions were employed to calibrate the biomarkers, including the existing one that divides the biomarker concentration by that of the creatinine. The AUC of the ROC of the calibrated biomarker with each function was chosen to evaluate the performance. Experimental results show that the best performance could be obtained by dividing the concentration of the biomarker by that of the creatinine raised to the power of the optimal exponent that was determined for the maximum AUC of the calibrated biomarker.

**Keywords** Biomarker · Ovarian cancer · Calibration · Logistic regression · Exponential · AUC

## 1 Introduction

To prevent the ovarian cancer or increase the possibility of survival, which is one of the most fatal malignant cancer, the development of early diagnosis method or detection of risk factors are paramount [1].

---

Y.-S. Kim · E.-S. Yang · K.-M. Nam · C.-Y. Park · H.-J. Song · J.-D. Kim (✉)  
Department of Ubiquitous Computing, Hallym University, 1 Hallymdaehak-gil,  
Chuncheon, Gangwon-do 200-702, Korea  
e-mail: kimjd@hallym.ac.kr

Y.-S. Kim · E.-S. Yang · K.-M. Nam · C.-Y. Park · H.-J. Song · J.-D. Kim  
Bio-IT Research Center, Hallym University, 1 Hallymdaehak-gil, Chuncheon,  
Gangwon-do 200-702, Korea



For the past several decades, considerable investment has been made in the early detection of cancer. However, biopsy is necessary to confirm the cancer, which opposes the goal of early diagnosis that should not be invasive. Biomarkers aim to achieve the early diagnosis, and are defined as markers that can objectively measure whether an organism is in a pathologically normal or abnormal state and the degree of reaction to certain drugs. More specifically, biomarkers can express a pathological state of illness, measure the degree of reaction that an organism shows when treated with certain drugs, and predict a viable treatment to an illness. An ideal tumor biomarker will be the protein fragments detected in the patient's urine or blood that cannot be found in a healthy people [2, 3]. Reference [4] reports the possibility of early detection of ovarian cancer using biomarkers found in urine.

Although urine samples are not useful after 24 h of collection, blood samples are disfavored over urine samples due to the invasiveness of collection and blood-borne diseases [5]. The American Conference of Governmental Hygienists (ACGIH) recommends random urine sampling on basis of the Biological Exposure Indices (BEIs). However random urines have drawbacks due to the variability of urinary output. When measuring the biomarkers in urine samples, the protein quantity in urine can change due to the digested food or the amount of water, and the concentration might also vary according to the time of collection or the sampling method. Much of this variability can be compensated for by adjustment of the concentration of the measured analyte based on the level of creatinine in the urine [5]. Creatinine is the metabolite of the muscle tissue and normally exists in urine. According to ACGIH, approximately 1.2 g of creatinine is produced per day. If the average daily urine volume is 1.2 L (range: 600–2500 ml), the mean creatinine concentration is approximately 1 g/L. Based on this assumption, the creatinine correction will adjust the urine concentration to an average concentration of 1 g/L. Some urines during a day will be above 1 g/L and others will be below 1 g/L, but the analyte concentration will be corrected to a value which would be theoretically equivalent to the value of a urine specimen which has a concentration of 1 g/L [5].

In Ref. [4], the biomarker concentration was calibrated by simply dividing it with the creatinine concentration. However the performance can be much more improved when using another function for calibration as will be shown in this paper. In order to find the best fit function for calibration, this paper uses the area under the curve (AUC) of the receiver operating characteristic curve (ROC) for the calibrated concentrations as the evaluation function to inquire the performance of the several calibration functions. The results show that, for most markers, the best performance could be achieved by dividing the biomarker concentration with the creatinine concentration raised to the power of the exponent smaller than one.

## 2 Experiments and Results

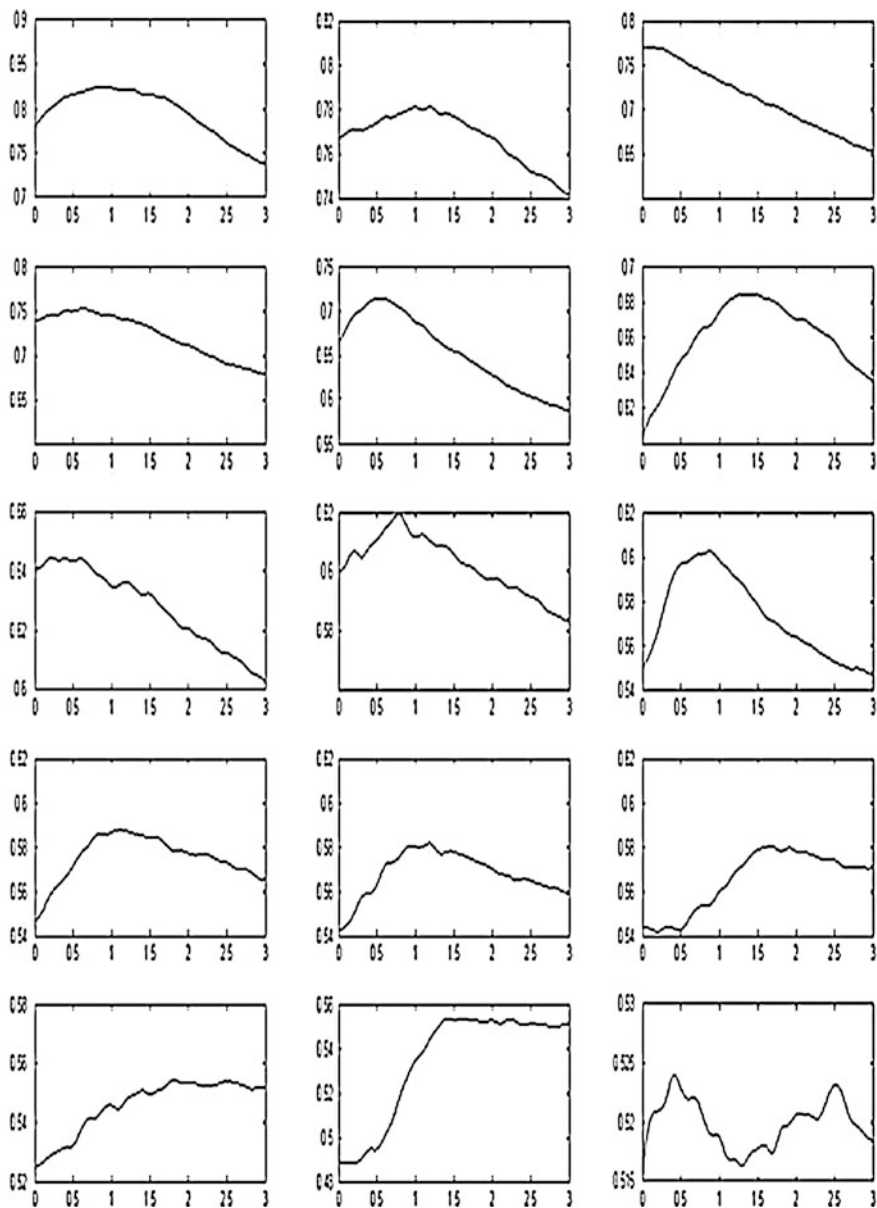
In this paper, the three functions,  $m/c^r$ ,  $e^m/e^c$ , and  $e^m/1 + e^c$  are compared, where  $m$ ,  $c$ , and  $r$  are the marker, creatinine concentration, and an exponent to be determined, respectively. When the exponent  $r$  in the first function equals to '1', the function will be the same as that used in Ref. [4]. Also note that the marker will not be calibrated with the exponent of '0'. The second and third function utilizes the exponential functions which were selected to increase the value at high concentrations, since the slope generally decreases significantly as the concentration increases in most of the pathological and biological phenomenon. However, these models showed lower performance than the first model for most biomarkers.

In the first model, the AUC of the ROC was calculated for each  $r$  from 0 to 15 with an incremental of 0.1, and the optimum exponent was determined when the AUC was the maximum.

In the paper, the proposed method was used to measure the concentration of fifteen biomarkers in 178 urine samples (57 patient samples and 121 healthy samples). In Table 1, the comparison of the AUC when various calibration functions were used for each biomarker is shown. The last two columns show the maximum AUC and the corresponding exponent when the biomarkers were calibrated by dividing with the creatinine value raised to the power of an exponent. The biomarkers in the table are sorted according to the values of this AUC, the highest on the top and the lowest on the bottom. The greatest AUC for each marker are bold-faced. The table is separated by the solid line between the 9 and 10th

**Table 1** AUC values after various calibration functions were applied. The greatest AUCs for each marker are bold-faced and the solid line is inserted between the 9 and 10th biomarker rows to distinguish the biomarkers with the AUC < 0.6

	Marker	$m$	$m/c$	$e^m/e^c$	$e^m/1 + e^c$	$m/c^r$	
						AUC	$r$
1	<b>HE4</b>	0.7791	0.8243	<b>0.8407</b>	0.7016	0.8249	0.9
2	<b>CRP</b>	0.7667	<b>0.7809</b>	0.7123	0.6149	<b>0.7809</b>	1.2
3	<b>TTR</b>	<b>0.7705</b>	0.7319	0.7283	0.6338	<b>0.7705</b>	0.0
4	<b>VCAM</b>	0.7379	0.7448	0.6910	0.6336	<b>0.7522</b>	0.6
5	<b>NCAM</b>	0.6654	0.6870	<b>0.7177</b>	0.6783	0.7134	0.6
6	<b>ApocIII</b>	0.6035	0.6745	<b>0.6946</b>	0.6910	0.6842	1.3
7	<b>MPO</b>	0.6414	0.6349	0.5939	0.5621	<b>0.6445</b>	0.2
8	<b>PDGF-AA</b>	0.6000	0.6117	0.5830	0.5782	<b>0.6194</b>	0.8
9	<b>CA 15-3</b>	0.5493	0.5982	0.5211	0.5205	<b>0.6021</b>	0.9
10	<b>CA 125</b>	0.5469	0.5866	0.5262	0.5447	<b>0.5876</b>	1.1
11	<b>CA 19-9</b>	0.5419	0.5801	<b>0.5817</b>	0.5750	0.5813	1.2
12	<b>Apo AI</b>	0.5434	0.5598	0.5607	0.5450	<b>0.5803</b>	1.7
13	<b>CEA</b>	0.5249	0.5457	0.5556	<b>0.5663</b>	0.5543	1.8
14	<b>OPN</b>	0.4891	0.5346	0.5224	0.5320	<b>0.5536</b>	1.4
15	<b>PAI-1</b>	0.5151	0.5188	<b>0.5341</b>	0.5224	0.5240	0.4



**Fig. 1** The AUC variations according to the exponent (From the *top-left*, HE4, CRP, TTR, VCAM, NCAM, ApocIII, MPO, PDGF-AA, CA 15-3, CA 125, CA 19-9, Apo AI, CEA, OPN, PAI-1)

markers, to distinguish the biomarkers with the  $AUC < 0.6$ . Note that the biomarkers with the  $AUC$  of 0.5 have no information on the disease.

As can be seen in the table, the calibration by the last function in the table shows the best  $AUC$  performance for the 6 markers among the 9 top markers. That is, 66 % of the valuable markers showed much better performance when they were calibrated by dividing their concentration with the creatinine concentration raised to the power of the optimal exponent. For 1st, 5th, and 6th markers, this calibration showed only about 2 % less performance than the best calibration method. The optimum exponent was  $< 1$  for the 7 markers among the top 9 markers, which opposed the fact that the conventional function is corresponding to the exponent of '1'.

Figure 1 illustrates the  $AUC$  variations according to the exponent when the last function was employed for the calibration. All of them showed distinct peaks except for the 3rd and the 15th markers, TTR and PAI-1. This implies that finding the optimum exponent is significant.

### 3 Conclusion and Discussion

This paper aims to improve the  $AUC$  of the ROC for the ovarian cancer diagnosis. The best fit calibration function was explored by comparing several functions that can calibrate the biomarker concentration which was extracted from urine samples. Among them, the calibration that divides the marker concentration with the creatinine concentration raised to the power of the optimum exponent worked best. According to the obtained exponent, this method covers no calibration as well as the existing calibration method where the marker concentration is simply divided by that of the creatinine. The  $AUC$  variation according to the exponent in this method was investigated and the distinct peaks were observed, showing that the search of the optimal exponent is more preferable.

**Acknowledgments** The research was supported by the Research & Business Development Program through the Ministry of Knowledge Economy, Science and Technology (N0000425) and the Ministry of Knowledge Economy (MKE), Korea Institute for Advancement of Technology (KIAT) and Gangwon Leading Industry Office through the Leading Industry Development for Economic Region.

### References

1. Choi SK, Cho TI, Lee TJ (2012) Immunohistochemical study of BRCA1, BRCA2, and poly(ADP-ribose) polymerase-1 in ovarian tumors. *Korean J Obstet Gynecol* 55:8–16
2. Hellstrom I, Heagerty PJ, Swisher EM, Liu P, Jaffar J, Agnew K, Hellstrom KE (2010) Detection of the HE4 protein in urine as a biomarker for ovarian neoplasms. *Cancer Lett* 296:43–48

3. Hellstrom I, Heagerty PJ, Swisher EM, Liu P, Jaffar J, Agnew K, Hellstrom KE (2010) Detection of the HE4 protein in urine as a biomarker for ovarian neoplasms. *Cancer Lett* 296:43–48
4. Nolen BM, Lokshin AE (2012) Multianalyte assay systems in the differential diagnosis of ovarian cancer. *Expert Opin Med Diagn* 6(2):131–138
5. Pacific Toxicology Laboratories <http://www.pactox.com/library/article.php?articleID=18>

# An Iterative Algorithm for Selecting the Parameters in Kernel Methods

Tan Zhiying, She Kun and Song Xiaobo

**Abstract** Giving a certain training sample set, the learning efficiency almost depends on the kernel function in kernel methods. This inspires us to learn the kernel and the parameters. In the paper, a selecting parameter algorithm is proposed to improve the calculation efficiency. The normalized inner product matrix is the approximation target. And utilize the iterative method to calculate the optimal bandwidth. The defect detection efficiency can be greatly improved adopting the learned bandwidth. We applied the algorithm to detect the defects on tickets' surface. The experimental results indicate that our sampling algorithm not only reduces the mistake rate but also shortens the detection time.

**Keywords** Kernel methods · Gaussian kernels · Iterative methods · Kernel PCA · Pre-image

## 1 Introduction

In kernel methods, different kernels have been used to solve a variety of tasks, such as the problems of classification, regression, image de-noising and dimensionality reduction et al. [1]. In support vector machines (SVMs), the three basic kernels polynomial kernels, radial basis functions and sigmoid kernels are successfully

---

T. Zhiying (✉)

1412 Huihong Building, 801 Changwu Mid Road, 213164 Changzhou, Jiangsu, China  
e-mail: tanzhiying1010@gmail.com

T. Zhiying · S. Kun

School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

S. Xiaobo

Institute of Advanced Manufacturing Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Changzhou, China

used [2]. Among the kernels, radial basis functions also named as Gaussian kernels are most widely used for its' stability.

To deal with more complex real problems, some new kernels are proposed. A combination kernel function was obtained by optimizing over a family of data-dependent kernels by Shao et al. [3]. Some regularization techniques are used for learning linear combinations of basic Kernels [4]. Using the learning kernels, the root mean squared error (RMSE) of some classification problems is lower than the single basic kernel. Recently, the non-linear combinations of kernels are also learning by solving the optimization problem. Learning kernels based on a polynomial combination of base kernels has been studied [5]. Multiple kernel learning (MKL) has been recently proposed, which aims at simultaneously learning a kernel and the associated predictor in supervised learning settings [6]. For solving the larger class of problems, the large scale MKL was proposed [7]. And more and more learning methods are constantly being proposed for solving the complex practical problems [8, 9].

Almost all the learning kernels methods are based on the basic kernels and training samples. However, there is little studying on the selection of basic kernels and the parameters. In the paper, we provide an algorithm for learning the parameters of kernels.

## 2 Kernel Methods

### 2.1 Kernel PCA

Denote the training set  $X = \{x_1, x_2, \dots, x_N\}$ , where the sample  $x_i \in R^n$  ( $i = 1, 2, \dots, N$ ). The nonlinear mapping  $\phi$  maps the samples  $x_1, \dots, x_N$  into the feature space  $F$  by [2]  $\phi : R^n \rightarrow F, x \mapsto Y$ . Define the kernel matrix  $K \in R^{N \times N}$  by  $K_{ij} := (\phi(x_i), \phi(x_j))$ .

The samples can be centered in feature space by  $K_c = (E - I_N)K(E - I_N)$ , where the unit matrix  $E \in R^{N \times N}, I \in R^{N \times N} (I(i, j) = 1, i, j = 1, 2, \dots, N)$ . The detail calculation of data centering can be found in Schölkopf et al.'s paper [10].

### 2.2 Pre-Image

To solve the optimization problem of minimizing the reconstruction error, the standard gradient ascent methods were used by the Mika et al [11]. And a modifying iteration method was also proposed to remove outliers from the data vectors by Takahashi and Kurita [12]. In the paper, the local linear property is used to calculate the pre-image by solving [13]

$$\min \rho(t_1, t_2, \dots, t_p) = \|\phi(\sum_{i=1}^p t_i x^i) - P_d \phi(x)\|^2$$

$$s.t. \quad \sum_{i=1}^p t_i = 1 \quad t_i \geq 0$$

The solution can be obtained by iterative formula

$$t = [[X^p]^T X^p]^{-1} [[X^p]^T X] B \tag{1}$$

where  $w_\ell = \sum_{j=1}^d \sum_{i=1}^N \alpha_i^j k(x, x_i) \alpha_\ell^j$ ,  $\ell = 1, 2, \dots, N$ ,  $\beta_\ell = w_\ell k(\sum_{i=1}^p t_i x^i, x_\ell)$ ,  $(\ell = 1, 2, \dots, N)$ ,  $B = [\beta_1 / \sum_{\ell=1}^N \beta_\ell, \beta_2 / \sum_{\ell=1}^N \beta_\ell, \dots, \beta_N / \sum_{\ell=1}^N \beta_\ell]^T$ . And the initial value of vector  $t$  can take  $t = [1/p, 1/p, \dots, 1/p]^T$ .

The process of calculating the pre-images can be simply summarized three steps. We firstly calculate the low coordinates  $y_i (i = 1, 2, \dots, N)$  in the feature space using the kernel PCA. Secondly find the  $p$  nearest neighbors  $x^i (i = 1, 2, \dots, p)$  by the coordinates  $y_i (i = 1, 2, \dots, N)$ . At last, obtain the pre-image by the linear expression of the  $p$  nearest neighbors.

### 3 Theoretical Results and Algorithm

#### 3.1 Theoretical Results

To calculate the optimal bandwidth of Gaussian kernel, establish the following optimization problem

$$\min_{\sigma} F(\sigma) = \sum_{i=1}^N \sum_{j=1}^N (G(i, j) - K(i, j))^2 \tag{2}$$

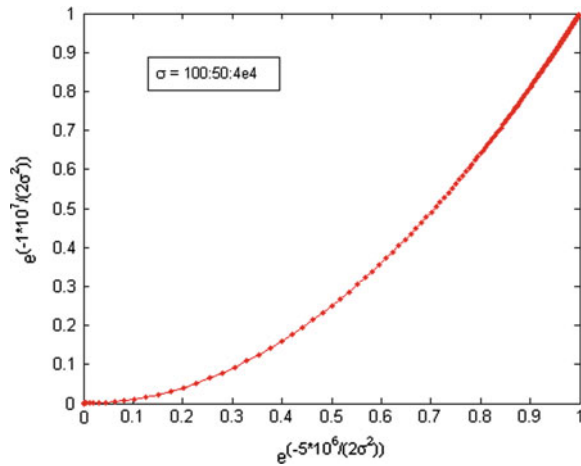
where

$$F(\sigma) = \sum_{i=1}^N \sum_{j=1}^N (\exp(-\|x_i - x_j\|^2 / 2\sigma^2) - G(i, j)) \tag{3}$$

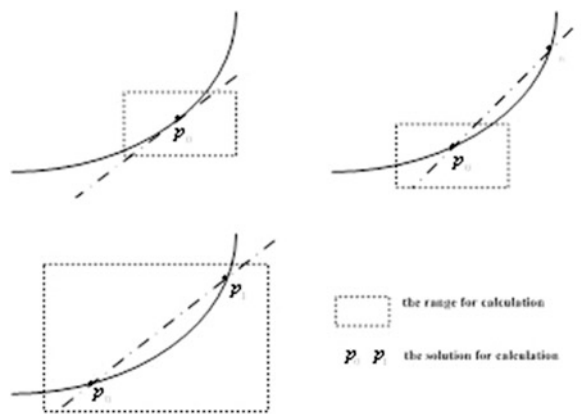
The necessary conditions of unconstrained optimization problems can be written as  $\frac{\partial F(\sigma)}{\partial \sigma} = R(\sigma) = 0$  in the extreme value. the cumulative sum can be seen as the inner product between vector  $a$  and vector  $b$ , where the vector  $b$  is constituted by the coefficients of the function  $k$ . The vector  $a$  is constituted by items  $k(x_i, x_j), k^2(x_i, x_j) (i, j = 1, 2, \dots, N)$ . Then the equation  $R(\sigma) = 0$  can be written as  $b'a = 0$ . In practice, vector  $b$  is known, and vector  $a$  changes with the variable  $\sigma$ . A plane can be determined by the linear equation  $b'x = 0$ . Solving the equation  $R(\sigma) = 0$  is equivalent to solve equations  $b'x = 0$  and  $x = a$ .



**Fig. 1** The curve determined by the parameter



**Fig. 2** Solutions' distribution with parameter in two dimension space



The Fig. 1 shows the changing of two dimensional the vector  $a$  following the parameter  $\sigma$ . In Fig. 2 shows three different kinds of solutions' distribution. In Algorithm 1, we will calculate the solution of equation by iterative method [14].

### 3.2 Algorithm

To calculate the optimal bandwidth  $\sigma$ , we should solve the equation  $R(\sigma) = 0$ , where 
$$R(\sigma) = \sum_{i=1}^N \sum_{j=1}^N (\exp(-\|x_i - x_j\|^2 / 2\sigma^2) - G(i, j)) \cdot \|x_i - x_j\|^2 \cdot \exp(-\|x_i - x_j\|^2 / 2\sigma^2).$$

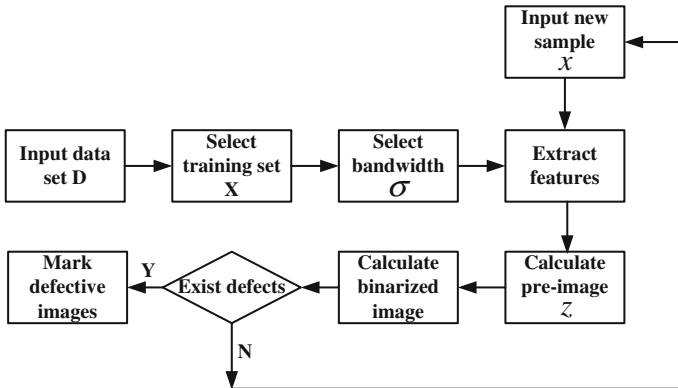
Based on the above analysis, the equation has at most two roots. We can select the optimal bandwidth  $\sigma$  by the iterative algorithm (Table 1).

**Table 1** Iterative algorithm

Algorithm 1	
1.	Input: $X = [x_1, x_2, \dots, x_N], \sigma_l(0), \sigma_r(0)$
2.	Calculate matrix $G, G(i, j) = \frac{\langle x_i, x_j \rangle}{\sqrt{\langle x_i, x_i \rangle} \sqrt{\langle x_j, x_j \rangle}}$
3.	Calculate $R_l(0) = R(\sigma_l(0)), R_r(0) = R(\sigma_r(0))$
4.	$nn = 0$
5.	while $( R_r(nn) - R_l(nn)  > 1 \times 10^{-6})$
6.	$nn = nn + 1, \sigma_c = \frac{\sigma_l(nn-1) + \sigma_r(nn-1)}{2}, R_c = R(\sigma_c)$
9.	if $R_l(nn - 1) \cdot R_c > 0$
10.	$R_l(nn) = R_c, R_r(nn) = R_r(nn - 1), \sigma_l(nn) = \sigma_c, \sigma_r(nn) = \sigma_r(nn - 1)$
14.	end
15.	if $R_r(nn - 1) \cdot R_c > 0$
16.	$R_r(nn) = R_c, R_l(nn) = R_l(nn - 1), \sigma_r(nn) = \sigma_c, \sigma_l(nn) = \sigma_l(nn - 1)$
20.	end
21.	end
22.	Output: $\sigma = \sigma_l(nn)$

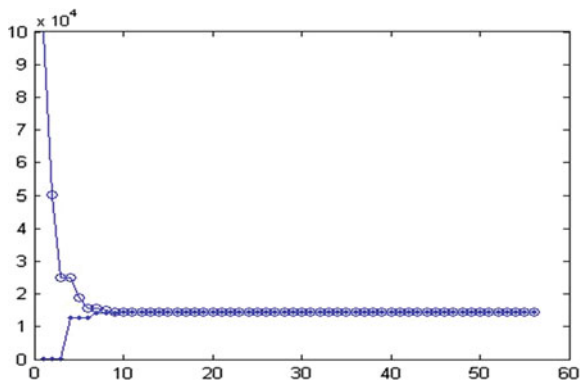
### 4 Experimental Results

To verify the learning kernel bandwidths' effect, we do some numerical experiments on two data set I and II. The two data sets are from some printed matters. Part samples can be found in Fig. 6. We respectively select 200 samples as the training samples from data sets I and II. We take the initial values  $\sigma_l(0) = 1 \times 10^2$  and  $\sigma_r(0) = 1 \times 10^5$ . In Figs. 3, 4 show the iterative results of bandwidths. The bandwidths can fast convergence to the optimal values 14494 and 18375 in the iterative process.

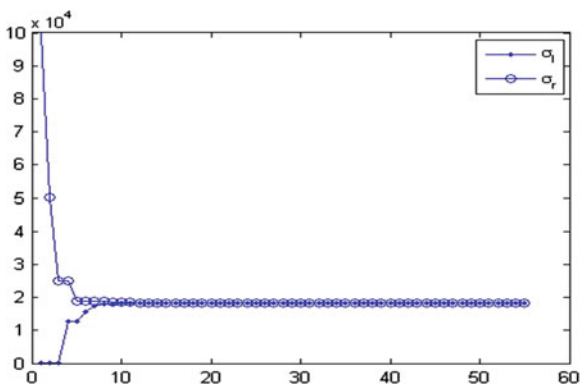


**Fig. 3** Solutions' distribution with the parameter

**Fig. 4** Solutions' distribution with the parameter



**Fig. 5** Solutions' distribution with the parameter



**Table 2** Margin specifications

	Test numbers	Bandwidth $\sigma$	MRSE
<i>Data 1</i>	1000	14494	1.1834e3
<i>Data 2</i>	724	18375	1.2897e3

Based on the selected 200 training samples, some experiments have been completed on the two data sets. In the experiments, 1000 and 724 test samples are used to verify the rationality of learning bandwidth. We choose 5 nearest neighbors and 80 principal components and the number of iterations is 20 (Figs. 5, 6 and 7), Table 2.

In Fig. 6, shows partial results which include the images with noise, the pre-images and the binary difference images. We can find that the defects can be easily detected by the reconstructed samples using the selected training samples.

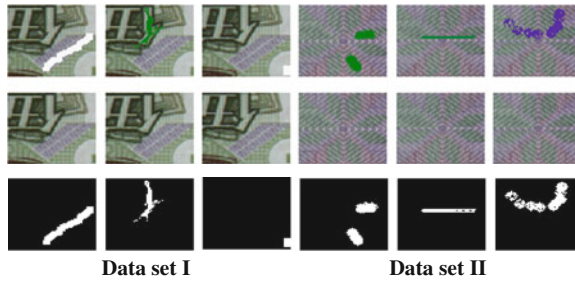


Fig. 6 Solutions' distribution with the parameter

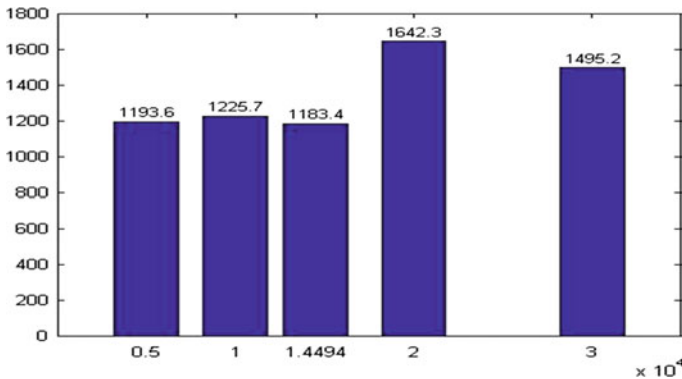


Fig. 7 Solutions' distribution with the parameter

## 5 Conclusions

We present a method of learning kernel parameters, including an iterative algorithm. The bandwidth of Gaussian kernel has a large influence in calculating the pre-images using the kernel PCA. The method not only can be used to select the parameters, but also can be used to learn the kernel. The sensible of selection theory is also corroborated by some of our empirical results.

## References

1. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University, London
2. Schölkopf B, Smola A, Müller KR (1997) Kernel principal component analysis. ICANN: artificial neural networks, pp 583–588
3. Shao JD, Rong G, Lee JM (2009) Learning a data-dependent kernel function for KPCA-based nonlinear process monitoring. Chem Eng Res Design 87:1471–1480

4. Cortes C, Mohri M, Rostamizadeh A (2009) L2 regularization for learning kernels. In: Conference uncertainty in artificial intelligence, pp 109–116
5. Cortes C, Mohri M, Rostamizadeh A (2009) Learning non-linear combinations of kernels. *Adv Neural Inf Proc Syst* 22:396–404
6. Rakotomamonjy A, Bach FR, Canu S et al (2009) SimpleMKL. *J Mach Learn Res* 9:2491–2521
7. Bach F (2008) Exploring large feature spaces with hierarchical multiple kernel learning. *arXiv preprint 0809:1–30*
8. Yi Y, Nan Y, Bingchao D et al (2012) Neural decoding based on Kernel regression. *JDCTA Int J Digit Content Technol Appl* 6:427–435
9. Shi WY (2012) The algorithm of nonlinear feature extraction for large-scale data set. *IJIPM Int J Inf Proc Manage* 3:45–52
10. Scholkopf B, Smola A, Muller KR (1998) Nonlinear component analysis as a Kernel eigenvalue problem, vol 10. pp 1299–1319
11. Mika S, Schölkopf B, Smola A et al (2001) Kernel PCA and de-noising in feature spaces. *Adv Neural Inf Proc Syst* 11:536–542
12. Takahashi T, Kurita T (2002) Robust de-noising by Kernel PCA. *Artificial neural networks-ICANN*, pp 789–789
13. Tan ZY, Feng Y (2011) A novel improved sampling algorithm. In: Conference communication software and networks, pp 43–46
14. Gerald CF, Wheatley PO (2006) *Applied numerical analysis*. Pearson Academic, America

# A Fast Self-Organizing Map Algorithm for Handwritten Digit Recognition

Yimu Wang, Alexander Peyls, Yun Pan, Luc Claesen  
and Xiaolang Yan

**Abstract** This paper presents a fast version of the self-organizing map (SOM) algorithm, which simplifies the weight distance calculation, the learning rate function and the neighborhood function by removing complex computations. Simplification accelerates the training process in software simulation and is applied in the field of handwritten digit recognition. According to the evaluation results of the software prototype, a 15–20 % speed-up in the runtime is obtained compared with the conventional SOM. Furthermore, the fast SOM accelerator can recognize over 81 % of handwritten digit test samples correctly, which is slightly worse than the conventional SOM, but much better than other simplified SOM methods.

**Keywords** Neural network · Self-organizing map · Handwritten digit recognition · Simplification

## 1 Introduction

The self-organizing map (SOM) also called Kohonen neural network is a competitive learning artificial neural network proposed by Kohonen in 1982 [1]. It is an unsupervised learning method which has both visualization and clustering properties by discovering the topological structure hidden in the data sets. Essentially the goal of a self-organizing map is to map continuous high-dimensional data onto a discrete low (typically one- or two-) dimensional feature map.

---

Y. Wang (✉) · Y. Pan · X. Yan  
Institute of VLSI Design, Zhejiang University, Hangzhou, People's Republic of China  
e-mail: wym85511@gmail.com

A. Peyls · L. Claesen  
EDM, Hasselt University, Diepenbeek, Belgium

As a clustering algorithm, the SOM has been applied widely in various fields including pattern recognition, defect inspection and as a data-mining tool to perform classification of high-dimensional data [2, 3]. Research on improving the performance of the SOM has been going on for decades. One of the key issues to overcome is the low speed learning process while obtaining a well trained map. A SOM is well trained if clustering is achieved in a short time and, at the same time, it creates a projection of data into the map strongly related to the distribution of data in the input space. One of the main reasons for this continued research effort is that the amount of data which is to be analyzed can be huge, for instance thousands of high-dimensional image vectors. The simulation of extensive networks with thousands of neurons, each with high-dimensional weights takes relatively much time on state of the art general purpose computers. To solve this problem, this paper presents a fast version of the SOM algorithm and software simulation proves that the SOM has been accelerated to some extent.

The remainder of this paper is organized as follows: [Sect. 2](#) gives a brief overview of related works. Next, [Sect. 3](#) presents the conventional self-organizing map. [Section 4](#) presents our proposed fast self-organizing map algorithm. [Section 5](#) discusses the experimental results on handwritten digit recognition and finally in [Sect. 6](#) the conclusions are drawn.

## 2 Related Work

To improve both the efficiency and effectiveness of the conventional SOM algorithm, many approaches have been proposed. A first possibility to reduce the runtime of the SOM is to compute initial values for the feature map instead of choosing them randomly in such a way that the training will be accelerated. In [4] the K-means clustering algorithm is used to select initial values for the weight vectors of the neurons, which subsequently reduces the required amount of training steps. Because the SOM offers multiple opportunities to exploit the parallel computing [5], a second way of handling the computational complexity is to transform the SOM algorithm into a distributed algorithm. Lobo et al. developed a distributed SOM in order to speed up the training of the SOM [6]. In order to shorten the processing time the batch version of the SOM has been used by Yu and Alahakoon [7], this version of the SOM is also more suitable for parallel implementation. A third way of accelerating the neural computations is to design simplified SOM algorithm. The weight update step is simplified by removing the non-linear functions in the following papers [8, 9] and therefore results in a more hardware-friendly version of the SOM algorithm. Nevertheless, these simplified methods suffer from a low recognition accuracy and are hardly effective in complex applications. In this context, a fast SOM algorithm is proposed in this paper which not only speeds up the training process but also promises a similar recognition accuracy with the conventional SOM.

### 3 Self-Organizing Map Algorithm

1. Initialization step: At the start of the SOM algorithm, typically all the weights  $w_j$  of the neurons are initialized with random values.
2. Compute the distance between the training vector  $X = \{x_1, \dots, x_M\}$  and each neuron  $N_j$  with weight  $w_j$ , using the Euclidean distance function:

$$D_j = \sqrt{\sum_{i=1}^M (x_i - w_{ji})^2} \quad (1)$$

3. Define the winning neuron as the neuron with the minimum distance.
4. Update each neuron according to the following update function:

$$w_j(t+1) = w_j(t) + \alpha(t) \cdot N_{j,I(X)}(t) \cdot (X - w_j(t)) \quad (2)$$

$w_j(t+1)$  is the updated weight vector,  $\alpha(t)$  the learning rate and  $N_{j,I(X)}(t)$  the topological neighbourhood value at training step  $t$ .  $I(X)$  is the winning neuron.

5. Update the neighborhood function and the learning rate.
6. Repeat steps 2–5 for the next training vector.

### 4 Fast Self-Organizing Map Algorithm

**Distance calculation** The conventional self-organizing map uses the Euclidean norm as the distance calculating function (see Eq. 1), however because it involves the squaring of values and a square root, the Euclidean distance computation is time-consuming for software prototype and also resource-intensive for hardware implementation. Following [8, 10], we use the Manhattan distance which is computationally simpler for calculating the distance between vectors.

$$D_j = \sum_{i=1}^M |(x_i - w_{ji})| \quad (3)$$

**Learning Rate Function** The learning rate is typically defined as the following exponential function.

$$\alpha(t) = \alpha_0 \cdot \exp\left(\frac{-t}{\tau_\alpha}\right) \quad (4)$$



Note that because of the multiplication, division and exponential function, this function will cost too much learning time. To reduce the computational complexity imposed by the exponential calculation of the conventional SOM, an alternative formula is selected to substitute the conventional Eq. 4. Actually, this alternative is the first term of the Taylor series expansion of Eq. 4.

$$\alpha = \alpha_0 \left(1 - \frac{t}{T}\right) \quad (5)$$

**Neighborhood Function** When using Kohonen's self-organizing map, the distance in the feature map between the neurons influences the learning process. A typical neighborhood function is shown in Eq. 6, which decreases not just over time, but also depends on the topological distance of the two neurons in the net.

$$N_{x,y}(t) = \exp\left(\frac{-d_{x,y}^2}{2\sigma^2(t)}\right) \quad (6)$$

Here  $d_{x,y}$  is the distance between node  $x$  and node  $y$ , more specifically it is the physical distance between the nodes in the feature map and  $\sigma(t)$  the time dependent value responsible for decreasing the neighborhood size over time.

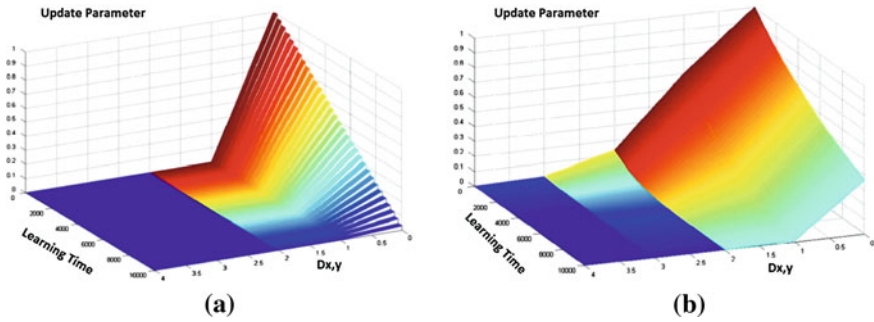
However our proposed neighborhood function ignores any influence of time and only depends on the topological distance between two neurons, which is computed by the Euclidean norm. For each neuron within the neighborhood size  $ns$ , the neighborhood parameter is calculated as shown in Eq. 7. Neurons outside this neighborhood area will not be updated.

$$N_{x,y} = \begin{cases} e^{-2d_{x,y}^2} & \text{if } d(x,y) \leq ns \\ 0 & \text{if } d(x,y) > ns \end{cases} \quad (7)$$

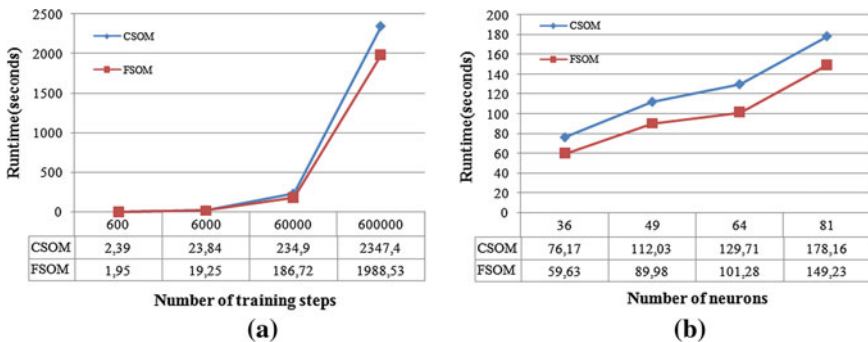
The weight update function depends on both the learning rate function and the neighborhood function. In Fig. 1, respectively the results of the conventional weight update function and our proposed weight update function are shown in the condition of  $\alpha_0 = 1, d_{x,y} = [0, 4], t = [0, 10000]$ . Note that the shapes of the 3D charts based on these functions are similar, which motivates the similarity in performance between both versions. The performance results will be given in the last section.

## 5 Case Study: Handwritten Digit Recognition

The performance of the SOM was tested in the field of handwritten digit recognition and the MNIST database was chosen to train and test the feature map [11]. We evaluated the proposed fast SOM by a software simulation on a PC with a general purpose processor clocked at 2.1 GHz and 2 GB of SDRAM. In Fig. 2, the runtime with varying amounts of iterations and varying amounts of neurons is



**Fig. 1** Comparison of weight update functions. **a** Proposed weight update function. **b** Conventional weight update function



**Fig. 2** Runtime comparison between conventional and fast SOM. **a** Runtime with different training steps. **b** Runtime with varying numbers of neurons

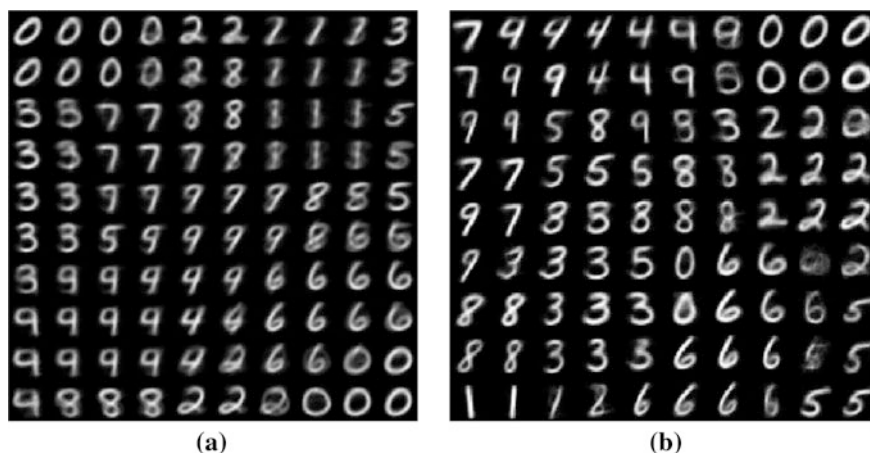
compared between the proposed fast SOM and the conventional SOM respectively. By using the Manhattan distance metric, a simplified learning rate and neighborhood function, no multiplications and exponential computations is required, due to this a reduction in time ranging from 15 to 20 % can be achieved.

Furthermore, it is also important to note that the proposed fast SOM algorithm is able to obtain this speed-up while maintaining an accuracy which is similar compared to the conventional SOM. Our fast SOM algorithm outperforms other simplified SOM algorithms such as [8, 9].

In Table 1 the recognition accuracy of the conventional, our fast version and also Pena’s [10] SOMs are shown. These were obtained by various numbers of iterations and each iteration equals training the feature map with 60,000 input vectors. Afterwards, the SOM is tested with 10,000 test samples. Finally the feature maps of the conventional and proposed SOMs are shown in Fig. 3. The neurons of both maps clearly organized themselves and clusters can be distinguished.

**Table 1** Recognition accuracy on MNIST database

Iterations	Conventional SOM (%)	Proposed fast SOM (%)	Pena's SOM (%)
1	82.34	81.83	64.96
10	84.93	83.24	66.03
100	85.01	83.79	66.81
200	85.74	84.13	67.24



**Fig. 3** Feature map after training with MNIST dataset. **a** Feature map of proposed SOM. **b** Feature map of conventional SOM

## 6 Conclusion

This paper proposes a fast SOM algorithm for handwritten digit recognition which simplifies the conventional SOM by removing complex computations in the weight distance calculation, the learning rate function and the neighborhood function. After evaluating the performance in software simulation, we conclude that the proposed fast SOM algorithm can reach the goal of accelerating to some extent, maintain similar recognition accuracy compared to the conventional SOM and performs much better than other simplified SOM methods.

## References

1. Kohonen T (1990) The self-organizing map. Proc IEEE 78(1):1464–1480
2. Kohonen T, Kaski S, Lagus K et al (2000) Self organization of a massive document collection. IEEE Trans Neural Netw 11(3):574–585
3. Silven O, Niskanen M, Kauppinen H (2003) Wood inspection with non-supervised clustering. Mach Vis Appl 3:275–285

4. Mu-Chun S, Hsiao-Te C (2000) Fast self-organizing feature map algorithm. *IEEE Trans Neural Netw* 11(3):721–732
5. Nordström T (1992) Designing parallel computers for self organizing maps. In: Fourth Swedish workshop on computer system architecture
6. Lobo VJ, Bandeira N, Moura-Pires F (1998) Distributed Kohonen networks for passive sonar based classification. In: International conference on multisource-multisensor information fusion, Las Vegas
7. Yaohua Y, Daminda A (2006) Batch implementation of growing self-organizing map. In: International conference on computational intelligence for modelling control and automation, and international conference on intelligent agents, web technologies and internet commerce
8. Pena J, Vanegas M (2006) Digital hardware architecture of Kohonen's self organizing feature maps with exponential neighboring function. In: IEEE international conference on reconfigurable computing and FPGA
9. Agundis R, Girones G, Palero C, Carmona D (2008) A mixed hardware/software SOFM training system. *Computaciny Sistemas* 4:349–356
10. Porrman M, Witkowski U, Ruckert U (2006) Implementation of self-organizing feature maps in reconfigurable hardware. In: FPGA implementations of neural networks. Springer, Heidelberg, pp 247–269
11. LeCun Y, Cortes C, The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>

# Frequent Graph Pattern Mining with Length-Decreasing Support Constraints

Gangin Lee and Unil Yun

**Abstract** To process data which increasingly become larger and more complicated, frequent graph mining was proposed, and numerous methods for this has been suggested with various approaches and applications. However, these methods do not consider characteristics of sub-graphs for each length in detail since they generally use a constant minimum support threshold for mining frequent sub-graphs. Small sub-graphs with a few vertices and edges tend to be interesting if their supports are high, while large ones with lots of the elements can be interesting even if their support are low. Motivated by this issue, we propose a novel frequent graph mining algorithm, Frequent Graph Mining with Length-Decreasing Support Constraints (FGM-LDSC). The algorithm applies various support constraints depending on lengths of sub-graphs, and thereby we can obtain more useful results.

**Keywords** Frequent graph mining · Length-decreasing support constraint · Sub-graph

## 1 Introduction

As data generated from the real world have been complicated and large increasingly, previous frequent pattern mining methods, which find frequent patterns from simple database composed of items, have been faced with limitations that cannot deal with these large and complex data. Thereafter, to overcome this problem, frequent graph mining has been proposed [2–5]. However, existing frequent graph mining methods

---

G. Lee · U. Yun (✉)

Department of Computer Science, Chungbuk National University,  
Cheongju-si, South Korea  
e-mail: yunei@chungbuk.ac.kr

G. Lee

e-mail: abcnaarak@chungbuk.ac.kr

extract frequent sub-graphs with only one minimum support constraint which is set in the early mining procedure regardless of sub-graphs' lengths. Therefore, they have the following problem. Small sub-graphs having a few vertices and edges tend to be interesting if they have high support values. In contrast, large sub-graphs having many vertices and edges can be interesting even though they have low supports. However, the previous methods cannot find interesting large sub-graph patterns with lower supports than a given minimum support threshold since the threshold is fixed regardless of patterns' lengths. To solve the problems, we propose a novel frequent graph mining algorithm, called Frequent Graph Mining with Length-Decreasing Support Constraints (FGM-LDSC).

## 2 Related Work

Frequent graph mining began from Broad First Search (BFS)-based methods, and thereafter, Depth First Search (DFS)-based mining methods have been studied actively. In addition, graph mining can be applied in other data mining area such as classification [7], and regression analysis [7], and so on. As fundamental graph mining algorithms, there are famous algorithms such as FFSM, gSpan, Gaston [4, 5], etc. Especially, Gaston is a state-of-the-art algorithm which has the fastest runtime performance among them. In addition, there are numerous graph mining algorithms such as applying weight conditions [1, 2], using abbreviated notations called maximal and closed sub-graphs [6], finding frequent sub-graphs with a strong correlation [3], and so on.

LPMiner/SLPMiner [8] is a fundamental frequent pattern mining algorithm applying length-decreasing support constraints. Thereafter, WSLPMiner [9] was proposed, which can mine weighted sequential frequent patterns in the same environment, where the length means the number of items belonging to any pattern (or a set of items). However, these algorithms deal with only itemset databases, and therefore, they are not suitable for mining frequent sub-graphs from databases consisting of graphs.

## 3 Frequent Graph Mining with Length-Decreasing Support Constraints

### 3.1 Length-Decreasing Support Constraints on Frequent Graph Mining

Previous frequent graph mining methods generally consider only one standard, a single minimum support threshold when they extract frequent sub-graphs. However, this is unsuited for determining whether all of the sub-graphs are actually

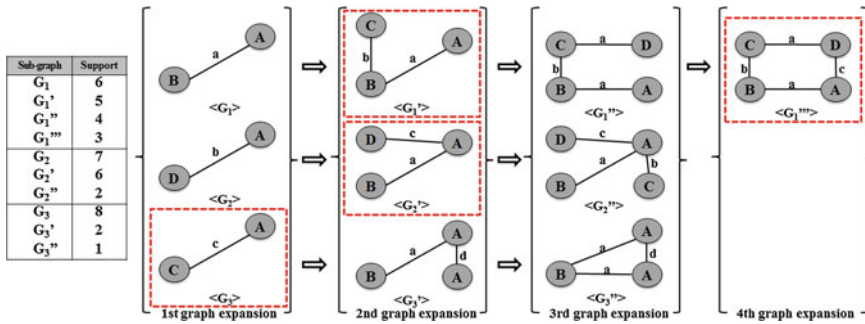


Fig. 1 An example of graph expansions

valid or not. Recall that certain sub-graphs with a small number of vertices and edges tend to be interesting if they have high supports. In contrast, other sub-graphs having lots of the elements can be interesting even though they have relatively low supports.

*Example 1* Figure 1 represents an example of graph expansions for mining frequent sub-graphs, where we assume that sub-graphs within the red dotted rectangles are interesting patterns and have to be extracted. In general frequent graph mining, a minimum support threshold has to be set as 3 to mine all of the interesting sub-graph patterns. However, since the method extracts not only interesting patterns but also all of the frequent but uninteresting ones with 3 or more supports such as  $\{G_1, G_1'', G_2\}$ , it eventually repeats mining operations many times to generate these meaningless sub-graphs.

To reduce these inefficient operations and find only interesting sub-graphs, we propose length-decreasing support constraints suitable for graph structures.

**Definition 1** A length of any sub-graph is determined by vertices and edges composing the sub-graph. If a certain sub-graph,  $G$  has a path or free tree structure, its length,  $length(G)$  is denoted as  $length(G) = \#$  of vertices included in  $G$ . On the other hand, consider that  $G$  has a cyclic graph form. Let  $G_{prev}$  be a graph such as a path or a free tree just before  $G$  is expanded as a cyclic graph. Then,  $length(G)$  is calculated by the equation,  $length(G) = \#$  of vertices in  $G_{prev} + \#$  of cyclic edges added to  $G$ .

FGM-LDSC assigns minimum support thresholds with respect to each sub-graph’s length depending on Definition 1, where these threshold values are gradually decreased from high to low values.

### 3.2 Pruning Strategy Retaining Anti-Monotone Property

As lengths of generated sub-graphs increase in FGM-LDSC, minimum support thresholds corresponding to each length are inversely decreased. In this environment, if we conduct mining process as a general method, fatal pattern losses can be caused since it does not satisfy Anti-monotone property (or downward closure property). To consider and overcome this problem, we propose measures and strategies for effectively removing unneeded sub-graphs as well as satisfying Anti-monotone property.

**Definition 2** Let  $GDB$  be a certain graph database and  $GL$  be a set of lengths for sub-graphs generated from  $GDB$ , denoted as  $GL = \{gl_1, gl_2, gl_3, \dots, gl_n\}$ . Then, a set of minimum supports for  $GL$ ,  $MS$  is denoted as  $MS = \{ms_1, ms_2, ms_3, \dots, ms_n\}$ , where subscripts means sub-graphs' lengths and the relation,  $ms_i \geq ms_j$  is satisfied for  $1 \leq i < j \leq n$  depending on the length-decreasing support constraint technique. Let  $\min(MS)$  be the lowest support among the values of  $MS$ , and then, we use the  $\min(MS)$  as a minimum support threshold since the value guarantee Anti-monotone property.

$$\min(MS) = \min_{1 \leq k \leq n}(ms_k) \quad (1)$$

Depending on the Definition 2, the minimum support for the sub-graph with the longest length is assigned as  $\min(MS)$ . Thus, if any sub-graph,  $G$  has a support less than  $\min(MS)$ , it means that  $G$  and all of the possible super patterns of  $G$  also have lower supports than  $\min(MS)$  since their supports become smaller as  $G$  is gradually expanded according to Definition 2, which satisfies Anti-monotone property. Consequently, it is certain that pruning  $G$  and  $G$ 's super patterns does not cause any problem.

*Example 2* Consider Fig. 1 again, where  $MS$  is set as  $MS = \{ms_1, ms_2, ms_3, ms_4\} = \{8, 5, 5, 3\}$  and we assume that  $G_3, G_1', G_2'$ , and  $G_1'''$  are interesting sub-graphs which FGM-LDSC has to extract. Then,  $\min(MS) = 3$  according to Eq. (1). In the first graph expansion ( $ms_1 = 8$ ),  $G_3$  is only a valid pattern since its support is larger than 8, while the others,  $G_1$  and  $G_2$  become invalid ones. However, FGM-LDSC does not prune them since their supports are higher than 3(=  $\min(MS)$ ). In the next expansion ( $ms_2 = 5$ ),  $G_1'$  and  $G_2'$  are interesting sub-graphs while  $G_3'$  is meaningless graph and also pruned permanently since its support is lower than not only 5(=  $ms_2$ ) but also 3(=  $\min(MS)$ ). Especially in here, we can show that  $G_1'$  and  $G_2'$ , which are infrequent sub-graphs at the first expansion step, are changed as frequent patterns in the current step due to the  $\min(MS)$ . In the third ( $ms_3 = 5$ ),  $G_1''$  is not pruned since its support is higher than 3 although it is lower than 5, while  $G_2''$  is permanently pruned since its support  $< \min(MS)$ .  $G_3''$  is not even considered due to pruning  $G_3'$ . In the last expansion ( $ms_4 = 5$ ),  $G_1'''$ , which grow



<b>input:</b> a graph database, $GDB$ , a set of supports for each length, $MS$ <b>output:</b> a set of frequent sub-graph patterns, $S$
<i>Mining_graph_patterns</i> ( $GDB, \delta, \lambda, \omega$ ) 1. calculate $\min(MS)$ // according to equation (1) 2. find all vertices and edges such that their support $\geq \min(MS)$ in $GDB$ 3. for each vertex, $v$ in a set of the found frequent vertices, $V$ do 4. a sub-graph, $SG \leftarrow v$ 5. a set of valid edge, $E' \leftarrow$ edges which can be attached to $v$ among the found frequent edges, $E$ 6. current graph state, $GS \leftarrow$ "path" 7. $S = S \cup \text{Expanding\_graph}(SG, E', GS)$
<i>Expanding_graph</i> (a sub-graph $SG$ , a set of edges $E$ , current graph state $GS$ ) 1. for each edge, $e$ in $E$ do 2. if $GS$ is "path" or "free tree" do 3. generate an expanded path or free tree, $SG'$ of $SG$ adding $e$ and a corresponding vertex, $v$ 4. calculate $\text{length}(SG')$ // depending on Definition 1 5. else generate an expanded cyclic graph, $SG'$ of $SG$ adding only $e$ 6. calculate $\text{length}(SG')$ // depending on Definition 1 7. select current minimum support for $\text{length}(SG')$ , $ms$ from $MS$ 8. if support of $SG' \geq \min(MS) \delta$ do 9. if support of $SG' \geq ms$ do 10. $S = S \cup SG'$ 11. else discard $SG'$ // however, $SG'$ is not pruned 12. else $e \leftarrow$ the next edge in $E$ and goto line 1 // $SG'$ is pruned 13. $E' \leftarrow$ a set of valid edges that can be attached to $SG'$ 14. $GS \leftarrow$ current graph state of $SG'$ 15. $S = S \cup \text{Expanding\_graph}(SG', E', GS)$

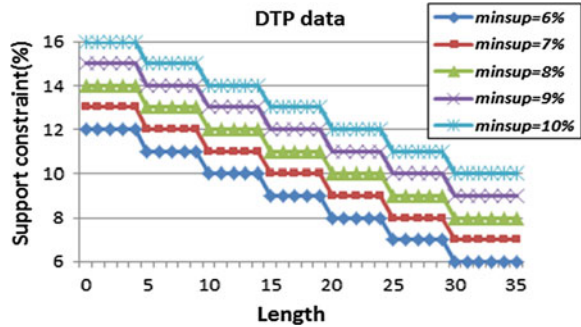
Fig. 2 FGM-LDSC algorithm

from the infrequent  $G_1''$ , becomes a frequent and interesting sub-graph again since it satisfies the  $\min(MS)$  condition.

### 3.3 FGM-LDSC Algorithm

Figure 2 presents frequent graph mining procedure performed by FGM-DLSC. In the function: *Mining\_graph\_patterns*, FGM-DLSC calculates  $\min(MS)$  with respect to the inputted  $MS$ . After that, it finds frequent vertices and edges satisfying the  $\min(MS)$  condition and extracts frequent sub-graph patterns by expanding graphs regarding the found elements. After the *mining\_graph\_patterns* function calls its sub-function, *Expanding\_graph*, FGM-LDSC conducts the graph expansion step and computes the length of the resulting graph for each edge, where it performs appropriate operations depending on whether the current graph state is a path, a free tree, or a cyclic graph. Thereafter, it finds the  $ms$  value corresponding to the calculated length from  $MS$ , and determines whether the currently expanded graph is frequent or has to be pruned. Then, it conducts a series of processes for  $SG'$  satisfying the condition of line 8, and continues the graph expansion steps with respect to  $SG'$  through recursive calls of this routine itself.

**Fig. 3** Support constraints of DTP and PTE datasets



## 4 Performance Analysis

### 4.1 Experimental Environment

In this section, performance evaluation results for the proposed algorithm, FGM-FDSC are presented. A target algorithm compared to FGM-FDSC is Gaston [4, 5], which is a state-of-the-art frequent graph mining algorithm. The two algorithms were written as the C++ language and ran with 3.33 GHz CPU, 3 GB RAM, and WINDOWS 7 OS environment. For these experiments, we used a real graph dataset, named DTP. Details of the dataset are available at [4, 5]. Figure 3 represents a distribution of length-decreasing support constraints for the dataset, DTP.

### 4.2 Experimental Results

Figure 4 shows the results for the number of frequent sub-graph patterns and runtime performance for the DTP dataset. As shown in the left figure, FGM-LDSC dramatically reduces sub-graph patterns, which are unnecessarily generated in mining process, by applying the proposed strategies and techniques. In contrast, Gaston extracts the enormous number of sub-graphs since it mines all of the patterns with higher supports than the single and fixed minimum support threshold. Especially, pattern results generated by Gaston are sharply increased as the threshold becomes low while the results by FGM-LDSC increase slightly and consistently since our algorithm selectively extracts actually interesting sub-graphs for each length. In the right part of Fig. 4, Gaston requires more time resources compared to FGM-LDSC in all of the cases since the Gaston finds and extracts all of the possible frequent sub-graphs not considering whether generated sub-graphs are really interesting or not. Especially when the minimum support threshold is lowered from 7 to 6 % in DTP, we can observe that corresponding runtimes of the two algorithms are greatly increased. However, our FGM-LDSC has an increasing

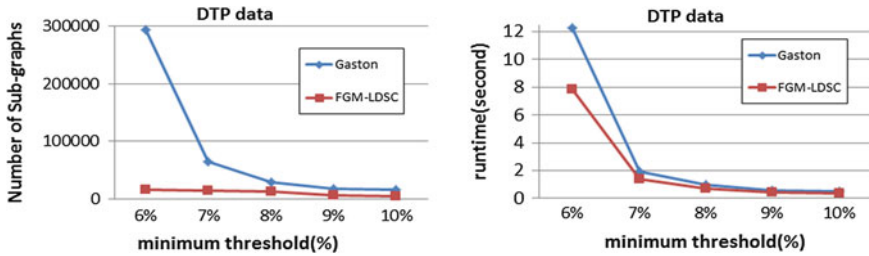


Fig. 4 The number of frequent sub-graphs and runtime results in DTP dataset

rate smaller than that of Gaston, and a gap between them becomes bigger whenever the threshold is lowered. This runtime interval occurs due to the meaningless sub-graphs pruned by Definition 2.

## 5 Conclusion

In this paper, we proposed a frequent graph mining algorithm with length-decreasing support constraints. Through the proposed algorithm, FGM-LDSC, we could obtain interesting sub-graphs having not only high supports and a few vertices and edges but also relatively low supports and a lot of the elements. Moreover, through the suggested pruning strategies and techniques, we demonstrated that our algorithm outperforms the previous method in terms of mining efficiency, as shown in the experimental results in this paper. Our algorithm can be applied to the other fields such as maximal/closed frequent graph mining, weighted frequent graph mining, and so on, and we expect that the strategies and techniques of our FGM-LDSC will contribute to advancing their mining performance in common with this paper.

**Acknowledgments** This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

## References

1. Günnemann S, Seidl T (2010) Subgraph mining on directed and weighted graphs. In: Proceedings of the 14th Pacific-Asia conference on knowledge discovery and data mining, pp 133–146
2. Hintsanen P, Toivonen H (2008) Finding reliable subgraphs from large probabilistic graphs. *Data Mining Knowl Discov* 17(1):3–23
3. Lee G, Yun U (2012) An efficient approach for mining frequent sub-graphs with support affinities. In: Proceedings of the 6th international conference on convergence and hybrid information technology, Korea, pp 525–532

4. Nijssen S, Kok JN (2004) A quickstart in frequent structure mining can make a difference. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, pp 647–652
5. Nijssen S, Kok JN (2005) The Gaston tool for frequent subgraph mining. *Electron Notes Theor Comput Sci* 127(1):77–87
6. Ozaki T, Etoh M (2011) Closed and maximal subgraph mining in internally and externally weighted graph databases. In: 25th IEEE international conference on advanced information networking and applications workshops, pp 626–631
7. Saigo H, Nowozin S, Kadowaki T, Kudo T, Tsuda K (2009) gBoost: a mathematical programming approach to graph classification and regression. *Mach Learn* 75(1):69–89
8. Seno M, Karypis G (2005) Finding frequent patterns using length-decreasing support constraints. *Data Min Knowl Disc* 10(3):197–228
9. Yun U, Ryu KH (2010) Discovering important sequential patterns with length-decreasing weighted support constraints. *Int J Inf Technol Decis Making* 9(4):575–599

# An Improved Ranking Aggregation Method for Meta-Search Engine

Junliang Feng, Junzhong Gu and Zili Zhou

**Abstract** A meta-search engine transmits the user's query simultaneously to several individual search engines and aggregate results into a single list. In this paper we conduct comparisons on several existing rank aggregation methods. Then based on those comparisons, an improved ranking aggregation method is proposed for meta-search engine. This method combines merits of the Borda's method and scaled footrule method. Extensive experiments show that this improved method outperforms the alternatives in most cases.

**Keywords** Search engine · Meta-search engines · Rank aggregation · Borda · Scaled footrule

## 1 Introduction

With the explosive growth of internet information, an effective search engine becomes more and more important for users to find their desired information from billions of web pages. Although the ranking algorithms (such as PageRank [1]) in search engines have been upgraded fast, but it's still impossible for one single search engine to cover all the web pages even for some famous general search engines, like Google and Bing. For specific queries, different search engines may

---

J. Feng (✉) · J. Gu · Z. Zhou  
Department of Computer Science and Technology,  
East China Normal University, Shanghai, China  
e-mail: jlfeng@ica.stc.sh.cn

J. Gu  
e-mail: jzgu@ica.stc.sh.cn

Z. Zhou  
e-mail: zlzhou@ica.stc.sh.cn

only search a subset of the internet. Meta-search engine [2] may help to improve this problem. Constructing a meta-search engine is quite desirable, while the most challenging problem for meta-search is the ranking aggregation method [3].

In this paper, a meta-search engine named ICASearch is implemented. Besides this, we make comparison among several ranking aggregation methods [4], and propose an improved ranking aggregation method. The proposed ranking method is evaluated on our meta-search engine system. The experiment results show that this optimal method has more precise results than the general methods.

## 2 Meta-Search Engine

A meta-search engine is a system, which fuses the search results from several individual search engines into a single result list. So it enables users to provide search criteria only once and access several search engines simultaneously [5]. When a query arrives, the meta-search engine forwards the query to several constituent components. Then the constituent components process the query and dispatch the query to several general search engines. Each engine responds to the query with a ranked result lists. Finally, the meta-search engine merges all the results lists, and returns the merged list to the user. Now meta-search systems have drawn attentions from both academic and commercial areas.

For web search engines, we only focus on the results in first 1 or 2 pages, that's to say, we only need to consider the partial list, the top 20–50 results from each engine, and merge these lists into a final result list for our user. This is different from merging the full list of the results. It is a more challenging task for aggregation.

Meta-search engine has some advantages over general search engines. Firstly, a more improved precision by merging multiple results, particularly for the web search engine. Secondly, it can provide a more consistent and reliable performance than individual search engines [6]. Thirdly, the architecture [7] of meta-search engine has a better modularity. It allows a single search engine system to be divided into smaller, special components.

## 3 Architecture of ICASearch

ICASearch, is a meta-search engine aiming to provide better search results for its users. The architecture of ICASearch is depicted in Fig. 1. There are three modules in the system: (1) search engine module; (2) controller module; (3) third party module. We will present the internal details in the following sections.

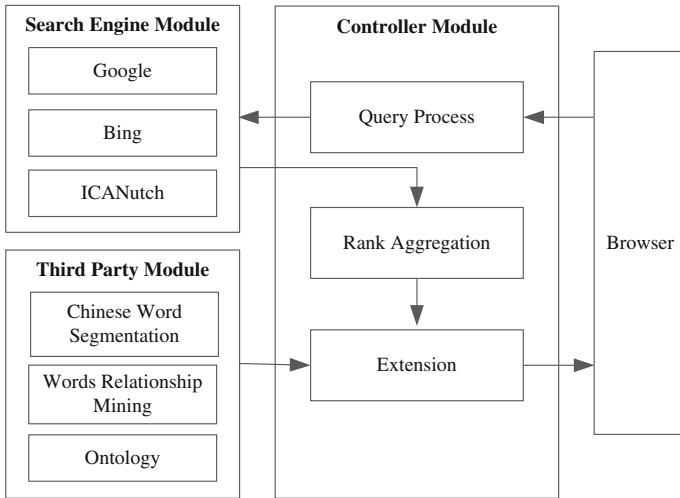


Fig. 1 The Architecture of ICASearch

### 3.1 Search Engine Module

The search engine module contains several individual search engines. In this system, we use three search engines: Google, Bing and ICANutch. ICANutch is a local search engine. It crawls web pages mainly on internet and mobile applications news. Our experiments are also conducted on these topics.

### 3.2 Controller Module

This module contains three parts: (1) Query processing. It performs word segmentation when user submits a query, and then dispatches it to the search engine module. (2) Rank aggregation. When relevant results are returned from Search Engine Module, it will merge the results into the final result. (3) Extension. After rank aggregation, we will use the third party module to get relationships and relevance words for the query. This information is shown on the web page to facilitate our user in further search.

### 3.3 Third Party Module

Three-third party APIs are invoked in this system. (1) Chinese Word Segmentation plugin. As our queries are mainly in Chinese currently, so it's appropriate for us to use Chinese word segmentation plugin to get a more precise understanding of

user's query. (2) Words Relationship Mining API. We use this API to extend our search results. (3) Ontology library. There is an ontology system which provides us with some ontologies to do query extensions on our search results.

## 4 Rank Aggregation Methods

There exist various methods for aggregate result from difference rank-ordered lists [3, 4]. In most case the methods can be classified with the following rule: (1) based on the score; (2) based on the rank; (3) required the training data or not [8]. In this paper, we will mainly discuss the methods based on the rank.

### 4.1 Preliminaries

Let  $U$  presents the universe, a set of items. An ordered list  $\tau$  with respect to  $U$  is an ordering of a subset  $S \subseteq U$ , i.e.,  $\tau = [x_1 \geq x_2 \geq \dots \geq x_k]$ , with each  $x_i \in S$ , and  $\geq$  is some ordering relation on  $S$ . If  $i \in \tau$ , let  $\tau(i)$  denotes the position or rank of  $i$  (a high ranked or preferred element has a low-numbered position in the list). For a list  $\tau$ ,  $|\tau|$  denotes the number of elements. With  $w^\tau(i)$  we will denote the normalized weight of item  $i \in \tau$  in ranked list  $\tau$  [5].

### 4.2 Borda's Method

In Borda's method, we use the Borda rank normalization [9] to calculate the  $w^\tau(i)$ , for an item  $c \in U$

$$w^\tau(c) = \begin{cases} 1 - \frac{\tau(c)-1}{|U|}, & \text{if } c \in \tau \\ \frac{1}{2} - \frac{|\tau|-1}{2 \cdot |U|}, & \text{if } c \notin \tau \end{cases} \quad (1)$$

For given ordered lists  $\tau_1, \tau_2, \dots, \tau_k$ , then for each element  $c \in S$  and list  $\tau_i$ , we assign the  $w^{\tau_i}(c)$  to each  $c$  in  $\tau_i$  as  $B_i(c)$ , so the total Borda score  $B(c)$  is defined as  $\sum_{i=1}^k B_i(c)$ . After calculated all the Borda score, we could sort the result in decreasing order by the total Borda score. The computation complexity of this method is  $O(n^2)$ ,  $n$  denotes the total size of the partial list results.



### 4.3 Scaled Footrule Optimization Method

The scaled footrule method use the footrule distances to rank the various results. In the full list scenario, the footrule optimal aggregation can be solve by construct a bipartite graph from the lists and compute the minimum cost perfect matching [6].

For partial lists  $\tau_1, \tau_2, \dots, \tau_k$ , we defined a weighted bipartite graph  $(C, P, W)$ .  $C$  denotes the set of nodes to be ranked.  $P = \{1, 2, 3, \dots, n\}$  denotes the set of available positions. The weight  $W(c, p)$  is the total footrule distance of ranking the element  $c$  in position  $p$ , given by

$$W(c, p) = \sum_{i=1}^k |\tau_i(c) / |\tau_i| - p/n| \quad (2)$$

So this problem has been converted to calculate the minimum cost perfect matching problem in a bipartite graph. In this paper, we use the Kuhn–Munkres algorithm to solve this matching problem. The computation complexity of the algorithm is  $O(n^3)$ ,  $n$  denotes the total size of the partial list results.

### 4.4 B-F-Rank Method

The Borda method focus the position on the initial return lists, and the scaled footrule optimization method will consider not only the original positions, but also consider the final rank positions (as the bipartite graph defined in Sect. 4.3). After research, we found that, in the first few results the Borda method are more accurate, but the precision declined quickly while the result size increase increasing. The precision change in scale footrule method is relatively stable. So we propose a method named B-F-Rank, which combines the two methods to rank the final result, and suppose it will get more accurate results. There are three steps for the method.

1. Get two aggregated result list  $L_B$  and  $L_F$ .  $L_B$  is the result list ordered by Borda's method, while  $L_F$  is ordered by scaled footrule method. As we know, the elements in the  $L_B$  and  $L_F$  is the same, the only difference is the ranking position.
2. Use the Eq. 1, to normalize the  $L_B$  and  $L_F$ , so for each element  $c$ , we get two weights,  $w_B(c)$  and  $w_F(c)$ , that are been normalized by  $L_B$  and  $L_F$  respectively.
3. The new weight of element  $c$  is given by

$$w_{B-F-Rank}(c) = \alpha \cdot w_B(c) + \beta \cdot w_F(c) \quad (3)$$

$\alpha$  and  $\beta$  is real value and  $\alpha + \beta = 1.0$ . Then rank the element list by  $w_{B-F-Rank}$  in decreasing order and we will get the final result ordered by B-F-Rank method.

## 5 Experimental Evaluation and Results

In our system, three search engines, Google, Bing, and ICANutch are used. The Borda's method and scaled footrule method are taken as benchmark methods. We prepared 10 queries, all those queries focus on the internet and mobile applications news. This experiment setting decreases the impact of the diverse result set and lets us focus on the rank aggregation method optimization. In each query round, we select the top 50 results from each engine, and after rank aggregation use specific method, our measurement is based on the precision of the top 50 of the final results. The precision is assessed by human judges. Average precision and precision in top-N results (P@N) are chosen as evaluation criteria. Figure 2 is the precision for the methods run in each query round, it shows the B-F-Rank result is more precise.

Figure 3 shows, at the rank N-th position, the average precision of 10 queries' results that use specific method. The B-F-Rank method shows a more stable and better precision curve than the other two methods.

Fig. 2 Precision of results

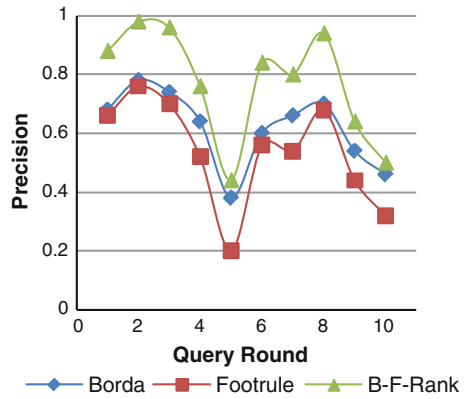
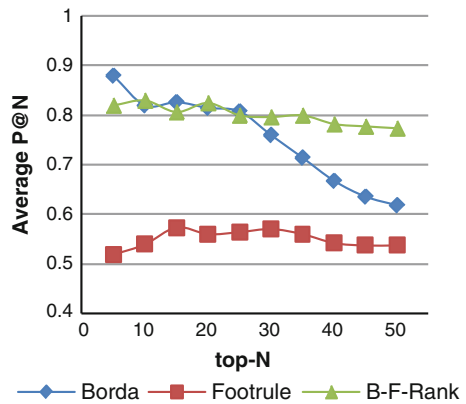


Fig. 3 P@N



**Table 1** Average precision of the three methods

Methods	Borda	Scaled footrule	B-F-Rank
Average precision	0.618	0.538	0.774

Table 1 presents the average precision of the three methods in top 50 results. The result shows that the B-F-Rank method outperforms the alternatives in most cases.

## 6 Conclusions and Future Work

In this paper, several rank aggregation methods are discussed and evaluated. An improved ranking aggregation method named B-F-Rank is proposed. The evaluation result shows that the proposed method outperforms classical methods, Borda's method and scale footrule optimization method, in most cases.

The future work involves, incorporating more search engines in our study and adding semantic and ontology extension in the queries. Furthermore we could also incorporate term similarities and correlation in our aggregation method.

**Acknowledgments** The work is supported by the Shanghai Scientific Development Foundation with Grant No. 11530700300, and Shandong Province Young and Middle-Aged Scientists Research Awards Fund with No. BS2010DX012.

## References

1. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. *Comput Netw* 30(1–7):107–117
2. Aslam JA, Montague M (2001) Models for metasearch. In: *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*. ACM Press, New York, pp 276–284
3. Dwork C, Ravi K, Moni N, Siva Kumar D: Rank aggregation methods for the web. In: *Proceedings of the tenth international conference on World Wide Web*. ACM Press, New York, pp 613–622
4. Farah M, Vanderpooten D (2007) An outranking approach for rank aggregation in information retrieval. In: *Proceedings of the 30th annual international ACM SIGIR conference*. ACM Press, New York, pp 591–598
5. Woodley Alan P, Geva S (2005) ComRank: metasearch and automatic ranking of XML retrieval systems. In: *International conference on cyberworlds*. IEEE Press, Singapore, pp 146–154
6. Montague M, Aslam JA (2001) Metasearch consistency. In: *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*. ACM Press, New York, pp 386–387

7. Gulli A, Signorini A (2005) Building an open source meta-search engine. In: Special interest tracks and posters of the 14th international conference on World Wide Web. ACM Press, New York, pp 1004–1005
8. Aslam JA, Montague M (2001) Models for metasearch. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. ACM Press, New York, pp 276–284
9. Renda ME, Straccia U (2003) Web metasearch: rank vs. score based rank aggregation methods. In: Proceedings of the 2003 ACM symposium on applied computing. ACM Press, New York, pp 841–846

**Part V**  
**Multimedia and Ubiquitous Computing**  
**Security**

# Identity-Based Privacy Preservation Framework over u-Healthcare System

Kambombo Mtonga, Haomiao Yang, Eun-Jun Yoon  
and Hyunsung Kim

**Abstract** The digitization of patient health information has brought many benefits and challenges for both the patient and doctor. But security and privacy preservation have remained important challenges for wireless health monitoring systems. Such concerns may result in reluctance and skepticism towards health systems by patients. The reason for this skepticism is mainly attributed to the lack of assurances about the way patient health information is handled and the implications that may result from it on patients' privacy. This paper proposes an identity-based privacy preservation framework over u-healthcare systems. Our framework is based on the concepts of identity-based cryptography and non-interactive key agreement scheme using bilinear pairing. The proposed framework achieves authentication, patient anonymity, un-traceability, patient data privacy and session key secrecy, and resistance against impersonation and replay attacks.

**Keywords** u-healthcare · Privacy preservation · Data privacy · Identity-base encryption · Non-interactive key agreement

---

K. Mtonga

Department of IT Convergence, Kyungil University, 712-701, Kyongsansi, Kyungpook Province, Korea

H. Yang

College of Computer Science and Engineering, UEST of China, Chengdu 610054, China

E.-J. Yoon · H. Kim (✉)

Department of Cyber Security, Kyungil University, 712-701, Kyongsansi, Kyungpook Province, Korea

e-mail: kim@kiu.ac.kr

E.-J. Yoon

e-mail: ejyoon@kiu.ac.kr

## 1 Introduction

Advances in telecommunication technology have made possible data transmission over the wireless system. This has enabled remote patient monitoring systems which collect disease-specific metrics via wireless biomedical devices used by patients. The collected health data is transmitted to a remote server for storage and later examination by the healthcare professionals. However, the different usage scenarios of remote monitoring systems e.g. in-hospital and home monitoring have resulted in diverse security and privacy concerns [1, 2]. Ensuring privacy and security of health information, including information in the electronic health record (EHR), is the key component to build the trust required to realize the potential benefits of electronic health information exchange [3]. Many protocols to enhance privacy and security of remotely collected patient health information have been put forward by researchers [4–7].

In this paper, we propose an identity-based privacy preserving framework over u-healthcare systems. In our framework; (1) Identity-based cryptography (IBC) is adopted to ensure the secure transmission, receiving, storing and access of patient data. (2) The doctor can give feedback directly to the patient on his/her health condition. (3) The patient and doctor can establish a secure channel directly by establishing session key with non-interactive manner.

## 2 Preliminaries

In this section we briefly review our threat model and present notations used throughout the paper. For details on Bilinear pairing, Bilinear Diffie-Hellman problem and non-interactive identity-based key agreement please refer to [8].

### 2.1 Threat Model

There are many threats to patient privacy and security in remote health monitoring system. Some of these threats include: data breach by insiders, insider curiosity, accidental disclosure and unauthorized intrusion of network system by outsiders [9]. In our framework, we aim to enhance patient data and identity privacy against both insider and outsider attacks i.e. attacks being provoked by an entity that is part of the network or by an outside entity who has, somehow, gained access to the network. For outside attackers, authentication and IBC-based data encryption have been adopted in order to prevent the attacker from gaining access to patient's data. In addition, a patient uses pseudo-ID when sending his/her health data and the data is stored encrypted, hence even if an insider accesses the patient's records, patient identity remains protected. Figure 1 illustrates our threat model.

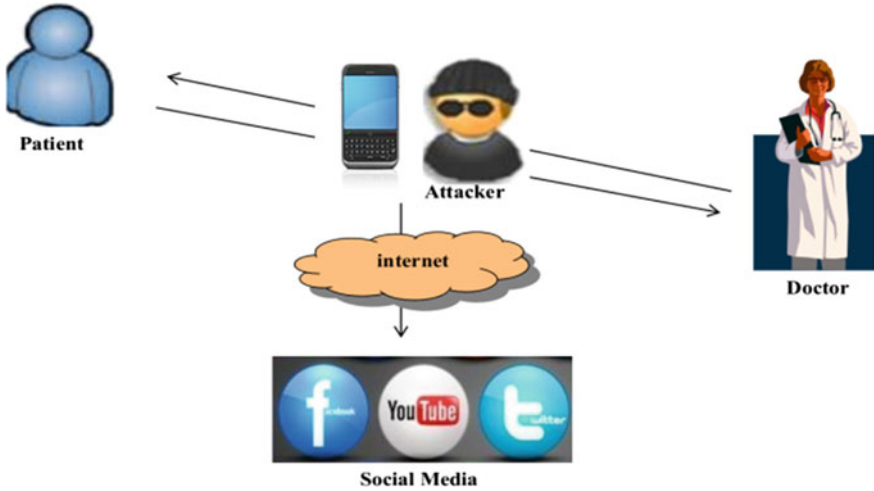


Fig. 1 Threat model to patient privacy

### 2.2 Notations

Table 1 introduces the notations used throughout the remainder of this paper.

Table 1 Notations

Notation	Meaning
$TA$	Trusted authority
$U_i$	Patient $i$
$D_l$	Doctor $l$
$S_x$	Private key for entity $x$
$PR_{U_i}$	Set of private keys for patient $i$
$Q_x$	Public key for entity $x$
$ID_x$	Identity of entity $x$
$SK_{x-y}$	Shared key between entity $x$ and $y$
$pid_j$	$j$ th pseudo-ID for patient $i$
$H_1(.)$	Hash function; $H_1: \{0, 1\}^* \rightarrow G_1$
$H_2(.)$	Hash function; $H_2: \{0, 1\}^* \rightarrow Z_q^*$
$M$	Patient health information
$M'$	Medical advice from doctor
$T_x$	Time stamp generated by entity $x$
$\hat{e}$	Bilinear map
$A$	Master secret key for health monitoring server
$\parallel$	Concatenation



### 3 Privacy Preservation Framework

In this section, we propose an identity-based privacy preservation framework over u-healthcare systems.

#### 3.1 System Initialization

In our framework, health monitoring server performs the role of trusted authority. To initialize the system, the health monitoring server first runs the set up for bilinearity as mentioned in Sect. 2.1A. The health monitoring server then chooses a random number  $a \in \mathbb{Z}_q^*$  as the master key and computes the corresponding public key  $Q_{TA} = aP$ . It also chooses two secure collision free hash functions  $H_1(\cdot)$  and  $H_2(\cdot)$ , where  $H_1(\cdot) : \{0, 1\}^* \rightarrow G_1$  and  $H_2(\cdot) : \{0, 1\}^* \rightarrow G_2$ . The server then publishes the public system parameters as  $\{G_1, G_2, q, P, Q_{TA}, H_1(\cdot), H_2(\cdot)\}$  and keeps the master key  $a$ , secret (Fig. 2).

#### 3.2 Registration

Let  $U_i$  be a patient seeking medical help from doctor  $D_j$ . Since each doctor of the hospital is registered with the health monitoring server and the server keeps a profile of the doctor,  $U_i$  registers directly with the health monitoring server and he/she is assigned a doctor depending on her/his health problem. To register,  $U_i$

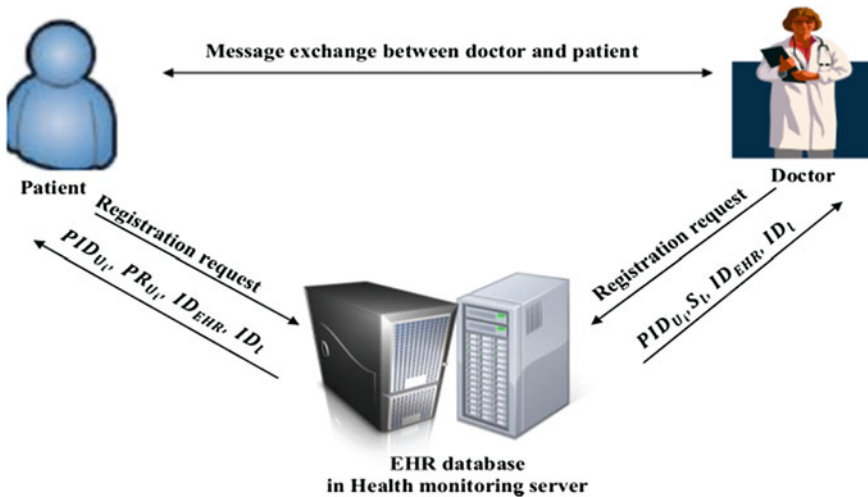


Fig. 2 Privacy preserving framework configuration

submits identity  $ID_i$  e.g. an email address or social security number to server. The server first validates  $ID_i$ . If the validation is successful, the server then chooses a family of  $n$  un-linkable pseudo-IDs given by

$$PID_{U_i} = \{pid_0, pid_1 \dots, pid_{j-1}, pid_j, pid_{j+1} \dots, pid_{n-1}\} \quad (1)$$

For each pseudo-ID  $pid_j$  in  $PID_{U_i}$ , the server computes the public key  $Q_j = H_1(pid_j)$  and the corresponding private key  $S_j = aH_1(pid_j)$ , such that the families of public and private keys are

$$PB_{U_i} = \{Q_0, Q_1, Q_2 \dots, Q_{j-1}, Q_j, Q_{j+1} \dots, Q_{n-1}\} \quad (2)$$

$$PR_{U_i} = \{S_0, S_1, S_2 \dots, S_{j-1}, S_j, S_{j+1} \dots, S_{n-1}\} \quad (3)$$

To allow revocation, the server adds an expiry date into  $pid_j$  such that each of the public keys  $Q_j = H_1(pid_j)$  is valid only before the specified expiry time  $t_j$ . After the specified time, the corresponding private key  $S_j = aH_1(pid_j)$  is revoked automatically. We also assume that the patient can only use the pseudo-IDs  $pid_j$ ,  $0 \leq j \leq n - 1$  sequentially. Finally, the server sends the whole tuples  $\{PID_{U_i}, aH_1(pid_j)\}$  for  $0 \leq j \leq n - 1$  to  $U_i$  via a secure channel. With these pseudo-IDs, the patient can constantly change his/her pseudo-ID to achieve anonymity and un-traceability during communication process in the u-healthcare environment.

Doctor  $D_i$  registers with the health monitoring server by providing his/her true identity  $ID_i$ . The server then computes  $S_i = aH_1(ID_i)$  the private key and  $Q_i = H_1(ID_i)$  the public key for doctor. It also computes  $S_{EHR} = aH_1(ID_{EHR})$  and  $Q_{EHR} = H_1(ID_{EHR})$ , the key pair for EHR.

### 3.3 Patient Health Information Transfer to EHR

To send her health data,  $U_i$  carries out the following steps:

- Pick a valid pseudo-ID  $pid_j$  in the sequence and corresponding private key  $S_j = aH_1(pid_j)$ .
- Using this private key, the patient computes a session key shared with doctor  $D_i$  as  $SK_{i-1} = \hat{e}(aH_1(pid_j), H_1(ID_i)) = \hat{e}(Q_j, Q_i)^a$ , and another session key shared with EHR as  $SK_{i-EHR} = \hat{e}(aH_1(pid_j), H_1(ID_{EHR})) = \hat{e}(Q_j, Q_{EHR})^a$ .
- Using  $SK_{i-1}$ , patient applies IBC-based encryption on her health data  $M$  as  $C = E_{SK_{i-1}}(M)$ .
- Use EHR's public key  $Q_{EHR}$  to encrypt message  $Y = E_{Q_{EHR}}(T_i || ID_i || pid_j || SK_{i-1} || SK_{i-EHR} || C)$  and send  $Y$  to EHR.

**Table 2** Patient health information table by EHR

Doctor ID	Patient ID	PHI
$ID_i$	$pid_j$	$C$
:	:	:

### 3.4 Patient Health Information Verification and Storing by EHR

When EHR receives patient health information, it carries out the following authentication steps:

- Use  $S_{EHR}$  to decrypt  $Y$  as  $\{T_i \| ID_i \| pid_j \| SK_{i-EHR} \| C\} = D_{SEHR}(Y)$ .
- Check if time stamp  $T_i$  is valid by verifying if  $T_i - T_{EHR} < \Delta T$  is satisfied, where  $T_{EHR}$  is time the message is received by EHR and  $\Delta T$  is predefined transmission delay. If the verification is not success, the message is rejected and HER generates a message asking patient to resend the message. This also helps to overcome replay attacks. Otherwise, EHR proceeds to verify the identity of doctor,  $ID_i$ .
- Once the verification of doctor is successful, EHR proceeds to compute  $SK_{EHR-i} = \hat{e}(aH_1(ID_{EHR}), H_1(pid_j))$  using the received  $pid_j$ .
- EHR then checks if  $SK_{EHR-i} = SK_{i-EHR}$ . Note that  $SK_{EHR-i}$  is equal to  $SK_{i-EHR}$  since;  $SK_{i-EHR} = \hat{e}(aH_1(pid_j), H_1(ID_{EHR})) = \hat{e}(Q_j, Q_{EHR})^a = \hat{e}(Q_{EHR}, Q_j)^a = SK_{i-EHR}$ .

If the above holds,  $U_i$  is authenticated by EHR, and EHR stores  $\{ID_i, pid_j, C\}$  in the database as shown in Table 2.

### 3.5 Patient Health Information Recovery by Doctor

To access patient's health data  $M$ , the doctor first gets her/himself authenticated to EHR by carrying out the following steps:

- Computes  $SK_{i-EHR} = \hat{e}(aH_1(ID_i), H_1(ID_{EHR})) = \hat{e}(Q_i, Q_{EHR})^a$ .
- Sends  $V = E_{Q_{EHR}}(T_1 \| pid_j \| ID_i \| SK_{i-EHR})$  encrypted with the public key of EHR,  $Q_{EHR}$ , as a request for the patient's health information. Since the doctor is aware that each of the pseudo-IDs has an expiry date and that they are used sequentially, when  $pid_j$  is chosen, the doctor chooses the one that is valid and current. Hence  $D_i$  can request for specific patient health information from EHR depending on the specified  $pid_j$ .
- When EHR receives the request  $V$ , it uses its private key  $S_{EHR}$  to decrypt the request, i.e.  $D_{SEHR}(V) = \{T_1 \| pid_j \| ID_i \| SK_{i-EHR}\}$ , and then checks if the received timestamp  $T_1$  satisfies the condition  $T_1 - T_{EHR} < \Delta T$ . If the verification is

success, EHR proceeds to check if the received pseudo-ID  $pid_j$  for  $U_i$  and the identity  $ID_l$  for  $D_l$  match the ones received from the patient.

- EHR then uses  $ID_l$  to compute a session key shared with  $D_l$ ,  $SK_{EHR-l} = \hat{e}(aH_1(ID_{EHR}), H_1(ID_l))$  and check if  $SK_{EHR-l} = SK_{l-EHR}$  holds. If the equation holds, EHR authenticates the doctor and sends  $\{C, pid_j\}$  to the doctor. Note that  $SK_{EHR-l} = \hat{e}(aH_1(ID_{EHR}), H_1(ID_l)) = \hat{e}(Q_{EHR}, Q_l)^a = \hat{e}(Q_l, Q_{EHR})^a = SK_{l-EHR}$ .

When  $D_l$  receives  $\{C, pid_j\}$ , he/she verifies  $pid_j$ . Since  $D_l$  already has  $pid_j$ , she/he can pre-compute  $SK_{l-i} = \hat{e}(aH_1(ID_l), H_1(pid_j))$  in advance or could compute after receiving the message from EHR. With this key,  $D_l$  successfully decrypts  $C$ , i.e.  $M = D_{SK_{l-i}}(C)$ .

The doctor can now analyze the patient's health information such that if there is need for immediate medical advice for the patient, the doctor generates medical advice  $M'$  and encrypts it with the session key  $SK_{l-i}$ . This session key is used to establish a secure channel for the subsequent communication sessions between the doctor and the patient till  $t_j$  the expiry date of  $pid_j$ .

For mutual authentication, after computing  $SK_{l-i}$ , the doctor can compute an authentication code,  $Auth = H_2(SK_{l-i} || pid_j || ID_l)$  and send it together with the response  $M'$  encrypted with  $SK_{l-i}$  as  $\{E_{SK_{l-i}}(M'), Auth\}$ . Upon receiving the doctor's response,  $U_i$  also generates a verification code  $Veri = H_2(SK_{l-i} || pid_j || ID_l)$  and checks if  $Auth = Veri$ . If the equation holds, then  $U_i$  believes that the medical advice  $M'$  came from legitimate doctor and that he/she has established a secure channel using the key  $SK_{l-i}$ . Otherwise the patient rejects the session. This protects the patient from bogus health advices from adversaries which could result in drug overdose or worse still unnecessary death.

## 4 Conclusion

In this article, we have presented a privacy preserving security framework over u-healthcare system. In our framework, patients are only pseudonymously identified hence protecting the patient from negative effects of identity theft such as fraudulent insurance claims by adversaries. However, since health monitoring server knows the patients' real identity, in case of apparent abuse via judicial procedure, this real identity can be revealed.

**Acknowledgments** This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (MEST) (NRF-2010-0021575) and was partially supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2012- H0301-12-2004) supervised by the NIPA (National IT Industry Promotion Agency).

## References

1. Varshney U (2003) Pervasive healthcare. *IEEE Comput* 36(12):138–140
2. Ng HS, Sim ML, Tan CM (2006) Security issues of wireless sensor networks in healthcare applications. *BT Technol J* 24(2):138–144
3. Appari A, Johnson ME (2010) Information security and privacy in healthcare: current state of research. *Int J Internet Enterp Manag* 6(4):279–314
4. Huang Q, Yang X, Li S (2011) Identity authentication and context privacy preservation in wireless health monitoring system. *Int J Comput Netw Inf, Security*, pp 53–60
5. Layouni M, Verslype K, Sandikkaya MT (2009) Privacy-preserving telemonitoring for eHealth data and applications security. *LNCS* 5645:95–110
6. Hasque MM, Pathan AK, Hong CS (2008) Securing u-Healthcare sensor networks using public key based scheme. In: *ICACT*, pp 17–20
7. Sax U, Kohane I, Mandl KD (2005) Wireless technology infrastructures for authentication of patients: PKI that rings. *J Am Med Informatics Assoc* 12(3):263–268
8. Kim H (2012) Non-interactive hierarchical key agreement protocol over hierarchical wireless sensor networks. *Commun Comput Inf Sci* 339(5):86–93
9. Dixon P (2006) Medical identity theft: the information crime that can kill you. The world privacy forum

# A Webmail Reconstructing Method from Windows XP Memory Dumps

Fei Kong, Ming Xu, Yizhi Ren, Jian Xu, Haiping Zhang and Ning Zheng

**Abstract** Retrieving the content of webmail from physical memory is one key issue for investigators because it may provide with useful information. This paper proposes a webmail evidence reconstructing method from memory dumps on Windows XP platform. The proposed method uses mail header format defined in RFC2822 and HTML frame based on specific webmail server to locate header and body respectively. Then webmail is reconstructed based on matching degree between FROM, TO(CC/BCC), DATE and SUBJECT fields of header and corresponding content extracted from body. The experiment results show that this method could reconstruct the webmail from memory dumps.

**Keywords** Digital forensics · Webmail · Memory dumps

---

F. Kong · M. Xu (✉) · Y. Ren · J. Xu · H. Zhang · N. Zheng  
College of Computer, Hanzhou Dianzi University, Hangzhou, China  
e-mail: 100061040@hdu.edu.cn

F. Kong  
e-mail: mxu@hdu.edu.cn

Y. Ren  
e-mail: renyz@hdu.edu.cn

J. Xu  
e-mail: jian.xu@hdu.edu.cn

H. Zhang  
e-mail: zhanghp@hdu.edu.cn

N. Zheng  
e-mail: nzheng@hdu.edu.cn

## 1 Introduction

Traditional digital forensic is mainly on permanent storage devices, but it could not work well to obtain volatile information. Webmail data in memory can provide wealthy of information such as crime plan of suspect and relationship of crime organization. And webmail reconstruction, for clues to solve the case and provide evidence in court, will play an increasingly important role.

To the best of our knowledge, there is no work focusing on webmail reconstructing from memory dumps. However, some past forensic research about web browser and email client has been done and achieved positive results. Rachid Hadjidj developed one e-mail forensic analysis software tool [1]. Pereira presented a method to recover Firefox remnants [2]. Oh et al., proposed a method [3] to collect information relevant to the case. But their methods cannot be directly used in webmail forensic from memory.

Here we study the reconstruction of webmail resided in memory, using matching degree between related mail fields. Our method scans memory dumps and locates web mail header and body using string match method without knowing process or kernel data structure.

## 2 Our Method

### 2.1 Basic Steps

The proposed method works in three steps: getting memory data, preprocessing dump file and reconstructing webmail.

We dump memory data and save them as a file following the rules [4]. Webmail accesses to mail server with HTTP, and there are three methods for HTTP data compressing. Specifically, we write a program to decompress dumps. If the traversed data is compressed data, it will be decompressed and then outputted, otherwise, the data will be directly outputted. Program read 4 KB data every time. If compressed data is separated in two blocks, program will read next block and decompress. The basic idea of reconstructing webmail is to locate and recover all headers and bodies of webmail, and then match them. According to RFC2822 [5], a complete webmail header must have FROM, TO(CC/BCC) and DATE fields. Every field has a field name, followed by a colon, a field body, and terminated by a CRLF. Since webmail content is processed in HTML form, so content of webmail are inserted with some HTML tags. Using these tags as fingerprint, we can find the body content.

## 2.2 Matching Degree Metrics

It is supposed that a full webmail includes a subject, which correlates with webmail body. Also, it contains a receiver's name, a sender's name and date at the greeting part and end part of message. Based on this assumption, the possibility of a header and body matched as one mail can be estimated based on the relation between the text of webmail body and DATE, FROM, TO(CC/BCC), SUBJECT field of the header.

Let  $D_H = \{d \mid d \text{ is a normalized date string from the DATE field in header } H\}$ ;  
 $E_H = \{e \mid e \text{ is an email address in FROM, TO(CC/BCC) field from header } H\}$ ;  
 $W_H = \{w \mid w \text{ is a word in SUBJECT field from header } H\}$ ;  
 $D_B = \{d \mid d \text{ is a normalized date string from text in mail body } B\}$ ;  
 $E_B = \{e \mid e \text{ is an email address from text in mail body } B\}$ ;  
 $W_B = \{w \mid w \text{ is a word from text in mail body } B\}$

**Definition 1** Matching degree (*MD*) between a mail header *H* and a body *B* denotes the probability of them belonged to a same webmail. It can be measured as:

$$MD(H, B) = \frac{|D_H \cap D_B| + |E_H \cap E_B| + |W_H \cap W_B|}{|D_H| + |E_H| + |W_H|} \times 100 \%$$

## 2.3 Related Algorithms

Idea of Go-Back-N in TCP protocol is taken to process length-fixed string in this paper. When reading data in dump file, it is not reading from the end but from the last N-1 bytes of last block. N is the length of longest string that would be located. This strategy could locate strings that across 2 blocks.

### (1) Webmail header recovering algorithms

MailHeaderRecovering is used to locate three required fields and recover webmail header. HeaderCompleteness is called to judge whether these fields are complete. If all three fields are complete, headers are saved to array *mh*[]. If one field terminates with non-ascii symbol or '\0', it is incomplete; otherwise it is separated.



**Algorithm 1:** MailHeaderRecovering( $F$ )

**Input:** the decoded memory dump files  $F$

**Output:** Mail Header Array  $mh[]$

```

read one block data  $b$  from dump file  $F$ ;
While(!EOF( $F$ ))
  if( $b \sim$  "From:" &&  $b \sim$  "To{cc|bcc}:" &&  $b \sim$  "Date:"
    && HeaderCompleteness( $b$ ) == CompleteHeader)
     $mh[i++] \leftarrow$  mail header in  $b$ ;
  read next block  $c$  from dump file  $F$ ;
  if(( $b \sim$  "From:" ||  $b \sim$  "To{cc|bcc}:" ||  $b \sim$  "Date:")
    && HeaderCompleteness( $b$ ) == SeparateHeader)
     $b \leftarrow b, c$ ;
  else  $b \leftarrow c$ ;

```

**Algorithm 2:** HeaderCompleteness( $b$ )

**Input:** a block data  $b$  from file  $F$ ,

**Output:** according the status of required fields, it returns CompleteHeader, SeparateHeader or IncompleteHeader.

```

 $Flag\{Date, From, To\} \leftarrow$  SeperateField;
foreach  $field$  in {Date, From, To}{
  if( $b \sim field$ )
    if( $field$  is terminated by CRLF)
       $Flag\{field\} \leftarrow$  CompleteField;
    else if( $field$  is terminated by Non-ASCII or '\0')
       $Flag\{field\} \leftarrow$  IncompleteField;
    else  $Flag\{field\} \leftarrow$  SeperateField;
if(all of  $Flag\{Date\}, Flag\{From\}$  and  $Flag\{To\}$  are CompleteField)
  return CompleteHeader;
else if( $Flag\{Date\}, Flag\{From\}$  or  $Flag\{To\}$  is IncompleteField)
  return IncompleteHeader;
else return SeparateHeader;

```

## (2) Webmail body recovering algorithms

There are some HTML tags before and after body message, and they will be called starting tags and ending tags for short. Variables of *startTagOffset* and *end-TagOffset* are used to record offset of starting and ending tag respectively. Complete webmail body will be saved to array  $mb[]$ . Details are listed in algorithm 3.

**Algorithm 3:** MailBodyRecovering( $F$ )

**Input:** the decoded memory dump files  $F$

**Output:** Mail Body Array  $mb[]$

```

startTag, endTag ← starting tag, ending tag;
startTagOffset, endTagOffset, startTagFlag ← 0;
read one block data  $b$  from dump file  $F$ ;
while (!EOF( $F$ ))
    if ( $b \sim$  startTag)
        startTagOffset ← offset of startTag;
        startTagFlag ← 1;
    else if ( $b \sim$  endTag)
        endTagOffset ← offset of endTag;
        if (startTagFlag == 1)
            startTagFlag ← 0;
             $ma[j++] \leftarrow$  Copy(startTagOffset, endTagOffset);
        read next block  $c$ ;
         $b \leftarrow c$ ;

```

### (3) Matching algorithms

MatchHeaderAndBody is used to match the recovered webmail headers and bodies. It saves results to  $ma[]$  according match degree.

**Algorithm 4:** MatchHeaderAndBody( $mh[], mb[]$ )

**Input:** the recovered mail header array  $mh[]$  and mail body array  $mb[]$ ;

**Output:** mail array  $ma[]$ ;

```

foreach  $B$  in  $mb[]$ 
     $DB, EB, WB \leftarrow$  GetNormalizedDateSet( $B$ ),
        GetEmailAddressSet( $B$ ), GetWordSet( $B$ );
     $MaxMD \leftarrow 0$ ;
    foreach  $H$  in  $mh[]$ 
         $D_H, E_H, W_H \leftarrow$  GetNormalizedDateSet( $H$ ),
            GetEmailAddressSet( $H$ ), GetWordSet( $H$ );

         $MD \leftarrow (|D_H \cap D_B| + |E_H \cap E_B| + |W_H \cap W_B|) / (|D_H| + |E_H| + |W_H|) \times 100\%$ 
        if ( $MaxMD < MD$ ) MatchingHeader =  $H$ ;  $MaxMD \leftarrow MD$ ;
    if ( $MaxMD >$  Threshold)  $ma[i++] \leftarrow$  MatchingHeader,  $B$ ;

```

## 3 Experiments

### 3.1 Experiment Preparations

VMWare is used in the experiment because its function of taking snapshot could minimize related interferences when imaging the system. The operating system of host is Windows XP, and version of VMWare is 7.1.5 build-491717 with 128 RAM. IE8 and chrome (ver. 20) are chosen according to StatCounter's statistical data of Oct 2012.

The same recipient and sender do not affect results once the webmail server is chosen. For simplicity, mem\_exp@126.com is chosen as test account. “mail.126.com” is a popular website of NTES in China, which provides free webmail service.

One message called “angel” is chosen in experiments, which has 254 words. Twenty copies of angel named angel01, angel02, angel03 ... angel20 and every word in the message will be added the same number as the title. Twenty messages are chosen for there are only latest twenty mails listed in first page of inbox.

Then we power on the virtual machine, open the web browser, IE or chrome only one is chosen in one snapshot, and IE image or chrome image is called for short. We sign in, select the seed emails “Angel” series and take snapshot as the ground truth image. Then we close web browser and take another snapshots without other system activities except screensaver.

### 3.2 Analyses

During preprocess step open source library zlib is used (NTES data takes gzip format) and KMP algorithm is used to locate gzip data header 0X1F8B08. Then we manually analyze the data and found that FROM field is like ‘from’:[‘xxx <mem\_exp@126.com>’]CRLF. The strings of TO and DATE field are ‘to’: [“some Chinese characters”<mem\_exp@126.com>]CRLF and ‘sentDate’:new Date (yyyy, MM, dd, hh, mm, ss)CRLF respectively. In experiment, the starting tag string is adjust to “<style>HTML{word-wrap:break-word;}” and the ending tag string is “<script language=“javascript”>try{parent.JS.modules[window.name].content.setHeight();}”.

Memory data will expire within definite time period [6]. We take one extra experiment to test time span and time limitation of this experiment gave was 5 min.

After preprocessing ground truth data, we manually analyze these data to find all headers and bodies, number is listed in total line (Tables 1 and 2) with pair H–B, which stand for number of copies of header’s and body’s respectively. Number of headers and bodies found by program is listed in located line.

This table indicates (1) copies of header are more than body, and (2) copies in memory vanished quickly at the first few minutes. The possible reasons are (1) another body format without HTML tag can not be found by program; browser has already requested the latest twenty mail headers’ information when users log in. (2) Some data is from network packets and vanished when they were flushed.

**Table 1** Copies of mail header and mail body located in chrome image over time

Result	0 min H–B	1 min H–B	3 min H–B	5 min H–B
Located	262-37	124-22	105-20	67-14
Total	268-37	126-23	106-21	70-15
Correctness	0.977-1	0.984-0.957	0.991-0.952	0.957-0.933

**Table 2** Copies of mail header and mail body located in IE image over time

Result	0 min H-B	1 min H-B	3 min H-B	5 min H-B
Located	248-34	111-22	87-20	49-13
Total	252-34	113-23	92-21	54-14
Correctness	0.984-1	0.982-0.957	0.946-0.952	0.907-0.929

In matching step, we extracted receiver's name from body before the first comma, sender's name and date from the last two lines. We extract names from three fields, which before symbol "@". Year, month and day is extracted from DATE field.

The highest matching degree in these twenty mails is 1/3, so threshold is set to 0.3. The matching result of twenty mails is complete matched. Given these twenty mails have high matching degree between subject and bodies, another 96 common English text mails are chosen to test the algorithm 4 and 78 mails are matched correctly.

## 4 Conclusion and Future Work

This paper solve the problem of webmail reconstruction in Windows XP memory dumps, there are still some future work and research needed to do. (1) Body content of webmail that without HTML tags. (2) Designed webmail that has irrelevant fields in header with body content. (3) Webmail with multi-media type files or attachment.

**Acknowledgments** This work is supported by NSFC (No. 61070212 and 61003195), Zhejiang Province NSF of China (No. Y1090114 and LY12F02006), Zhejiang Province key industrial projects in the priority themes of China (2010C11050), and the science and technology search planned projects of Zhejiang Province (No. 2012C21040).

## References

1. Hadjidj R, Debbabi M, Lounis H et al (2009) Towards an integrated e-mail forensic analysis framework. Proc Digital Invest 5:124–137
2. Pereira MT (2009) Forensic analysis of the Firefox 3 Internet history and recovery of deleted SQLite records. Proc Digital Invest 5:93–103
3. Oh J, Lee S, Lee S (2011) Advanced evidence collection and analysis of web browser activity. Proc Digital Invest 8:62–70
4. Vömel S, Freiling FC (2012) Correctness, atomicity, and integrity: defining criteria for forensically-sound memory acquisition. Proc Digital Invest 9:125–137
5. <http://www.ietf.org/rfc/rfc2822.txt>
6. Solomon J, Huebner E, Bem D, Szezynska M (2007) User data persistence in physical memory. Proc Digital Invest 4:68–72

# On Privacy Preserving Encrypted Data Stores

Tracey Raybourn, Jong Kwan Lee and Ray Kresman

**Abstract** Bucketization techniques allow for effective organization of encrypted data at untrusted servers and for querying by clients. This paper presents a new metric for estimating the risk of data exposure over a set of bucketized data. The metric accounts for the importance of bucket distinctness relative to bucket access. Additionally, we review a method of controlled diffusion which improves bucket security by maximizing entropy and variance. In conjunction with our metric we use this method to show that the advantages of bucketization may be offset due to a loss of bucket security.

**Keywords** Privacy · Trust · Bucketization · Encryption · Multimedia databases

## 1 Introduction

Data is a valuable asset in modern enterprise, and the need to facilitate a variety of multimedia types such as voice, video, text, and images is ever more imperative. The Database As a Service (DAS) model is one system promoted to minimize the overall costs of asset ownership [7]. Private clouds have used DAS for very large, non-relational and multimedia databases, such as search engines [13]. Medical institutes have also used the service to store a variety of digitized patient images. Despite its advantages, outsourcing raises concerns over data confidentiality when

---

T. Raybourn · J. K. Lee (✉) · R. Kresman  
Department of Computer Science, Bowling Green State University, Bowling Green,  
OH 43403, USA  
e-mail: leej@bgsu.edu

T. Raybourn  
e-mail: traybou@bgsu.edu

R. Kresman  
e-mail: kresman@bgsu.edu

the service provider is untrusted [3]. Database encryption is a typical solution, where the client downloads and decrypts records from the server for further processing [4]. Most encryption ciphers, however, do not support SQL queries, resulting in query methods that return unwanted records and perform unnecessary decryption [3]. Balancing privacy and efficiency is the focus of much research on querying encrypted databases [1, 2, 4, 8, 12, 14, 15].

Bucketization is one technique for executing range queries over encrypted data on a DAS server [3, 4]. Encrypted records are divided into buckets, each of which has an ID and a range defined by its minimum and maximum values. In a multimedia environment, the DAS may maintain a grid of nodes, with each node housing a particular type of multimedia data (voice, video, text, images, etc.) [10, 13]. The client holds indexing information about the range of each bucket on the server. Client queries are mapped to the set of buckets that contain any value satisfying the query. The relevant buckets are then requested from the server.

To illustrate, suppose a film production company outsources its video clip database to Amazon's Relational Database Service (RDS), and Amazon uses two buckets: bucket B1 holds a total of 500 clips for every year between 1990 and 2001; and B2 holds a total of 400 clips for 2002 through 2012 of which 100 are made after 2007. A client query for clips shot after 2007 consults the local bucket index, and sends a request for B2 to the server. All records from B2 are downloaded, decrypted, and processed as needed. A client query for clips shot before 2008, will return and decrypt both B1 and B2. In either case, the returned records include clips beyond the desired range, known as false positives, which must be filtered out at the client. False positives are especially problematic for multimedia databases, where retrieval and decryption of non-text data (video, audio, images, etc.) means high computational overhead for the client [13].

While false positives are considered an acceptable cost of bucketization, there are strategies to minimize them [1, 4, 11, 15]. We briefly discuss three of these bucketization methods: Query Optimal Bucketization, Controlled Diffusion, and Deviation Bucketization. Some detail is given to facilitate the reader's understanding of our experiments, for which these algorithms are implemented.

Hore et al. [4] presented a Query Optimal Bucketization (QOB) algorithm that minimizes bucket cost, where cost is a function of the value range and value frequencies in the bucket (Eq. 1). QOB generates an optimal solution to the problem of bucketizing a set of values,  $V = v_1, \dots, v_n$ , using at most  $M$  buckets, where each value,  $v_1 < \dots < v_n$ , occurs at least once in  $V$ . Each bucket covers the values  $(v_i, v_j]$  and the bucket cost  $BC$  is given by,

$$BC(i, j) = (v_j - v_i + 1) * \sum_{i \leq t \leq j} f_t, \quad (1)$$

where  $f_t$  is the frequency of each distinct bucket value. The algorithm computes the summed cost of every two bucket combination over the data set, partitioning the bucket pair that returns the minimum cost. QOB reduces false positives by minimizing the total cost of all buckets.

Hore et al. [4] introduced a method of controlled diffusion, which allows a bounded performance degradation ( $K$ ) in order to improve optimal (QOB) bucket privacy, as measured by entropy and the variance. Controlled diffusion creates a new set of  $M$  approximately equi-depth buckets, called composite buckets ( $CB$ ), and redistributes (*diffuses*) elements from optimal buckets into the  $CB$ s. Diffusion is *controlled* by restricting the number of  $CB$ s into which elements from a given optimal bucket are diffused. Given a maximum performance degradation factor  $K$ , for an optimal bucket  $B_i$  of size  $|B_i|$ , its elements diffuse into no more than  $\frac{K * |B_i|}{f_{CB}}$  composite buckets, where  $f_{CB}$  is the size of data set  $D$  over bucket size  $M$ , i.e.,  $\frac{|D|}{M}$ . The resulting set of  $M$  composite buckets is stored in encrypted form on the server, with  $CB$ -specific bucket IDs.

Yao et al. [15] proposed a Deviation Bucketization (DB) scheme that extends QOB by further reducing false positives at the cost of at most  $M^2$  buckets. First, DB generates a set of  $M$  QOB buckets. For each QOB bucket, DB computes an array comprising the deviations of each distinct data point from the bucket mean. QOB then bucketizes the deviation arrays, creating a set of second level buckets by subdividing each QOB bucket according to its deviation values. Higher frequency values are more likely to be queried, and DB buckets tighten the grouping of these values, greatly reducing false hits over QOB.

On average, lowering bucket width (i.e., increasing the number of buckets) reduces false positives by allowing queries more granular access to bucket domains, but is not without risk. Bucketization is susceptible to estimation and linking attacks [1, 3], as well as query access pattern attacks [5]. A tighter estimate of the underlying data distribution does not ensure that an adversary can determine precise plaintext values, but may be damaging to the extent an adversary does have particular knowledge of data values [1, 3].

In the following section, we introduce a new metric for estimating the risk of an adversary discovering information about the value distribution of a bucketized data set. Later, we present experiments demonstrating (1) the efficacy of our metric and (2) when controlled diffusion is applied to the set of DB buckets, the advantages afforded by DB's decreased bucket width are diminished. The authors in [4] acknowledged that QOB buckets normally lack sufficient entropy and variance, so it is reasonable to expect that subdividing them, as DB does, may cause additional and possibly substantial loss of privacy.

## 2 Metrics for Evaluation

We propose a new metric that considers the risk of exposure as a function both of the distinctness of the data distribution within a bucket, and the frequency with which queries access the bucket. For a set of  $M$  buckets over data set  $D$  with query distribution  $Q$ , we express the risk  $R$  of an adversary reliably estimating the data distribution as:

$$R(D, Q, M) = \sum_{i=1}^M \left(1 - \frac{d_i}{|D|}\right) * \frac{q_i}{|Q|} * \left(1 - \frac{1}{M}\right), \quad (2)$$

where  $|D|$  is the size of  $D$  and  $d_i$  is the number of distinct values in the  $i$ th bucket, yielding the proportion of distinct bucket values  $\frac{d_i}{|D|}$ ;  $|Q|$  is the size of query distribution  $Q$ , and  $q_i$  is the number of query values accessing the  $i$ th bucket, yielding the proportion of query values  $\frac{q_i}{|Q|}$ . We add a normalization term,  $1 - \frac{1}{M}$ , which imposes a penalty for increasing the number of buckets over  $D$ . The frequency of bucket access is important as it reveals something about the preference for values a bucket contains. As the number of buckets increases, the distinct values per bucket must decrease on average, and some values are likely to be queried more often than others. Thus, our metric conveys that fewer distinct values per bucket, relative to a high rate of query access, may disclose not only *intra*-bucket distribution, but data distribution across buckets. If an adversary has some knowledge of one or more buckets, the access pattern across all buckets may help the adversary to extrapolate *inter*-bucket probabilities.

To measure how well a bucketization method minimizes false hits, we use a query precision metric [4]. Query precision is the number of values in the set of buckets satisfying a range query (i.e., positives), over the total number (superset) of values in those buckets. Returning to our previous video clip illustration, a query for clips shot before 2008 yields a query precision, positives/superset = 800/900, of 0.89. An increase in query precision is equivalent to a decrease in false positives.

We also consider a well-known measure of bucket privacy: entropy [9]. Entropy entails distributing a bucket's values as widely and uniformly over a large domain as possible, in order to decrease the probability (increase the uncertainty) of estimating the true value distribution. Bucket entropy is given by:

$$H(X) = - \sum_{i=1}^n p(x_i) * \log_2 p(x_i), \quad (3)$$

where  $p(x_i)$  gives the probability mass function of outcome  $x_i$ . While entropy as a measure of bucket security is relatively static (i.e., varies only as buckets change), our metric,  $R$ , affords a dynamic query access component, which provides a mechanism to account for bucket privacy in terms of access frequency.

### 3 Experiments

We created a data set comprising  $10^5$  integer values generated randomly from a uniform distribution over the domain [1, 1000]. The query set comprised  $10^4$  range queries corresponding with the range of the data set, where values were also drawn from a uniform distribution.

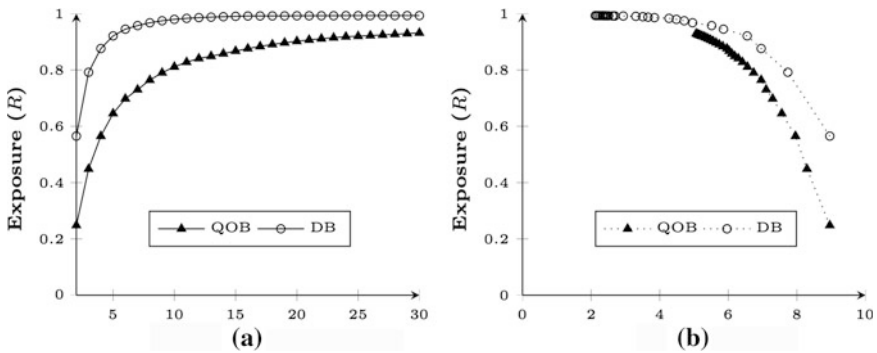


We conducted three experiments. For each, the data set was bucketized using both QOB and DB algorithms at bucket sizes  $M = 2, \dots, 30$  ( $M^2$  for DB). In experiment 1, we applied our exposure metric to respective bucket sets. For experiments 2 and 3, both QOB and DB bucket sets were rebucketized using controlled diffusion with degradation factors of  $K = 2, 4, 6$ . Experiment 2 calculated the average query precision, for optimal (QOB), deviation (DB), and composite buckets (controlled diffusion). Experiment 3 measured entropy for optimal, deviation, and composite buckets.

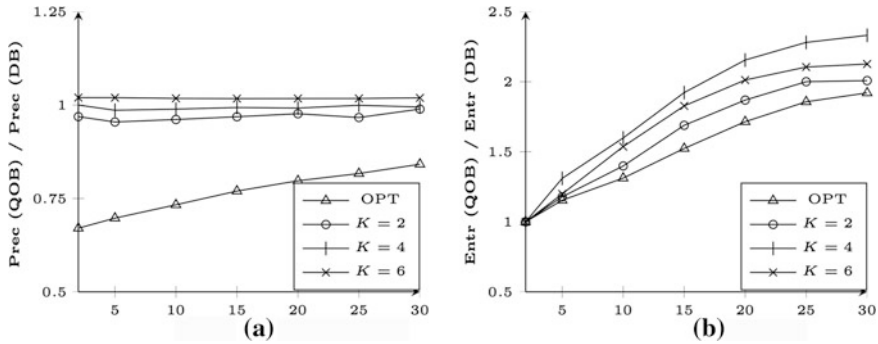
All experiments were conducted in a simulated database environment, i.e., with data structures representing the various database tables necessary for each algorithm to operate. We ran our experiments on a 2 GHz i3-CPU PC, with 2 GB RAM.

### 3.1 Performance Evaluation

Figure 1a shows the risk of exposure  $R$  for optimal buckets (QOB) and second level deviation buckets (DB). Bucket exposure increases proportionally with bucket size  $M$  (i.e., with a decrease in bucket width). In terms of  $R$ , exposure risk is notably higher for DB than QOB. Put in perspective, for  $M = 10$  optimal buckets, DB has generated at most 100 s level buckets. Recall that the data set contains a maximum 1000 distinct values, meaning the deviation buckets contain approximately  $1000/100 \approx 10$  distinct values ( $1000/10 \approx 100$ , for optimal). This also means that the query access pattern gives a more precise reading of which data are requested. Put differently,  $R$ , shows that as buckets decrease in width, they have a proportionally higher risk of losing one of the primary advantages that bucketization affords, namely privacy. QOB also shows relatively high exposure; around 64 %, for example, at  $M = 5$ . This is consistent with [4] who pointed out that optimal buckets may warrant privacy concerns due to low variance and entropy.



**Fig. 1** **a** Estimated exposure risk for optimal (QOB) and deviation buckets (DB) and **b** comparison of  $R$  with entropy showing agreement of security assessment. **a** Size of  $M$ , **b** Average entropy



**Fig. 2** Performance ratios for QOB, DB, and controlled diffusion bucket sets: **a** Average Query Precision. **b** Average Entropy;  $K$  = degradation factor for composite buckets. **a** Size of  $M$ , **b** Size of  $M$

Figure 1b gives  $R$  as a function of average entropy for decreasing values of  $M = 30, \dots, 2$ . Due to the nature of the comparison, data points for DB and QOB do not align one-to-one over the  $x$ -axis, with DB originating at entropy  $\approx 2$  ( $M^2 = 900$ ), and QOB at entropy  $\approx 4.5$  ( $M = 30$ ). The figure highlights a general agreement between  $R$  and entropy, i.e., exposure risk decreases as entropy (uncertainty) increases. Thus, our exposure metric,  $R$ , confirms that additional security mechanisms are necessary to increase bucket privacy.

In Fig. 2 we give results—query precision (a), entropy (b)—as ratios of QOB performance over DB performance for optimal and deviation buckets (OPT), and their corresponding composite buckets. QOB/DB composite bucket ratios are indicated by their respective degradation factors ( $K = 2, 4, 6$ ). Thus, results with a value of 1.0 indicate no performance differences. We plot results for bucket sizes  $M = 2, 5, 10, 15, 20, 25, 30$ .

Figure 2a shows average query precision for optimal, deviation, and corresponding composite bucket sets. Consistent with previous findings [15], precision increases much faster for DB than QOB, as the number of DB buckets is increasing by a factor of  $M$  over QOB. This is indicated by ratio values less than 1.0. The increasing ratio (OPT) shows the advantages of DB decreasing as it approaches a ceiling on additional buckets around  $M = 22$ . That is, precision increases more slowly as bucket granularity rapidly approaches a maximum, while QOB precision is relatively constant increasing over sizes of  $M$ . When controlled diffusion is applied to each bucket set, however, advantages in query precision are minimized across all sizes of  $M$ . For a degradation factor of 2 ( $K = 2$ ), DB composite buckets marginally outperform their QOB counterparts, while increasing factors of  $K$  reveal little to no difference (i.e., ratio is always close to 1.0).

Figure 2b shows average entropy for optimal, deviation, and corresponding composite bucket sets. Overall, QOB-based buckets have higher entropy than their DB counterparts, indicated by ratios greater than 1.0. The figure shows a relatively

consistent relationship between QOB and DB. Even as entropy increases with controlled diffusion, QOB maintains its initial advantage in terms of bucket security due primarily to the greater width of its buckets.

Our results show that decreasing bucket width, even with controlled diffusion, may increase the risk of an adversary discovering information about the underlying data distribution. The diffusion of values from second level deviation buckets to composite buckets minimizes the advantages of DB; query precision substantially decreases, while minimal to no benefits over QOB are found in terms of entropy.

## 4 Concluding Remarks

In modern enterprise, it is critical to manage and protect the variety of sensitive multimedia data. In this paper, we introduced a new metric for estimating the risk of data exposure over a bucketized database. Our metric differs from established measures (e.g., entropy and variance) by accounting for the importance of bucket distinctness *relative to bucket access*. While entropy as a measure of bucket security is relatively static (i.e., varies only as buckets change), our metric,  $R$ , provides a mechanism to account for bucket privacy in terms of access frequency. Bucketized data that is never queried reveals only what can be discovered through standard inference attacks [3], while queries introduce information about the distribution of values across the data store. Our metric, then, may lend itself to modeling different scenarios of query access to determine whether the risk of exposure is acceptable for a given bucket set.

This paper also highlights the importance of bucket width in evaluating the security of bucketization methods. Our metric demonstrates that the advantage of decreasing bucket width, i.e., reducing false positives, can be offset due to a proportional loss of bucket privacy. In their presentation of Deviation Bucketization, Yao et al. [15] did not discuss the implications of bucket security. Thus, we used DB to both test its robustness with respect to data privacy concerns, as well as to highlight possible limitations of bucketization techniques that opt for a narrow bucket strategy, as this arguably undermines the purpose of bucketization.

As follow up to this work, we plan to investigate how the distribution of queries on a DAS server, relative to the bucketization method, impacts performance and security. QOB, for example, assumes that query distribution is uniform, which may be problematic. Our metric,  $R$ , shows that query access can enhance the likelihood of estimating the true data distribution. If queries are uniformly distributed, there is less risk of exposure. Optimal buckets may be less secure, however, if the query distribution is non-uniform, which is the most likely scenario for real-world databases. It may be that a more efficient bucketization strategy is one sensitive to the access pattern. Thus, we will explore bucketization based on the properties of query distributions more likely to reflect real-world access patterns, and their implications for existing bucketization methods. This may also provide insight as to the precise importance of the query distribution in modeling secure databases.

## References

1. Agrawal R, Kiernan J, Srikant R, Xu Y (2004) Order preserving encryption for numeric data. In: 2004 ACM SIGMOD international conference on management of data, Paris, pp 563–574
2. Alwarsh M, Kresman R (2011) On querying encrypted databases. In: 2011 international conference on security and management, Las Vegas, pp 256–262
3. Damiani E, De Capitani di Vimercati S, Jajodia S, Paraboschi S, Samarati P (2003) Balancing confidentiality and efficiency in untrusted relational DBMSs. In: 10th ACM conference on computer and communication security, Washington, DC, pp 93–102
4. Hore B, Mehrotra S, Tsudik G (2004) A privacy-preserving index for range queries. In: 30th international conference on very large databases, Toronto, pp 720–731
5. Hore B, Mehrotra S, Canim M, Kantarcioglu M (2012) Secure multidimensional range queries over outsourced data. *VLDB J* 21(3):333–358
6. Huet B, Chua TS, Hauptmann A (2012) Large-scale multimedia data collections. *IEEE MultiMedia* 19(3):12–14 (IEEE Computer Society)
7. Li J, Omiecinski ER (2005) Efficiency and security trade-off in supporting range queries on encrypted databases. In: 19th annual IFIP WG 11.3 working conference on data and applications security, Storrs, CT, pp 69–83
8. Liu D, Wang S (2012) Programmable order-preserving secure index for encrypted database query. In: 2012 IEEE 5th international conference on cloud computing, Honolulu, pp 502–509
9. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
10. Smith JR, Döller M, Tous R, Gruhne M, Yoon K, Sano M, Burnett IS (2008) The MPEG query format: unifying access to multimedia retrieval systems. *IEEE Multimedia* 15(4):82–95
11. Sun W, Rane S (2012) A distance-sensitive attribute based cryptosystem for privacy-preserving querying. In: 2012 IEEE international conference on multimedia and expo, Melbourne, pp. 386–391
12. Wang J, Du X, Lu J, Lu W (2010) Bucket-based authentication for outsourced databases. *Concurr Comput Pract Experience* 22(9):1160–1180
13. Weis J, Alves-Foss J (2011) Securing database as a service: issues and compromises. *IEEE Secur Privacy* 9:49–55
14. Win LL, Thomas T, Emmanuel S (2011) A privacy preserving content distribution mechanism for DRM without trusted third parties. In: 2011 IEEE international conference on multimedia and expo, pp 1–6, Barcelona
15. Yao Y, Guo H, Sun C (2008) An improved indexing scheme for range queries. In: 2008 international conference on security and management, Las Vegas, pp 397–403

# Mobile User Authentication Scheme Based on Minesweeper Game

Taejin Kim, Siwan Kim, Hyunyi Yi, Gunil Ma  
and Jeong Hyun Yi

**Abstract** The latest boom in the prevalence of smartphones has been encouraging various personal services to store and utilize important data such as photos and banking information. Thus, the importance of user authentication has also been growing rapidly. Nevertheless, many problems have arisen as a result of the common method of using a four-digit personal identification number (PIN) because of its potential for being breached by a brute force attack or shoulder-surfing attack. Various authentication schemes have been developed to overcome these problems. In this paper, we also propose a new password-based user authentication scheme that utilizes the well-known Minesweeper game, providing better usability as well as greater security. The proposed scheme provides its users a simple method for memorizing their passwords and usable security by allowing them to enter calculated values rather than the password itself.

**Keywords** Password · Usable security · Authentication · Shoulder-surfing attack

---

T. Kim · S. Kim · H. Yi · G. Ma · J. H. Yi (✉)  
School of Computer Science and Engineering, Soongsil University, Seoul, Korea  
e-mail: jhyi@ssu.ac.kr

T. Kim  
e-mail: tjkim@ssu.ac.kr

S. Kim  
e-mail: kimsiwan@ssu.ac.kr

H. Yi  
e-mail: hyunyii@ssu.ac.kr

G. Ma  
e-mail: gima@ssu.ac.kr

## 1 Introduction

The recent increase in the use of smartphones has been replacing PCs in handling applications as their scope and area of application expands, providing their users with greater comfort. However, there are still many concerns about personal information leakage, viruses, and malware. Thus, there has been a gradual increase in the importance of secure user authentication methods to protect the personal data stored in smartphones. The current password-based authentication measures are user-friendly, but highly vulnerable to shoulder-surfing attacks, brute force attacks, and smudge attacks.

Much research has been conducted to resolve these concerns. However, the most of the methods developed [1–4] have been unsuitable for mobile devices, have compromised user-friendliness, or have continued to remain vulnerable to shoulder-surfing attacks. In this paper, we propose a user-friendly security method for mobile devices that provides defense against shoulder-surfing attacks. In addition, we conduct a security analysis through a comparison with previous techniques against brute force and shoulder-surfing attacks.

This paper is organized as follows. In [Sect. 2](#), we discuss the suggested password authentication method, which is followed by a safety analysis in [Sect. 3](#). [Section 4](#) concludes the paper.

## 2 Proposed Scheme

Current password authentication schemes do not satisfy both security and usability requirements at the same time. For example, the PIN-Entry [1], DAS [2], and Passfaces [4] methods are secure against shoulder-surfing attacks but are vulnerable to recording attacks, while the Dementor-SGP scheme [3] has good security but lacks usability. Thus, in this paper, we propose a new scheme [5] that satisfies both the security and usability requirements by applying the idea of the Minesweeper game.

The password in the proposed scheme involves the locations and number of mines. First, the user will be prompted to set the mine locations at will. Then, the set-up is completed by pressing the OK button. After this, every time authentication is required, the user will be prompted to fill randomly selected cells in the entry interface with the number of mines in the  $3 \times 3$  grid with the selected cell as the center. Authentication is carried out by entering the correct number of mines around each entry cell and then pressing the OK button. If each of the entries matches the correct number of mines, the user is successfully authenticated. The example shown in [Fig. 1](#) depicts the mechanism with four entries in a  $4 \times 5$  grid interface. Let us assume that we arrange the mines in an H shape, as shown in [Fig. 1\(left\)](#). Let us denote a coordinate in the grid as  $(x, y)$ , where the grid consists of columns from 1 to  $y$  and rows from 1 to  $x$ . First, in the user authentication

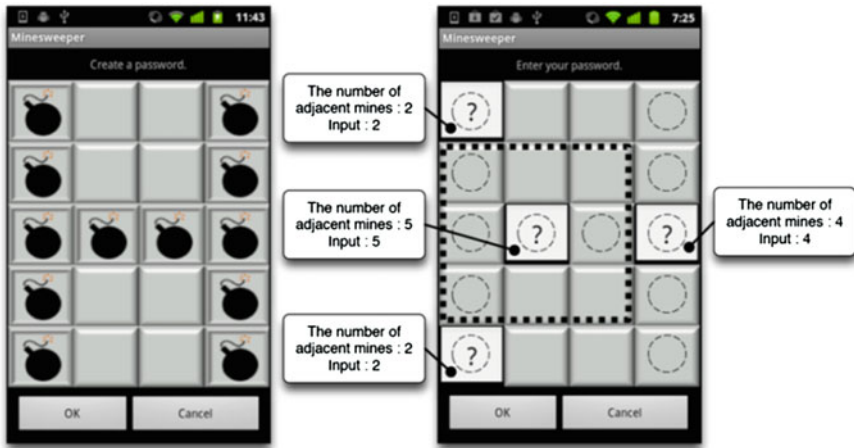


Fig. 1 The proposed scheme: password set-up (left) and user authentication (right)

process, the password is entered via randomly selected cells in the entry interface, as shown in Fig. 1(right).

The user fills in the entry interface cell located at (1,1) with the number '2' which is the total of the mines located at (1,1) and (1,2). For the entry at (2,3), the user fills in the number '5' because mines are planted at (1,2), (1,3), (1,4), (2,3), and (3,3). For the entry at (4,3), the number '4' is entered for the mines at (3,3), (4,2), (4,3), and (4,4). Finally, for the last entry at (1,4), the number '2' is entered because there are two mines around it. Once all the requested entries are filled, the user can confirm with the OK button for authentication. The authentication succeeds if all of the entered numbers are correct.

### 3 Security Analysis

This section describes a security analysis of the proposed password authentication scheme.

#### 3.1 Password Space

The password space indicates the number of possible combinations for a given password length. Table 1 lists the password composition types and space values of the proposed and previous schemes.

The password space of the proposed scheme is equal to  $2^{XY}$ , which is greater than those of the PIN-based password schemes. Similar to the Dementor-SGP and

**Table 1** Password composition, space, and probability of a successful brute force attack

Scheme	Password composition	Password space	Probability of A successful brute force attack
PIN	$N$ -digits	$10^N$	$10^{-N}$
DAS	$N$ -digits	$10^N$	$10^{-N}$
Dementor-SGP	$N$ user images (incl. hole image)	${}^T P_N$	$1/({}^T P_N)$
Passfaces	$N$ facial images	${}^T P_N$	$9^{-N}$
PIN-entry	$N$ -digits	$10^N$	$10^{-N}$
Proposed scheme	$N$ mine locations	$2^{XY}$	$1/(2^{XY})$

Note  $T$  number of elements in the password;  $N$  password length,

$X$  X-axis units/length of grid;  $Y$  Y-axis units/length of grid,

$M$  entry number;  $S$  number of entries exposed to attacker,

${}^T P_N$  permutation to obtain an ordered subset of  $k$  elements from a set of  $n$  elements

Passfaces schemes that are allowed to expand the password space by increasing  $T$ , the new scheme can also expand the space by increasing  $X$  and  $Y$  to increase security against brute force attacks.

### 3.2 Brute Force Attack

A brute force attack attempts to hack a password by guessing every possible combination of the password. Table 1 lists the probabilities of successfully carrying out a brute force attack on the proposed and existing schemes.

The probability of a successful brute force attack depends on the password space. In a case where  $X = 4$  and  $Y = 5$ , the probability of success is  $1/(2^{20})$ . This indicates that the proposed scheme is about 1.59 times more secure than Dementor-SGP with  $T = 30$  and  $N = 4$ ; 104 times more secure than PIN, DAS, or PIN-Entry with  $N = 4$ ; and approximately 159 times more secure than Passfaces with  $N = 4$ .

### 3.3 Shoulder-Surfing Attack

The user authentication method proposed in this paper can conceal the locations of the mines because the user enters the number of adjacent mines instead of the actual password. Thus, even if an attacker obtains the correct entry values, the probability of success for an attacker is very low because the locations of the entry interface cells randomly change at each authentication attempt. In order to calculate the probability of success, the entry range of each grid coordinate must be known. Figure 2 depicts the division of a  $3 \times 3$  grid into  $A$ ,  $B$ , and  $C$  cells. In this



Fig. 2 Grid division

A	B	A
B	C	B
A	B	A

figure, the four corners are marked as A. The other outer cells are B, and C is used for the cells that are not marked as A or B. Counting itself and the adjacent cells, the total number of cells associated with an A cell is 4, with 6 for B and 9 for C, leading to digit ranges of 0–4, 0–6, and 0–9, respectively.

Thus, the probability of success of a brute force attack is 1/5 for A, 1/7 for B, and 1/10 for C. Equation (1) depicts the relationship between  $X \times Y$  and the number of cells for each area:

$$\begin{aligned}
 A &= 4 \\
 B &= 2(X + Y) - 8 \\
 C &= XY - 2(X + Y) + 4
 \end{aligned}
 \tag{1}$$

Consequently, the probability of entering the correct value in one random entry cell from an  $X \times Y$  grid, based on Eq. (1), is as follows.

$$\frac{4}{5XY} + \frac{2(X + Y) - 8}{7XY} + \frac{XY - 2(X + Y) + 4}{10XY}
 \tag{2}$$

Let us assume that the number of cells seen by the attacker is  $S$ , and  $M$  is the number of random entries in the authentication when the attack occurs. In this case, only a portion of  $M$  can be identical to the cells seen by the attacker. Let us define the number of unidentified cells from  $M$  as  $n$ . The number of combinations for  $n$  can be calculated by  ${}_{(XY-S)}C_n$ , and for the identified cells within  $M$ , the number can be calculated by  ${}_S C_{(M-n)}$ . Equation (3) shows the proportion of selected combinations that include  $n$  among the possible combinations  $M$ :

$$\frac{{}_S C_{(M-n)} {}_{(XY-S)} C_n}{XY C_M}
 \tag{3}$$

Only the correct combination for the randomly selected entry interface cells of  $n$  would allow the attacker to succeed in their attack. Finally, the overall probability of a successful brute force attack can be obtained by combining all cases of  $n$ . Equation (4) shows the attacker’s probability of success in authentication.

$$\sum_{n=0}^M \left\{ \frac{{}_S C_{(M-n)} {}_{(XY-S)} C_n}{XY C_M} \left( \frac{4}{5XY} + \frac{2(X + Y) - 8}{7XY} + \frac{XY - 2(X + Y) + 4}{10XY} \right)^n \right\}
 \tag{4}$$

As discussed earlier, if we assume that the parameters are  $X = 4$ ,  $Y = 5$ ,  $M = 4$ , and  $S = 4$ , then in a single attempt, the proposed scheme will allow a probability of success of as little as approximately  $6.51 \times 10^{-3}$ . Thus, the proposed scheme provides considerably more security against a shoulder-surfing attack than previous methods.

### 3.4 Recording Attack

A recording attack is a type of shoulder-surfing attack where the entire authentication process may be recorded using an electronic device such as a camera. In the case of the proposed scheme, the number of entries acquired by each recording attack is  $M$ . However, because the locations of the entry interface cells are randomly selected for each trial, the attacker may need the original mine locations or the correct digits for all of the entry cells. For an attacker, the latter option seems more feasible than the former. This is because the locations of the password mines can be relocated by only a change of one digit for every entry value, and the original mine locations can vary for the same digit entry combination [6].

In order to obtain the correct digits for all of the entries, the recording needs to be performed at least  $XY/M$  times. However, if a recording attack is performed twice or more, there will be considerable duplication in the obtained information. Thus, the chance of obtaining every entry from  $XY/M$  recording attacks is extremely low. Equation (5) shows the average of the numbers of newly obtained entries from recording attacks when  $S$  is the number of exposed entries.

$$\begin{cases} f_{(0)} = M \\ f_{(S)} = \sum_{n=0}^M n \frac{{}^S C_{(M-n)} (XY-S) C_n}{XY C_M} \end{cases} \quad (5)$$

Equation (6) shows the number of obtainable entries from multiple recording attacks based on Eq. (5).

$$\begin{cases} g_{(0)} = f_{(0)} \\ g_{(a)} = g_{(a-1)} + f_{(g_{(a-1)})} \end{cases} \quad (6)$$

Thus,  $g(a) > XY$  has to be met for the number recording attacks to obtain the correct digits for all of the entries. For example, with the proposed scheme, an average of 17 recording attacks are required to obtain all of the entry values with parameters  $X = 4$ ,  $Y = 5$ , and  $M = 4$ . Because, in reality, a recording attack can rarely be performed more than 10 times on the same user, the proposed scheme is secure against recording attacks.

## 4 Conclusion

This paper proposes a new password authentication method that adopts the Minesweeper game and can satisfy both security and usability requirements. Its suitability was proven using both a model test and user test with implementation in actual Android phones. The results proved that the suggested scheme can ensure user security against a shoulder-surfing attack, brute force attack, and especially, a recording attack. Although not very remarkable, the user test results showed improved usability over former schemes, which might be because it was inspired by the well-known Minesweeper game. We conclude that the presented authentication scheme ensures considerably better security than current mobile authentication schemes while providing proper usability for its users.

**Acknowledgments** This work was supported by a grant from the KEIT funded by the Ministry of Knowledge Economy (10039180).

## References

1. Roth V, Richter K, Freidinger R (2004) A PIN-entry method resilient against shoulder surfing. In: Proceedings of the 11th ACM conference on computer and communications security, USA, pp 236–245
2. Park SB (2004) A method for preventing input information from exposing to observers. Patent application no.: 10-2004-0039209, Korea
3. MinInfo Co., <http://www.mininfo.co.kr>
4. Passfaces, <http://www.passfaces.com>
5. Yi JH, Kim T, Ma G, Yi H, Kim S (2012) Method and apparatus for authenticating password. Patent application no.: US 13/623,409
6. Kaye R (2000) Minesweeper is NP-complete. *Math Intell* 22:9–15
7. Olson JR, Olson GM (1990) The growth of cognitive modeling in human-computer interaction since GOMS. *Hum Comput Interact* 5:221–265
8. Lee S, Myung R (2009) Modified GOMS-model for mobile computing. *J Soc Korea Ind Syst Eng* 32:85–93

# Design and Evaluation of a Diffusion Tracing Function for Classified Information Among Multiple Computers

Nobuto Otsubo, Shinichiro Uemura, Toshihiro Yamauchi  
and Hideo Taniguchi

**Abstract** In recent years, the opportunity to deal with classified information in a computer has increased, so the cases of classified information leakage have also increased. We have developed a function called “diffusion tracing function for classified information” (tracing function), which has the ability to trace the diffusion of classified information in a computer and to manage which resources might contain classified information. The classified information exchanged among the processes in multiple computers should be traced. This paper proposes a method which traces the diffusion for classified information among multiple computers. Evaluation results show the effectiveness of the proposed methods.

**Keywords** Prevention of information leaks · Network security · Log management

## 1 Introduction

The improvement in computer performance and propagation in various services has increased the opportunity to deal with classified information, such as customer information. According to the analysis [1] of personal information leakage incidents, it has been reported that leaks often happen by inadvertent handling and mismanagement, which account for approximately 57 % of all known cases of information leakage. In addition, several employees often share classified information. To trace the status of classified information in a computer and to manage the resources that contain classified information, we proposed a diffusion

---

N. Otsubo · S. Uemura · T. Yamauchi (✉) · H. Taniguchi  
Graduate School of Natural Science and Technology, Okayama University, Okayama, Japan  
e-mail: yamauchi@cs.okayama-u.ac.jp

H. Taniguchi  
e-mail: tani@cs.okayama-u.ac.jp

tracing function for classified information (tracing function), which manages any process that has the potential to diffuse classified information [2].

In this paper, we propose a method that uses the tracing function to trace the classified information being exchanged among multiple computers in internal network and to prevent information leakage outside internal network.

## **2 Requirements of Diffusion Tracing Function for Classified Information Among Multiple Computers**

By tracing how classified information is diffusion, a computer can know which resources contain classified information. However, the tracing function [2] only traces the status of classified information in a computer. Thus, we propose a method that uses tracing function to trace classified information being exchanged among multiple computers in internal network and to prevent information leakage outside internal network.

In order to prevent information leakage outside the internal network, function needs to centrally manage the classified information that exists in the client computers in network. Moreover, it is necessary to determine whether the client computers are installed the tracing function or not, in order to prevent the diffusion of classified information to not installed computers.

We split the computers in network into a managed network and a non-managed network. Computers that need to handle classified information would be in the managed network, and the tracing function would be installed on them. A diffusion tracing function for classified information among multiple computers (tracing function for networks) must meet the following requirements:

- (1) Computers that are installed the tracing function can be distinguished from computers that are not.
- (2) The location and the flow of classified information in the managed network can be managed.
- (3) The diffusion of classified information in the managed network can be instantly traced.
- (4) Leakage of classified information out of the managed network can be detected.
- (5) Leakage of classified information out of the managed network can be stopped in advance.

The tracing function for networks must also meet the following requests:

- (1) Accurately trace the diffusion of classified information in a managed network.
- (2) The processing overhead of tracing function for networks should be small.

### 3 Design

#### 3.1 Overview of the Proposed Function

Figure 1 shows an overview of the function design to handle classified information in a managed network. A managed network consists of a single management server and multiple client computers installed the tracing function. The following describes what is achieved by using the tracing function for networks.

- (1) Before sending data to another client computer, a process managed by tracing function (hereafter, managed process) in the client computer queries the management server to determine whether the tracing function is installed on the receiving computer. If the tracing function is installed, the transmission of classified information is allowed. If the function is not installed, the transmission is disallowed. This makes it possible to satisfy function requirements (1), (4), and (5).
- (2) The tracing function writes a log in the client computer and transfers that log to the management server. The collected logs in the management server are used to manage the location and diffusion of classified information in the managed network. This satisfies function requirement (2).
- (3) Before a managed process sends classified information, a managed process in the sending computer notifies sending classified information to the receiving computer. Then tracing function in the receiving computer mark a process receiving the classified information as managed process. This satisfies function requirement (3).

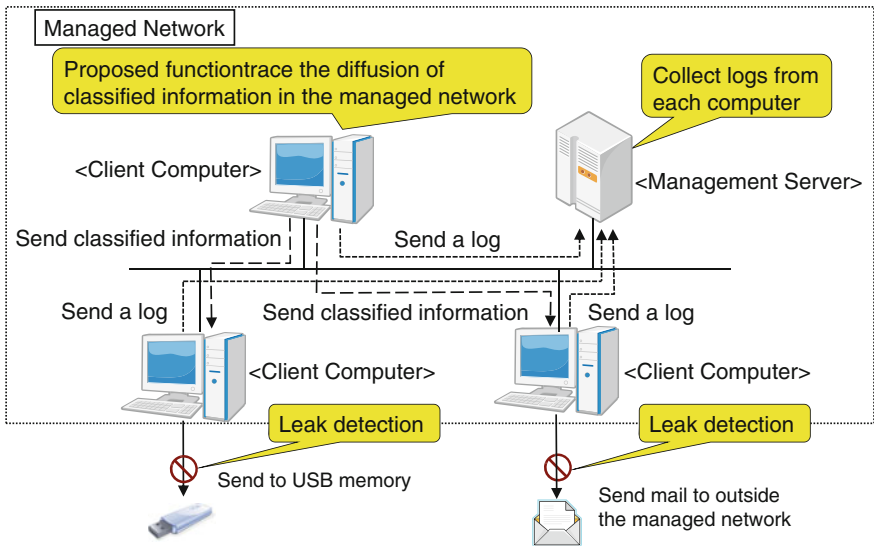


Fig. 1 Overview of diffusion tracing function for classified information among multiple computers

In order to satisfy requirements (1) and (3), communication functions must ensure that the receiving client computer is installed the tracing function before transmitting classified information. In the following sections, we propose a design of the function.

### 3.2 Communication Method to Pass the Information to be Managed

In order to control the communication of classified information, an application program, which monitors the communication of classified information (hereafter, communication monitor), is developed. Before computers exchange classified information, the communication monitor in the sending computer sends the destination port number and IP address of the receiver socket to the receiving computer. The receiving computer uses port number to mark a receiver socket as managed socket. Then the receiving computer marks a process receiving classified information from managed socket as managed process.

The management server maintains the list of IP address of client computers installed the tracing function. The communication monitor in the sending computer queries whether a receiving computer has been registered in the management server list, and, if so, the transmission process is initiated.

Figure 2 shows the flow of process using the communication monitor to send the classified information from the sending computer to the receiving computer as described below:

- (1) Send system call of managed process is invoked.
- (2) The processing of the send system call is suspended, and the IP address and port number of the receiver socket is sent to the communication monitor.
- (3) The communication monitor queries the management server whether the receiving computer has been installed tracing function.

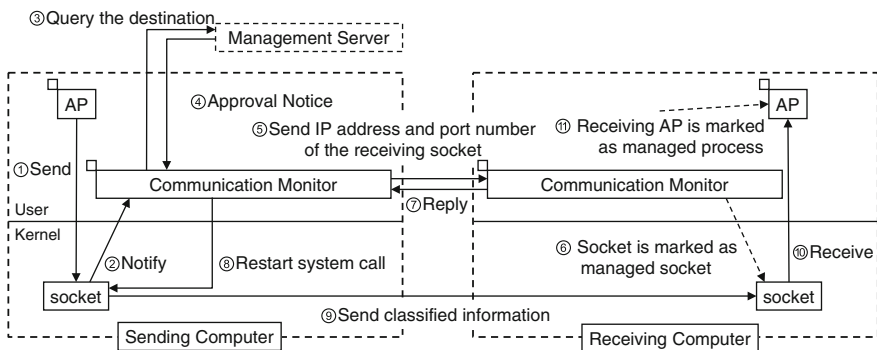


Fig. 2 Sending classified information to other computer AP

- (4) If the receiving computer has been installed, the management server notifies the communication monitor that it has been allowed to send classified information. If it is not installed, the management server notifies the communication monitor that it is not allowed to send classified information.
- (5) If transmission of classified information is allowed, the communication monitor in the sending computer sends the port number of the receiver socket to the receiving computer. If transmission of classified information is not allowed, the function terminates the processing of the send system call.
- (6) The communication monitor in the receiving computer marks a socket using the receiving port number as managed socket.
- (7) The communication monitor in the receiving computer replies to the communication monitor in the sending computer that it is ready to receive the classified information.
- (8) After receiving the confirmation, the communication monitor in the sending computer restarts the send system call that was suspended in step (2).
- (9) Classified information is sent.
- (10) The application program in the receiving computer initiates a receiving system call and receives the classified information from the managed socket.
- (11) The receiving AP is marked as a managed process.

## 4 Evaluation

### 4.1 Overhead

In order to evaluate overhead of the diffusion tracing function for classified information among multiple computers, we measured the time required to transfer managed files to other computers. We measured the time it takes to upload a managed file (1–10 MB) from an FTP client to an FTP server.

We used the following configuration as our measurement environment: the computer of FTP client has a Celeron D 2.8 GHz CPU and 768 MB memory. The computer of FTP server has a Pentium 4 3.0 GHz CPU and 512 MB memory. The two computers are connected using Ethernet 100Base-TX. OS used on both computers is Linux-2.6.0.

The measurement results are shown in Table 1. We found that the transfer time using the managed kernel function is a maximum of about 2.8 times of the transfer time using the original kernel function. This is because the communication monitor communicates with the management server when the send system call is called. Therefore, in order to reduce overhead, it is necessary to reduce the exchange of information between the communication monitor and the management server.



**Table 1** File transfer time by FTP (ms)

	Data size of a transmitting file	
	1 MB	10 MB
Function before implementation	83	885
Function after implementation	233	2010
Overhead	150 (181 %)	1125 (127 %)

### 4.2 Evaluation of Diffusion Tracing of Classified Information from Logs

The diffusion of classified information among multiple computers can be analyzed by tracing the flow of classified information, using data in collected logs during FTP file transfer of the classified information. Logs from two the computer of FTP client and the computer of FTP server were evaluated after uploading files with classified information from the client to the server. ProFTP was used for the FTP server; LFTP was used for the FTP client.

Figure 3 shows the log output from the computer of FTP client. From the first line of the log, we find that the process PID 2907 of the FTP client “lftp” is a managed process, because it read the classified information file secret.txt. From lines 2 and 4–6 of the log, we see that the process PID 2907 has sent the data to the computer with the IP address 192.168.8.201.

Figure 4 shows the log output from the computer of FTP server. From the first line of the log, we find that a socket receiving classified information from the computer with the IP address 192.168.8.166 is marked as managed socket. From the second line of the log, the process PID 2796 of the FTP server “proftpd” is marked to be managed. Lines 2–4 and 6 indicate that the managed process PID 2796 will exchange data with a computer with the IP address 192.168.8.166. Line 7 of the log shows that the process PID 2796 writes data to a managed file called secret.txt; this file is marked as a managed file.

```

1: Thu Feb 4 15:40:41 2010 PID: 2634 Socket Marked (by remote) DOMAIN:INET
SRC-IP:192.168.8.166 SRC-Port:32769, SRC-PID:2907
2: Thu Feb 4 15:40:41 2010 PID: 2796 Process Marked PID:2796
PNM:/usr/local/sbin/proftpd (socket)
3: Thu Feb 4 15:40:41 2010 PID: 2796 Send to remote machine. DOMAIN:INET
PID:2796 PNM:/usr/local/sbin/proftpd DST-IP:192.168.8.166 DST-Port:32769
4: Thu Feb 4 15:40:41 2010 PID: 2796 Send to remote machine. DOMAIN:INET
PID:2796 PNM:/usr/local/sbin/proftpd DST-IP:192.168.8.166 DST-Port:32769
5: Thu Feb 4 15:40:41 2010 PID: 2634 Socket Marked (by remote) DOMAIN:INET
SRC-IP:192.168.8.166 SRC-Port:32771, SRC-PID:2907
6: Thu Feb 4 15:40:41 2010 PID: 2796 Send to remote machine. DOMAIN:INET
PID:2796 PNM:/usr/local/sbin/proftpd DST-IP:192.168.8.166 DST-Port:32769
7: Thu Feb 4 15:40:41 2010 PID: 2796 File Marked. FILE:/home/s-uemura/secret.txt
INODE:426211 PID:2796 PNM:/usr/local/sbin/proftpd MODE:2 NO:3 DEV:300004
    
```

**Fig. 3** Log output from the computer of FTP client

```
1: Thu Feb 4 15:40:41 2010 PID: 2907 Process Marked PID:2907 PNM:/usr/bin/lftp  
FILE:secret.txt  
2: Thu Feb 4 15:40:41 2010 PID: 2907 Send to remote machine. DOMAIN:INET  
PID:2907 PNM:/usr/bin/lftp DST-IP:192.168.8.201 DST-Port:21  
3: Thu Feb 4 15:40:41 2010 PID: 2871 Socket Marked (by remote) DOMAIN:INET  
SRC-IP:192.168.8.201 SRC-Port:21, SRC-PID:2796  
4: Thu Feb 4 15:40:41 2010 PID: 2907 Send to remote machine. DOMAIN:INET  
PID:2907 PNM:/usr/bin/lftp DST-IP:192.168.8.201 DST-Port:21  
5: Thu Feb 4 15:40:41 2010 PID: 2907 Send to remote machine. DOMAIN:INET  
PID:2907 PNM:/usr/bin/lftp DST-IP:192.168.8.201 DST-Port:21  
6: Thu Feb 4 15:40:41 2010 PID: 2907 Send to remote machine. DOMAIN:INET  
PID:2907 PNM:/usr/bin/lftp DST-IP:192.168.8.201 DST-Port:32777
```

**Fig. 4** Log output from the computer of FTP server

From these logs, the proposed methods can trace the flow of classified information from the computer with the IP address 192.168.8.166 to the computer with the IP address 192.168.8.201.

## 5 Related Work

By using another connection to send address range of the data when sending tainted data, received data is to be tainted, to trace the diffusion of classified information among multiple computers [3]. However, there is a problem that unable to trace the diffusion of classified information by UDP. On the other hand, by storing the address range of the data to improve the header of the packet, taint is imparted to the header, to trace the diffusion of classified information among multiple computers [4]. However, there is a problem that when send the packet to a computer that does not eliminate the header of the packet, improved header is treated as data. Multiple virtual machines can be processed on a trusted virtual machine monitor, providing isolated virtual environments on a per-machine basis to serve as mechanisms that prevent other users from accessing files [5]. To limit access to files by isolating environmental in the program unit, the mechanism used is to assign the domain name of the group to the files, and same domain user only access the files [6].

## 6 Conclusion

We proposed a diffusion tracing function of classified information among multiple computers to prevent information leakage outside internal networks. The classified information exchanged among the processes in multiple computers should be traced. Therefore, before computers exchange classified information, the proposed

function sends IP address and the port number of the receiver socket. Then, the diffusion tracing function for classified information in the receiving computer traces the receiving classified information. An evaluation of logs shows that the proposed function traces a diffusion of classified information among multiple computers.

In future work, we will reduce the overhead of the proposed function.

## References

1. Japan Network Security Association (2008) Information Security Incident Survey Report, [http://www.jnsa.org/result/incident/data/2008incident\\_survey\\_e\\_v1.0.pdf](http://www.jnsa.org/result/incident/data/2008incident_survey_e_v1.0.pdf)
2. Tabata T, Hakomori S, Ohashi K, Uemura S, Yokoyama K, Taniguchi H (2009) Tracing classified information diffusion for protecting information leakage. *IPSJ J* 50(9):2088–2012 (in Japanese)
3. Kim CH, Keromytis DA, Covington M, Sahita R (2009) Capturing information flow with concatenated dynamic taint analysis. 2009 International conference on Availability, Reliability and Security (ARES 2009), pp 355–362
4. Zavou A, Portokalidis G, Keromytis DA (2011) Taint-Exchange: A generic system for cross-process and cross-host taint tracking. *The 6th International Workshop on Security (IWSEC 2011)*, vol 7038. LNCS, pp 113–128
5. Garnkel T, Pfaff B, Chow J, Rosenblum M, Boneh D (2003) Terra: A virtual machine-based platform for trusted computing. In: *Proceedings of 19th ACM SIGOPS Symposium on Operating System Principles (SOSP 2003)*, pp 193–206
6. Katsuno Y, Watanabe Y, Furuichi S, Kudo M (2007) Chinese-Wall process confinement for practical distributed coalitions. *Proceedings of 12th ACM Symposium on Access Control Models and Technologies (SACMAT2007)*, pp 225–234

# DroidTrack: Tracking Information Diffusion and Preventing Information Leakage on Android

Syunya Sakamoto, Kenji Okuda, Ryo Nakatsuka  
and Toshihiro Yamauchi

**Abstract** An app in Android can collaborate with other apps and control personal information by using the Intent or user's allowing of permission. However, users cannot detect when they communicate. Therefore, users might not be aware information leakage if app is malware. This paper proposes DroidTrack, a method for tracking the diffusion of personal information and preventing its leakage on an Android device. DroidTrack alerts the user of the possibility of information leakage when an app uses APIs to communicate with outside. These alerts are triggered only if the app has already called APIs to collect personal information. Users are given the option to refuse the execution of the API if it is not appropriate. Further, by illustrating how their personal data is diffused, users can have the necessary information to help them decide whether the API use is appropriate.

**Keywords** Android · Malware · Preventing information leakage · API control

## 1 Introduction

In recent years, adoption of the smartphone has been rapidly spreading, and Android [1] is one of the popular operating systems (OS) for smartphones. An app developer can make the app available through a Web site, such as Google Play Store [2]. However, an app [3] can hijack administrative privileges in order to exploit vulnerability in the Android OS and send out illegally collected personal information.

---

S. Sakamoto (✉) · K. Okuda · R. Nakatsuka · T. Yamauchi  
Graduate School of Natural Science and Technology, Okayama University,  
Okayama, Japan  
e-mail: sakamoto@swlab.cs.okayama-u.ac.jp

T. Yamauchi  
e-mail: yamauchi@cs.okayama-u.ac.jp

Malware that target the Android OS are usually intended to illegally collect personal information. A mobile device contains a large amount of personal information, such as name, address, phone number, etc. and their information can be easily obtained by apps using the Android API. In addition, many users are unaware that mobile phones are not secure and usually do not come with any anti-malware software. For this reason, there is a possibility of information leakage while user did not notice the infection of malware.

An Android app is executed in sandbox, and communication with other apps is severely restricted, except using Intent [4]. Key features such as external communications and the acquisition of personal information require permissions from the user. However, the user cannot detect when the personal information is obtained by the app and whether that personal information was leaked.

In this paper, we propose DroidTrack: a method for tracking information leakage diffusion and preventing information leakage on Android, tracks information diffusion after the app has obtained personal information. DroidTrack alerts the user if there is a possibility of information leakage, and allows the user to limit the use of the API. DroidTrack monitors any app that uses the information-gathering API and displays a warning when the app also uses the API that sends information outside of the device. Personal information can also be leaked when one app obtains personal information and then sends it to another app, which sends the information out of the device. For this reason, DroidTrack manages both apps using the Intent. In addition, the user is allowed to decide whether to limit the use of the API, which sends information out of the device, thereby preventing information leakage.

## 2 Android Component and Security Issues

### 2.1 Android Component

In the Android OS, all apps operate on the Applications layer. If an app requires resources, it must use the API provided by the Application Framework. Android apps have individual user IDs (UID), and communication with an app with a different UID is highly restricted, except when using the Intent.

Apps cannot use the Android API to access a protected resource or to gather personal information without user's permission. It is necessary to obtain the user's permission [5] for app to use these APIs. An app can request specific permissions, e.g., permission to connect to the Internet (INTERNET) or permission to read the status of the unit (READ\_PHONE\_STATE). The safety of the resources is preserved by granting only minimum permissions required by the app.

Each Android app runs in sandbox. By default, apps cannot communicate with another app because they are strictly separated from each other. However, it is possible to enable communication between apps by using the Intent, which allows

an app to communicate with another app and receive the results of process. In addition, an app can pass data both as a string or as an object.

## ***2.2 Security Issues in Android***

The following are the problems associated with malware and malware infections in the main security areas in the Android OS:

- (1) Problems obtaining administrator authority.
- (2) Problems with development tools (Android Debug Bridge (ADB) [6]).
- (3) Permission abuse.
- (4) Difficult detection of information leakage.
- (5) WebKit abuse.

Problems (1), (2), and (3) are cause of the infection to malware. Problems (4) and (5) are related to malware behavior. Problem (4) makes it very difficult to inform the user when the app gathers information and what kind of personal information it gathers. Therefore, if a user installs malware by mistake, user cannot detect the leakage of personal information. In this work, we deal with problem (4) by preventing the transmission of information out of a smartphone.

## **3 Design Principles of the Proposed Method**

### ***3.1 Requirements and Challenges***

In order to deal with the problem, we propose the following requirements:

- (1) Detect all APIs with the possibility of information leakage.
- (2) User can judge the risk of information leakage.
- (3) Information leakage can be prevented by disallowing the execution of API.

In order to satisfy these three requirements, we propose the following challenges:

- (A) Clarifying the condition of information leakage.
- (B) Detecting all uses of APIs that have a possibility of information leakage.
- (C) Controlling the operations of apps that have a possibility of information leakage.
- (D) Allowing the user to decide whether app can receive sensitive information.

## 3.2 Solution

### 3.2.1 Solution for Challenge (A)

Information leakage can occur when an app uses the Android API to send information to out of the device (diffuse information) after obtaining personal information. The app can also obtain personal information without using the information-gathering API by using the Intent instead. In another scenario of information leakage, one app uses the information-gathering API and then communicates with another app that uses the information-diffusing API.

In this work, we address the following scenarios of information leakage:

- (1) A single app uses the information-gathering API and the information-diffusing API or the Intent.
- (2) One app uses the information-gathering API and then communicates with another app using either the Intent or the information-diffusing API.

### 3.2.2 Solution for Challenges (B) and (C)

As mentioned in Sect. 3.2.1, information leakage can occur when an app uses the API that obtains personal information, the API that diffuses the information, or the Intent. Therefore, to deal with challenges (B) and (C) (detecting all use of the API and controlling the operations of apps), we propose a method to control the behavior of the app as follows:

- by intercepting calls to the information-gathering API, information-diffusing API, or Intent,
- by determining the user's preference to control the use of API if either scenario described in Solution for challenge (A) is true, and
- by controlling the use of the APIs or the Intent based on the user's preference.

## 4 Method for Tracking Diffusion of Information and Preventing Information Leakage on Android

### 4.1 Design Principle

To address challenges (B) and (C), we propose the following requirements:

- (1) To inform the user if there is potential for information leakage.
- (2) To limit the use of APIs and the Intent with accuracy based on user's preference.

### 4.2 Basic Method

We change the framework of the Android as follows:

- (1) “Hook” or intercept calls to the information-gathering API and the information-diffusing API and inform the user of both the name of the app using the API and the name of the API used.
- (2) “Hook” or intercept calls to the Intent and inform the user of both the name of the app that uses the Intent and the name of the app called by the Intent.
- (3) Execute a process based on the user’s preference regarding the use of the APIs or the Intent.

The API is used differently depending on the type of personal information that is gathered by the app. Therefore, because of the change described in (2), the user can be informed when and what kind of personal information which the app obtains and attempts to transmit out of the device. Furthermore, the change described in (3) can be used to prevent information leakage according to the user’s preferences. In the following section, we describe a method to track and prevent information diffusion by using the modified framework described above.

### 4.3 Control of API in the Framework

Figure 1 shows the flow of control of the API in the framework. DroidTrack consists of two “Control Aps” at the Applications layer and one “Calling Control AP Unit” at the Application Framework layer.

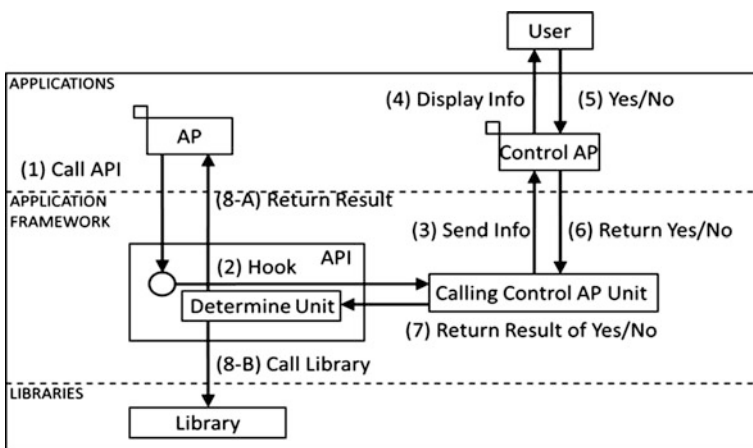


Fig. 1 Control of APIs



In Fig. 1, “Control AP” is an app that provides information about the API to the user and prompts the user to choose whether to limit the use of the API. The “Calling Control AP Unit” informs the “Control AP” whenever an app calls the information diffusing APIs and transmits back to the API engine the user’s preferences regarding the use of the API.

The following describes details of the flow of the process in the framework:

- (1) The app “AP” calls the information-diffusing API.
- (2) The “Hook” intercepts the call to the information-diffusing API in the framework.
- (3) The “Calling Control AP Unit” passes the information about the intercepted call to the “Control AP.”
- (4) The “Control AP” displays a warning dialog to the user if it suspects that an information leak is possible.
- (5) The user replies to the dialog to indicate whether user allows the use of the API.
- (6) The “Control AP” forwards the user’s preference to the “Calling Control AP Unit.”
- (7) The “Calling Control AP Unit” returns the result to the “Determine Unit.”
- (8) The “Determine Unit” handles API based on the user’s preference, as follows:
  - (A) Error handling is carried out if the user disallows the API to be called by the app.
  - (B) API process returns to normal mode if the user permits the API to be called by the app.

Like the above-mentioned procedure, we satisfy the requirement 4.1-(2) by requiring user’s preference before use of APIs and processing according to the preference.

## **4.4 Control AP**

### **4.4.1 Basic Mechanism**

Figure 2 illustrates the mechanism of “Control AP.” The Search-Leakage function triggers the Info Diffusion Manage Unit to check the possibility of information leakage. If there is a possibility, it passes the process to the Control-Write-Out function. If there is no possibility, it allows the API calls to be processed. The Info Diffusion Manage Unit updates the Information Diffusion data structure, examines the possibility of information leakage, and returns the result of the test to the Search-Leakage function. The Control-Write-Out function displays a warning dialog, and then accepts and returns the user’s preference regarding the use of the APIs.

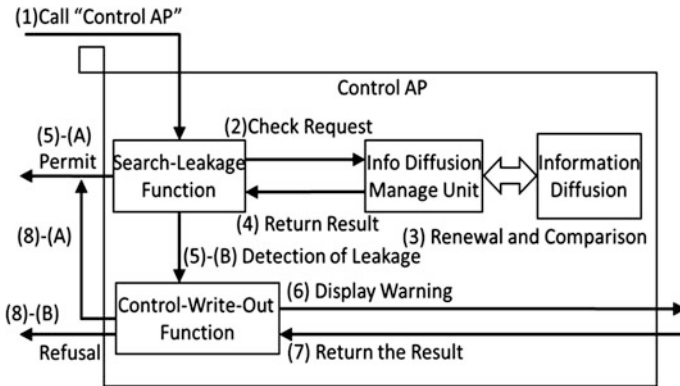


Fig. 2 Basic mechanism of control AP

### 4.4.2 Information Leakage Determination Method

Control AP monitors apps that use the information-gathering API by wrapping the app in a managed object. If the managed AP communicates with another app that uses the Intent, the second app is added to the managed object. Control AP also warns the user of the risk of information leakage if the managed app uses the information-diffusing API, thereby satisfying the requirement described in Sect. 4.1-(1).

## 5 Experiment of the Operation of DroidTrack

Prevention of information leakage by apps was tested using the following procedure:

- (1) Obtain the phone number of the mobile device by using “getLineNumber,” which serves as the information-gathering API.
- (2) Transmit the personal information out of the device by using “sendText-Message,” which serves as the information-diffusing API.

Figure 3 shows the dialog displayed by DroidTrack when the example app runs. The user can detect the use of the API by various information from the dialog. In this case, the user presses “Yes” to allow the use of the API and “No” to disallow the use of the API. DroidTrack could prevent information leakage by pressing “No”.

**Fig. 3** Warning dialog



## 6 Related Work

MockDroid [7] allows the user to provide fake or “mock” data to prevent a real personal information leakage. This method needs user’s set up of permissions for each apps. Therefore, DroidTrack needs no user’s set up, but needs user’s input only when there is a possibility of information leakage. Furthermore, DroidTrack tracks all transmitted information, including transmissions without leakage, in order to check all API communications in and out of the device. AppFence [8] that using TaintDroid [9] provides a similar approach, but it modifies framework and Dalvik VM, and uses the policy which is made except on Android. On the other hand, DroidTrack modifies only the framework of the Android. Therefore, DroidTrack is easier to implement.

## 7 Conclusion

We have proposed DroidTrack, a method to warn of the risk of information leakage by monitoring apps that obtain personal information and by keeping track

of information diffusion. DroidTrack prevents personal information leakage by controlling the behavior of the API, based on the user's preference, which they indicate when a warning is displayed. DroidTrack can also detect information leakage in scenarios where the Intent is used and where the information-diffusing API is used. In addition, DroidTrack can inform the user of the risk of information leakage and display a list of information-gathering APIs that could have diffused information to other apps.

## References

1. Android. <http://www.android.com/>
2. Google play store. <https://play.google.com/store>
3. Droid dream. Google android market kills droid dream malware in Trojans. <http://blogs.computerworld.com/17929>
4. Intent. <http://developer.android.com/reference/android/content/Intent.html>
5. Access permissions. <http://developer.android.com/intl/ja/reference/android/Manifest.permission.html>
6. Android debug bridge. <http://developer.android.com/guide/developing/tools/adb.html>
7. Beresford AR, Rice A, Skehin N, Sohan R (2011) MockDroid: trading privacy for application functionality on smartphones. In: Proceedings of the 12th workshop on mobile computing systems and applications, pp 49–54
8. Hornyack P, Han S, Jung J, Schechter S, Wetherall D (2011) These aren't the droids you're looking for: retrofitting android to protect data from imperious applications. In: Proceedings of the 18th ACM conference on computer and communications security (CCS2011)
9. Enck W, Gilbert P, Chun B, Cox LP, Jung J, McDaniel P, Sheth AN (2010) TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones. In: Proceedings of the 9th USENIX symposium on operating systems design and implementation (OSDI'10)

# Three Factor Authentication Protocol Based on Bilinear Pairing

Thokozani Felix Vallent and Hyunsung Kim

**Abstract** Secure authentication mechanism is a pre-requisite to remote access of server's resources particularly when done over the Internet. This paper presents a three factor authentication protocol which is based on verification of user's: biometrics, knowledge proof of a password and possession of token to pass authentication. The proposed protocol utilizes bilinear mapping for session key establishment and elliptic curve discrete logarithm problem for security.

**Keywords** Authentication · Bilinear pairing · Three factor authentication

## 1 Introduction

Information security is concerned with the assurance of confidentiality, integrity and availability of information in all forms. There are many tools and techniques that can support the management of information security one of which is the use of tokens that store client identifying information like smart card [1–5]. Smart card authentication falls short of password sharing among colleagues, password guessing and smart card breaching [3]. In applications with strict user identification smart card flaws can be dealt with by employing the biometric authentication besides the password. Biometrics is hard to forge hence outworks impersonation attack resilience thus provides a reliable means of authentication [3–5].

---

T. F. Vallent

Department of IT Convergence, Kyungil University, Kyungpook,  
Kyungsansi 712-701, Korea  
e-mail: tfvallent@gmail.com

H. Kim (✉)

Department of Cyber Security, Kyungil University, Kyungpook,  
Kyungsansi 712-701, Korea  
e-mail: kim@kiu.ac.kr

A biometric system is a pattern recognition system that extracts an individual's unique features set for authentication by comparing these features' template pre-stored in the database [2, 4, 6]. Three factor authentication involves knowledge proof by checking user's knowledge of correct password, token possession and biometrics matching before authentication. Biometric authentication can be applied for identification and non-repudiation and for preserving the integrity like in passport, medical records access control among others [2, 3].

In 1981, Lamport first proposed a remote password authentication scheme for insecure communication. The protocol uses verification table hence it's at the edge of a huge security risks once the system is compromised [7]. Hwang and Li (2000) proposed a remote user authentication scheme using smart cards based on ElGamal's public key cryptosystem the protocol suffers from man-in-the middle attack [2]. In 2010 Li and Hwang proposed another remote user authentication based on biometrics verification, smart card, one-way hash function but still bears a problem of man-in-the-middle attack [2, 4].

## 2 Preliminaries

This section introduces mathematical background necessary for the proposed protocol's description.

### 2.1 Bilinear Pairing

Consider two groups  $G_1$  and  $G_2$  both of order  $q$ , a cyclic additive group and cyclic multiplicative group respectively. A map  $\hat{e}: G_1 \times G_1 \rightarrow G_2$  has following properties [6].

- Bilinearity;  $\forall P, Q \in G_1$  and  $\forall a, b \in \mathbb{Z}_q^*$   

$$\hat{e}(aP, bQ) = \hat{e}(P, Q)^{ab}.$$
- Non-Degeneracy;  $\hat{e}(P, P) \neq 1$ , where  $P \neq 0$  to avoid everything totally mapped to the identity element.
- Computability:  $\hat{e}: G_1 \times G_1 \rightarrow G_2$  is efficiently computable.

Notice that  $\hat{e}(aP, bQ) = \hat{e}(bP, aQ)$  by the bilinear property.

### 2.2 Computational Discrete Logarithm Problem

Given  $Q = kP$ , where  $P, Q \in G_1$ , it is relatively easy to compute  $Q$  given  $k$  and  $P$ . However, it is relatively hard to determine  $k$  given  $Q$  and  $P$  [8–10].

### 3 Three Factor Authentication Protocol

The major contributions of the proposed protocol are: (1) achieving mutual authentication and session key agreement in an efficient way; (2) no need of directory or password verification table (3) offline password guessing attack resilience (4) insider attack resilience and replay attack resilience. The network environment involves one registration server, many users subscribing to different service servers under the same registration server. The users access services from the service servers over an insecure channel by means of handy devices. Entities mutual authentication is a pre-requisite before granting access to service server's resources. The protocol takes four phases: *set up phase*, *registration phase*, *login phase* and *password change phase* as discussed below.

#### 3.1 Set Up Phase

The registration server,  $RS$  selects two groups  $G_1$  and  $G_2$ , a bilinear mapping  $\hat{e}: G_1 \times G_1 \rightarrow G_2$  and a cryptographic function  $H():\{0, 1\}^* \rightarrow G_1$ .  $RS$  selects a master secret key  $s$  and computes a corresponding public key as  $P_s = sP$ , where  $P$  is the generator of the group  $G_1$ . Then  $RS$  publishes  $\{G_1, G_2, q, P, P_s, \hat{e}, H()\}$  while keeping the master key  $s$  secret.

#### 3.2 Registration Phase

Any user registers with a registration server ( $RS$ ) and the procedures is as follows.

- R1. A user,  $U_i$ , submits  $H(PW_i||b)$ ,  $H(f_i||ID_i)$  and  $ID_i$  to  $RS$  where  $PW_i$  is a chosen password,  $b$  is a random number,  $f_i$  is hashed biometrics ( $B_i$ ) and  $ID_i$  is identity.
- R2.  $RS$  computes  $U_i$ 's public key  $Q_u = H(ID_i)$  and  $U_i$ 's private key  $P_u = sQ_u$ .
- R3. Further  $RS$  computes  $V_i = P_u \oplus y_i$  and  $t_i = H(P_u)$ , where  $y_i = H(PW_i||b) \oplus H(f_i||ID_i)$ .
- R4.  $RS$  sends the smart card to  $U_i$  securely stored with  $\{ID_i, H(), t_i, V_i\}$ .
- R5. Upon receipt of the smart card,  $U_i$  inserts the random number  $b$  on the smart card so the information stored in the card is  $\{ID_i, H(), Q_u, t_i, V_i, b\}$ .

In a similar manner, a service server,  $SS_j$  registers with  $RS$  and is issued a pair of a public key,  $Q_{ssj} = H(ID_{ssj})$  and a private key,  $P_{ssj} = sH(ID_{ssj})$ .

### 3.3 Login and Verification Phase

To login to the service server  $SS_j$  user inserts the smart card into a terminal and inputs identity  $ID_i$ , password  $PW_i$  and his/her biometrics  $B_i$  into a device and then:

- A1. The smart card computes  $H(PW_i||b)$ ,  $f_i = H(B_i)$  and  $H(f_i||ID_i)$  then cross-checks the correctness of  $ID_i, f_i$  and  $H(PW_i||b)$  and if found valid the processes continues as below otherwise terminates.
- A2. The smart card computes  $P_u^* = V_i \oplus y_i$ .
- A3. Then the smart card checks if  $t_i = H(P_u^*)$ . If the result holds, it means  $U_i$  inputted correct password  $PW_i$ , identity  $ID_i$  and the biometrics  $B_i$  hence is authenticated.
- A4. Now the smart card selects a random number  $r_i$  and computes  $r_iP$ .
- A5. Smart card computes  $K_a = H(r_iP||P_s||Q_u||Q_{ssj}||\hat{e}(sQ_uP, r_iQ_{ssj}))$ ,  $SK = H(K_a||ID_i||ID_{ssj})$  and  $auth = H(r_i||y_i||ID_i||ID_{ssj}||SK)$  and sends  $\{E_{K_a}(r_i||auth||ID_i), r_iP, y_i\}$  to  $SS_j$ .
- A6. In turn  $SS_j$  computes  $K_a = H(r_iP||P_s||Q_u||Q_{ssj}||\hat{e}(Q_u r_i P, sQ_{ssj}))$  and uses it to decrypt  $E_{K_a}(r_i||auth||ID_i)$ , the message from smart card  $\{E_{K_a}(r_i||auth||ID_i), r_iP, y_i\}$ .
- A7. Then  $SS_j$  cross-checks if  $ID_i$  and  $auth$  are valid then continues to compute the session key as  $SK = H(K_a||ID_i||ID_{ssj})$  and verifies  $auth = H(r_i||y_i||ID_i ||ID_{ssj}||SK)$ . When it computations hold  $SS_j$ 's accepts  $U_i$ 's request otherwise rejects.

### 3.4 Password Change Phase

At will  $U_i$  has the right to change his/her password from  $PW_i$  to  $PW_i^{new}$  for security reasons without involving  $RS$ . Below is the procedure for password change phase.

- C1.  $U_i$  submits identity,  $ID_i$ , current password,  $PW_i$  and his/her biometrics,  $B_i$  into a device before submitting the new password  $PW_i^{new}$ .
- C2. Smart card computes  $H(PW_i||b)$ ,  $f_i = H(B_i)$  and  $H(f_i||ID_i)$  and checks the validity of the results. The process continues if the calculations holds otherwise terminates.
- C3. Then smart card computes  $y_i = H(PW_i||b) \oplus H(PW_i^{new}||b)$ .
- C4. Smart card replaces  $y_i$  with  $y_i^{new} = H(PW_i^{new}||b) \oplus H(f_i||ID_i)$ .

## 4 Security and Performance Analysis

This section points out the security properties and computational performance.



## 4.1 Security Analysis

This section shows how the proposed protocol achieves notable security requirements.

**Proposition 1** *The proposed protocol supports mutual Authentication.*

*Proof* Mutual authentication is satisfied in the sense that each party verifies the counterpart's legitimacy. This property is implicitly embedded in the ability to establish a session key  $SK$  between the two. When  $U_i$  sends the message  $\{E_{K_a}(r_i||auth||ID_i), r_iP, y_i\}$  to  $SS_j$ , h/she is assured that only the intended service server will be able to compute the session key  $SK = H(r_iP||P_s||Q_u||Q_{ssj}||\hat{e}(sQ_uP, r_iQ_{ssj}))$ , because no one else could compute  $\hat{e}(Q_u r_iP, sQ_{ssj}) = \hat{e}(Q_u r_iP, sQ_{ssj})$  but  $SS_j$  only by bilinear pairing property. While on the hand  $SS_j$  knows only  $U_i$  could form a valid pair  $\hat{e}(sQ_u r_iP, Q_{ssj})$ .

**Proposition 2** *The proposed protocol supports offline password guessing attack resilience.*

*Proof* Offline password guessing attack resilience is met by the fortification of the smart card stored information,  $\{ID_i, H(), Q_u, t_i, V_i, b\}$ . Even if an adversary skims the information the tough task of guessing  $PW_i$  awaits. The attempt would mean computing  $H(PW_i||b) \oplus H(f_i||ID_i) = P_u \oplus V_i$ , which would require knowledge of the private key  $P_u$ , the random umber  $b$  and the hashed biometrics  $f_i$ . So guessing the four parameters,  $PW_i, P_u, b$  and  $f_i$  is extremely hard besides  $b$  is a high entropy number.

**Proposition 3** *The proposed protocol supports replay attack resilience.*

*Proof* Replay attack is resisted by usage of fresh random number  $r_i$  in the message  $\{E_{K_a}(r_i||auth||ID_i), r_iP, y_i\}$ . Even if an adversary would try to replay an old authenticated message  $\{E_{K_a}(r_i||auth||ID_i), r_iP, y_i\}^*$ , h/she will fail to get authenticated obviously because  $auth \neq auth^*$  as a result of the randomness of  $r_i$  in each session.

**Proposition 4** *The proposed protocol supports insider attack resilience.*

*Proof* Insider attack is overcome by securing the message  $\{H(PW_i||b), H(f_i||ID_i)\}$  and  $ID_i\}$  during registration phase. In order to fabricate a correct login message,  $P_u^{**} = V_i^* \oplus y_i^*$  an adversary has to guess  $PW_i, b$ , and  $f_i$  simultaneously, which is impossible within polynomial time. Thus the protocol copes up with insider attack.

## 4.2 Performance Analysis

This section gives computational cost of the protocol in comparison with related protocols depicted in Table 1.

**Table 1** Comparison with related protocols on computational cost

Protocols	Computational operations				
	No. of pairing	No. of hash function	No. of point addition in EC	No. of scalar multiplication in EC	No. of symmetric encryption or decryption
Ours	2	9	1	2	1
Juang et al's	2	10	0	3	2
Das et al's	2	2	1	2	0

Table 1 is gives communication cost of the shown protocols with an exception of symmetric encryption/decryption, the operations' computational load in ascending order are: bilinear pairing operation is heaviest followed by hash function then EC point addition operation and finally EC scalar multiplication with reference to [6]

Compared with communication load of other smart card based authentication protocols shown in Table 1 and presented in [6, 11] our protocol is moderately heavier but achieves high security level. Therefore the proposed protocol is relatively efficient and strong against smart card well known attacks.

## 5 Conclusion

A secure three factor smartcard biometric authentication protocol has been proposed that is facilitated by the properties of bilinear pairing and computational discrete logarithm problem. The proposed protocol does not require a directory or password verification table and supports offline password change without involving the server. Further the protocol resists lost or stolen password guessing attack. Though has comparatively higher computation load the proposed protocol achieves high security.

**Acknowledgments** This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (MEST) (NRF-2010-0021575).

## References

1. Sanchez-Reillo R (2001) Smart card informations and operations using biometrics. IEEE AESS systems magazine
2. Jeon IS, Kim HS, Kim MS (2011) Enhanced biometrics-based remote user authentication using smart cards. *J Secur Eng* 16(1):9–19
3. Arkantonakis KM, Tunstall M, Hancke G, Askoxylakis I, Mayes K (2009) Attacking smart cards systems: theory and practice. *Inf Secur Tech Rep* 14:46–56
4. Li X, Niu JW, Ma J, Wang WD (2011) Cryptanalysis and improvement of a biometric-based remote user authentication scheme. *J Comput Appl* 34:73–79

5. Chen TH, Hsiang HC, Shih WK (2011) Security enhancement on an improvement on two remote user authentication schemes using smart cards. *Furth Gener Comput Syst* 27:377–380
6. Juang WS, Nien WK (2008) Efficient password authenticated key agreement using bilinear pairing. *Math Comput Model* 47:1238–1245
7. Lao YP, Hsiao CM (2012) A novel multi-server remote user authentication scheme using self-certified public keys for mobile clients. *Furth Gener Comput Syst* 29(3):886–900
8. Giri D, Srivastava PD (2006) An improved remote user authentication scheme with smart cards using bilinear pairings. In: *IACR Cryptology ePrint Archive*, 2006, p 274
9. Lao YP, Hsiao CM (2011) The improvement of ID-based remote user authentication scheme using bilinear pairings. In: *Consumer electronics, communications and networks international conference (CECNet)*, pp 865–869
10. Sarier ND (2010) Improving the accuracy and storage cost in biometric remote scheme. *J Netw Comput Appl* 33:268–274
11. Das ML, Saxena A, Gulati VP, Phatak DB (2006) A novel remote user authentication scheme using bilinear pairings. *J Comput Secur* 25:184–189
12. Wang D, Ma CG, Wu P (2012) Secure password-based remote user authentication scheme with non-tamper resistant smartcards. In: *Proceedings of the 26th annual IFIP WG 11.3 conference on data and application security and privacy*, pp 112–121
13. Pippal RS, Jaidhar CD, Tapaswi S (2012) Security issue in smart card authentication scheme. *Int J Comput Theory Eng* 4(2):206–211
14. Chang CC, Lee CY, Chiu YC (2009) Enhanced authentication scheme with anonymity for roaming service in global mobility networks. *Comput Commun* 32:611–618
15. Li X, Ma J, Wang W, Xiong Y, Zhang J (2012) A novel smart card and dynamic ID based remote user authentication scheme for multi-server environment. *Math Comput Model*. doi: [10.1016/j.mcm.2012.06.033](https://doi.org/10.1016/j.mcm.2012.06.033)

# A LBP-Based Method for Detecting Copy-Move Forgery with Rotation

Ning Zheng, Yixing Wang and Ming Xu

**Abstract** Copy-move is the most common tampering manipulations, which copies one part of the image and pastes into another part in the same image. Most existing techniques for detecting tampering are sensitive to rotation and reflection. This paper proposed an approach to detect Copy-Move forgery with rotation. Firstly the suspicious image is divided into overlapping blocks, and then LBP operator are used to produce a descriptor invariant to the rotation for similar blocks matching. It is effective to solve the mismatch problem caused by the geometric changes in duplicated regions. In order to make the algorithm more effective, some parameters are proposed to remove the wrong matching blocks. Experiment results show that the proposed method is not only robust to rotation, but also to blurring or noise adding.

**Keywords** Copy-move · Image forgery · LBP · Rotation invariant

## 1 Introduction

With the wide application of powerful digital image processing software, such as Photoshop, digital image tampering becomes increasingly easy. At the same time, digital image forensic cause more and more attention. Copy-move is the most common tampering manipulations, which copies one part of the image and pastes into another part in the same image. The first method for detecting copy-move

---

N. Zheng · Y. Wang (✉) · M. Xu  
College of Computer, Hangzhou Dianzi University, Hangzhou 310018, China  
e-mail: star\_rats@163.com

N. Zheng  
e-mail: nzheng@hdu.edu.cn

M. Xu  
e-mail: mxu@hdu.edu.cn

forgery was proposed by Fridrich [1]. They lexicographically sorted quantized discrete cosine transforms (DCT) coefficients of small blocks and then checked whether the adjacent blocks are similar or not. In [2] Farid proposed a new method by adopting the PCA-based feature, which can endure additive noise, but the detection accuracy is low. Nevertheless, these methods are sensitive to the geometric changes in the copied part.

To solve the above problem, log-polar transform (LPT) may be performed on image blocks followed by wavelet decomposition [3]. There are some approaches [4] that extracted interest points on the whole image by scale-invariant feature transform (SIFT). However, these schemes still have a limitation on detection performance since it is only possible to extract the key points from particular points of the image. This paper presents a comprehensive novel method to detect duplicated regions that have undergone geometric changes, particularly reflection and rotation.

## 2 Rotation Invariant Feature

### 2.1 LBP Operator

LBP operator is an effective texture description operator. It has been successfully applied in image processing areas these years. Next, introduce how to calculate the LBP value. In  $3 \times 3$  window, the gray value of the center point of the window as a threshold value, other pixels in the window do binarized processing, generates an 8-bit binary string. Then, according to the different positions of the pixels, get the LBP value of the window by weighted summing. It can be computed by

$$LBP = \sum_{i=0}^7 s(g_i - g_c)2^i, \text{ where } s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (1)$$

Here  $g_c$  is the center pixel of the window,  $g_i$  represents surrounding pixels. Generally the order of the neighboring pixels is started by the pixel to the right of the center pixel, counterclockwise marked. The LBP value can reflect the texture information for the region. LBP can be expanded to a circular neighborhood. Using  $(P, R)$  to describe the neighborhood, where  $P$  represents the number of sampling points,  $R$  is the radius of the neighborhood. The gray values of neighbors which do not fall exactly in the center of pixels are estimated by interpolation.

### 2.2 Rotation Invariance

The  $LBP_{P,R}$  operator produces  $2^P$  different output values, corresponding to  $2^P$  different binary patterns that can be formed by the  $P$  pixels in the neighbor set. When the image is rotated, the gray values  $g_i$  will move along the perimeter of the

circle. After rotation, a particular binary pattern results in a different  $LBP_{P,R}$  value. This does not apply to patterns comprising of only 0 (or 1) which remain constant at all rotation angles. To remove the effect of rotation, assign a unique identifier to each rotation invariant local binary pattern, we define:

$$LBP_{P,R}^{ri} = \min\{ROR(LBP_{P,R}, i), i = 0, 1, \dots, P - 1\} \tag{2}$$

where  $ROR(x,i)$  performs a circular bit-wise right shift on the P-bit number  $x$   $i$  times, superscript  $ri$  means rotation invariant.  $LBP_{P,R}^{ri}$  quantifies the occurrence statistics of individual rotation invariant patterns corresponding to certain features in the image, hence, the patterns can be considered as feature detectors. In the case of  $P = 8$ ,  $LBP_{P,R}^{ri}$  will generate 36 different values or 36 patterns. Let vector  $V$  represents the occurrence number of individual patterns. When block is rotated,  $V'$  is extracted. It is expected that  $V$  and  $V'$  are similar, the correlation coefficients between them is close to 1. Compare the similarity between  $V$  and  $V'$ , it is easy to identify the duplicated blocks.

$$corr2(V, V') = \left( \sum_m \sum_n (v_{mn} - \bar{v})(v'_{mn} - \bar{v}') \right) / \sqrt{\left( \sum_m \sum_n (v_{mn} - \bar{v})^2 \right) \left( \sum_m \sum_n (v'_{mn} - \bar{v}')^2 \right)} \tag{3}$$

### 3 The Proposed Method

The framework of the proposed method is given as follow: (1) Dividing the suspicious image into blocks; (2) Extracting appropriate features from each block; (3) Searching similar block pairs; (4) Finding correct blocks and output them. The following is the implement details.

Step1: An  $M \times N$  color image is first split into overlapping blocks of  $B \times B$  pixels. Adjacent blocks have one different row or column. Thus  $(M - B + 1)(N - B + 1)$  blocks would be getting. Let  $A_i$  denote the  $i$ -th block of pixels,  $i = 1, 2, \dots, N_{blocks}$ , where  $N_{blocks} = (M - B + 1)(N - B + 1)$ .

Step2: Then feature vectors are extracted from each block  $A_i$ . The mean of each of three channels has been proved effective against JPEG compression and blurring. The fourth feature is entropy. It can be calculated as,  $v = - \sum_{i=0}^{255} p_i \log_2 p_i$

where  $p_i$  is the proportion of the number of pixels which gray value is  $i$  to total pixels numbers. In previous detection methods, uniform areas in the image may lead to false matches. The entropy can be used to identify blocks with insufficient textural information. Thus, blocks whose entropy is lower than a defined threshold  $e_{min}$  could be discarded. However, this could prevent the system from detecting duplicates in areas with scarce textural information. Finally, a  $1 \times 4$  eigenvectors  $v$  can be gotten from each block.

Step3: The extracted eigenvectors are arranged a matrix  $L$  with the size of  $(M - B + 1)(N - B + 1) \times 4$ , then  $L$  is lexicographically sorted. To search for

the similar block pairs, the  $corr2(V_i, V_j)$  is computed using (3), for every  $|j - i| \leq \varphi_a$ , when the following conditions are satisfied:  $|v_i^k - v_j^k| \leq \varphi_b, k = 1, 2, 3, |v_i^4 - v_j^4| \leq \varphi_c$  and  $d_{ij} \leq \varphi_d$ . Where  $\varphi_a, \varphi_b, \varphi_c, \varphi_d$  are threshold discussed in next section and  $d_{ij}$  is the Euclidean distance between two blocks. It can be calculated as  $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ ,  $(x_i, y_i)$  is the top-left corner's coordinate of the  $i$ -th block.

Let  $c_{ip}$  be the higher correlation coefficient computed for  $V_i$ . If  $c_{ip}$  is greater than a user-defined similarity threshold  $\varphi_s$ , records the offsets between  $i$ -th block and  $p$ -th block as well as their locations. Whenever we find a pair of matching blocks, create a record. Finally, form a list  $Q$  with all the created records.

Step4: Initialize a black map image  $P$  with the size  $M \times N$ . According to the  $Q$ , mark the suspicious blocks in  $P$ . Morphologic operations are applied to  $P$  to fill the holes in the marked regions and remove the isolated points, then output the final result.

## 4 Experimental Results and Analysis

The experiments were carried out on the Matlab R2009a. All tampered images are generated from three datasets. The first dataset are the Uncompressed Color Image Database (UCID) [5]. The second is several uncompressed color PNG images of size  $768 \times 512$  pixels released by the Kodak Corporation [6]. The last is a tampered image sets including original color images and their copy-move forged versions [7]. Two standard metrics detection accuracy rate (DAR) and false positive rate (FPR) will be adopted to quantify the accuracy and robustness. These two parameters indicated how precisely our method could locate the duplicated regions. The more *DAR* is close to 1 and *FPR* is close to 0, the more precise the method would be. Unless otherwise noted, all the thresholds in the experiment are set as:  $B = 8, \varphi_a = 35, \varphi_b = 2, \varphi_c = 0.3, \varphi_d = 40, \varphi_s = 0.95, e_{\min} < 4$ . The specific thresholds would be given in case of the default parameters are not well working.

### 4.1 Effectiveness Testing

In order to test the effectiveness of the proposed method, for the first experiment, we choose some images with the size of  $200 \times 200$  pixels from the third dataset. The detection results can be seen from Fig. 1. The first line is forgery images, the second line is detection results. All the duplication regions are non-regular and the detection algorithm can find the tampered regions precisely, though each image has complex texture background.

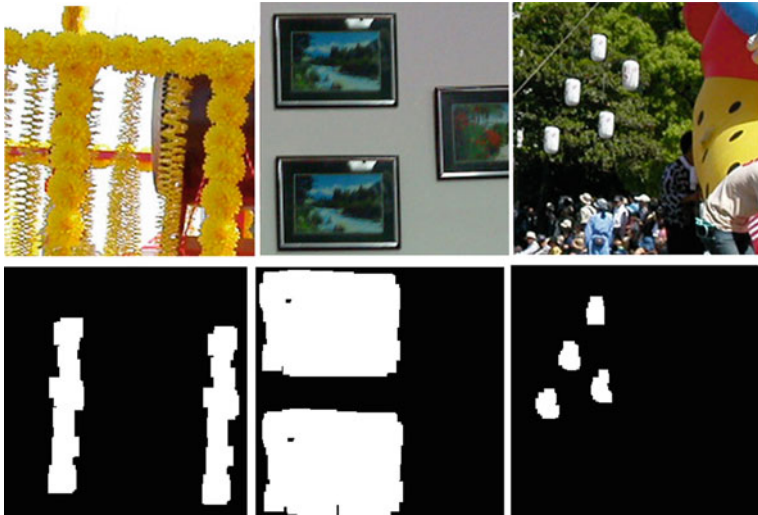


Fig. 1 The detecting results for non-regular copy-move forgery

The basic motivation of our scheme is to detect regions that have undergone geometric changes, particularly reflection and rotation. Following experiments are designed to detect the duplicative region when it is rotated with different degrees. This kind of forgery can not be detected by the other existing methods. In Fig. 2 the duplicative regions have been horizontal reflection, vertical reflection and rotated. Detecting result shows the method can effectively solve this situation.



Fig. 2 The detecting results when the regions are reflected and rotated



## 4.2 Robustness and Accuracy Test

In real life, some evil people often handle the tampered images with post-processing operation, such as noise adding, blurring or mixture operations. Experiments on images that distorted by blur and additive white Gaussian noise were also performed to test how precise our method was in these two cases. 100 images were chosen from the datasets, for each image, copying a square region at a random location and pasting onto a non-overlapping region. The square region's size was fixed as  $48 \times 48$  in this part and the parameters were set as:  $B = 16$ ,  $\varphi_a = 30$ ,  $\varphi_b = 2$ ,  $\varphi_c = 0.3$ ,  $\varphi_d = 40$ ,  $\varphi_s = 0.98$ ,  $e_{\min} < 3$  for AWGN distortion, while default parameters were used for blurred images. Results of tampered images distorted by AWGN with different power and blur were shown in Table 1. The values indicated that the algorithm had the ability to locate tampering regions in the case of processing distorted image. All tampering images had been detected and the detecting precision was good.

## 4.3 Computational Complexity

The number of blocks and the dimension of the vector are two important aspects of reducing computational complexity. Assuming a  $256 \times 256$  image and the size of block is  $8 \times 8$ . Table 2 displays the comparison results between the different methods. Table 2 shows the time complexity of algorithm [1, 2, 8] has higher than our method.

**Table 1** The result of additive white Gaussian noise and Gaussian blurring

SNR	20 dB	30 dB	40 dB
DAR	0.924	0.961	0.982
FPR	0.085	0.092	0.087
$w, \sigma$	5, 0.5	5, 1	5, 1.5
DAR	0.988	0.928	0.895
FPR	0.041	0.106	0.213

**Table 2** Computation complexity comparisons

Methods	Numbers of blocks	Feature dimension	Detection time (s)
Fridrich's	58081	64	55.6
Farid's	58081	32	50.8
Wang's	58081	4	20.2
Our proposed	58081	4	4.3

## 5 Conclusions

This paper present an automatic duplication image region detection algorithm based on LBP. It works in the absence of digital watermarking and does not need any prior information about the tested image. Compared with previous works, our algorithm used less features to represent each blocks, and was more effective. The experiment results prove that the proposed method have nice robustness to post-processing and rotation. Thus, our method could be useful in some areas of forensic science.

**Acknowledgments** This paper is supported by NSFC (No. 61070212, No.61003195), NSF of Zhejiang Province, China (No. Y1090114), the State Key Program of Major Science and Technology (Priority Topics) and the science and technology search planned projects of Zhejiang Province, China (No 2010C11050, No. 2012C21040).

## References

1. Fridrich J, Soukal D, Lukas J (2003) Detection of copy–move forgery in digital images. In: Proceedings of digital forensic research workshop, IEEE Computer Society, Cleveland, OH, USA, August, pp 55–61
2. Popescu A, Farid H (2004) Exposing digital forgeries by detecting duplicated image regions, Technical Report TR2004-515, Department of Computer Science, Dartmouth College
3. Myna AN, Venkateshmurthy MG, Patil CG (2007) Detection of region duplication forgery in digital images using wavelets and log-polar mapping. In: Proceedings of the international conference on computational intelligence and multimedia applications, Washington, DC, USA, pp 371–377
4. Huang H, Guo W, Zhang Y (2008) Detection of copy–move forgery in digital images using sift algorithm. In: Proceedings of the 2008 IEEE Pacific-Asia workshop on computational intelligence and industrial application, IEEE Computer Society, Washington, DC, USA, pp 272–276
5. Schaefer G, Stich M (2004) UCID-An uncompressed colour image database. In: Proceedings of the SPIE, storage and retrieval methods and applications for multimedia, pp 472–480
6. <http://r0k.us/graphics/kodak/>
7. <http://faculty.ksu.edu.sa/ghulam/Pages/ImageForensics.aspx>
8. Wang JW, Liu GJ (2009) Detection of image region duplication forgery using model with circle block. In: International conference on multimedia information networking and security

# Attack on Recent Homomorphic Encryption Scheme over Integers

Haomiao Yang, Hyunsung Kim and Dianhua Tang

**Abstract** At CDCIEM 2012, Yang et al. proposed a new construction of somewhat homomorphic encryption scheme over integers, which is quite efficient in the perspective of the key size. In this paper, we present an effective lattice reduction attack on Yang et al.'s scheme, where it is easy to recover the plaintext by applying LLL algorithm.

**Keywords** Homomorphic encryption · LLL algorithm · Lattice · Cloud computing

## 1 Introduction

Fully homomorphic encryption (FHE) can operate the arbitrary plaintext information homomorphically, just by operating ciphertexts, without decryption. However, how to construct an efficient FHE scheme has been still an open problem for over 30 years. In 2009, the old open problem was solved by the breakthrough work of Gentry [1]. At the same time, Gentry still gave a construction framework that a fully homomorphic scheme could be transformed from a “somewhat” homomorphic scheme.

---

H. Yang

College of Computer Science and Engineering, UEST of China, Chengdu 610054, China

H. Kim (✉)

Department of Cyber Security, Kyungil University, Kyungpook, Kyungsansi 712-701, Korea

e-mail: kim@kiu.ac.kr

D. Tang

Science and Technology on Communication Security Laboratory, Chengdu 610041, China

Gentry's somewhat scheme originally worked with ideal lattices. At 2010, Dijk et al. proposed a very simple somewhat homomorphic scheme only over the integers, which had owned merit of conceptual simplicity [2]. However, this simplicity came at the cost of public key size in  $O(\lambda^{10})$ . Although at 2011, Coron et al. reduced the public key size to  $O(\lambda^7)$ , it was still too large for practical applications [3]. At 2012, Yang et al. further reduce the public key size to  $O(\lambda^3)$  by encrypting with a new form [4].

In this paper, based on LLL algorithm, we present an effective attack on Yang et al.'s scheme. Our attack shows that it is easy to recover the plaintext by using lattice reduction: it is a matter of applying LLL in a lattice of dimension 3.

## 2 Recent Somewhat Homomorphic Encryption Scheme

For convenience, the same notations are used as in [4]. The construction of Yang et al.'s somewhat homomorphic encryption scheme is as follows.

- **KG**( $\lambda$ ): Choose randomly an odd  $\eta$ -bit integer  $p \in [2^{\eta-1}, 2^\eta)$ . Choose randomly four integers  $l_0, l_1 \in \mathbb{Z} \cap (0, 2^\gamma/p)$ ,  $h_0, h_1 \in (-2^\rho, 2^\rho)$ . Compute  $x_i = pl_i + 2h_i$ ,  $i = 0, 1$ . Assume that  $|x_0| > |x_1|$ . Set public key  $pk = \langle x_0, x_1 \rangle$ , and secret key  $sk = p$ .
- **Enc**( $pk, m$ ): To encrypt a bit  $m \in \{0, 1\}$ , choose randomly two integers  $r \in (-2^{\rho'}, 2^{\rho'})$ ,  $r_1 \in (-2^\rho, 2^\rho)$  and compute the ciphertext  $c = m + 2r + r_1x_1 \pmod{x_0}$ .
- **Eval**( $pk, C, \vec{c}$ ): The function is the same as in [2].
- **Dec**( $sk, c$ ): Output  $m = (c \pmod p) \pmod 2$ , where  $(c \pmod p)$  is the integer in  $(-p/2, p/2)$ .

To foil various attacks, a convenient parameter set is  $\rho = \lambda, \rho' = 2\lambda, \eta = O(\lambda^2), \gamma = O(\lambda^3)$ , where  $\lambda$  is a security parameter.

*Remark 1* According to the parameter set in Yang et al.'s scheme, we have  $r = O(2^{2\lambda}), r_1 = O(2^\lambda)$  and  $c = O(2^{\lambda^3})$ .

*Remark 2* In Yang et al.'s scheme, the message is encrypted with a new form. However, the new form of encryption results in the lattice reduction attack by LLL algorithm. Before describing the attack, we first give a brief introduction of LLL algorithm.

## 3 Lattice and LLL Algorithm

**Definition 1** (*Lattice*) Let  $B = (b_1, b_2, \dots, b_n)^T$ , where  $b_1, b_2, \dots, b_n \in \mathbb{R}^m$  are  $n$  linearly independent row vectors, the lattice generated by  $B$  is

$$\mathcal{L}(B) = \{x_1b_1 + x_2b_2 + \dots x_nb_n | x_i \in \mathbb{Z}\},$$

where, we refer to  $b_1, b_2, \dots, b_n \in \mathbb{R}^m$  as a *basis* of the lattice and  $m$  as its *dimension* [5].

One basic parameter of a lattice is the length of the shortest nonzero vector in the lattice. This parameter is denoted by  $\lambda_1$ . By *length* of a vector  $v$ , the Euclidean norm of  $v$ , or the  $\|v\|$  norm, defined as following

$$\|v\| = \sqrt{\sum_{i=1}^n |v_i|^2}.$$

**Definition 2** (*LLL Algorithm of Lattice*) LLL algorithm can transform a basis to a reduced basis [6].

**INPUT:**  $b_1, \dots, b_n \in \mathbb{Z}^n$

**OUTPUT:**  $\delta$ -LLL reduced basis for  $\mathcal{L}(B)$

**Start:** Compute Gram-Schmidt orthonormal basis  $\tilde{b}_1, \tilde{b}_2, \dots, \tilde{b}_n$

**Reduction Step:**

for  $i = 2$  to  $n$  do

for  $j = i - 1$  to  $1$  do

$b_i \leftarrow b_i - c_{i,j}b_j$  where  $c_{i,j} = \lceil \langle b_i, \tilde{b}_j \rangle / \langle \tilde{b}_j, \tilde{b}_j \rangle \rceil$

**Swap Step:**

if  $\exists i$  s.t.  $\delta \|\tilde{b}_i\|^2 > \|\mu_{i+1,i}\tilde{b}_i + \tilde{b}_{i+1}\|$  then

$b_i \leftrightarrow b_{i+1}$

goto **start**

**Output:**  $b_1, \dots, b_n$

One important property of LLL-reduced basis is that its first vector is relatively short, as shown in the next claim. Our attack is based on the following claim.

*Claim 1.* Let  $b_1, b_2, \dots, b_n \in \mathbb{R}^n$  be a  $\delta$ -LLL reduced basis. Then

$$\|b_1\| \leq \left( \frac{2}{\sqrt{4\delta - 1}} \right)^{n-1} \lambda_1(\mathcal{L}).$$

## 4 Attack to Yang et al.'s Somewhat Scheme

This section provides an effective lattice reduction attack on Yang et al.'s scheme, where it is easy to recover the plaintext by applying LLL algorithm.

### 4.1 Basic Idea

A ciphertext in Yang et al.'s scheme is of the form:  $c = m + 2r + r_1x_1 \pmod{x_0}$ , where  $m$  is the message (0 or 1),  $x_0$  and  $x_1$  are known large integers, and  $r$  and  $r_1$  are random small integers. But it is very easy to recover the unknown integers and hence  $m$  using lattice reduction: it is a matter of applying LLL algorithm in a lattice of dimension 3.

### 4.2 Construction of Lattice

By using **Enc** algorithm, we have  $c = m + 2r + r_1x_1 \pmod{x_0}$ . Since  $|x_0| > |x_1|$ , we have that

$$c = m + 2r + r_1x_1 + ax_0, \quad |a| < r_1.$$

Let  $k_1 = m + 2r$ ,  $k_2 = r_1$ ,  $k_3 = a$  and  $k = (k_1, k_2, k_3)$ , we have that

$$(1, k_2, k_3) \begin{pmatrix} c & 0 & 0 \\ -x_1 & 1 & 0 \\ -x_0 & 0 & 1 \end{pmatrix} = (k_1, k_2, k_3),$$

where  $k_1 = O(2^{2\lambda+1})$ ,  $k_2 = O(2^\lambda)$  and  $k_3 = O(2^\lambda)$ . Let

$$B = \begin{pmatrix} c & 0 & 0 \\ -x_1 & 1 & 0 \\ -x_0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

Therefore, we have a lattice of dimension 3:  $\mathcal{L}(B)$ , and the lattice basis is  $b_1 = (c, 0, 0)$ ,  $b_2 = (-x_1, 1, 0)$  and  $b_3 = (-x_0, 0, 1)$ .

### 4.3 Attack by Using LLL Algorithm

On one hand, we would like to show how can find a short vector  $v$  in this lattice. First of all, we compute

$$\det(\mathcal{L}(B)) = c.$$

Then by using the LLL algorithm, we find a short lattice vector  $v$  whose length satisfies

$$\|v\| \leq O(\lambda_1(\mathcal{L}(B))) \leq O\left((\det(\mathcal{L}(B)))^{1/3}\right) = O(c^{1/3}) = O\left(2^{\lambda/3}\right),$$

where the second inequality follows from Minkowski's theorem [6] and the second equality follows from Yang et al.'s parameter set.

On the other hand, we have  $\|k\| = O(2^{2\lambda+1})$ , which is much smaller than  $O(2^{\lambda^3/3})$ . So the vector  $k$  is just the shortest vector. We know that for a lattice of dimension 3, there exists a probabilistic polynomial time algorithm [7] to find the shortest vector  $k$ . Therefore, from  $k_1 = m + 2r$ , we have  $m = k_1 \bmod 2$ , and the message  $m$  is recovered.

## 5 Experiment of Attack

We run it on a Thinkpad Notebook, featuring an Intel CPU P8400 (2.26 GHz), with 3 GB of RAM. Our implementation uses Shoup's NTL library [8] version 5.5.2 for high-level numeric algorithms.

The parameters  $\rho = \lambda, \rho' = 2\lambda, \eta = O(\lambda^2), \gamma = O(\lambda^3)$  is set, where  $\lambda$  is security parameter. For convenience, first let  $\lambda = 10$ . The running result is as follows.

[The secret key]

$p = 814105364630556351240736280183$

[The public key]

$x_0 = 78346958740686353860440815564658914756426973101293533788204$   
 213865118944356453715800823786425964931810026856009990255869548506  
 029792032338238257897021458714393582082905984639667197072688343270  
 317287337116407175580467867712338785238976504175547934667309204186  
 47299829548252839709977921008223739649015191

$x_1 = 49904705158886425782744562622451521600687523551167430388079$   
 095620911235428047145543612591989134670946735716126048747386820180  
 748497914316405897990865092507125241729118662840060489456988543515  
 007765223327056952653894471517794934268922493112843996227517667057  
 68350604728270811758268128643502622618273548

[The random numbers for encrypting]

$r = 101223, r_1 = 17$

[The ciphertext of 0 bit]

$c = 134365584464806541581914066295721951090088037443825550729017$   
 269608173856441893995676475868703248438157882419670641089890904936  
 032478123417205710225294732372002935169485627553108470307665362183  
 581519117205107362689405290332127550570591630126793454726009060705  
 8337844650177436919198944150916551628314339

**Table 1** Attack Yang' et al. scheme with LLL algorithm (Intel CPU P8400 (2.26 GHz) with 3 GB of RAM)

Security parameter $\lambda$	Number of run	Number of success	Ratio of success (%)	Average time per run (s)
6	100	96	96	0.00078
10	100	95	95	0.02481
20	100	98	98	6.7111
30	100	92	92	210.962

As a result, the lattice basis:  $b_1 = (13436558446480654158191406629572$   
 $195109008803744382555072901726960817385644189399567647586870324843$   
 $815788241967064108989090493603247812341720571022529473237200293516$   
 $948562755310847030766536218358151911720510736268940529033212755057$   
 $059163012679345472600906070583378446501774369191989441509165516283$   
 $14339, 0, 0)$ ,  $b_2 = (-499047051588864257827445626224515216006875235511$   
 $674303880790956209112354280471455436125919891346709467357161260487$   
 $473868201807484979143164058979908650925071252417291186628400604894$   
 $569885435150077652233270569526538944715177949342689224931128439962$   
 $2751766705768350604728270811758268128643502622618273548, 1, 0)$ ,  $b_3 =$   
 $(-783469587406863538604408155646589147564269731012935337882042138$   
 $651189443564537158008237864259649318100268560099902558695485060297$   
 $920323382382578970214587143935820829059846396671970726883432703172$   
 $873371164071755804678677123387852389765041755479346673092041864729$   
 $9829548252839709977921008223739649015191, 0, 1)$ .

The first vector of LLL-reduced basis:  $\tilde{b}_1 = (202446, -17, 11)$ . It is easy to verify  $k = (k_1, k_2, k_3) = (202446, -17, 11) = (2r + 0, k_2, k_3)$ . So, the message 0 is recovered. Then we run experiments with different security parameters. The result is shown in Table 1.

## 6 Conclusion

In this paper, we pointed out that the somewhat homomorphic encryption scheme over integers presented by Yang et al. is vulnerable to lattice reduction attack. By using LLL algorithm, we could recover the plaintext easily. As shown in experiments, there was more than 90 % of recovering plain texts for different security parameters.

**Acknowledgments** This work was supported by National Research Foundation of Korea (NRF) under "2011 Korea-China Young Scientist Exchange Program" and National Science Fund of China under Grant No. 61103207 and also was partially supported by the National Research Foundation of Korea Grant funded by the Korean Government (MEST) (NRF-2010-0021575).



## References

1. Gentry C (2009) Fully homomorphic encryption using ideal lattices. In: STOC 2009, pp 169–178
2. Dijk M, Gentry C, Halevi S, Vaikuntanathan V (2010) Fully homomorphic encryption over the integers. In: Advances in cryptology—EUROCRYPT 2010. LNCS, vol 6110, pp 24–43
3. Coron JS, Mandal A, Naccache D, Tibouchi M (2011) Fully homomorphic encryption over the integers with shorter public keys. In: Advances in cryptology—CRYPTO 2011. LNCS, vol 6841, pp 487–504
4. Yang H, Tang D, Xia Q, Wang X (2012) A new somewhat homomorphic encryption scheme over integers. In: Proceedings of CDCIEM 2012, pp 61–64
5. Regev O (2005) On lattices, learning with errors, random linear codes, and cryptography. In: Proceedings of STOC 2005, pp 84–93
6. Nguyen PQ, Vallée B (2009) The LLL algorithm: survey and applications., Information security and cryptographySpringer, Heidelberg
7. Lenstra HW, Lenstra AK, Lovasz L (1982) Factoring polynomials with rational coefficients. Math Ann 261:515–534
8. Shoup V. NTL: A library for doing number theory. <http://shoup.net/ntl/>, Version 5.5.2

# A New Sensitive Data Aggregation Scheme for Protecting Data Integrity in Wireless Sensor Network

Min Yoon, Miyoung Jang, Hyoung-il Kim and Jae-woo Chang

**Abstract** Since wireless sensor networks (WSNs) are resources-constrained, it is very essential to gather data efficiently from the WSNs so that their life can be prolonged. Data aggregation can conserve a significant amount of energy by minimizing transmission cost in terms of the number of data packets. Many applications require privacy and integrity protection of the sampled data while they travel from the source sensor nodes to a data collecting device, say a query server. However, the existing schemes suffer from high communication cost, high computation cost and data propagation delay. To resolve the problems, in this paper, we propose a new and efficient integrity protecting sensitive data aggregation scheme for WSNs. Our scheme makes use of the additive property of complex numbers to achieve sensitive data aggregation with protecting data integrity. With simulation results, we show that our scheme is much more efficient in terms of both communication and computation overheads, integrity checking and data propagation delay than the existing schemes for protecting integrity and privacy preserving data aggregation in WSNs.

**Keywords** Sensor network · Data aggregation · Integrity · Data privacy · Signature

---

M. Yoon · M. Jang · H. Kim · J. Chang (✉)  
Department of Computer Engineering, Chonbuk National University, Jeonju,  
Republic of Korea  
e-mail: jwchang@jbnu.ac.kr

M. Yoon  
e-mail: myoon@jbnu.ac.kr

M. Jang  
e-mail: brilliant@jbnu.ac.kr

H. Kim  
e-mail: melipion@jbnu.ac.kr

## 1 Introduction

Recently, due to the advanced technologies of mobile devices and wireless communication, wireless sensor networks (WSNs) have increasingly attracted much interest from both industry and research. Since a sensor node has limited resources (i.e., battery and memory capacity), data aggregation techniques have been proposed for WSNs [1]. Another issue of WSNs is how to preserve sensitive measurements of everyday life where data privacy becomes an important aspect. In many scenarios, confidentiality of transported data can be considered critical, for instance, data from sensors might measure patients' health information such as heartbeat and blood pressure details. So, maintaining data privacy of a sensor node even from other trusted participating sensor nodes of the WSN is critical issue [2]. Although the existing data aggregation schemes have been proposed to preserve data privacy, they have the following limitations. First, the communication cost for network construction and data aggregation and data integrity is considerably expensive. Secondly, the existing schemes do not support data integrity due to communication loss. However, since the existing privacy-preserving schemes do not support privacy preservation and integrity protection simultaneously, it is required to carefully design a good data aggregation scheme for recent applications of WSNs, where both the privacy of sensed data and the integrity of the data should be provided [3].

To reserve these problems, in this paper, we propose a new and resource efficient scheme that can aggregate sensitive data protecting data integrity in WSNs. Our scheme utilizes complex numbers, which is an algebraic expression and can use arithmetic operations, such as addition (+), to aggregate and hide data (for data privacy) from other sensor nodes and adversaries during transmissions to the data sink. In our scheme, the real unit of a complex number is used for concealing sampled data whereas the imaginary unit is used for providing data integrity checking. Thus, our scheme not only prevents recovering sensitive information even though private data are overheard and decrypted by adversaries or other trusted participating sensor nodes but also provides data integrity checking. For data security, our scheme can be built on the top of the existing secure communication protocols like [4].

The rest of the paper is organized as follows. In Sect. 2, we present some related work. Section 3 describes our integrity protecting sensitive (private) data aggregation scheme in detail. Simulation results are shown in Sect. 4. Along with some future research directions, we finally conclude our work in Sect. 5.

## 2 Related Work

In this section, we present the existing data aggregation schemes for supporting data privacy and data integrity in WSNs. There are privacy preserving data aggregation schemes, such as *i*PDA and *i*CPDA. He et al. proposed *i*PDA [3] and

*i*CPDA [5] schemes for WSNs to support integrity. In the *i*PDA scheme, they protect data integrity by designing node disjoint two aggregation trees rooted at the query server where each node belongs to a single aggregation tree. In this technique, first, every sensor node slices its private data randomly into  $L$  pieces and  $L-1$  pieces are encrypted and sent to the randomly selected sensor nodes of the aggregation tree keeping one piece at the same sensor node. The same process is independently done for each sensor node using another aggregation tree. Then, all the sensor nodes which received data slices from multiple sensor nodes decrypt the slices using their shared keys and sum the received data slices including its own. After that, each sensor node sends the sum value to its parent from the respective aggregation tree. In the same way, the sum data from another set of sensor nodes are transmitted to the query server through another aggregation tree. In the end, the aggregated data from two node-disjoint aggregation trees reach to the base station where the aggregated data from both aggregation trees are compared. If the difference of the aggregated data from the two aggregation trees doesn't deviate from the predefined threshold value the query server accepts the aggregation result, otherwise, it rejects the aggregated result by considering them as polluted data. However, there are some shortcomings in the *i*PDA. First of all, during protecting data privacy it generates high traffics in the WSN. As a result, communication cost is significantly increased in the *i*PDA. Secondly, all sensor nodes use secret keys to encrypt their all data slices before sending to their respective  $2(L-1)$  number of sensor nodes. So, every sensor node has computation overhead of decrypting all the slices they received before aggregating them. In the *i*CPDA, three rounds of interactions are required: Firstly, each node sends a seed to other cluster members. Next, each node hides its sensory data via the received seeds and sends the hidden sensory data to each cluster member. Then, each node adds its own hidden data to the received hidden data, and sends the calculated results to its cluster head which calculates the aggregation results via inverse and multiplication of matrix. To enforce data integrity, cluster members check the transmitted aggregated data of the cluster head. There are some disadvantages of *i*CPDA. Firstly, the communication overhead of *i*CPDA increases quadratically with the cluster size. Secondly, the computational overhead of CPDA increases quickly with the increase of the cluster size which introduces large matrix, whereas lower cluster size introduces lower privacy-preserving efficacy.

### 3 Integrity-Protecting Sensitive Data Aggregation Protocol

To overcome previously mentioned shortcomings of the *i*PDA and *i*CPDA, in this section, we propose a new, efficient scheme in order to support data privacy and integrity in data aggregation for WSNs. Our scheme is based on the algebraic properties of the complex numbers and it not only ensures that no trend about sensitive data of a sensor node is released to any other nodes and adversaries but also provides data integrity of the aggregated value of sensor data. In particular,

we apply the additive property of complex number for data aggregation. We know that other aggregation functions, such as Average, Count, Variance, Standard Deviation and any other Moment of the measured data, can be reduced to the additive aggregation function Sum [4].

Our privacy and integrity preserving scheme is performed through five step. In the first step, we assign a special type of positive integer  $2^n$  (where,  $n = 0$  to  $Bn \times 8 - 1$ , such that  $Bn$  is the number of free bytes available in the payload) to every sensor node as node ID. This is because the binary value of every integer of  $2^n$  type has only one high bit (1). In addition, the position of the high bit for all integers of this type is unique. The sink node knows a data contributing sensor node through the signature of Node-ID. The Node-ID of a sensor node is used to generate a signature of a fixed length. A signature is a fixed size bit stream of binary numbers for a given integer. Signature of a sensor node ID can be generated by using the technique presented in the work [6]. We can determine the length of the signature based on the size of a given WSN. When the size of the WSN increases we can increase the length of the signature up to the  $Bn$  bytes. In other words, different size WSNs can have signatures of different lengths. The detail of using signatures has been presented in our previous work [7].

When the network receives a SQL like query for SUM aggregation function, in the second step, the sampled sensitive data  $ds$  of each sensor node is, first, concealed in a by combining with a unique seed ( $sr$ ) which is a private real number. The seeds can be selected from an integer range (i.e., space between lower bound–upper bound). By increasing the size of the range, we can further increase the level of the data privacy. Hence, our approach can support data privacy feature strongly. To support data integrity, an integer value  $b$ —the difference of the previous sensed value and the current sensed value of the sensor node—with  $i$  is appended to the  $a$  by using  $genCpxNum()$  function to form a complex number  $C = a + bi$ . We assumed that any sensor node cannot be compromised before sending first round data to the sink node. Every source sensor node keeps the original sensed value  $d$  of the current round to deduce  $b$  in the next round which is updated in each round of data transmission. Next, the source node encrypts the customized data  $R'_1$ , i.e.,  $R_1 = a + bi$ , and the signature of the node by using a secret key  $K_{x,y}$  [8] and transmits the cipher text  $C_j$  to its parent. The term  $K_{x,y}$  denotes a pairwise symmetric key shared by nodes  $x$  and  $y$  where the node  $x$  encrypts data by using a key  $K_{x,y}$  and the node  $y$  decrypts the data by using the key  $K_{x,y}$ . In this way, our algorithm converts the sampled data into an encrypted complex number form. Hence, it not only protects the transmitting trend of private data but also doesn't let neighboring sensor nodes and adversaries to recover sensitive data even though they overheard and decrypted the sensitive data.

In the third step, the parent sensor node (i.e., data aggregator) decrypts the received data by using respective pairwise symmetric keys of its child sensor nodes. For each child node, the parent node computes the difference value ( $b'$ ) of the two real units by using the stored previous data and received current data of the child node. For the first round, the value of  $b'$  is also zero. For this, the parent node always keeps the record of the previously received data from each of the child

nodes and it updates the previous data by current one in every round. To support local integrity checking, the parent node first compares just computed difference value with the currently received difference value (imaginary unit) from the child node and then compares the difference value with local threshold  $\delta$ . If the imaginary unit of the child's current data is equal to the computed difference value and the imaginary unit is not greater than  $\delta$  then the parent node accepts the data of the child node. Otherwise, the parent node rejects the data of the child sensor node considering as polluted data. After that the parent node adds the data of child nodes including its own by using additive property of complex number to produce an intermediate result  $R'$ . At the same time, it superimposes signatures ( $SSig$ ) of the contributed nodes by performing bitwise *OR* operation on the bit-streams of the node IDs and forwards the encrypted intermediate result ' $Cr$ ' towards the sink node. Since this approach needs just one bit to carry an ID of a sensor node it is 16 times scalable than the existing work CMT [4] where plaintexts (2-byte each) are used for carrying IDs of sensor nodes by simply concatenating them. Note: Different types of application can have different value for the threshold  $\delta$ . Thus, our algorithm supports local integrity checking which enforces to provide consistent data from child nodes. Above process continues at all nodes of the upper levels of the network until the whole partially aggregated data of the network reach to the sink node.

In the fourth step, when the sink node receives all intermediate result sets  $C_{rs}$  (partially aggregated encrypted customized data with superimposed signature) from the 1-hop child nodes, it decrypts them by using respective pairwise symmetric keys and computes the final aggregation  $SUM_2$  from  $C_{rs}$ . Since  $SUM_2$  is of complex number form and the sensed data has been concealed in the real unit by using private seeds identifying the information of the contributed sensor nodes is necessary to deduce actual  $SUM$  value.

In the last step, the sink node first knows data contributing nodes by checking the high bits (1 s) of the received superimposed signature by performing bitwise AND operation with the pre-stored signature files or superimposed signature of the Node-IDs of the all nodes of the network. For this, it separates  $SUM_2$  into real unit  $SUM_{2R}$  and imaginary unit  $SUM_{2IM}$ . Because the sampled data of sensor nodes has been concealed within the real unit, the sink node computes the actual aggregated result  $SUM$  by subtracting (an inverse operation of masking, step 2)  $SUM_{1R}$  (a freshly computed sum value of the private seeds of the contributed source nodes) from  $SUM_{2R}$ . The final result  $SUM$  is always accurate and reliable because of the following two reasons. First, a complex number is an algebraic expression and hence the underlying algebra gives the accurate result of the aggregated sensor data. Second, since the private seeds are fixed integer values (i.e., seeds are not random numbers) after collecting data by the sink node it subtracts exactly the same values that have been added to the sensor data during data hiding process by every source node. At the same time, before accepting the  $SUM$ , the sink node performs global integrity checking of  $SUM$  to assure whether the  $SUM_2$  has been polluted by an adversary in transit or not. For this, like parent nodes, the sink node also computes the difference value ( $B'$ ) of the two real units

by using the stored previous data and received current data from the network. The sink node first compares just computed difference value  $B'_i$  with the currently received difference value i.e.,  $SUM_{2IM}$ , from the network and then compares the difference value ( $SUM_{2IM}$ ) with global threshold  $\Delta$  (for every application, the maximum value for  $\Delta = \delta \times N$ , where  $N$  is the total number of nodes in a network). If the imaginary unit  $SUM_{2IM}$  of the current data from the network is equal to the just computed difference value  $B'_i$  and the  $SUM_{2IM}$  is not larger than  $\Delta$  then the sink node accepts the data of the network and returned the actual  $SUM$  to the query issuer. Otherwise, the sink node rejects the  $SUM$  considering it as forged/polluted data by adversary or other nodes.

## 4 Performance Evaluation

In this section, we present simulation results of our scheme by comparing it with iPDA and iCPDA schemes in terms of communication overhead and integrity checking. For this, we use TOSSIM simulator running over TinyOS operating system and GCC compiler. We consider 100 sensor nodes distributed randomly in  $100 \times 100$  m area.

Figure 1 shows communication overhead in terms of energy dissipation by the iPDA, iCPDA and our schemes with respect to varying number of sensor nodes in the WSN. The power consumption by our scheme is always lower than that of iPDA and iCPDA schemes. The reason is that the iPDA and iCPDA schemes generate too many unnecessary messages in the WSN while achieving integrity protecting and privacy preservation in data aggregation. And Fig. 2 compares integrity checking feature of all the three schemes. It is shown that our scheme can detect every polluted message but the iPDA and iCPDA has very low rate of polluted message detection. The reason is that every node in our scheme performs local integrity checking of the coming data from the lower level nodes. But, only sink node checks the integrity in iPDA and so does the cluster heads in iCPDA.

Fig. 1 Energy consumption

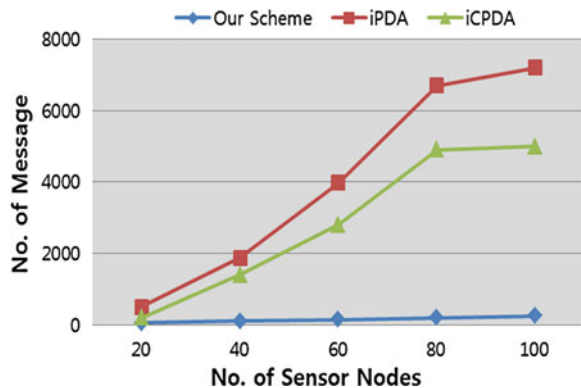
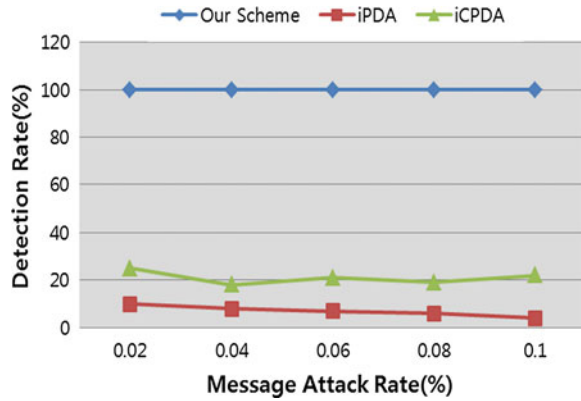


Fig. 2 Integrity checking



## 5 Conclusion

In this paper, we proposed an efficient and general scheme in order to aggregate sensitive data protecting data integrity for private data generating environments such as patients’ health monitoring application. For maintaining data privacy, our scheme applies the additive property of complex numbers where sampled data are customized and given the form of complex number before transmitting towards the sink node. As a result, it protects the trend of private data of a sensor node from being known by its neighboring nodes including data aggregators in WSNs. Moreover, it is still difficult for an adversary to recover sensitive information even though data are overheard and decrypted. Meanwhile, data integrity is protected by using the imaginary unit of complex-number-form customized data at the cost of just two extra bytes. Through simulation results, we have shown that our scheme is much more efficient in terms of communication and computation overheads, data propagation delay and integrity checking than the iPDA and iCPDA schemes.

As future work, we will provide more simulation results by designing data integrity and sensitive data preserving scheme under collusive attacks.

**Acknowledgments** This research was supported by Basic Science Research program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (grant number 2010-0023800).

## References

1. Considine J, Li F, Kollios G, Byers J (2004) Approximate aggregation techniques for sensor databases. In: Proceedings of ICDE, April 2004
2. Conti M, Zhang L, Roy S, Di Pietro R, Jajodia S, Mancini L-V (2009) Privacy-preserving robust data aggregation in wireless sensor networks. *Secur Commun Netw* 2:195–213



3. He W, Liu X, Nguyen H, Nahrstedt K, Abdelzaher T (2008) iPDA: an integrity-protecting private data aggregation scheme for wireless sensor networks. In: Proceedings of the IEEE MILCOM
4. Castelluccia C, Mykletun E, Tsudik G (2005) Efficient aggregation of encrypted data in wireless sensor networks. In: The second annual international conference on mobile and ubiquitous systems: networking and services, pp 109–117
5. He W, Liu X, Nguyen H, Nahrstedt K (2009) A Cluster-based protocol to enforce integrity and preserve privacy in data aggregation. In: Proceedings of the 29th IEEE international conference on distributed computing systems workshops, pp 14–19
6. Zobel J, Moffat A, Ramamohanarao K (1998) Inverted files versus signature file for text indexing. *ACM TDS* 23(4):453–490
7. Bista R, Chang JW (2010) Energy efficient data aggregation for wireless sensor networks. *Sustainable wireless sensor networks*, ISBN, pp 978–995
8. Blaß E-O, Zitterbart M (2006) An efficient key establishment scheme for secure aggregating sensor networks. In: Proceedings of the 2006 ACM symposium on information, computer and communications security, March 2006, pp 303–310

# Reversible Image Watermarking Based on Neural Network and Parity Property

Rongrong Ni, H. D. Cheng, Yao Zhao, Zhitong Zhang and Rui Liu

**Abstract** Reversible watermarking can recover the original cover after watermark extraction, which is an important technique in the applications requiring high image quality. In this paper, a novel image reversible watermarking is proposed based on neural network and parity property. The retesting strategy utilizing the parity detection increases the capacity of the algorithm. Furthermore, the neural network is considered to calculate the prediction errors. Experimental results show that this algorithm can obtain higher capacity and preserve good visual quality.

**Keywords** Reversible watermarking · Neural network · Parity property · Retesting strategy

## 1 Introduction

Reversible watermarking can completely restore the original digital contents after data extraction. For this characteristic, reversible watermarking is very useful for some applications where the availability of the original data is essential, such as military image processing and medical image sharing.

Early reversible watermarking algorithms mainly focus on lossless compression until the difference expansion algorithm is proposed by Tian [1]. The method divides the image into pairs of pixels and uses each legitimate pair for hiding one bit of information. It has high embedding capacity and high quality, and becomes the basic idea of some reversible watermarking methods. Later, prediction error

---

R. Ni (✉) · Y. Zhao · Z. Zhang · R. Liu  
Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China  
e-mail: rni@bjtu.edu.cn

H. D. Cheng  
Department of Computer Science, Utah State University, Logan, UT, USA

expansion (PEE) method is proposed by Thodi and Rodriguez [2]. Their method uses PEE to embed data, and suggests incorporating expansion embedding with histogram shifting to reduce the location map. Then, several PEE-based methods have been proposed [3–6]. In [6], Sachnev et al. propose a method which combines sorting and two-pass-testing with prediction error expansion method. The algorithm has higher capacity and lower distortions than most of other existing reversible watermarking methods.

In this paper, a novel image reversible watermarking is proposed based on neural network and parity property. Because the real embedded data is not always identical with the testing bit, some ambiguous pixel cells are generated. A retesting strategy utilizing the parity detection activates the capacity of the ambiguous pixel cells. As a result, the capacity is increased. Furthermore, considering the global feature, the neural network is used to predict the prediction errors. The experimental results show that the proposed algorithm can obtain higher capacity and preserve good visual quality.

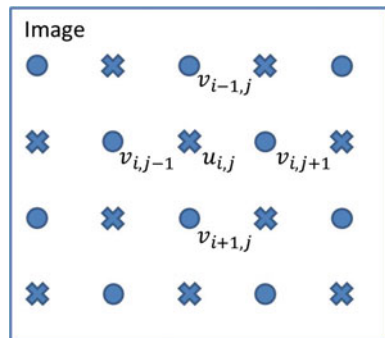
## 2 Proposed Algorithm Based on Neural Network and Parity Detection

In the proposed algorithm, all pixels of the image are divided into two sets: the “Cross” set and the “Dot” set (Fig. 1) as suggested in [6]. The watermark bits are embedded in the “Cross” set first, and then embedded in the “Dot” set.

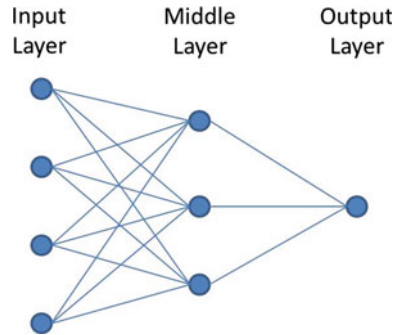
### 2.1 Prediction Based on Neural Network

During “Cross” embedding, the “Cross” set is used for embedding data while the “Dot” set works as the reference signals. And vice versa.

**Fig. 1** “Cross” set and “Dot” set



**Fig. 2** Structure of neural network



A center pixel of a cell is predicted by the four neighboring pixels. In this paper, neural network is used to predict the pixel values considering the global feature. Since four pixels in the neighboring region are utilized to calculate the prediction values, a neural network with four inputs is designed here. As shown in Fig. 2, the input layer has four neurons and the middle layer is created with three neurons. The output layer has one neuron which refers to the central pixel value.

After the construction of the neural network, the corresponding weights and parameters can be determined by training. Considering the global influence and generalization, a great number of pixel cells from many natural images are input to the neural network for obtaining a common model of prediction. This model can be shared by encoder and decoder in advance. Thus, the prediction value  $u'_{i,j} = \lfloor nnpredict(v_{i,j-1}, v_{i-1,j}, v_{i,j+1}, v_{i+1,j}) \rfloor$ . Where,  $nnpredict(.)$  is the prediction model based on neural network.

### 2.2 Data Embedding and Extraction

The combination of difference expansion and histogram shifting method [2] is also utilized in this paper.

If the prediction error  $e_{i,j} = u_{i,j} - u'_{i,j}$  inside the region  $[T_n, T_p]$ ,  $e_{i,j}$  is expanded to  $E_{i,j} = 2 \times e_{i,j} + b$ .  $T_n$  is the negative threshold and  $T_p$  is the positive threshold. Otherwise, the pixel does not carry any data and the prediction error is simply shifted. That is,

$$E_{i,j} = \begin{cases} 2 \times e_{i,j} + b & \text{if } e_{i,j} \in [T_n, T_p] \\ e_{i,j} + T_p + 1 & \text{if } e_{i,j} > T_p \text{ and } T_p \geq 0 \\ e_{i,j} + T_n & \text{if } e_{i,j} < T_n \text{ and } T_n < 0 \end{cases} \quad (1)$$

The watermarked value is computed by  $U_{i,j} = u'_{i,j} + E_{i,j}$ .

During extraction, if  $E_{i,j} \in [2T_n, 2T_p + 1]$ , the watermark can be extracted. Otherwise, shifting is used to recover the image. That is,

$$e_{i,j} = \begin{cases} \lfloor E_{i,j}/2 \rfloor & \text{if } E_{i,j} \in [2T_n, 2T_p + 1] \\ E_{i,j} - T_p - 1 & \text{if } E_{i,j} > 2T_p + 1 \\ E_{i,j} - T_n & \text{if } E_{i,j} < 2T_n \end{cases} \quad (2)$$

Then,  $u_{i,j} = u'_{i,j} + e_{i,j}$ .

### 2.3 Improved Classification Using Parity Property

To ensure  $U_{i,j}$  without overflow or underflow problems, two-pass-testing [6] is used here. If a pixel can be modified twice based on Eq. (1), it belongs to Class A; if the pixel is modifiable once owing to overflow or underflow errors during the second embedding test, it belongs to Class B; and if the pixel cannot be modified even once, the pixel belongs to Class C. During the testing process, bit “1” is used as an embedding bit for positive prediction errors, and bit “0” is for negative prediction errors. The locations of Class B and Class C are marked in a location map, which is also embedded with the payload.

In the decode phase, use once-embedding-test to distinguish Class A, and Class B (or Class C). And further discriminate Class B and Class C using the location map. However, some pixel cells belonging to Class B will be misclassified to Class A if the actually embedded bit does not coincide with the testing bit.

We utilize the parity characteristic and retesting strategy to activate the capacity of Class B. After once-embedding-test during the extraction phase, the pixel cells are assigned into two parts: Part one contains the cells without overflow or underflow, and Part two contains the overflow or underflow cells. As a result, Part one is the set consisting of Class A and partial Class B, while Part two is the set containing Class C and part of Class B. It is obvious that the elements of Class B which are attributed in Part one are problem pixel cells. Since they will cause the wrong localization in the location map, these problem pixel cells should be identified further.

For the cells in Part one, a retesting detection is designed to distinguish the ambiguous cells belonging to Class B. As for the expandable pixel cells,  $U_{i,j} = u'_{i,j} + e_{i,j} = u'_{i,j} + 2e_{i,j} + b$ . Thus,  $U_{i,j} - u'_{i,j} = 2e_{i,j} + b$ . Due to  $2e_{i,j}$  is an even number,  $b = \text{LSB}(U_{i,j} - u'_{i,j})$ , here  $\text{LSB}(x)$  means the LSB of  $x$ . For the positive prediction errors, if  $U_{i,j} - u'_{i,j}$  is an even number, the embedded bit is not consistent with the testing bit. Thus, add one to the pixel value  $U_{i,j}$  and retest the corresponding prediction error using the testing bit “1”. For the negative prediction errors, if  $U_{i,j} - u'_{i,j}$  is odd, subtract one from the pixel value  $U_{i,j}$  and retest the corresponding prediction error using the testing bit “0”. If the retesting result shows the pixel value is overflow, it belongs to Part two. Otherwise, it still belongs

to Part one. After the retesting, Part one only contains Class A, and Part two contains Class B and Class C. Further classification is conducted to distinguish Class B and Class C with the help of the location map.

### 3 Encoder and Decoder

#### 3.1 Data Embedding

We first embed data in “Cross” set, then embed in “Dot” set. For recovering data, threshold values  $T_n$  (7 bits),  $T_p$  (7 bits), payload size  $|P_{cross}|$  (17 bits) or payload size  $|P_{dot}|$  (17 bits), and the length of location map (7 bits) should be known first. We will embed these 38 bits into the first 38 pixels’ LSB. The original 38 LSB should be recorded with the payload. The “Cross” embedding method is designed as follows:

- Step 1: Calculate the prediction errors. For each pixel  $u_{i,j}$ , compute the prediction value and the corresponding prediction error  $e_{i,j}$  based on the common neural network.
- Step 2: Sort the prediction errors. For each pixel  $u_{i,j}$ , compute the variance  $Var_{i,j}$  of the four neighbor pixels which is used as the sorting parameter. Skip the first 38 pixels. Sort the pixel cells according to the ascending order  $\{Var_{i,j}\}$  to produce a sorted row of prediction errors  $e_{sort}$ .
- Step 3: Determine the threshold. According to the two-pass-testing, all pixels are classified in one of classes A, B and C. Although the shiftable pixels can be modified, they cannot carry watermark bits. Only the expandable pixels in Class A and Class B are capable of carrying data. Let set of expandable pixels in class A be  $EA$ . Let set of expandable pixels in class B be  $EB$ .

In the sorted vector  $e_{sort}$ , create the location map  $L$ . If a pixel belongs to Class B, the corresponding element in the location map is marked as “0”; while if the pixel belongs to Class C, it is marked as “1”. If  $|P_{cross}| \leq |EA| + |EB| - |L| - 38$  and  $|EA| \geq |L|$  are satisfied, the to-be-embedded bits can be successfully embedded. Otherwise, increase the threshold  $T_p$  or decrease  $T_n$ , and repeat Step 3.

- Step 4: Embed data. The location map  $L$ , the true payload  $P_{cross}$ , and the first 38 LSBs will be embedded in the image by using the embedding method described in Sect. 2.2. The location map  $L$  is first embedded in Class A. The elements belonging to Class A and Class B are all used to improve the capacity. Use the auxiliary data to modify the first 38 LSB values of the pixels by simple binary replacement. If the last to-be-embedded bit is processed, the “Cross” embedding phase is completed.

After 4 steps, the “Cross” embedding process is finished. The “Dot” embedding scheme uses the modified pixels from the “Cross” set for computing the

predicted values. The original pixels from the “Dot” set are used for embedding data, and the embedding procedure is similar to the “Cross” embedding. After the “Dot” embedding, the watermarked image is obtained.

### 3.2 Data Extraction

Double decoding scheme is the inverse of the double encoding scheme. We only describe the “Cross” decoding method.

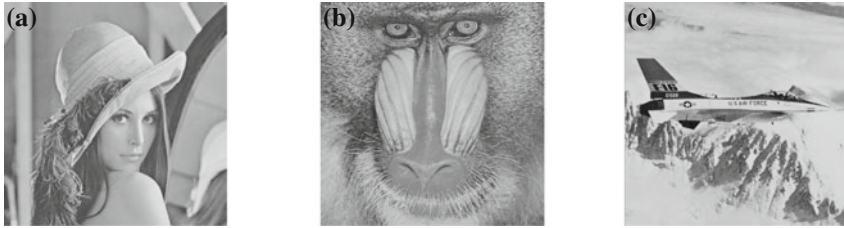
- Step 1: Calculate the prediction values. For each pixel  $U_{i,j}$ , compute the prediction value based on the common neural network. Then, the prediction errors  $E_{i,j}$  are obtained afterwards.
- Step 2: Sort the prediction errors. Skip the first 38 pixels. Sort the pixels according to  $Var_{i,j}$  to get the set of sorted prediction errors  $E_{sort}$ . Read the first 38 LSB values to recover the values of  $T_n$ ,  $T_p$ , payload size  $P_{cross}$ , and the length of location map.
- Step 3: Extract the watermark. Skip the first 38 sorted cells. Test every pixel cell to classify it into Class A, Class B and Class C according to Sect. 2.3. Extract location map from Class A firstly. Further classification is conducted to distinguish Class B and Class C based on the location map. Then, extract data from Class A and Class B, meanwhile recover the original prediction errors using the method in Sect. 2.2. The extracted data is the cascading of the true payload, and the 38 LSBs.
- Step 4: Restore the original image. Computer the original pixel values based on  $u_{i,j} = u'_{i,j} + e_{i,j}$ .
- Step 5: Recover the rest pixels. Replace the first 38 LSB values of the pixels with the extracted 38 LSBs.

When the “Dot” and “Cross” decoding are both finished, the entire watermark is obtained and the original image is restored.

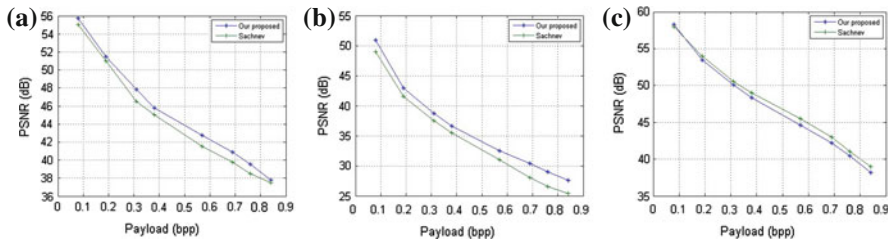
## 4 Experimental Results

Several 8-bit gray images “Lena”, “Baboon” and “Plane” with size  $512 \times 512$  are used in the experiments. Figure 3 shows the watermarked images with payload 50000 bits:

Figure 4 shows the performances of Capacity versus Visual quality in terms of payload and Peak Signal-to-Noise Ratio (PSNR). The horizontal axis represents the capacity in terms of bpp (bits per pixel). The vertical axis represents the corresponding PSNR. The results show that our method has both high visual



**Fig. 3** The watermarked *gray* images. **a** Lena (PSNR = 51.4 dB). **b** Baboon (PSNR = 43.0 dB). **c** Plane (PSNR = 53.5 dB)



**Fig. 4** Capacity versus PSNR for testing images. **a** Results for Lena. **b** Results for Baboon. **c** Results for Plane

quality and high capacity. Compared with [6], our method can achieve better results for “Lena” and “Baboon”, and get comparable result for “Plane”. The reason is that the neural network is not trained enough.

## 5 Conclusions

A high capacity image reversible watermarking based on neural network and parity property is proposed. A retesting strategy utilizing the parity detection activates the capacity of the ambiguous pixel cells. In addition, the prediction errors are obtained by using the neural network to consider the global feature. The experimental results show that this algorithm can obtain higher capacity and preserve good visual quality. In the future, we will further research and discuss the effectiveness of the neural network.

**Acknowledgment** This work was supported in part by 973 Program (2011CB302204), National Natural Science Funds for Distinguished Young Scholar (61025013), National NSF of China (61073159, 61272355), Sino-Singapore JRP (2010DFA11010), Fundamental Research Funds for the Central Universities (2012JBM042).



## References

1. Tian J (2003) Reversible data embedding using a difference expansion. *IEEE Trans Circuits Syst Video Technol* 8:890–896
2. Thodi DM, Rodriguez JJ (2007) Expansion embedding techniques for reversible watermarking. *IEEE Trans Image Process* 3:721–730
3. Tsai WL, Yeh CM, Chang CC (2009) Reversible data hiding based on histogram modification of pixel differences. *IEEE Trans Circuits Syst Video Technol* 6:906–910
4. Tsai PY, Hu C, Yeh HL (2009) Reversible image hiding scheme using predictive coding and histogram shifting. *IEEE Signal Process Mag* 6:1129–1143
5. Luo LZ, Chen N, Zeng X, Xiong Z (2010) Reversible image watermarking using interpolation technique. *IEEE Trans Inf Forensics Secur* 1:187–193
6. Sachnev V, Kim HJ, Nam J, Shi YQ, Suresh S (2009) Reversible watermarking algorithm using sorting and prediction. *IEEE Trans Circuits Syst Video Technol* 7:989–999

# A Based on Single Image Authentication System in Aviation Security

Deok Gyu Lee and Jong Wook Han

**Abstract** An image protection apparatus includes an information collecting unit for collecting personally identifiable information to be embedded in images captured by an image capturing instrument; and an information processing unit for extracting personal information from the collected personally identifiable information. Further, the image protection apparatus includes an information embedding unit for embedding the extracted personal information into a captured image; and an image signature unit for writing a signature on the captured image by using the extracted personal information.

**Keywords** Aviation Security · Surveillance System · Authentication · Authorization

## 1 Introduction

As the national airspace system grows increasingly interconnected to partners and customers both within and outside the Rep. of Korea government, the danger of cyber-attacks on the system is increasing. Because of low-cost computer technology and easier access to malware, or malicious software code, it is conceivable

---

This research was supported by a grant (code# 07aviation-navigation-03) from Aviation Improvement Program funded by Ministry of Construction & Transportation of Korean government.

---

D. G. Lee (✉) · J. W. Han  
Electronic and Telecommunications Research Institute, 161 Gajeong-dong, Yuseong-gu,  
Daejeon, Republic. of Korea  
e-mail: deokgyulee@etri.re.kr

J. W. Han  
e-mail: hanjw@etri.re.kr

for individuals, organized crime groups, terrorists, and nation-states to attack the Rep. of Korea air transportation system infrastructure.

Consider an airport in which passengers and employees can enter common areas, like transportation facilities, and waiting areas. However, secured areas, like luggage transport and service stations, are available for authorized employees only. The highest security areas, such as the air traffic control room, are accessible to specialized personnel who are appropriately authorized. The keyword here is “authorization”, meaning that people who are not authorized to access a physical location should not be allowed physical or electronic access to that location. In the surveillance world, the exact same rules apply and the potential recipient of the surveillance data must have the same authorization that an ordinary person of any trade would have to be physically or electronically present at that location. However, during emergency operations, controlled dissemination of sensitive data may become necessary in order to obtain support services or to prevent panic. It has been shown that during crisis people require clear instructions so that their maximum cooperation is obtained. However, these instructions should not release unauthorized information or reveal the existence of such information.

This paper relates to an apparatus and a method for processing image information, and more particularly, to an image information processing apparatus and method capable of adding information on an image capturing device and signature information to image data and storing the image data to maintain security of the image data and use the image data as digital proof.

## 2 Related Work

With the development of image photographing technology, techniques for maintaining security of image data captured by an image capturing device and protecting copyright are proposed. For example, captured images are transmitted to a limited image information output device and reproduced or identification information such as watermarking is embedded in image data to protect copyright of image information [1].

In the case of embedding watermarking in image information, it is possible to confirm the copyright holder of the image information even though the image information is displayed at or transmitted to an undesired place and prevent the image information from being illegally copied. Furthermore, users can watch the image information without having any difficulty and track the source of the image information and image information copy routes when watermarking is embedded in the image information.

However, watermarking does not have legal force capable of preventing the image information from being illegally copied or transmitted although it can show the copyright holder or the source of the image information and allow users to confirm image information copy routes and the source of the image information. Accordingly, security of image information cannot be efficiently maintained only

with watermarking when the image information includes personal information related to privacy or data requiring the maintenance of security [2. 3].

A distributed architecture for multi-participant and interactive multimedia that enables multiple users to share media streams within a networked environment is presented in “An architecture for distributed, interactive, multi-stream, multi-participant audio and video”. In this architecture, multimedia streams originating from multiple sources can be combined to provide media clips that accommodate look-around capabilities. SMIL has been the focus of active research “The use of smil: Multimedia research currently applied on a global scale” and “About the semantic verification of SMIL documents”, and many models for adaption to real world scenarios have been provided. A release control for SMIL formatted multimedia objects for pay-per-view movies on the Internet that enforces DAC is described in “Regulating access to smil formatted pay-per-view movies”. The cinematic structure consisting of acts, scenes, frames of an actual movies are written as a SMIL document without losing the sense of a story. Here access is restricted to the granularity of an act in a movie. A secure and progressively updatable SMIL document “Sputers: A secure traffic surveillance and emergency response architecture” is used to enforce RBAC and respond to traffic emergencies. In an emergency response situation, different recipients of the live feeds have to be discriminated to people playing different roles [1–7].

While most models addresses the need of multimedia, their approach does not incorporate semantics of multimedia. None of the approaches are completely satisfactory for surveillance multimedia. They primarily address textual documents and exploit the granular structure of XML documents. Multimedia for various reasons as discussed above has to be treated differently because there is a sense of temporal synchrony and continuity involved. Synchronization and integration of different and diverse events to produce sensible information is nontrivial when compared to textual data. The process of retrieval without losing the sense of continuity and synchronization needs sophisticated techniques and algorithms which all of the above models do not completely address. Although our approach to provide controlled information flow in real-time multimedia systems is based in concepts similar to MLS, the developed methods and techniques are also applicable in other security models, like Role-Based or Discretionary Access Control models.

### **3 ACRS (Aviation Control Room Surveillance)**

It is an object of the paper to provide an image information processing apparatus and method for adding information on an image capturing device and predetermined signature information to image data obtained using the image capturing device to protect the image data from infringement of security such as illegal copy and transmission and adding information on the place and time at which the image data is obtained to the image data to use the image data as digital proof.

An apparatus for processing image information according to the paper comprises: an image capturing unit for generating image data and collecting information on the image capturing unit; an image processing unit for adding at least one of the information on the image capturing unit and signature information to the image data using the image data and the information on the image capturing unit transmitted from the image capturing unit; and an image storage unit for storing the image data output from the image processing unit.

A method for processing image information according to the paper comprises: an image capturing step of generating image data and collecting information on the image capturing step; an image processing step of adding at least one of the information on the image capturing step and signature information to the image data; and an image storing step of storing the image data.

### ***3.1 Security Framework for Physical Environment***

The Security framework for physical environment contains a few essential components, such as an authentication, an authorization, and a security policy. They work at each smart door and often cooperate with a smart surveillance established by a smart image-unit in the physical environment. Since the smart door is installed at the border of each physical domain and every physical environment must pass through it, it is supposed to be a core component and suitable in providing security functions described in security framework for physical environment. Whenever a new access to physical environment is found, it should be able to authenticate and authorize it and enforce security policy based on security rules set by the corresponding smart security administrator [8].

Figure 1 depicts the overall architecture of secure physical environment.

Figure 1 is a view illustrating a configuration of an image photographing system to which an image information processing apparatus according to the paper is applied.

In view of the above, the present invention provides an image protection apparatus that embeds personal information and signature information in an image to thereby protect the image from others. In accordance with an embodiment of the present invention, there is provided an image protection apparatus including: an information collecting unit for collecting personally identifiable information to be embedded in images captured by an image capturing instrument; an information processing unit for extracting personal information from the collected personally identifiable information; an information embedding unit for embedding the extracted personal information into a captured image; and an image signature unit for writing a signature on the captured image by using the extracted personal information. It is preferable that the information processing unit extracts device information of the image capturing instrument, and the information embedding unit embeds the device information into the captured image. Further, it is preferable that the information embedding unit verifies validity of the extracted

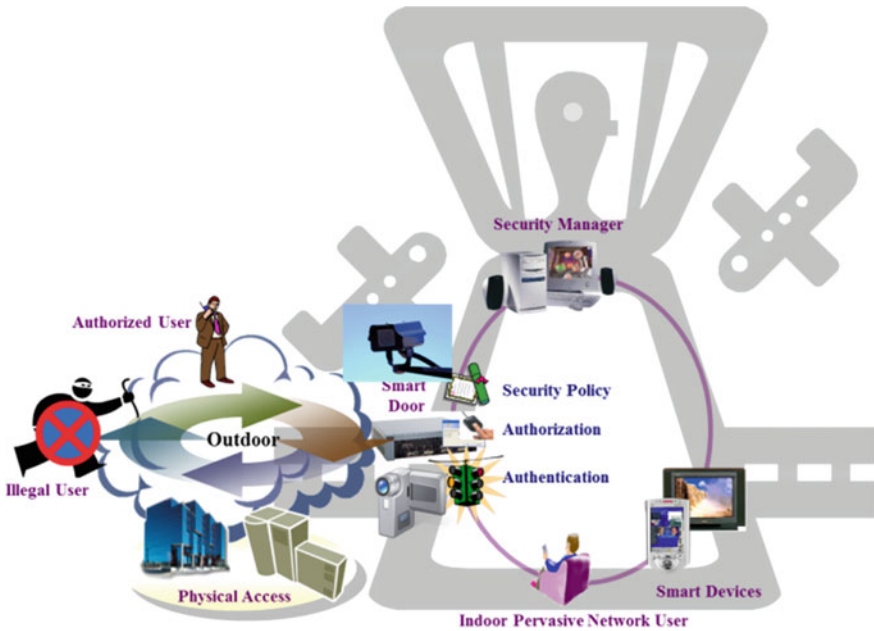


Fig. 1 Architecture of secure physical environment

personal information in cooperation with an external personal information server through a network, and embeds the verified personal information in the captured image. It is preferable that the information collecting unit collects personally identifiable information by using one or more method of image recognition barcode, personal identification tag, radio frequency identification (RFID), sensor, face recognition, and iris recognition. It is preferable that the image signature unit writes signatures on frames in the captured image by using the personal information. Further, it is preferable that the image signature unit writes signatures on each or parts of the units of preset number of frames in the captured images. Furthermore, it is preferable that the information collecting unit and information processing unit are installed in the image capturing instrument, and the information embedding unit and image signature unit are installed in a server connected through a network to the image capturing instrument. In a feature of the present invention, because both personal information and signature information are embedded in images, a user cannot identify an image of another user without validating relevant personal information and additionally access control is enforced on the administrator administrating the images. When a particular case occurs, personal information embedded in an associated image can be used as digital evidence.

Further, because signature information are embedded into images in the form of various cryptographic signatures including public key certificates, particular types of facts such as place and time can be verified using the signature information.

Therefore, illegal acts related to the images can be prevented before the fact, and legal measures can be taken after the fact as the image source and illegal act associated with the images can be identified. Hereinafter, an embodiment of the present invention will be described below with reference to the accompanying drawings. The image capturing device and the image information server illustrated in Fig. 1 according to an embodiment of the present invention, which shows a case in which the image processor adds the information on the image capturing unit or the signature information to the image data. The image photographing device includes the image capturing unit and the image processor. The image processor includes an information receiver, a device information processor, a signature information processor, and an image transmitter. The image information server includes an image receiver and a storage unit. The image capturing unit is a device capable of capturing an image, obtaining image data from the captured image and collecting information on the image capturing unit. For example, a CCTV, a digital camera, a video camera or a communication terminal including a camera module can be used as the image capturing unit. The image data obtained by the image capturing unit is transmitted to the image processor and undergoes a data processing operation of adding signature information or the information on the image capturing unit thereto. The image processor adds the information on the image capturing unit or the signature information to the image data generated by the image capturing unit. The information receiver included in the image processor receives the image data generated by the image capturing unit and the information on the image capturing unit and the device information processor adds the information on the image capturing unit to the image data transmitted from the information receiver. The signature information processor adds the signature information to the image data transmitted from the device information processor and the image transmitter transmits the image data having the information on the image capturing unit or the signature information added thereto to the image information server. Here, positions of the device information processor and the signature information processor can be changed each other. That is, the signature information can be added first, and then the information on the image capturing unit can be added.

The information receiver receives the image data generated by the image capturing unit and transmits the information to the device information processor. The information on the image capturing unit can include an identifier given to the image capturing unit or information on the place and time at which the image capturing unit obtains the image data. When the image capturing unit is a CCTV, for example, the information on the image capturing unit can include the place where the CCTV is installed and the time when the CCTV records the image. When the image capturing unit is a communication terminal, the information on the image capturing unit can include an identification number given to the communication terminal. The device information processor adds the information on the image capturing unit to the image data transmitted from the information receiver. That is, the place and time at which the image data is obtained or the identification number of the image capturing unit that captures the image data can

be added to the image data as the information on the image capturing unit. The image data can be used as digital proof of a specific event when the place and time at which the image data is captured is added thereto and the source of the image data can be easily detected when the identification number of the image capturing unit is added thereto. The device information processor transmits the image data having the information on the image capturing unit added thereto to the signature information processor. The signature information processor can embed signature information including a predetermined encryption key in the image data transmitted from the device information processor. According to an embodiment, the signature information processor can add public key based signature information, symmetric key based signature information or public key and symmetric key based signature information to the image data.

The public key based signature information can be generated according to Rivest Shamir Adleman (RSA) algorithm and the symmetric key based signature information can be generated according to Vernam or data encryption standard (DES) algorithm. The symmetric key based signature information requires transmission of an additional secret key and is difficult to authenticate with safety although it is encrypted at a high speed. On the other hand, the public key based signature information does not require transmission of the additional secret key and is easily authenticated with safety while it is encrypted at a low speed. Accordingly, an algorithm of generating the signature information can be selected according to a degree to which maintenance of security of the image data is required. When the signature information is added to the image data, the image data can be accessed only using a predetermined decryption key. Accordingly, the possibility that the image data is exposed to hacking or illegal copy according to arbitrary access can be reduced when the signature information is added to the image data. The application server that provides the image data to communication subscribers can provide the decryption key to only an authenticated communication subscriber through a text message to maintain security of the image data. The signature information can be added to the image data at regular intervals.

The image data having the information on the image capturing unit and the signature information added thereto is transmitted to the image transmitter. That is, at least one of the information on the image capturing unit and the signature information is added to the image data transmitted from the image capturing unit and transmitted to the image transmitter. The image transmitter transmits the image data received from the signature information processor to the image information server.

The image receiver receives the image data from the image transmitter. The storage unit stores the image data transmitted from the image receiver. The image data transmitted from the image receiver has at least one of the information on the image capturing unit and the signature information added thereto. The image information server can determine whether the decryption key transmitted from the application server corresponds to the encryption key embedded in the image data, extract the image data stored in the storage unit and transmit the image data to the



application server when the application server requests the image information server to transmit the image data through the communication network.

The image capturing device and the image information server illustrated in Figure 1 according to another embodiment of the present invention, which shows a case in which the image information server adds the information on the image capturing unit or the signature information to the image data. The image capturing device includes the image capturing unit and the image processor and the image information server includes a receiver, a device information processor, a signal information processor and a storage unit. The image capturing unit is a device capable of recognizing an object through a lens and a sensor, obtaining image data from the recognized object and collecting information on the image capturing unit. The image processor transmits the image data and the information on the image capturing unit received from the image capturing unit to the image information server. The receiver included in the image information server receives the image data and the information on the image capturing unit transmitted from the image processor and the device information processor receives the image data and the information on the image capturing unit from the receiver and embeds the information on the image capturing unit in the image data. The signature information processor adds predetermined signature information to the image data having the information on the image capturing unit added thereto and the storage unit stores the image data including the signature information. The receiver receives the image data and the information on the image capturing unit from the image processor.

The device information processor embeds the information on the image capturing unit in the image data and transmits the image data to the signature information processor. The information on the image capturing unit depends on the type of the image capturing unit. For example, when the image capturing unit is a CCTV, the information on the image capturing unit can include the place where the CCTV is installed and the time when the CCTV obtains the image data. When the image capturing unit is a communication terminal including a camera module, the information on the image capturing unit can include the identification number of the communication terminal.

The source of the image data can be easily searched when the identification number of the image capturing unit is embedded in the image data and the image data can be used as digital proof when the place and time at which the image capturing unit captures the image data is added thereto. The information on the image capturing unit can be recorded in a meta-data region of the image data. The signature information processor can embed signature information including a predetermined encryption key in the image data transmitted from the device information processor. The signature information processor can generate the signature information according to a predetermined algorithm and embed the signature information in the image data. The signature information can be generated according to a public key based algorithm or a symmetric key based algorithm. In general, the case that the image information server embeds the information on the image capturing unit and the signature information in the image

data requires a data processing speed and available capacity greater than the data processing speed and available capacity required for the case that the image photographing device embeds the information on the image capturing unit and the signature information in the image data. Accordingly, it is desirable to generate the signature information using the public key based algorithm that easily performs safe authentication and does not require an addition secret key to be transmitted while having a low encryption speed. The image data to which the signature information has been added is stored in the storage unit of the image information server.

The signature information can be added to the image data at regular intervals. The storage unit stores the image data transmitted from the signature information processor. The image data transmitted from the signature information processor has at least one of the information on the image capturing unit and the signature information added thereto. The image information server can receive a predetermined decryption key from the application server and compare the decryption key with the encryption key included in the image data to determine whether the image data is transmitted when the application server requests the image information server to transmit the image data through the communication network.

## 4 Conclusion

We have presented a surveillance framework for audio–video surveillance of multi-level secured facilities during normal and pre-envisioned emergencies. This paper relates to an apparatus and a method for processing image information, and more particularly, to an image information processing apparatus and method capable of adding information on an image capturing device and signature information to image data and storing the image data to maintain security of the image data and use the image data as digital proof. However, it is also important to address data integrity and source authentication issues. These issues, along with the development of a complete and comprehensive prototype system are part of our future work.

## References

1. Kodali N, Farkas C, Wijesekera D (2003) Multimedia access control using rdf metadata. In: workshop on metadata for security, WMS 03
2. Kodali N, Wijesekera D (2002) Regulating access to smil formatted pay-per-view movies. In: 2002 ACM workshop on XML security
3. Rutledge L, Hardman L, Ossenbruggen J (1999) The use of smil: multimedia research currently applied on a global scale
4. Kodali N, Wijesekera D, Michael. Sputers: A secure traffic surveillance and emergency response architecture. *J Intell Transp Syst*

5. Pihkala K, Cesar P, Vuorimaa P (2002) Cross platform smil player. In: International conference on communications, internet and information technology
6. Rutledge L, Ossenbruggen J, Hardman L, Dick CA (1999) Bulterman. Anticipating SMIL 2.0: the developing cooperative infrastructure for multimedia on the Web. *Computer Networks*, Amsterdam, Netherlands, 31(11–16):1421–1430
7. Schmidt BK (1999) An architecture for distributed, interactive, multi-stream, multi-participant audio and video. In: Technical report no CSL-TR-99-781, stanford computer science department
8. Kodali N, Wijesekera D, Farkas C (2004) SECRETS: a secure real-time multimedia surveillance system. In: Proceedings of the 2nd Symposium on intelligence and security informatics
9. Damiani E, di Vimercati SDC (2003) Securing xml based multimedia content. In: 18th IFIP international information security conference
10. Damiani E, di Vimercati SDC, Paraboschi S, Samarati P (2000) Securing XML documents. *Lect Notes Compt Sci* 1777:121–122
11. FAA’S NEXTGEN AIR TRAFFIC CONTROL SYSTEM A CIO’s Perspective on Technology and Security Georgetown University Institute for Law, Science, and Global Security & Billington CyberSecurity, 28 Feb 2011
12. Damiani E, di Vimercati SDC, Paraboschi S, Samarati P (2002) A fine grained access control system for xml documents. *ACM Trans Info Syst Security* 5:121–135
13. Gu X, Nahrstedt K, Yuan W, Wichadakul D, Xu D (2001) An xml-based quality of service enabling language for the web. Kluwer Academic Publishers, Norwell
14. Kodali N, Farkas C, Wijesekera D (2003) Enforcing integrity in multimedia surveillance. In: IFIP 11.5 working conference on integrity and internal control in information systems

**Part VI**  
**Multimedia and Ubiquitous Services**

# A Development of Android Based Debate-Learning System for Cultivating Divergent Thinking

SungWan Kim, EunGil Kim and JongHoon Kim

**Abstract** Six Thinking Hats which is designed by Edward de Bono enhances the excellence of the thinking and has an effect on cultivating divergent thinking. In particular, it is effective in seeking a reasonable solution by analyzing some problems from a variety of views in debate-learning. In this paper, we developed a system sharing voice and images based on Six Thinking Hats, using sensors of android device. We analyzed tools and guidelines by making design structural model for designing the system. We developed debate-learning system, verified its utility and analyzed improvements through a demonstration and a practice for educational experts.

**Keywords** Android · Divergent thinking · Six thinking hats · Debate-learning system · Mobile learning

## 1 Introduction

Thinking also can be improved through practicing and many methods were studied as a learning method. There are brainstorming, Six Thinking Hats, Attribute Listing, Morphological Synectics, Forced relation, Synectics in typical creative methods [1]. It has been studied that Six Thinking Hats makes people practice each field of thinking and a change in attitude by looking problems in a different ways and improve different abilities of thinking.

However, it does not guarantee time for activity of thinking which is a prerequisite of creative thinking method due to lots of contents of curriculum compared to time

---

S. Kim · E. Kim · J. Kim (✉)

Department of Computer Education, Teachers College, Jeju National University,  
Jeju-si, Korea

e-mail: jkim0858@jejunu.ac.kr

S. Kim

e-mail: kswandrea@naver.com

E. Kim

e-mail: computing@korea.kr

for each studying. In addition, it is difficult to study because there is a shortage of some examples of thinking methods and programs for students studying the method.

Thus, in this paper, we developed a debate-learning system based on android among smart devices which are embedded portable and different sensors. Using our developed system, students' thinking progress with sync function of phonic and image data and provide them other learners through the sharing server so that it is possible to share and evaluate the opinion. We analyzed the result by conducting expert evaluation to verify its benefits and find some future improving way.

## 2 Six Thinking Hats

Edward de Bono argued that thinking is a function to manipulate the intelligence and it can be improved through the process of practicing [2]. Six Thinking Hats, designed by him, is a method to intend people to think one thing at a time. If you decouple emotion and logic, and information and creativity, you can come up with more ideas. You can use some, not six at once according to the six thinking colors. You need to be aware of the rules that if you wear a certain color's hat, you have to think in a way of applying to the color. Mental activities by each hat's color are same as Table 1 as follows.

The purpose of this method is to simplify the thinking by dealing with each field at a time, and guide a change in attitudes by changing the color of hat. Simplification of thinking lessens difficulties of thinking from a variety of fields at the same time and improves the ability of having different perspectives and positions. A change in attitude also requires people to think in a broad ways including a negative way and a creative way.

People should take a neutral attitude using objective information to illuminate the problem and its background which is white hat supposed to discuss. Red hat provides an opportunity to express people's emotions and feelings. Emotions are supposed to be excluded in a process of solving the logic problems generally; however, it is very hard to obstruct the involvement of emotion. Therefore, we can expect some creative discoveries such as insight by suggesting emotions and feelings as the whole truth without the logical reasons. Black hat is based on critical thinking and it needs legitimate grounds. Critical thinking is one of important thinking abilities to prevent illegal or expected damages in advance and understand weakness of some countermeasures. Yellow hat serves as a role to find

**Table 1** Mental activities by hat's colors

Types of hat	Mental activities
White hat	Facts and objective information
Red hat	Emotions and feelings, intuition and sixth sense
Black hat	Negative judgement, impossible reason
Yellow hat	Positive judgement, constructive decision and opinion
Green hat	New and creative ideas
Blue hat	Organization of all ideas

positive value and it requires people to be sensitive to an object’s value all the time. Optimistic value can be distinguished from mental activity of red hat which presents simple emotions and feelings just founded on facts and truth. Green hat suggests new ideas. It makes an entirely new up-to-date idea or modifies or improves established ideas. Blue hat is a manager to comment other people’s thinking and synthesize a final conclusion.

### 3 Design of Debate-Learning System

We present the design structural model of our debate-learning system using Six Thinking Hats with help of procedures of web debate-learning from Korea Education & Research Information Service [3] as follows as Table 2.

In the stage of pre-learning, students study characteristic and roles of Six Thinking Hats and look into concrete examples from examples. In the stage of discussion, it is necessary to choose the topic of debate and students collect opinion and grounded data according to the color of hat from online and offline. They think steadily presenting their data with spoken languages. They also have peer review about their opinion and this would make the atmosphere of discussion more active through choosing a great participant and rewarding in the stage of after-learning.

### 4 A Development of Debate-Learning System

The map of overall debate-learning system developed by suggested design structural model is as in the following Fig. 1.

It makes log-in function with procedures of user confirmation to distinguish one’s opinion from others in the discussion. People would become a member of teacher and learner largely and teacher obtains certification to acquire the authority to register the topic of discussion. Learner can join without difficulty and use the system.

**Table 2** Design structural model of debate-learning system

Stage	Details	Tool and guide line
Pre-learning	Six Thinking Hats	Guiding characteristics and roles of Six Thinking Hats looking into examples for each theme
Discussion activities	Choose the topic of discussion	Guiding themes in diverse criteria choosing a topic
Discussion activities	Wear the hat and collect the grounded data	Choosing one type of thinking hats individually (log-in function) collecting data for an argument (a camera, WebView)
Discussion activities	Discussion activities and peer review	Presenting opinion and its data with voice and images (voice and images sync) listening to other people’s opinion and evaluation (a developed player and a separate point)
After-learning	Offer feedback	Selecting a good participant with peer review vitalizing atmosphere of a debate

Before discussion, people have to learn about the Six Thinking Hats. They examine the multi-media data about thinking method and study the opinion according to the hat's color from examples using the player developed from this system.

When you finish learning of the method, you participate in the discussion with the procedures of choosing the topic and the hat. You collect grounded data for your opinion from Internet, a embedded camera on smart device and so on. Learner also evaluates other learner's opinion besides debating.

### 4.1 Membership and User Authentication

Our developed application starts like Fig. 2.

When you start the android application, you can check the condition of communication through Thread in Main Activity and process the update of topics for discussion. Topics for discussion are defined as a language of XML and it updates when teacher register a topic or modify the topic. If there are communication problems, the application is closed after noticing the user.

Information of log-in and membership is written as a form of a request message in the mode of POST and transmitted to the server. User authentication in the process of log-in maintains and Session starts when it corresponds with the information from DB information.

### 4.2 A Guide to Six Thinking Hats

It is so essential to understand about Six Thinking Hats before a debate that we organize specific characteristics and roles of Thinking Hats to make people learn them from image data and examples. We present guided-materials using ScrollView to overcome the small screen of smart devices and examples with images and voice using our own player.

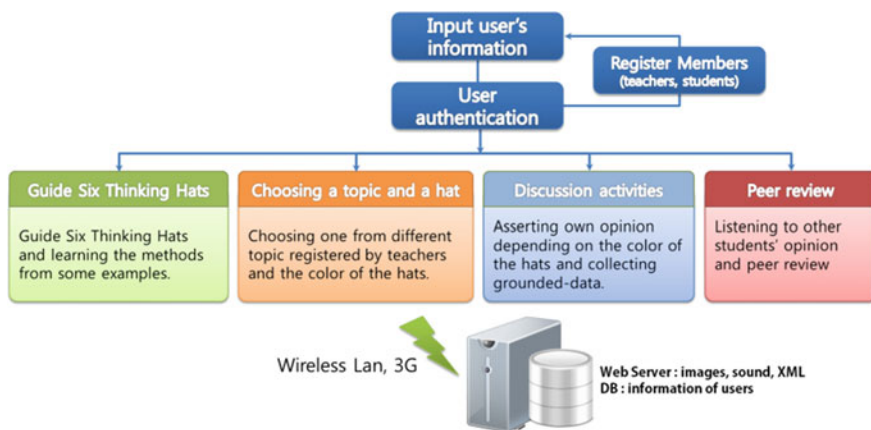
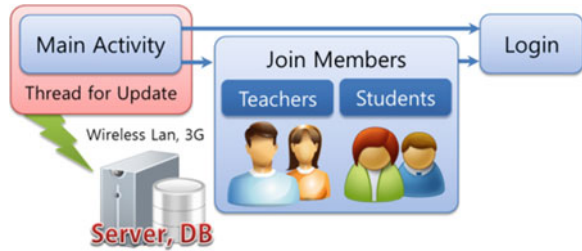


Fig. 1 Overall debate-learning system



**Fig. 2** Topic update and user authentication



### 4.3 Choice of a Topic for a Debate and One of Thinking Hats

One of topic for a debate and classified opinion are defined as XML language and then they communicate with server. XML files are supposed to feed the relevant information to DB and renew themselves when teacher register a topic or learner record own opinion. A XML file which has been already downloaded need not to be reused because it has been expressed on ListView the first one time parsing. Therefore, we have intended that parsing would be fulfilled by SAX Parser which rarely use its memory, considering features of mobile device [4]. Information from parsing has re-defined to express the attributes of the topic and the opinion more clearly.

After parsing a XML file which all the contents about the topic for a debate is defined, it would be expressed on re-defined ListView. A user choose the topic for a debate on the ListView and download a XML file through Thread, which defines the registered contents of opinion according to the selected topic. Once you download it, Handler sends the message, summon serve activity from Intent of android and print it on the ListView by parsing the downloaded XML file. If the user click the Thinking Hats button arranged at the bottom, registered opinion would be renewed, and you can put your opinion using the registration of opinion button.

### 4.4 Discussion Activities

We record learner’s argument in a voice using the microphone sensor of the smart device depending on Six Thinking Hats. The voice would be compressed in a AMR-NB way which has been developed from 3GPP (3rd Generation Partnership Project), a joint research project [5]. AMR-NB, designed for voice recorder and communication in mobile device, has a compressive force of 4.75–12.2 kbps capacity [6].

We also collect well-grounded data in a image form using WebView and the camera sensor. This improvement item aims for students who do not have good power of word-painting to make their thinking more easily. Collected image data would be printed on GalleryView at the bottom and we can check them moving from side to side with a user’s touch input. Besides, if you choose something in the middle of recording an argument, it would be inserted into the opinion. This point of time that something inserted for Sync modulation between voice and images would be remembered by millisecond and we write it with XML elements.

## 4.5 Peer Review

Android provides verbal or video-typed playback player, however, we have developed our own player which has sync function of voice and images because a debate in our system can be progressed with them. Learner can give five grades after listening to other learner's opinion, judging validity and suitability of Six Thinking Hats. We provide interface which learner can give separate grades with touch input to overcome the limit of input in smart device.

## 5 Results

We have conducted expert evaluation to examine utility of our developed Six Thinking Hats-based debate-learning system and improvements. Evaluators are consisting of twenty teachers who have level-one qualification for a licensed teacher and career for ten years. Expert evaluation has been progressed by responding some questions through a demonstration and practice experience about the application and by checking level-five Likert criterion at intervals of 2.5 grades, choosing or describing opinion. Contents of the survey have been made in four fields like Table 3.

Ninety percentages of teachers have agreed that Six Thinking Hats is effective on debate-learning from the following assessment's results. However, few of them have used this method in the educational practice. After analyzing the reasons why the Six Thinking Hats rarely has been used, we can find out lack of guideline and activity programs for Six Thinking Hats consists fifty percentages and the lack of school hours in school forms thirty-five percentage, like Fig. 3. For such these reasons, we expect our debate-learning system would have high effectiveness because it does not have restrictions from time and space.

The result from expert evaluation about contents and construction of the debate-learning system is as follows, Fig. 4. We can see constructive results generally. In particular, there is a high percentage on the response that expressing learner's thinking into voice and images would be effective.

Viewed from the functional side, sync accuracy of voice and images of our developed player has acquired 8.5 points and the response that peer review in a touch input way would be positive in attracting people to learn also has formed same points (Fig. 5).

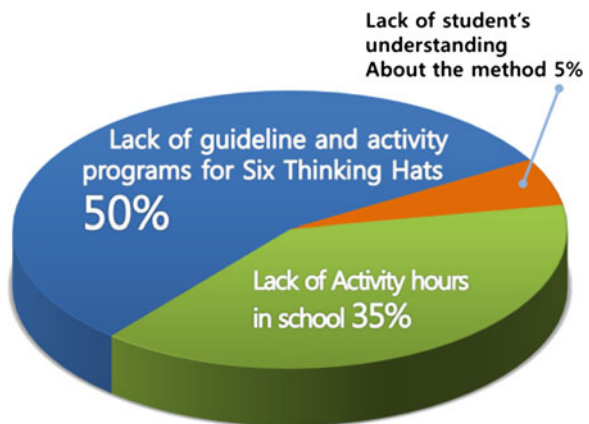
We have looked into benefits and improvements of our developed debate-learning system, comparing with existing web-based system and other system in offline. Above all, the response that expressing learner's thinking into voice and images is more effective than its of established text-centered, web-based debate-learning system acquires positive 7.375 points.

We can obtain some improvements in the narrative that it would be better to provide more materials with responded opinion and to improve a speed-control function of spoken languages. We consider these comments are due to the difference between a small screen and listening ability of speech.

**Table 3** Expert evaluation contents for verification of the system

Field	Contents details
Values and utility	Is Six Thinking Hats effective on improving debate skills? Do you use the thinking method in the school? If you do not, why is the reason?
Contents and construction	Is a guide for Six Thinking Hats suitable for the level of learners? Is the method of participating in a debate appropriate to the level of learners? Is using sync function of voice and images effective for learners to express their opinion? Is overall user's interface construction convenient?
Functions	Is the sync function of the player accurate? Is peer review effective to attract learners to participate in a debate? Are you satisfied with operating time of the system?
Benefits and improvements	Is the new system more effective than text-centered and web-based debate-learning system? What is the merit of this system compared to the web-based debate-learning system and mental activities offline? What is the improvements of developed debate-learning system?

**Fig. 3** Reasons of unused the method in the educational practice



## 6 Conclusion

This research is based on Edward de Bono's Six Thinking Hats and we developed android-based debate-learning system. Six Thinking Hats lessens difficulty of mental activities thinking one thinking field at a time and makes it possible to contemplate problems from different perspectives by experiencing various thinking fields. Therefore, thinking depending on Six Thinking Hats is one of effective learning methods for learners to solve diverse problems differently and reasonably.

We sometimes face into the limit that it is difficult to show thinking in case of learner whose language expression is not good. To solve this problem, it would be better to use multi-media as data to support learner's thinking. It is very convenient to make analog data to the digital and to collect enormous data from Internet.

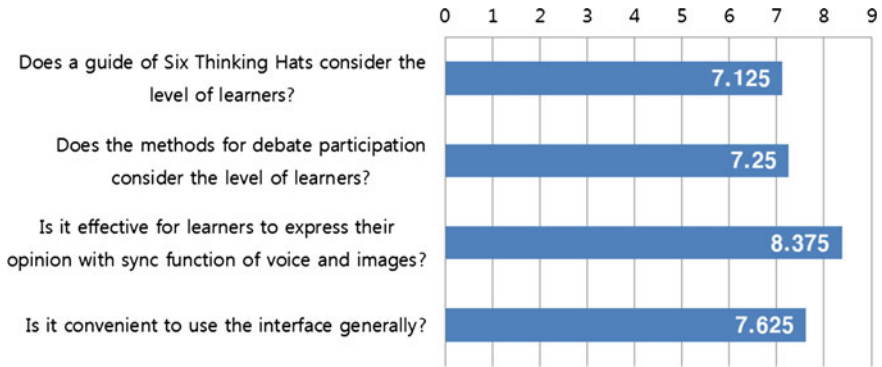


Fig. 4 Results of responding about contents and construction

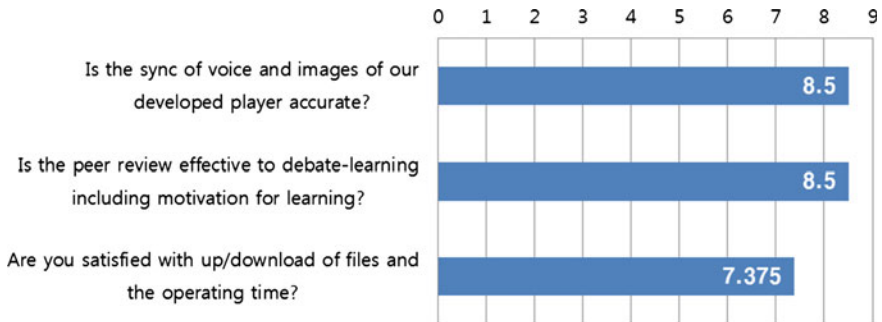


Fig. 5 Results in a functional part

Using smart device with high portability, we solve problems offline because we would gather data without restrictions from time and space and guarantee the time for learners to organize their thinking. As you see in the results of expert evaluation, we could find out that it is very positive to express learner’s thinking into voice and images with sync function in an educational way and, if we consult some future improvements, our system would serve as a debate-learning system with high effectiveness and it would be helpful to improve learner’s mental activities.

## References

1. Pike RW (1994) Creative training techniques handbook: tips, tactics, and how-to’s for delivering effective training. Lakewoods Publications, Minneapolis
2. de Bono E (1985) Six thinking hats. Penguin Book, London
3. Korea Education & Research Information Service (2001) Guide for the use of information & communication technology education. Korea Education & Research Information Service, Seoul
4. SAX. <http://www.saxproject.org>
5. Google. <http://developer.android.com/guide/basics/what-is-android.html>
6. Wikipedia. [http://en.wikipedia.org/wiki/Adaptive\\_Multi-Rate\\_audio\\_codec](http://en.wikipedia.org/wiki/Adaptive_Multi-Rate_audio_codec)

# Development of a Lever Learning Webapp for an HTML5-Based Cross-Platform

TaeHun Kim, ByeongSu Kim and JongHoon Kim

**Abstract** With the advent of smart devices, educational apps for smart learning are actively being developed, but the existing native apps run only on specific devices and are not compatible with other devices. Webapp, a new app development method, which is written using HTML5, supports a cross-platform. In this study, a Webapp for learning about levers as mentioned in elementary school textbooks was developed using HTML5. The proposed Webapp was tested by a group of incumbent expert elementary school teachers, revealing that the proposed contents and Webapp offer high educational value.

**Keywords** HTML5 · Webapp · Cross-platform · Smart learning

## 1 Introduction

Amid the new information technologies that are being developed at alarming speeds, diverse inventions that encompass new information technologies are being developed. Of them, representative inventions are smart devices such as smart-phones and smart pads, the advent and spread of which have enormously changed modern-day people's lifestyles. This is true for education, for which smart devices have provided new paradigms in terms of methods. Smart-learning has the same root as e-learning, m-learning, and u-learning, but uses smart devices, which

---

T. Kim · B. Kim · J. Kim (✉)

Department of Computer Education, Teachers College, Jeju National University, Jeju, Korea

e-mail: jkim0858@jejunu.ac.kr

T. Kim

e-mail: gtranu@naver.com

B. Kim

e-mail: pigpotato79@naver.com

differentiates it from other learning methods and is drawing attention to it as a new education method.

Current operating systems for smart devices are dominated by Apple's IOS and Google's Android. Large numbers of smart-device-driven apps are now hitting the market. IOS and Android are different operating systems, however, which require developers to make apps using these two platforms or give up on one OS-based app for the other. This inter-device incompatibility problem, if not resolved, will adversely affect digital textbook projects, app development, and other smart learning initiatives.

Recently, the web standardization organization W3C proposed the next-generation web-standard advanced HTML5 based on the web standard and Web 2.0. HTML5 pursues web standards and can support cross-platforms. A Cross-platform refers to the capability of a computer program, operating system, computer language, programming language, computer software, etc. to operate on various kinds of computer platforms. HTML5, which is enabling the implementation of the cross-platform feature, is drawing attention as a new method of developing smart device apps.

This study developed the lever learning Webapp that can support cross-platforms to eliminate inter-smart device incompatibility problems. It also investigated methods of implementing smart learning using HTML5.

## **2 Related Researches**

### ***2.1 Smart Learning***

Smart learning is a method of electronic learning that the learner can access the learning content easily using smart device and its related technologies. It is self-initiated learning method that enables user-tailored learning and self-directed learning and the interaction among the learners and between the learners and the teachers [1, 2].

Smart learning is similar to electronic learning that encompasses e-learning (electronic learning), m-learning (mobile learning), and u-learning (ubiquitous learning), and is no different from these methods. Each learning method is able to classify by their characteristic.

E-learning is used to access the learning content using desktop computers and it is distributed in many different forms of educational applications including online courses and web-based learning. It can be usually accessed at fixed locations with internet connections such as computer labs or from homes [3].

M-learning is an advanced stage of e-learning where in the learner is equipped with handheld mobile device to access the learning content using various wireless technologies. M-learning has the benefits of mobility and its supporting platform, which can be summarized as being ubiquity, convenience, localization and

personalization. The major advantage of m-learning is learning can happen anytime and anywhere, so it can support continuous learning [3–5].

U-learning is equivalent to some form of simple m-learning. But it is context aware and also provides anywhere, anytime learning using various mobile and sensor technologies. Besides the domains of e-learning or m-learning, u-learning may use more context awareness to provide most adaptive content for learners. The main characteristics of u-learning are permanency, accessibility, immediacy, interactivity, situating of instructional activities [3, 6].

## 2.2 HTML5

HTML5 changed the concept of Web from documents to a platform for Webapps.

The HTML5 specifications encompass the HTML5 grammar, available elements, attributes, and relevant APIs, and the various surrounding APIs are related to HTML5 but are basically individual, independent specifications. Although these APIs do not actually belong to the HTML5 specifications, all of them are generally referred to as “HTML5” in a broad sense, and HTML5, CSS, and Javascript are collectively called “Open Web Platform” [7, 8].

## 3 System Design and Content Selection for the Implementation of Webapp

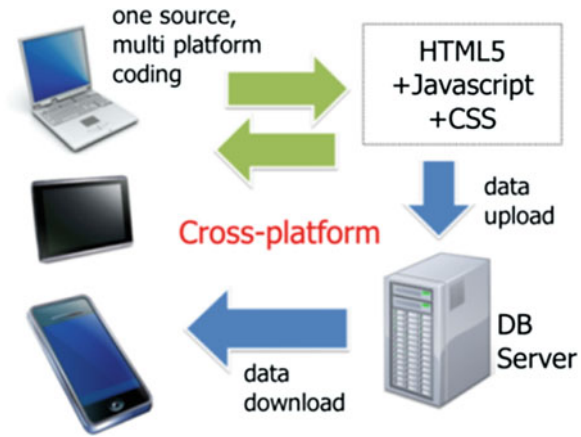
### 3.1 System Design

The proposed Webapp system is illustrated in Fig. 1. Because it is intended to be compatible with all platforms, it was implemented using HTML5, CSS, and Javascript so that it could be used on diverse devices (PCs, smartphones, and smart pads). PCs or smartphones enable the use of the Internet, unlike smart pads, so Webapp was designed to store and load data using DB servers. Specifically, in a non-networked situation, data are internally stored; and in a networked situation, data are uploaded onto a DB server and downloaded from it to enable learning with different devices.

The proposed Webapp content comes from the Elementary School Science Subject Sixth-year Second Semester Unit 3 lesson “What are the benefits of using levers?” This unit helps students understand the concept, forms, and transformation of energy and the benefits of using levers and pulleys [9].

We proposed the integrated subject content shown in Table 1 that is suitable to theme-oriented project learning tailored to elementary students’ cognitive levels and for performing experiments and tasks using STEAM-based apps, and that enables convergence of textbooks. Levers were selected as the learning content to

**Fig. 1** Overall of the proposed Webapp system



**Table 1** Analysis of textbooks encompassing the subject of steam learning

Subject (grade)	Unit	Description
Science (4)	Weighing	Weighing by balancing Weighing using the scales that I made
Science (6)	Energy and tools	Knowing about the benefits of levers Knowing about the benefits of pulleys
Mathematics (6)	Proportional expression	Knowing about and using the features of proportional expressions
Fine Arts (5 and 6)	Designs that deliver laughter	Making everyday products based on interesting ideas

enable STEAM-based learning using Webapp, in line with the design that encompasses weighing in the science subject, proportional formulae in mathematics, and laughter in fine arts.

## 4 Webapp Development

### 4.1 Development Environment

To implement native apps, the integrated development environment should be used to enable utilization of programming languages and SDK; but for Webapp, HTML5, CSS, and Javascript can all be written in general text editors and executed and confirmed in browsers. The web programming environment for storing and sharing learning data was crafted in line with Apache, PHP, and MySQL. To use the Webapp offline by installing smart devices such as native apps, Webapp was ported in the form of a hybrid app using Appspresso 1.0, and Android-SDK and JDK are needed to port it to hybrid apps for Android-based devices.



## 4.2 Implementation

To enable the lever learning Webapp to support diverse devices, a large view of letters on smartphones was enabled using the viewport function, in consideration of the small display of smartphones. The viewport-declared metatag does not affect desktop PCs and thus, can adjust the display according to the Webapp developer's intention.

The learning content offered units, problems for study, and activities to help students understand the learning activities. HTML5 has adopted new tags designed to divide the semantic markup into logical structures, and has eliminated mere modification tags, thereby expressing document structures more distinctively and data embedded in documents.

To help students understand the lever's point of force, supporting point, and point of action, a related game was developed using everyday items. The game, using HTML5 Canvas, enables the implementation not only of diagrams but also of animations without plug-ins.

Clicking the Start button will make three gray points appear on everyday items. The learner should guess what point each location refers to. When s/he answers correctly, s/he confirms the explanation of the Nos. 1, 2, and 3 levers, and can go to the next problem.

To receive the user inputs, event listeners were used. Desktops and smart devices have different input methods, namely, a mouse or a touchpad, so event listeners should be separately registered. The mouse-input-based desktop PCs used mousedown, mousemove, and mouseup events, and the touchpad-based input smart devices used touchstart, touchmove, and touchend events. The event listeners should be selected depending on the device used, so the method of receiving the userAgent character string that identifies the relevant browser that sent that string was used to determine the content to be sent to the system from the server.

To store learning content and share it with other learners, text-based note pads and image-based picture boards were used. The text-based note pads were designed to be stored in the local storage, among the Web storage types supported by HTML5. Data, once stored in the local storage, remain even when the window is closed, which allows the user to later access the page and continue to use the data. Data can also be stored in the local storage during offline learning and can be uploaded in an online environment to be effectively used for learning.

The image-based picture board was implemented using Canvas. If the canvas.toDataURL() method, which functions with the conversion of the canvas into png or jpg files, is used, the relevant image can be stored as a text file in the DB server to be shared and used as backup data and as learning data. All the note pads and picture boards were produced using the web program that ran on PHP and MySQL, in line with the databases.

## 5 Experts' Evaluation and Analysis of the HTML5 Webapp

### 5.1 Experts' Evaluation and Method

The education potential of the proposed Webapp was evaluated by eight experts who specialized in elementary school education and computer education. They had taught for more than 5 years and had experience of app development.

The evaluation items were developed and used to accurately assess the educational value and potential for use of the proposed Webapp, and to define what must be improved. The use of the Webapp was demonstrated to the experts, and the experts were allowed to use it on smartphones and desktop browsers. The evaluation items were rated on a five-step Likert interval scale with 2.5-point intervals, or else the experts were allowed to select or state their opinions.

### 5.2 Results of the Experts' Evaluation and Analysis

First, to determine the experts' knowledge of smart learning, their interest and development experience in smartlearning were analyzed. All of them showed much interest in smartlearning, and seven had experience in using smartphones and developing native apps.

On the question that pertained to the difficulty of developing educational apps using Android or IOS, 37.5 % of the experts highly rated the high-level maintenance difficulty and the inter-OS incompatibility (Fig. 2).

Regarding the educational value of Webapp and its possible use in the education field, 87.5 % of the experts gave it a high education value, and all of them stated that it is worth using in the education field (Fig. 3).

In addition, the experts indicated that diverse educational contents should be secured and that the browser compatibility should be improved (Fig. 4).

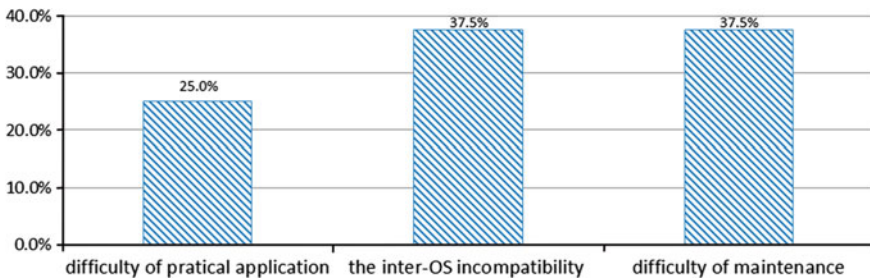


Fig. 2 Difficulty of developing educational native app

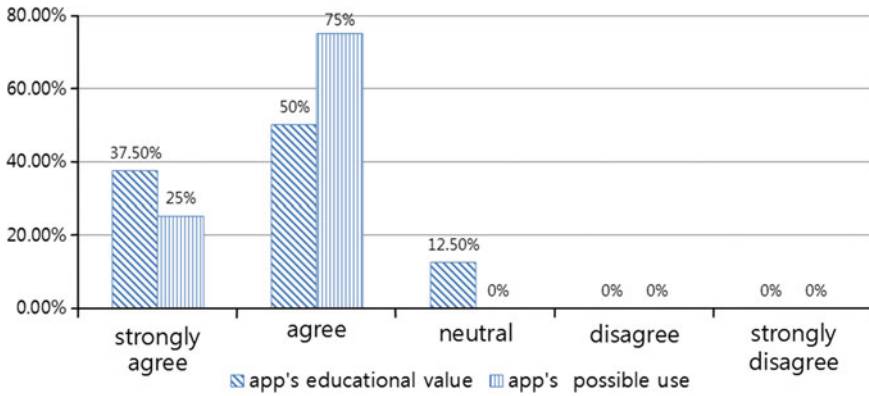


Fig. 3 Educational value and possible use in the education field of the lever learning Webapp

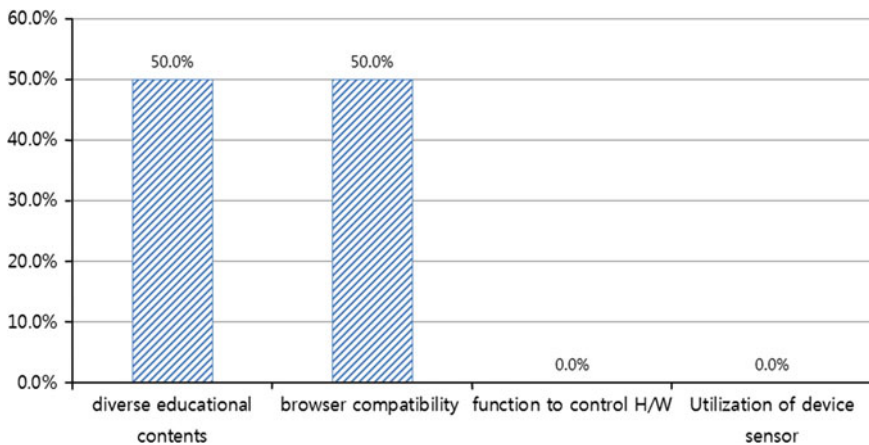


Fig. 4 Expert's advice for improvements

Thus, Webapp for lever learning is worth using in the actual educational field, and has excellent strengths compared with native apps, making it worthwhile to use for developing Webapps.

## 6 Conclusion

To develop e-learning as a new educational methodological approach, this study used the latest HTML5, developed the lever learning Webapp, and examined the possibility of its educational application.

The HTML5-based Webapp, similar to existing apps, have different strengths that enable the pursuit of diversified learning and promote motivation for learning.

Furthermore, it can support multiple platforms regardless of the device, which allows learners in diverse learning situations to easily access learning contents. Notably, it allows learners even without smartphones to use the same content on desktop PCs, thus enabling easy access to the latest education contents. Also, for teachers, Webapp can be developed in a short period with an easy method, and can be easily maintained in a short period, thereby boosting its educational value.

The HTML5 standard is still popular and is expected to be further improved, and yet it has to be perfectly supported by all browsers. Native apps have the advantage of controlling the hardware levels, which Webapps cannot do. Nonetheless, Webapps, if produced in the form of hybrid apps, enable access to hardware. If the requirements are accurately analyzed and if diverse HTML5 functions are designed and developed, Webapps can have tremendous education value.

## References

1. Kim S, Yon YI (2011) A model of smart learning system based on elastic computing. In: Proceedings of 9th international conference on software engineering research, management and applications, Baltimore, USA, pp 184–185
2. Shin DH, Shin YJ, Choo H, Beom K (2011) Smartphones as smart pedagogical tools: implications for smartphones as u-learning devices. *Comput Hum Behav* 27(6):2207–2214
3. Mandula K, Meda SR, Jain DK, Kambham R (2011) Implementation of ubiquitous learning system using sensor technologies. In: Proceedings of IEEE international conference on technology for education, pp 142–148
4. Parsons D, Ryu H (2006) A framework for assessing the quality of mobile learning. In: Proceedings of international conference for process improvement, research and education, Kerkrade, The Netherlands
5. Seppälä P, Alamäki H (2003) Mobile learning in teacher training. *J Comput Assist Lear* 19(3):330–335
6. Ogata H, Yano Y (2004) Context-aware support for computer-supported ubiquitous learning. In: Proceedings of 2nd IEEE international workshop on wireless and mobile technologies in education, Taiwan, pp 27–34
7. Lubbers P, Albers B, Salim F (2010) *Pro HTML5 programming*. Appress, New York
8. Hogan BP (2010) *HTML5 and CSS3: develop with tomorrow's standards today*. Pragmatic Bookshelf, Dallas
9. Ministry of education science and technology of South Korea (2011) *Elementary school science 6th 2nd semester teacher's guide*. Kumsung, Seoul

# Looking for Better Combination of Biomarker Selection and Classification Algorithm for Early Screening of Ovarian Cancer

Yu-Seop Kim, Jong-Dae Kim, Min-Ki Jang, Chan-Young Park and Hye-Jeong Song

**Abstract** This paper demonstrates and evaluates the classification performance of the optimal biomarker combinations that can diagnose ovarian cancer under Luminex exposed environment. The optimal combinations were determined by T Test, Genetic Algorithm, and Random Forest. Each selected combinations' sensitivity, specificity, and accuracy were compared by Linear Discriminant Analysis (LDA) and k-Nearest Neighbor (k-NN). The 8 biomarker data used in this experiment was obtained through Luminex-PRA from the serum of 297 patients (cancer 81, benign 216) of two hospitals. In this study, the results showed that selecting 2–3 markers with Genetic Algorithm and categorizing them with LDA shows the closest sensitivity, specificity, and accuracy to those of the results obtained through complete enumerations of the combination of 2–4 markers.

**Keywords** Biomarker · Ovarian cancer · Marker · T Test · Genetic algorithm · Random forest · LDA · Logistic regression

---

Y.-S. Kim · J.-D. Kim · C.-Y. Park · H.-J. Song (✉)

Department of Ubiquitous Computing, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do 200-702, Korea  
e-mail: hjsong@hallym.ac.kr

Y.-S. Kim  
e-mail: yskim01@hallym.ac.kr

J.-D. Kim  
e-mail: kimjd@hallym.ac.kr

C.-Y. Park  
e-mail: cypark@hallym.ac.kr

M.-K. Jang  
Department of Computer Engineering, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do 200-702, Korea  
e-mail: wscang@gmail.com

Y.-S. Kim · J.-D. Kim · M.-K. Jang · C.-Y. Park · H.-J. Song  
Bio-IT Research Center, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do 200-702, Korea

## 1 Introduction

Ovarian cancer is a malignant tumor frequently arising in the age between 50 and 70. Early diagnosis is very closely associated with a 92 % 5-year survival rate, yet only 19 % of ovarian cancers are detected early [1]. Therefore, early detection of ovarian cancer has great promise to improve clinical outcome. It is evident that the development of a biomarker for early detection of the ovarian cancer has become paramount [2].

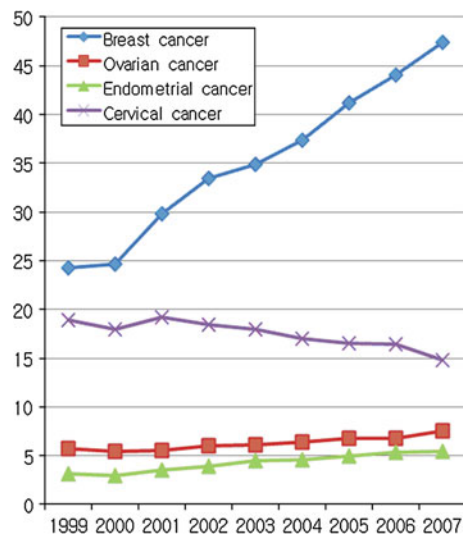
Biomarker consists of molecular information based on the pattern of a single or multiple molecules originating from DNA, metabolite, or protein. Biomarkers are indicators that can detect the physical change of an organism due to the genetic changes.

Along with the completion of the genome project, various biomarkers are being developed, providing critical clues for cancers and senile disorders (Fig. 1).

The early stages of research focused on a single biomarker for cancer diagnosis. Recent researches focus on combining multiple biomarkers to diagnose cancer more efficiently. Researches tend to focus especially on improving the sensitivity and specificity in order to increase the accuracy of the diagnosis, and the commercialization of multi-biomarkers seems to be close at hand. However, a new technology to find the right biomarker combinations is required, since the sensitivity and quantity has not yet reached a satisfactory level [3].

In this research, the mass value of the biomarkers was obtained using Luminex [4]. Luminex follows the panel reactive antibody (PRA), a solid phase-based method of Luminex corp. This paper determines the optimal marker combinations for ovarian cancer diagnosis from the combinations selected with T Test [5], Genetic Algorithm [6], and Random Forest [7] from all the possible combinations

**Fig. 1** Changes in gynecologic cancer causes in Korea (1999–2007)



of the 8 markers, based on their the florescence data measured. Linear Discriminant Analysis [5] and k-Nearest Neighbor [6] were used to evaluate the sensitivity, specificity, and classification accuracy of the optimal combinations. The research aims to determine the optimal marker combination and categorization algorithm by comparing the experimental results with all the possible combinations.

Methods for the collection of data are illustrated in Sect. 2, and the experimental details are demonstrated in Sect. 3. The results of the marker combinations and its classification performance are discussed in Sect. 4, and Sect. 5 presents the conclusion.

## 2 Data Collection

The serum samples from 81 patients with ovarian cancer, 216 patients with benign pelvic masses were used. Sera were provided by Hallym University Medical Center (HUMC) and ASAN Medical Center. These samples were reacted with Lumindex-beads attached with 8 biomarkers, and the florescence from the antibodies on the beads was measured. In order to equalize the range of the biomarker florescence, the florescence values of each biomarker were normalized to 0–1 based on their maximum and minimum values.

## 3 Methods

This paper conducts two experiments: (1) determination of biomarkers with T Test, Genetic Algorithm (GA), and Random Forest (RF), and (2) performance comparison of the selected markers using Linear Discriminant Analysis (LDA) and k-Nearest Neighbor (k-NN).

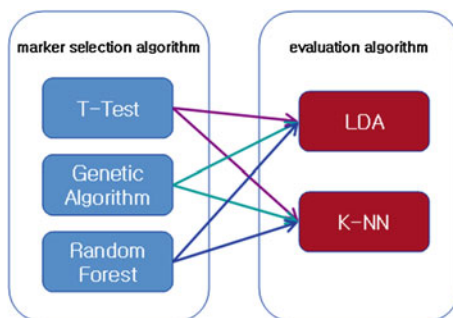
The number of randomly created tree for the RF was 50, and the k for k-NN was 3. The combination of the biomarkers consisted of 2–4 markers, and leave-one-out cross validation was conducted for the evaluation.

The algorithms for the marker combination and combinations of evaluation algorithms used in the experiment are shown in Fig. 2. As a control, the optimal marker combination gained through a complete enumeration survey of 2–4 markers was used.

## 4 Results

The experiment compares the difference in performance of the selected 2–4 multi-biomarkers by T Test, GA, and RF to that of the optimal combination amongst the total possible combinations of the markers. The sensitivity, specificity, and

**Fig. 2** The algorithms for the marker combination and combinations of evaluation algorithms



accuracy of each optimal combination for both cancer and benign group was measured and compared with LDA and k-NN.

The markers that ought to be combined was limited to four, because the high cost to combine more than 4 markers will make it difficult to realize and commercialize the use of multi-biomarkers. Also to avoid the infringement of patent, the names of the markers are concealed.

#### ***4.1 Optimum Combination Results According to Classification Algorithm***

Table 1 shows the optimal marker combination and their performance when applying LDA and k-NN to the combine the 2–4 markers. *Mnumber* means an individual bio marker. As seen in the Table 1, M1 and M7, M6 are most frequent, having the appearance of 6, 4, and 3 times respectively. The best accuracy of 80.5 % was seen in the 4-marker combinations.

Table 2 compares the accuracy of the selected markers through T Test with accuracy of optimum marker combinations. From the selected four markers through T test, M6 is the most probable marker as the high frequency marker. This might have been the cause of the large performance difference of about 5 %.

**Table 1** Performance test of the classification algorithm through all the possible marker combinations of 2–4 markers and the formation of the optimal combination

Classifier	Marker 1	Marker 2	Marker 3	Marker 4	Sensitivity	Specificity	Accuracy
LDA	M1	M7			0.531	0.894	0.795
	M1	M4	M7		0.556	0.894	0.801
	M1	M2	M5	M7	0.543	0.903	0.805
k-NN	M1	M6			0.519	0.830	0.785
	M1	M6	M7		0.506	0.912	0.801
	M1	M2	M6	M8	0.494	0.921	0.805



**Table 2** Classification performance comparison of the marker combinations obtained through T test

Classifier	Marker 1	Marker 2	Marker 3	Marker 4	Sensitivity	Specificity	Accuracy
LDA	M8	M3			0.543	0.796	0.727
	M8	M3	M6		0.568	0.815	0.748
	M8	M3	M6	M5	0.531	0.815	0.737
k-NN	M8	M3			0.333	0.847	0.707
	M8	M3	M6		0.457	0.847	0.741
	M8	M3	M6	M5	0.469	0.889	0.774

**Table 3** Classification performance comparison of the marker combinations obtained through Genetic Algorithm

Classifier	Marker 1	Marker 2	Marker 3	Marker 4	Sensitivity	Specificity	Accuracy
LDA	M1	M8			0.531	0.880	0.785
	M1	M8	M2		0.543	0.884	0.791
	M1	M8	M2	M7	0.556	0.866	0.781
k-NN	M1	M8			0.432	0.903	0.774
	M1	M8	M2		0.395	0.917	0.774
	M1	M8	M2	M7	0.407	0.889	0.758

The four markers selected with Genetic Algorithm (Table 3) and Random Forest (Table 4) includes the two most frequent markers (M1, M7). It shows 3 and 2 % difference for the accuracy.

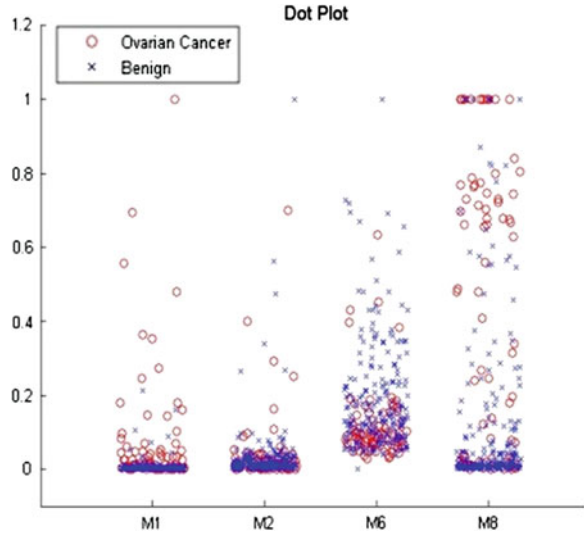
### 4.2 Dot Plot and ROC Curve for Optimum Marker Combination

As illustrated in Fig. 3, the cancer and benign are not distinguishable in one dimension. The results were easier to analyze when it was projected in two dimensions using a marker combination. The ROC curve of Fig. 4 demonstrates the aforementioned statement.

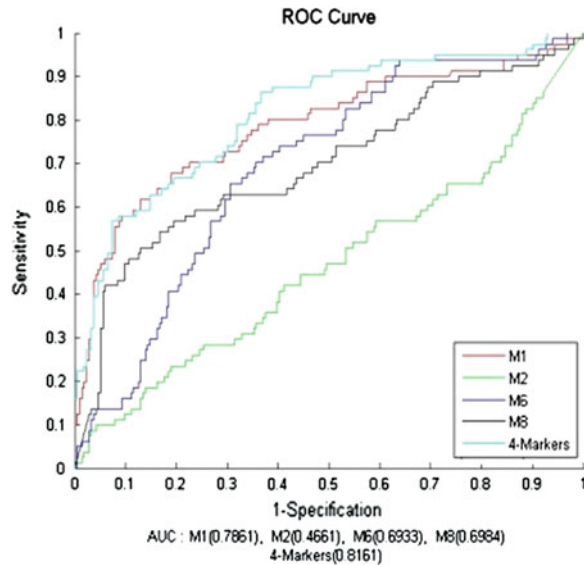
**Table 4** Classification performance comparison of the marker combinations obtained through Random Forest

Classifier	Marker 1	Marker 2	Marker 3	Marker 4	Sensitivity	Specificity	Accuracy
LDA	M1	M6			0.543	0.852	0.768
	M1	M6	M5		0.494	0.866	0.764
	M1	M6	M8	M5	0.543	0.866	0.778
k-NN	M1	M6			0.494	0.875	0.771
	M1	M6	M5		0.432	0.861	0.744
	M1	M6	M8	M5	0.407	0.857	0.734

**Fig. 3** Dot plot of individual markers in optimum marker combination



**Fig. 4** ROC curve of individual markers and 4-markers combination



## 5 Conclusion

This paper searches for the biomarker combination that can easily distinguish between malignant and benign tumors. Comparing the classification performance of ovarian cancer, selecting 2 or 3 markers through GA and classifying with LDA shows the most similar sensitivity, specificity, and accuracy to that of the marker combinations of 2–4 markers derived from the complete enumeration survey.

However, except for the LDA 3-marker combination of GA, the marker combination of 2–4 markers obtained with the marker selection algorithm showed significantly low accuracy compared to the marker combinations obtained from the complete enumeration survey.

**Acknowledgments** The research was supported by the Research & Business Development Program through the Ministry of Knowledge Economy, Science and Technology (N0000425) and the Ministry of Knowledge Economy (MKE), Korea Institute for Advancement of Technology (KIAT) and Gangwon Leading Industry Office through the Leading Industry Development for Economic Region.

## References

1. American Cancer Society (2012). <http://www.cancer.org/Cancer/OvarianCancer>
2. Brian N, Adele M, Liudmila V, Denise P, Matthew W, Elesier G, Anna L (2009) A serum based analysis of ovarian epithelial tumorigenesis. *Gynecol Oncol* 112:47–54
3. ChiHeum C (2008) Biomarkers related to diagnosis and prognosis of ovarian cancer. *Korean J Obstet Gynecol* 39:90–95
4. SunKyung J, EunJi O, ChulWoo Y, WoongSik A, YongGu K, YeonJun P, KyungJa H (2009) ELISA for the selection of HLA isoantibody and comparison evaluation of Luminex panel reactive antibody test. *J Korean Soc Lab Med* 29:473–480
5. David F, Roger P, Robert P (1998) *Statistics*, 3rd edn. W. W. Norton & Company, New York
6. Tom MM (1997) *Machine learning*. The McGraw-Hill, New York
7. Leo B (2001) Random forest. *Mach Learn* 45:5–32
8. Suraj DA, Greg PB, Tzong-Hao C, Katharine JB, Jinghua Z, Partha S, Ping Y, Brian CM (2009) Development and preliminary evaluation of a multivariate index assay for ovarian cancer. *PLoS ONE* 4:e4599

# A Remote Control and Media Sharing System Based on DLNA/UPnP Technology for Smart Home

Ti-Hsin Yu and Shou-Chih Lo

**Abstract** The remote control and media sharing of consumer devices are key services for smart living. The involving of mobile devices into these services has become a technology trend. Existing solutions to these services restrict these devices to be located in the same local network. In this paper, we design and implement an integrated architecture that supports the outdoor remote control to home devices and the sharing of digital media among indoor and outdoor devices. By following the digital home related standards, we show our system design with the details of hardware and software components.

**Keywords** DLNA · UPnP · Smart home · Media sharing · Remote control

## 1 Introduction

With the popularity of digital consumer products, digital content can be seen everywhere. For the easy sharing of digital media such as videos, photos, and music between these consumer products, the Digital Living Network Alliance (DLNA) was initiated in 2003 to define interoperability guidelines [1, 2]. The underlying technology is Universal Plug and Play (UPnP) [3] for media management, discovery, and control. The UPnP, which includes a set of standard network protocols such as TCP/IP, HTTP, and Simple Object Access Protocol (SOAP), enables digital devices having networking capability to be connected.

DLNA compliant devices can seamlessly discover each other's presence on the same home network and share functional services or media content with each other. Four types of DLNA devices are defined in the standard: DMS, DMC, DMP,

---

T.-H. Yu · S.-C. Lo (✉)

Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien 974, Taiwan, Republic of China

e-mail: scllo@mail.ndhu.edu.tw

and DMR. A Digital Media Server (DMS) stores and provides media content to other types of devices. A Digital Media Controller (DMC) can discover media content and command a Digital Media Renderer (DMR) to play the content. A Digital Media Player (DMP) can discover and play media content directly, which can be considered as the combination of DMC and DMR.

The media sharing between DLNA supported devices contributes significantly to our comfortable life. By integrating sensor technology and automatic control to these home appliances, some research efforts [4–6] have been done for achieving a smart living environment. With the popularity of smart phones, using a mobile phone to monitor and control the living environment becomes a potential service model. For example, a mobile user when being out of home can retrieve and play media streaming from an indoor DMS by connecting the mobile phone to the home network using the Session Initial Protocol (SIP) [7].

In this paper, we provide a service platform using smart phones that can remotely access and control any home appliances. The home appliances include any DLNA or UPnP supported devices and traditional electronic devices such as air conditioners, TVs, and desk lamps. Our prototype system can remotely monitor and switch the power of a home device and share media content between an indoor device and a remote device.

Our developed platform is based on the Android system which provides open sources and many free libraries. To enable a smart phone to have extra control functions to other devices, we use the Arduino programmable platform [8]. Arduino is an open-source single-board microcontroller which can be easily programmed to control robots, lighting, etc. An Arduino board also provides the linkage to other external modules such as sensor modules and wireless communication modules.

The remainder of this paper is organized as follows. [Section 2](#) explains some issues about our system design. [Section 3](#) shows the design of each hardware/software component. Finally, the concluding remarks are given in [Sect. 4](#).

## 2 System Design

We mainly extend smart living services from an indoor environment to an outdoor one. This platform provides the following services:

1. A user no matter in home or out of home can monitor and control home appliances and share digital content with home DLNA compliant devices through a mobile phone (For example, a remote device turns off the desk lamp and downloads a music file from an indoor DLNA device).
2. Two outdoor users can share digital contents as if they are in the same home network (For example, a remote device accesses image files from another remote device).

3. An outdoor user can redirect digital contents from a home DMS to a local DMR (For example, a remote device commands a local DMR to retrieve and play a video from an indoor DLNA device).

There are two types of home devices in our system: DLNA devices and UPnP devices. DLNA devices provide or play media contents and UPnP devices provide remote control services. To enable traditional devices such as desk lamps and televisions to support for UPnP functions, we externally equip each of these devices with a control board. Consequently, these traditional home devices become so called external UPnP devices (in contrast to internal UPnP devices with built-in UPnP functions). The control board works as a ZigBee End Device (ZED) that can be controlled by a ZigBee Coordinator (ZC) via the ZigBee wireless communication.

The core component in our system is the DroidHome which discovers and maintains DLNA and UPnP devices. Moreover, DroidHome plays the role of a proxy server for remote devices and informs these remote devices about the status of an indoor device by a message push mechanism. The message push mechanism is provided by an existing push service on the cloud called Cloud to Device Message (C2DM) [9, 10]. C2DM is released with the Android operation system by Google. Any Android device can be notified in background by the C2DM if the corresponding user has registered itself using a Google account.

## 2.1 Challenges and Solutions

Both in DLNA and UPnP networks, the service discovery and invocation follow the same method of Simple Service Discovery Protocol (SSDP). A control point or a DMC multicasts a discovery message (ssdp:discover) into the local home network. A device in the house responds its presence by sending back a replay message (ssdp:alive). A new joining DMS can also multicast this ssdp:alive message to inform other existing devices of its presence. This ssdp:alive message contains an Uniform Resource Locator (URL) to locate the file about the device profile. This device profile includes meta-information in XML format such as the device name, production factory, and a URL list to locate those service functions provided by the device. The retrieval of the device profile is based on the HTTP protocol. The service invocation is by sending a SOAP message to the device.

When we integrate any non-UPnP devices and remote devices into the above service platform, some problems are encountered. In the following, we discuss the new challenges and our proposed solutions.

1. **Remote DMS:** If we allow an indoor DLNA device to directly access digital contents from a remote device configured as a DMS, this remote DMS should join to the same home network. One simple way is to connect this remote device back to the home network using the Virtual Private Network's (VPN) Point-to-Point Tunneling Protocol (PPTP). However, this remote device cannot

show its presence by simply multicasting an `ssdp:alive` message into the home network due to a different setting of the subnet mask. The provided solution is that this remote device first unicasts the message to our DroidHome which in turns multicasts the message into the home network.

2. **Remote DMP:** If we configure a remote device as a DMP and connect it to the home network using the PPTP, this remote DMP can directly play the digital content of an indoor DMS. However, the drawback of this approach is that the remote DMP should stay connected to the home network for keeping the new status of an indoor DMS. Here, we use DroidHome to decouple this strong connection. DroidHome acts as a DMC and stores the service state of each indoor DLNA or UPnP device. A table to record the URLs to these device profiles is also maintained. A remote DMP simply contacts with DroidHome to retrieve this URL table for invoking a certain service. When DroidHome detects any state change (e.g., the joining or leaving of an indoor DMS), it will notify those remote devices using the C2DM push service.
3. **Private home network:** Almost all home networks are configured as private networks where all indoor devices use private IP addresses and an Internet Gateway Device (IGD) uses the only public IP address. This IGD provides the Network Address Translation (NAT) function to enable an indoor device to be reachable from the Internet. The core technique is by the port mapping. However, most NAT devices maintain dynamic port mapping, which makes an indoor device unreachable from the Internet. Here, we use an UPnP supported IGD which can be automatically discovered and remotely configured with a static port mapping. DroidHome would first register a mapping port to this IGD such that any other remote devices can connect to the DroidHome using the IGD's public IP address and the registered port number. Moreover, to enable a remote device to access the profile of an indoor device, DroidHome translates the private address in the URL table to the IGD's public address. Each address translation needs DroidHome to register a new mapping port to the IGD.
4. **External UPnP device:** In our system, an external UPnP device actually refers to a non-UPnP device. These devices naturally have no networking capability and service functions. We first equip each of these devices with a control board. This control board can switch the power or enable certain functions of the associated device by a programmable current relay unit or an infrared emitter. These control boards can communicate with our DroidHome using the ZigBee wireless communication. Second, we create a software device object for each external UPnP device in DroidHome. A device object stores the corresponding service profile and service functions about the physical device. Then, DroidHome records the access path to this device object into the URL table. Consequently, a remote device can look up services provided by external UPnP devices and can invoke these functions through DroidHome.

## 3 Component Design

In the following, we explain the function and the design of each hardware component in our system.

### 3.1 *Zed*

The ZED is responsible for the control of the associated external UPnP device. Each ZED is implemented by an Arduino Uno platform and has the following types: infrared ZED and relay ZED. An infrared ZED can emit infrared signal to switch the power or the channel of the associated device. The relay ZED can switch the power of the associated device.

An infrared ZED contains the following hardware components inside. A current transformer (CT) sensor detects the current of a power line and is used to detect the power status of a device. An XBee module provides ZigBee communication and is configured as ZED mode. An IR module sends and receives infrared signal. A button module is used to set the infrared signal on the IR module with the same frequency as the remote controller of a home device. A relay ZED contains a CT sensor and an XBee module, and additionally a current relay module which can open or close the power. Beside these hardware components, each ZED maintains some data: power status (on or off) of the associated device, type of the ZED (infrared or relay), and identification (a unique sequence number) of the ZED.

### 3.2 *DroidHome*

This hardware component is composed of an Android device (a smart phone in our test system) and an Arduino Mega ADK platform through USB communication. The former subcomponent discovers and maintains indoor DLNA devices and internal UPnP devices in the home network (WiFi networking environment in our test system). The latter subcomponent discovers and controls indoor ZEDs using the ZigBee network.

The Arduino platform contains an XBee module which is configured with ZC mode. This ZC would periodically multicast a discovery message (performed by ZED discovery module) such that all surrounding ZEDs will respond this message with their types and ID data. These response data are recorded in the ZED table. The control command (e.g., power switch) to each ZED is issued from the ZED control module.

In the Android device, a registration table keeps the identification information for each authorized remote device, which includes an USN (Unique Service Name used to identified a device in our system), a C2DM token (a certification code



given by the C2DM server), and a password. A DMC module is responsible for device discovery. The access path to the profile of a discovered device is stored in the URL table. This table keeps two versions of URLs (private URL and public URL) with the primitive and the translated IP addresses and port numbers, respectively, for each access path. An external device object table stores UPnP device objects for all discovered ZEDs. Each device object keeps the status and profile of the associated device. The request and response messages to and from the IGD for port mapping are handled by a port mapping module. A remote presence module is responsible to notify indoor devices about the joining or leaving of a remote DMS by multicasting messages into the home network. A server push module handles the contact to the C2DM server for pushing a notification to a remote device. An HTTP server module handles the request of device profile downloading and the request of service invocation.

### ***3.3 Remote Device***

This device is a smart phone installed with some software components. A register module processes the registration to the C2DM server and to the DroidHome. A DMP module can discover DMSs and render the output of service content. If this remote device is out of the home network, this device can discover indoor services by retrieving device profiles from the URL table of DroidHome. This task is performed by an import devices module. The notification of the state change of any indoor device is listened by a push receiver module. A DMS module is activated when the remote device would like to share digital contents. In this situation, the remote device waits for incoming requests to access its profile via a HTTP server module.

## **4 Conclusions**

We have demonstrated a home service platform that integrates media sharing and power control to home appliances. Four distinguish features are highlighted. First, the restriction of media sharing among only indoor DLNA devices is broken. Mobile phones can be involved no matter in home or out of home. Second, traditional home appliances without automatic control and networking capability can be involved too in our platform by equipping with a simple external device. Third, our implementation is based on an open-source software stack, some existing equipments, and low-price hardware chips. Fourth, a variety of home services such as temperature sensing and video surveillance can be easily integrated into our service platform.

## References

1. Digital Living Network Alliance. <http://www.dlna.org/>
2. Digital Living Network Alliance (2006) DLNA networked device interoperability guidelines v1.5, Mar 2006
3. UPnP. <http://www.upnp.org/>
4. Horng M-F, Chang B-C, Su B-H (2008) An intelligent intrusion detection system based on UPnP technology for smart living. In: Proceedings of 8th international conference on intelligent systems design and applications, Cairo. pp 14–18
5. Leu J-S, Lin W-H, Tzeng H-J (2009) Design and implementation of an OSGi-centric remote mobile surveillance system. In: Proceedings of IEEE international conference on systems, man and cybernetics, San Antonio, Oct. 2009. pp 2498–2502
6. Chen Y-S, Chen I-C, Chang W-H (2010) Context-aware services based on OSGi for smart homes. In: Proceedings of 3rd IEEE international conference on Ubi-media computing July 2010, China, pp 38–43
7. Oh Y-J, Lee H-K, Kim J-T, Paik E-H, Park K-R (2007) Design of an extended architecture for sharing DLNA compliant home media from outside the home. *IEEE Trans Consum Electron* 53(2):542–547
8. Arduino. <http://www.arduino.cc/>
9. C2DM. <https://developers.google.com/android/c2dm/>
10. Hansen J, Gronli T-M, Ghinea G (2012) Cloud to device push messaging on Android: a case study. In: Proceedings of 26th international conference on advanced information networking and applications workshops, Mar 2012, Japan, pp 1298–1303

# A New Distributed Grid Structure for k-NN Query Processing Algorithm Based on Incremental Cell Expansion in LBSs

Seungtae Hong, Hyunjo Lee and Jaewoo Chang

**Abstract** To manage the frequent updates of moving objects' locations on road networks in an efficient way, we propose a new distributed grid scheme which utilizes node-based pre-computation technique to minimize the update cost of the moving objects' locations. Because our distributed grid scheme manages spatial network data separately from the POIs (Point of Interests) and moving objects, it can minimize the update cost of the POIs and moving objects. To process k-nearest neighbor (k-NN) query in our distributed grid scheme, we propose a k-NN query processing algorithm based on Incremental cell expansion which minimize the number of accesses to adjacent cells during POIs retrieval in a parallel way. Finally, we show from our performance analysis that our algorithm is better on retrieval performance than the k-NN algorithm of the existing work.

**Keywords** Distributed grid scheme · Query processing algorithm · Road network · Moving objects

## 1 Introduction

With the advancements on GPS and mobile device technologies, it is required to provide location-based services (LBS) to moving objects which move into spatial networks. Several types of location-dependent queries are significant in LBS, such

---

S. Hong · H. Lee · J. Chang (✉)

Department of Computer Engineering, Chonbuk National University, Chonju,  
Chonbuk 561-756, South Korea

e-mail: jwchang@jbnu.ac.kr

S. Hong

e-mail: dantehst@jbnu.ac.kr

H. Lee

e-mail: o2near@jbnu.ac.kr

as range queries [1], k-nearest neighbor (k-NN) queries [1–3], reverse nearest neighbor queries [4], and continuous queries [5]. Among them, the most basic and important queries are k-NN ones. The existing k-NN query processing algorithms use pre-computation techniques for improving performance [6–8]. However, when POIs need to be updated, they are inefficient because distances between new POIs and nodes should be re-computed. To solve it, S-GRID [9] divides a spatial network into two-dimensional grid cells and pre-compute distances between nodes which are hardly updated. However, S-GRID cannot handle a large number of moving objects which is common in real application scenario. As the number of moving objects increases, a lot of insertions and updates of location data are required due to continuous changes in the positions of moving objects. Because of this, a single server with limited resources shows low performance for handling a large number of moving objects. Therefore, we, in this paper, propose a new distributed grid scheme which manages the location information of a large number of moving objects in spatial networks. Based on our distributed grid scheme, we propose a new k-NN query processing algorithm based on incremental cell expansion which minimize the number of accesses to adjacent cells during POIs retrieval in a parallel way.

The rest of the paper is organized as follows. In Sect. 2, we present related works. In Sect. 3, we describe the details of our distributed grid scheme. Section 4 presents a new k-NN query processing algorithm based on our grid scheme. In Sect. 5, we provide the performance analysis of our k-NN query processing algorithm. Finally, we conclude this paper with future work in Sect. 6.

## 2 Related Work

In this section, we describe some related works on k-NN query processing in spatial networks. First, VN3 [6], PINE [7], and islands [8] were proposed to pre-compute the distance between POIs and nodes (or border points) in road networks. However, when POIs need to be updated, they are inefficient because distances between new POIs and nodes should be re-computed. To resolve the problem of the VN3, PINE and Island approaches, Huang et al. [9] proposed S-GRID (Scalable Grid) which represents a spatial network into two-dimensional grids and pre-computes the network distances between nodes and POIs within each grid cell. To process k-NN query, they adopt the INE algorithm [1] which consists of inner expansion and outer expansion. The inner expansion starts a network expansion from the cell where a given query point is located and continues processing until the shortest paths to all data points inside the cell have been discovered or the cell holds no data points. Whenever the inner expansion visits a border point, the outer expansion is performed from that point. The outer expansion finds all POIs in the cells sharing the border point. This process continues until k nearest POIs are found. In S-GRID, the updates of the pre-computation data are local and POI independent. However, S-GRID have a critical problem that it is not efficient in

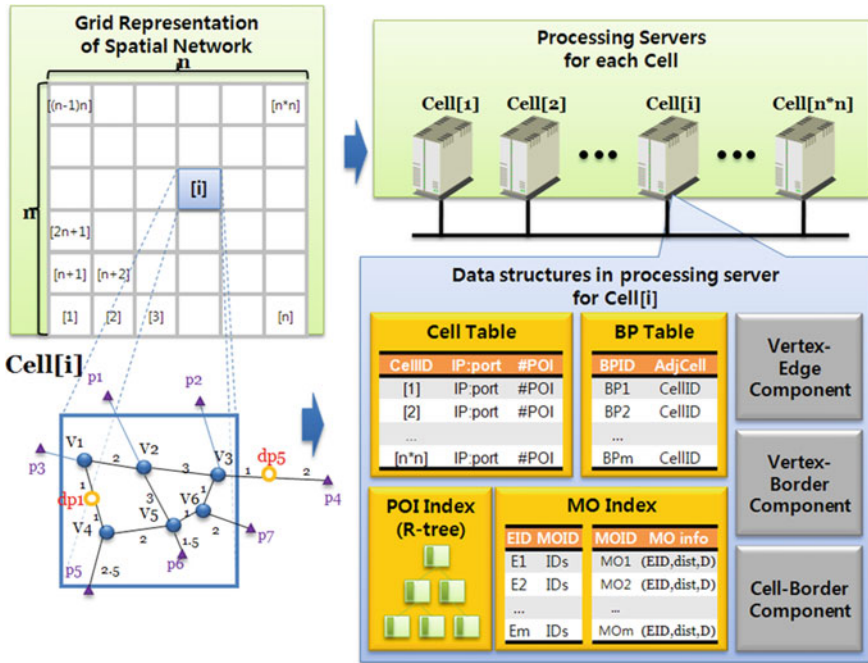


Fig. 1 Overall structure of distributed grid scheme

handling a large number of moving objects, which are common in real application scenario, because it focuses on a single server environment. That is, when the number of moving objects is great, a lot of insertions and updates of location data are required due to continuous changes in the positions of moving objects. Thus, a single server with limited resources shows bad performance for handling a large number of moving objects (Fig. 1).

### 3 Distributed Grid Scheme

To support a large number of moving objects, we propose a distributed grid scheme by extending S-GRID. Similar to S-GRID, our distributed grid scheme employs a two-dimensional grid structure for a spatial network and performs pre-computations on the network data, such as nodes and edges, inside each grid cell. In our distributed grid scheme, we assign a server to each cell for managing the network data, POIs and moving objects. Each server stores the pre-computed network data and manages cell-level two indices, one for POIs and the other for moving objects. Figure 2 shows an overall structure of our distributed grid scheme. We describe each component of our distributed grid scheme.

```

InnerExpansion Algorithm(q, k, Qv, Qdp, BPList)
01. edge = findEdge(q)
02. for each POI∈findPOI(edge)      Qdp.update(POI, di st(q, POI))
03. for each v∈{edge.start_node, edge.end_node }  Qv.update(v, dist(q, v))
04. for each bp∈Cell-Border Component  Qv.update(bp, dist(q, bp))
05. dMax=Qdp.dist(k)
06. do
07.   vx=Qv.dequeue, mark vx as visited
08.   if (vx is a vertex)
09.     for each adjacent vertex vy of vx in Vertex-Edge Component
10.       for each POI∈findPOI(ex,y)
11.         Qdp.update(POI, dist(q,vx)+dist(vx,POI))
12.       Qv.update(vy, dist(q,vx)+dist(vx,vy))
13.       if (all POI in myCell is discovered or Qdp.maxdist()<dist(q,vx))
14.         break
15.     else      BPList.update(vx, dist(q,vx))
16.     dMax=Qdb.dist(k)
17. while( d(q, vx) < dMax && Qv≠∅ )

```

**Fig. 2** Inner expansion algorithm

## 4 K-NN query processing algorithm

In this paper, we propose a new k-NN query processing algorithm based on our distributed grid scheme, namely Incremental Cell Expansion (ICE) algorithm. First, our ICE algorithm finds all the border points and creates a list of cells containing the border points by doing the inner expansion. Next, a coordinate (server) sends the query to all the cells in the cell list. Secondly, servers receiving the query retrieve both POIs and other border points by doing outer expansion. Then, they send the retrieved POIs and border points to the coordinate from which query is originated. Thirdly, the algorithm checks whether or not there is a border point being nearer than the k-th POI. If true, the process is repeated until no border point is nearer than the k-th POI. To find k nearest neighbors, our ICE algorithm performs both inner expansion and outer expansion by using two priority queues, Qv and Qdp. Qv stores both relevant nodes and the distance between a query point and the nodes while Qdp stores both retrieved POIs and their distances from a query point. Thus, our ICE algorithm can improve retrieval performance by minimizing unnecessary visiting of adjacent cells. Figure 2 shows an inner expansion algorithm.

In addition, our ICE algorithm performs outer expansion in two cases; (i) the number of retrieved POIs is less than k and (ii) there remains a border point in the cell list. Figure 3 shows the outer expansion algorithm. First, the algorithm sends a query to the servers managing respective adjacent cells of the cell list. Then, by using a cell-border component, the servers insert both the border points of related cells and their distances from a query point into Qv. Secondly, by using a vertex-border component, the servers insert POIs within the related cells and their

```

OuterExpansion Algorithm(q, k, BPlist)
Qdp= $\emptyset$ , Qv= $\emptyset$ 
01. for each bpi $\in$ BPlist
02.   for each bpj $\in$ Cell-Border Component
03.     if (bpi bpj)   Qv.update(bpj, dist(q,bpi)+dist(bpi+bpj))
04.     for each POI $\in$ myCell
05.       Qdp.update(POI, dist(q,bpi)+dist(bpi+POI))
06.     dMax=Qdp.dist(k)   bp=Qv.deque
07. return POIs in Qdp, bps in Qv

```

**Fig. 3** Outer expansion algorithm

distances from a query point into Qdp At last, they return both the retrieved POIs and the border points to the coordinator where the query is originated.

## 5 Performance Analysis

We present performance analysis of k-NN query processing algorithm for our distributed grid scheme. We implement our grid under HP ML 150 G3 server with Intel Xeon 3.0 GHz dual CPU, 2 GB memory. In our experiments, we used multiple processes in a single server and each process manages a single cell. To provide an environment appropriate to a distributed grid scheme, we let each process use a different port number to communicate with other processes by using TCP/IP protocol. For spatial network data, we use San Francisco Bay map consisting of 220,000 edges and 170,000 nodes, and generate four sets of POIs (i.e., 2,200, 4,400, 11,000, 22,000) by using Brinkhoff algorithm [10]. These POIs are indexed by using R-trees. Moreover, we randomly select 100 nodes from San Francisco Bay map as query points. To measure the retrieval performance of k-NN queries, we average response times for all the 100 query points. Because the existing works VN3 [6], PINE [7], island [8] are very inefficient for the update of POIs due to their POI-based pre-computation techniques, they are not appropriate for dealing with a large number of mobile objects in spatial networks. Thus we compare our algorithm with S-GRID algorithm in terms of POI retrieval time.

Figure 4a first shows the performance of k-NN query processing with the different number of grid cells when  $k = 20$  and POI density = 0.01. The performance of our algorithm is better than that of S-GRID when the number of grid cells is more than  $10 \times 10$ . This is because our algorithm performs outer expansion in a parallel way. Figure 4b shows the retrieval time of k-NN query with the varying value of k when the density of POI is 0.01 and the number of grid cells equals  $20 \times 20$ . We can say from the performance result that our ICE algorithm is better because it can minimize the number of accesses to adjacent cells during POIs retrieval.

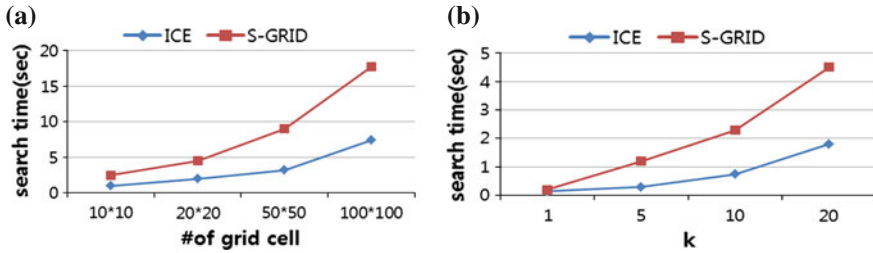


Fig. 4 Retrieval performance **a** with different number of grid cells **b** in terms of  $k$

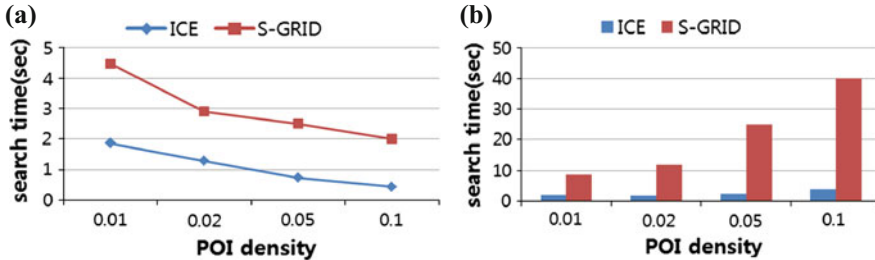


Fig. 5 Retrieval performance **a** in terms of the density of POIs **b** after updating POIs

Figure 5a shows the retrieval time of  $k$ -NN query with the varying density of POIs where the number of grid cells equals  $20 \times 20$  and  $k = 20$ . As a result, we can reduce the cost of inner expansion within a cell and the number of adjacent cells to be visited. Figure 5b shows the retrieval time of  $k$ -NN query after updating POIs. For this experiment, we measure the search time of  $k$ -NN query when the 10 % of POIs is updated. In the case of S-GRID, the retrieval time is exponentially increased as the density of POIs increases. This is because S-GRID uses one R-tree to index all the POIs of the network and so the update of POIs in a cell affects the whole system. Whereas, because our grid scheme uses a separate R-tree per each grid cell to index POIs within it, the update of POIs in a cell does not affect all the grid cells globally. As a result, even though the number of updated POIs increases, the retrieval performance of our grid scheme is not dramatically increased.

## 6 Conclusion and Future Work

In this paper, we proposed a new distributed grid scheme to manage the location information of a large number of moving objects in spatial networks. Our distributed grid scheme makes use of a node-based pre-computation technique so that it can minimize the update cost of the moving objects' locations. Our distributed



grid scheme splits a spatial network into two-dimensional grid cells so that it can update network data locally. Based on our grid scheme, we proposed a new k-NN query processing algorithm. Our ICE algorithm improves the retrieval performance of K-NN queries because it decreases the number of adjacent cells visited by transmitting a query to all the shared border points. Our experimental results show that our algorithm is better on retrieval performance than that of S-GRID. As a future work, we need to extend our grid scheme to handle a spatial network with dense and sparse regions in an efficient manner by using non-uniform grid cells.

**Acknowledgment** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0023800).

## References

1. Papadias D, Zhang J, Mamoulis N, Tao Y (2003). Query processing in spatial network databases. In: Proceedings of VLDB, pp 802–813
2. Shahabi C, Kolahdouzan MR, Sharifzadeh M (2003) A road network embedding technique for K-nearest neighbor search in moving object databases. In: Proceedings of GeoInformatica, vol 7, no 3, pp 255–273
3. Jensen CS, Pedersen TB, Speicys L, Timko I (2003) Data modeling for mobile services in the real world. In: Proceedings of SSTD, pp 1–9
4. Benetis R, Jensen CS, Karčiauskas G, Šaltenis S (2006) Nearest and reverse nearest neighbor queries for moving objects. In: Proceedings of VLDB, pp 229–250
5. Huang YK, Chen C-C, Lee C (2009) Continuous K-Nearest neighbor query for moving objects with uncertain velocity. In: Proceedings of GeoInformatica, vol 13, no 1, pp 1–25
6. Kolahdouzan MR, Shahabi C (2004) Voronoi-based nearest neighbor search for spatial network databases. In: Proceedings of VLDB, pp 840–851
7. Safar M (2005) K Nearest Neighbor Search in Navigation Systems. *Mobile Inf Syst* 1(3):207–224
8. Huang X, Jensen CS, Šaltenis S (2005) The islands approach to nearest neighbor querying in spatial networks. In: Proceedings of SSTD, LNCS 3633, pp 73–90
9. Huang X, Jensen CS, Lu H, Šaltenis S (2007) S-GRID: a versatile approach to efficient query processing in spatial networks. In: Proceedings of SSTD, LNCS 4605, pp 93–111
10. Brinkhoff T (2002) A framework for generating network-based moving objects. In: Proceedings of GeoInformatica, pp 153–180

# A New Grid-Based Cloaking Scheme for Continuous Queries in Centralized LBS Systems

Hyeong-Il Kim, Mi-Young Jang, Min Yoon and Jae-Woo Chang

**Abstract** Recent development in wireless communication technology and mobile equipment is making location-based services (LBSs) more popular day by day. However, because users continuously send queries to a server by using their exact locations in the LBSs, private information can be in danger. Therefore, a mechanism for users' privacy protection is required for the safe and comfortable use of LBSs. For this, we, in this paper, propose a grid-based cloaking area creation scheme in order to support continuous queries in LBSs. Our scheme creates a cloaking area rapidly by using grid-based cell expansion to efficiently support the continuous LBSs. In addition, to generate a cloaking area which lowers the exposure probability of a mobile user to a minimum level, our scheme computes a privacy protection degree by granting weights to the mobile users. Finally, we show from our performance analysis that our cloaking scheme shows better performance than the existing cloaking scheme.

**Keywords** Privacy protection · Continuous Queries · Cloaking scheme

---

H.-I. Kim · M.-Y. Jang · M. Yoon · J.-W. Chang (✉)  
Department of Computer Engineering, Jeonbuk National University,  
Jeonju, Jeonbuk, South Korea  
e-mail: jwchang@jbnu.ac.kr

H.-I. Kim  
e-mail: melipion@jbnu.ac.kr

M.-Y. Jang  
e-mail: brilliant@jbnu.ac.kr

M. Yoon  
e-mail: myoon@jbnu.ac.kr

## 1 Introduction

Recent development in wireless communication technology and mobile equipment, location-based services (LBSs) become more popular. A location-based service is a service which is accessible by mobile devices through the communication network and utilizing the ability to make use of the geographical position of the mobile device. By using LBS, we can get various services such as finding the nearest Point of Interest (POI) like an ATM or a restaurant, and receiving the warning of traffic jam. However, we must send the exact location information to a LBS server when using these services. However, in this case, users' privacy may be leaked to unauthorized users and illegally used by them. For example, attackers can analyze users' leaked data and identify their life style. To solve these problems, a mechanism for users' privacy protection is required for the safe and comfortable use of LBSs.

There are many existing studies on the cloaking method of  $k$ -anonymity to protect users' privacy. The cloaking method makes a cloaking area, which includes a query issuer and  $k - 1$  other users, while sending a query to LBS server. So, exact location information can be hidden with  $1/k$  leaking probability. However, the existing cloaking methods have a problem when a user continuously request queries. While making a cloaking region for each time, the methods include different group of  $k - 1$  users so that the unauthorized users are able to detect the query issuer by comparing the  $k - 1$  users of successive time frames. To solve the problem, Xu et al. proposed Advanced KAA [1]. Advanced KAA calculates privacy degree of generated cloaking region. But it has two problems. At first, because it considers all candidate areas to generate minimal sized cloaking region, it takes much processing time. Secondly, because the Advanced KAA calculates privacy degree with random sampling of user data, it does not guarantee high similarity between  $k - 1$  other users of current cloaking area and those of previous cloaking area. As a result, the privacy protection level may be reduced.

For this, we, in this paper, propose a grid-based cloaking area creation scheme in order to support continuous queries. Our scheme creates a cloaking area rapidly by using grid-based cell expansion to efficiently support the continuous queries. In addition, to generate a cloaking area which lowers the exposure probability of a mobile user to a minimum level, it computes a privacy protection degree by granting weights to mobile users. The rest of the paper is organized as follows. In Sect. 2, we introduce related works. In Sect. 3, we propose a grid-based cloaking area creation scheme supporting continuous queries. We present our performance analysis in Sect. 4. Finally, we conclude this paper with brief summary and future work in Sect. 5.

## 2 Related Work

The existing cloaking methods [2-5] don't protect user privacy to support continuous queries. While making cloaking region for each time, the methods include a different group of  $k - 1$  users. Therefore, the unauthorized users are able to detect the query issuer by comparing the  $k - 1$  users of successive time frames. There exists only work by Xu and Cai [1] to deal with this kind of problem. They proposed Advanced KAA (K-anonymity Area) method that calculates anonymity degree of newly added users, and finds minimal sized circle for generating the cloaking area with satisfying k-anonymity. Advanced KAA uses entropy for measuring privacy degree. The entropy is the amount of data that is needed to identify the query issuer in a cloaking area. If we assume that A is the cloaking area of the query issuer N, and there exist the group of m users, the entropy of A is measured by expression (1).

$$H(A) = - \sum_{i=1}^m p_i \log p_i \quad (1)$$

Here,  $H(A)$  is the value of entropy for A,  $p_i$  ( $1 \leq i \leq m$ ) is the detecting probability that the selected user is the query issuer N. After calculating the value of entropy for A, we can measure the Anonymity Degree of A by using expression (2).

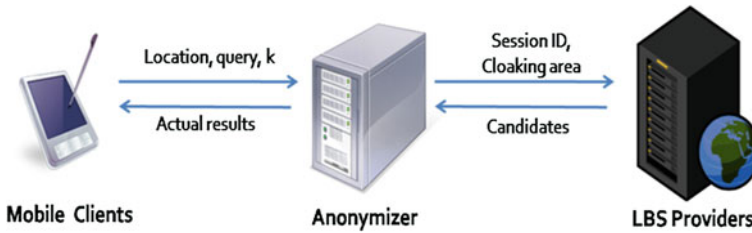
$$D(A) = 2^{H(A)} \quad (2)$$

After  $T = 0$ , we use following way to calculate the detecting probability. At first, we make the transition matrix M with  $\alpha$  number of sample data which includes the information of users' movement. In M, the value in each cell means the number of sample data which is in both previous cloaking area and current cloaking area. The probability of users at  $T = i$ ,  $p_i$ , is calculated by the product of the  $p_{i-1}$  and M.

But this method has a problem of computational overhead to find the minimum circle. With the many mobile objects, which mean the LBS users, the processing time of this method becomes larger with polynomial basis. Therefore, we propose a grid-based cloaking scheme to reduce the cloaking region processing time.

## 3 Grid-Based Cloaking Scheme for Continuous Queries

We use a centralized approach with the trusted third party called an anonymizer which creates the cloaking region. Figure 1 shows the system architecture consisting of three components, a mobile user, anonymizer, and LBS server. The query processing step is as follows. First, a mobile user sends a query to the anonymizer with user location information and he/she periodically updates the location data during service time. Secondly, the anonymizer generates a cloaking



**Fig. 1** System architecture

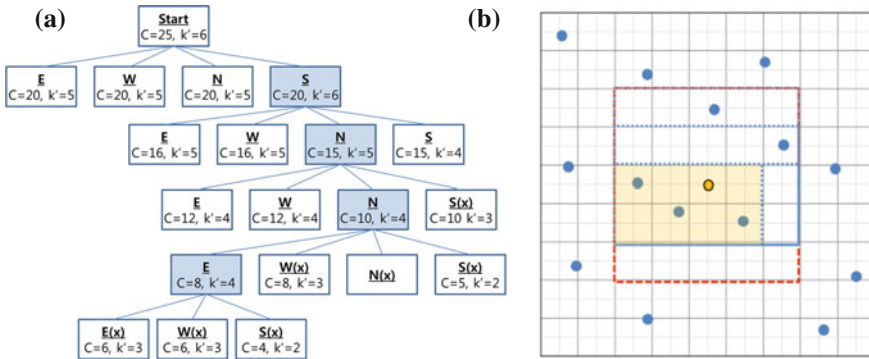
area with  $k - 1$  other users and sends the cloaking region to LBS server with a session ID. Thirdly, LBS server processes the query based on the cloaking region. Finally, the anonymizer filters out the query result based on the exact location of the query issuer and returns the exact result to the user.

On the other hand, we assume that the user location information is sent to LBS server by using GPS system. This is because, while sending the location data, an adversary can get the exact user location so that the privacy of query issuer can be in danger. Here, an adversary means an unauthorized user who leaks and uses the privacy data illegally. In serious cases, the service provider can be an adversary.

### 3.1 Generating the Initial Cloaking Region

Our algorithm first expands cells around the cell which the query issuer is located. During expansion, if the number of mobile users in expanded area ( $=k'$ ) is greater than the value of user given  $k$ -anonymity, it sets the minimal boundary rectangle of the expanded cells as a temporary cloaking region. Here, the algorithm counts the number of cells in temporary area as the value of  $C$ .

Secondly, to generate minimal sized cloaking region, our algorithm sets the initial limitation number of cells to  $C$  and the initial value of  $k$  to  $k'$ . After that, it measures  $C$  and  $k'$  in every cases while deleting some rows or columns for each directions. With this, the scheme can reduce the size of the temporary cloaking region. If the scheme finds an area which has same  $k'$  and less number of cells, it sets the area to the temporary cloaking region and changes the current values of limitations. Or if the scheme finds an area which has same number of cells but has larger  $k'$ , it sets the area to the temporary cloaking region and changes the values of limitations. The algorithm will run until there is no cloaking region which satisfies the value of  $k$ -anonymity. By using this step, we can generate the minimal sized cloaking region and reduce the query processing time in LBS server. For example, Fig. 2 shows the example of this step. In Fig. 2a, by using width-based search information tree, the algorithm finds the minimal sized temporary cloaking region. Here, all nodes in the tree contain the  $C$  and  $k'$  per each cases. Child nodes of each tree node include the information which can be deleted by one row or



**Fig. 2** Example of minimal sized cloaking region. **a** Information tree. **b** Minimal sized cloaking area

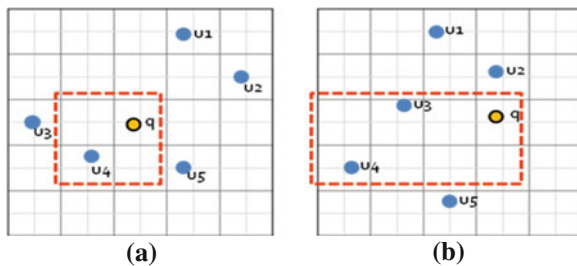
column for 4 directions. As a result, the algorithm finds the minimal sized cloaking area as shown in Fig. 2b (the colored area). Based on the initial cloaking area which was created in the step 1, the algorithm generates a cloaking area which can guarantee the privacy degree during the service time.

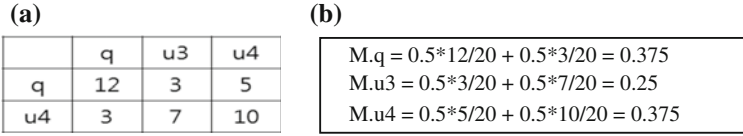
### 3.2 Generating a Cloaking Area with Guaranteeing Privacy Degree

In this step, the algorithm sets the weights for each user in the previous service time. During the service time ( $T$ ), users in cloaking area can move to another places based on the networks. Therefore, at first, the algorithm generates temporal cloaking area ( $TC(i), i = \text{current time}$ ), which contains all the users in the previous cloaking area. For example, Fig. 3a shows the cloaking region of  $T = i$  with  $k = 2$ . The region at  $T = i$  includes  $q$  and  $u_4$ . Then at  $T = i + 1$ , the algorithm generates minimal boundary rectangle  $TC(i + 1)$ . As shown in the Fig. 3b,  $TC(i + 1)$  includes  $q, u_4$  and  $u_3$ .

Secondly, the algorithm calculates the entropy for guaranteeing privacy degree. For this, it makes the transition matrix ( $M$ ). However, a transition matrix in the

**Fig. 3** Example of setting temporal cloaking area. **a**  $T = i$ . **b**  $T = i + 1$





**Fig. 4** Example of calculating user probability. **a** Weight-based transition matrix. **b** User probability

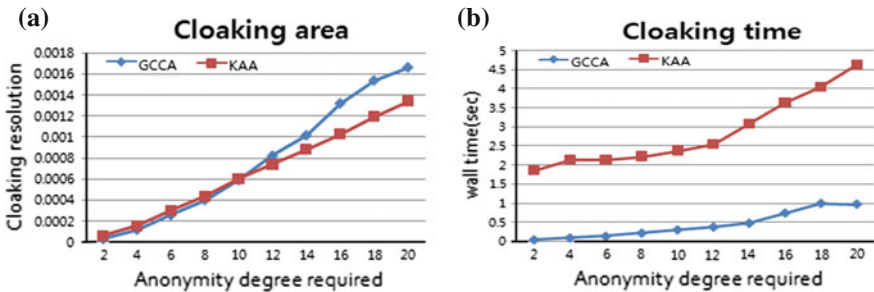
existing Advanced KAA method has a problem that  $M$  does not reflect previous users' information. So in our work, we consider how many times each user in  $TC(i + 1)$  have been a member of previous cloaking regions (i.e. from  $T = 0$  to  $T = i$ ). Based on that, the algorithm creates  $\alpha$  number of samples and makes  $M$ . For example, we assume that  $u4$  and  $q$  in  $TC(i + 1)$  are members of the previous cloaking regions in Fig. 5, and the value of  $\alpha$  equals to 20. Figure 4a shows the transition matrix. The probability of users at  $T = i$ ,  $p_i$ , is calculated by the product of the  $p_{i - 1}$  and  $M$  as shown in Fig. 4b. Here, we assume that the probability of  $q$  and  $u4$  in  $T = i$  are 0.5.

By considering weight of each user, we can increase the probability of previous members. As a result, the privacy degree can be improved. Thirdly, based on the cell containing the query issuer, the algorithm expands the cells for generating a cloaking region when  $T = i + 1$ . Here, the algorithm calculates the entropy by using the formula (1), and then measures the privacy degree by the formula (2). If the measured privacy degree is larger than a given  $k$ -anonymity, the algorithm sets the minimal boundary rectangle to a cloaking area. For example, if the algorithm selects the  $u3$  and  $q$  in Fig. 5b as members of candidate cloaking area, the entropy and the privacy degree can be calculated as follows.

$$H(A) = -(0.375 \times \log 0.375 + 0.25 \times \log 0.25) = 1.03064$$

$$D(A) = 2^{1.03064} = 2.04293$$

Here, the privacy degree is larger than  $k = 2$ , so the algorithm sets the minimal boundary rectangle including  $q$  and  $u3$  to a cloaking region.



**Fig. 5** Performance according to the value of  $k$ . **a** Size of cloaking area. **b** Cloaking time

## 4 Performance Evaluation

In this section, we show the performances of our Grid based Continuous Cloaking Algorithm (GCCA). We implemented GCCA by using MS Visual Studio.NET 2003 running on the Window XP system with 2.20 GHz Intel Core2 Duo CPU and 2 Gb memory. We generated moving objects data by using the network-based moving object generator [6]. We use real road network data of Oldenburg, Germany ( $15 \times 15 \text{ km}^2$ ). For measuring cloaking area easily, we set the total size of the map as 1. We compare our GCCA with the existing advanced KAA (KAA) which was proposed by Xu [1]. We evaluated the performance by changing the anonymity level from 2 to 10. The value of session life time is set to 5, whereas the grid cell size is set to  $1,000 \times 1,000$ .

Figure 5a shows the size of cloaking area according to the value of  $k$ -anonymity. As the value of  $k$  is increased, the size of cloaking areas of both GCCA and KAA is increased. However, it is shown that GCCA generates 15 % larger area than KAA. This is because GCCA includes more users in cloaking region than KAA, by considering the users in the previous cloaking area for continuous query. Figure 5b shows the cloaking time according to the value of  $k$ . The average cloaking times of GCCA and KAA are 0.4273 and 2.8673, respectively. It is shown that GCCA achieves 50 % better performance than KAA. This is because GCCA can reduce the computational overhead by using grid cell expansion.

## 5 Conclusion

In this paper, we propose the grid-based cloaking area creation scheme to support continuous queries for LBSs. For reducing computational overhead while generating cloaking area, we design a grid cell expansion approach. In addition, to guarantee user privacy, we consider the weight-based privacy degree. From our performance analysis, it is shown that our GCCA algorithm is better than the existing scheme, in terms of cloaking time, service time and privacy degree. Meanwhile, both methods show similar performance in terms of the size of the cloaking area. As a future work, we need to enhance our GCCA algorithm to apply it to a distributed computing environment.

**Acknowledgments** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0023800).



## References

1. Xu T, Cai Y (2007) Location anonymity in continuous location-based services. In: ACMGIS, pp 221–238
2. Mokbel M, Chow C, Aref W (2006) The new casper: query processing for location services without compromising privacy. In: Proceedings of the international conference on very large data bases, pp 763–774
3. Xu T, Cai Y (2009) Location anonymity in continuous location-based services. In: Proceedings of ACM conference on computer and communications security, pp 348–357
4. Yang L, Wei L, Shi H, Liu Q, Yang D (2011) Location cloaking algorithms based on regional characteristics. In: Proceedings of the international conference on computer science and automation engineering, pp 93–98
5. Jang M, Chang J (2012) New cloaking method based on weighted adjacency graph for preserving user location privacy in LBS. In: Proceedings of computer science and its applications, pp 129–138
6. Brinkhoff T (2002) A framework for generating network-based moving objects. *GeoInformatica* 6(2):153–180

# New Database Mapping Schema for XML Document in Electronic Commerce

Eun-Young Kim and Se-Hak Chun

**Abstract** This paper considers a relational data system to store XML document efficiently. Also this paper proposes a data model to rewrite XML documents from data storage by representing data view and structure view at the same time and introduce a mapping schema to relational data base system from the data for electronic commerce.

**Keywords** XML document management • Relational database • Multi-format information retrieval • Electronic commerce

## 1 Introduction

Extensible Markup Language (XML) is a simple and flexible markup language. As XML has been used for data transaction in EDI from 1998, it supports all kinds of electronic commerce transaction [1]. There are two methods to map XML documents to RDBMS which are an approach for XML document only and an approach for definition of XML document structure with DTD. The element of DTD is been mapped as a relation of relational database system or as an attribute of relation according to element type, the number of repetition and whether an attribute is contained or not. The attribute of DTD is been mapped as the attribute of relation in the relational database system.

---

E.-Y. Kim

Department of Multimedia Contents, Sin Ansan University, 671 Chosi-dong, Ansan City, Kyunggi-do 425-792, Republic of Korea  
e-mail: key@sau.ac.kr

S.-H. Chun (✉)

Department of Business Administration, Seoul National University of Science and Technology, Kongneung-gil 138, Nowon-gu, Seoul 139-743, Republic of Korea  
e-mail: shchun@seoultech.ac.kr

**Fig. 1** Order element

```
<!ELEMENT order (product,quantity,price)>
<!ATTLIST order no ID #REQUIRED>
<!ELEMENT product (#PCDATA)>
<!ELEMENT quantity (#PCDATA)>
<!ELEMENT price (#PCDATA)>
```

**Fig. 2** Order structure from DTD

```
<order no="120806">
  <product>m234-2t</product>
  <quantity>3</quantity>
  <price>50000</price>
</order>
```

Figure 1 shows order elements in XML document, Fig. 2 shows a structure from DTD. Figure 3 shows RDMS which defines elements of no, product, quantity and price.

The existing method has difficulty to search according to attribute and element and to reproduce original XML document from stored RDBMS. There is a problem that changes a relational structure of RDBMS when attributes in element are added as normalization is needed when there are many duplicate attributes in relational DBMS.

This paper extends Kim and Chun’s [2] model which considers a relational data system to store XML document efficiently. This paper proposes a data model to rewrite XML documents from data storage by representing data view and structure view at the same time and introduce a mapping schema to relational data base system from the data for electronic commerce. The paper is organized as follows. In Sect. 2, we describe the data model used in the study. In Sect. 3, we show the results of our analyses. Section 4 concludes this study.

## 2 The Extended Model Considering DTD

In the basic model, the XML document is represented as a graph that all nodes and edges are labeled and that is ordered among nodes and edges and that edges are directed on. When the XML document is represented as a graph  $G = \{V,E,A\}$  where  $V$  is a vertex, i.e. set of nodes,  $E$ , a set of edges, and  $A$ , a set of attributes defined on start tag of element [2].

no	product	quantity	price
120806	m234-2t	3	50000

**Fig. 3** Mapping results in RDBMS

```

<?xml version="1.0" encoding="euc-kr"?>
<!DOCTYPE purchaseOrder [
  <!ELEMENT purchaseOrder(c_INFO,c_orderList,shipTo)>
  <!ATTLIST purchaseOrder no ID #REQUIRED>
  <!ELEMENT customer ANY>
  <!ATTLIST customer id ID #REQUIRED>
  <!ELEMENT c_PHONE (#PCDATA)>
  <!ELEMENT orderList (order)+>
  <!ELEMENT order (product,quantity,price)>
  <!ATTLIST order no ID #REQUIRED>
  <!ELEMENT product (#PCDATA)>
  <!ELEMENT quantity (#PCDATA)>
  <!ELEMENT price (#PCDATA)>
  <!ELEMENT shipInfo (s_NAME,s_PHONE,s_ADDR1,s_ADDR2,s_MSG)>
  <!ELEMENT s_NAME (#PCDATA|EMPTY)>
  <!ATTLIST s_NAME cID IDREF #IMPLIED>
  <!ELEMENT s_TEL (#PCDATA)>
  <!ELEMENT s_PHONE (#PCDATA)>
  <!ELEMENT s_ADDR1 (#PCDATA)>
  <!ELEMENT s_ADDR2 (#PCDATA|EMPTY)>
  <!ELEMENT s_MSG (#PCDATA)>
]>

```

Fig. 4 An example of purchase order with IDFEF type

DTD means a mutual agreement on the XML document transmitted when XML documents are exchanged. Attributes referring elements in DTD are presented by defining the type of IDREF or IDREFS. Also the value of attributes of IDREF or IDREFS must be identical to the value of the specified property in ID type of

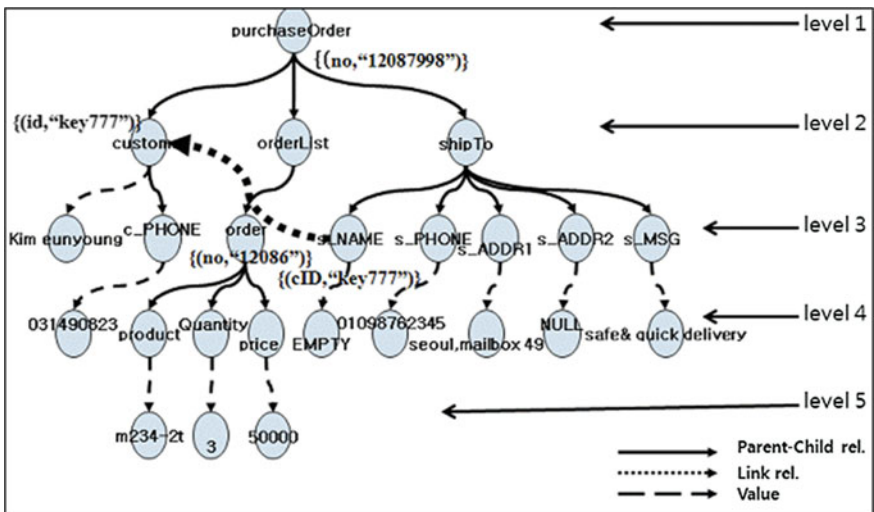


Fig. 5 A graph of the data model with reference information

different type. Figure 4 shows an example of purchase order with IDFEF type. Figure 5 shows the graph of the XML document with reference information between elements by using DTD.

When accompanied by DTD, the XML document can get the reference information between elements from the data of a DTD document. The XML document is extended like below if all the reference edge sets that are added on the graph of the XML document are  $R(G)$ .

$$G = (V, E, R, A)$$

### 3 The Mapping Schema to RDBMS

#### 3.1 Four Objects of XML document

XML documents are represented as a graph. The graph consists of nodes, edges, attributes, reference edges, etc. and each component has its own inherent attributes as an object. When tuples are represented as attributes that contain V, E, A, R, each component object of a graph G, they are:

$$\textit{Vertex Object} = (\textit{label}, \textit{level})$$

$$\textit{Edge Object} = (\textit{from}, \textit{to}, \textit{relation}, \textit{order})$$

$$\textit{Attribute Object} = (\textit{node}, \textit{name}, \textit{value}, \textit{type})$$

$$\textit{Reference Edge Object} = (\textit{refFrom}, \textit{refTo}, \textit{refAttr})$$

This tuple structure offers a unique structure regardless of the data and structure of the XML document.

#### 3.2 Relational Structure

Figure 6 represents the graph G when VID(Vertex ID) is added in Vertex Object tuples, it is represented as follows:

$$\textit{Vertex Object} = (\textit{VID}, \textit{label}, \textit{level})$$

$$\textit{Edge Object} = (\textit{VID of from}, \textit{VID of to}, \textit{relation}, \textit{order})$$

$$\textit{Attribute Object} = (\textit{VID of node}, \textit{name}, \textit{value}, \textit{type})$$

$$\textit{Reference Edge Object} = (\textit{VID of refFrom}, \textit{VID of refTo}, \textit{refAttr})$$

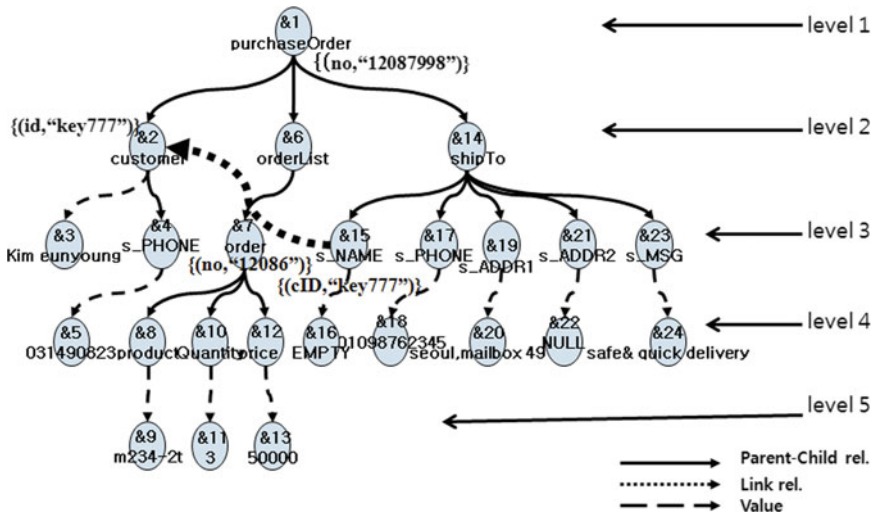


Fig. 6 The data model of the XML document

Figure 7 shows a mapping schema from the proposed data model to RDBMS. This proposed model is an applicable model in the case of designing a new database system for the XML document and in the case of using an existing database system like a relational database system. Because the data view and the structure view of the original XML document are lost in the data model for mapping to the existing relational database system [3-5] or the object-oriented system, not only can't XML generate again from the stored data, but also XML

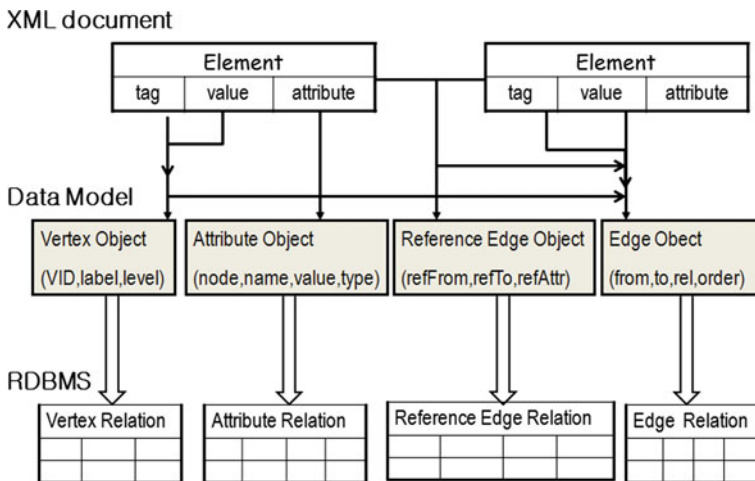


Fig. 7 The mapping schema to RDBMS

sub-graph corresponding to element as search result can't be returned [6]. In the proposed model, this problem is solved because the data view and structure view of the original XML document are stored.

## 4 Conclusion

In the proposed data model defines the data structure, depending on the specific properties because it does not attribute with a value of NULL data does not exist. The graph data model can be mapped naturally into existing RDBMS. Also the proposed model recreates the XML document from stored data when a graph data model is mapped to RDBMS all the data and structure view of the XML source document is stored. For a future study, we expect to do research on speed and accuracy valuation for RDBMS data stored in this way through keyword search queries.

## References

1. Wills B (2009) The business case for environmental sustainability (Green). A 2009 HPS white paper. MicroSoft, "XML Scenarios". <http://msdn.microsoft.com/xml/scenario/inro.asp>
2. Kim E, Chun S-H (2012) New hybrid data model for XML document management in electronic commerce. *Lect Notes Electr Eng* 203:445–451
3. Du F, Sihem A-Y, Freire J (2004) ShreX: managing XML documents in relational databases. In: *Proceedings of the 30th VLDB conference*, Toronto, Canada
4. Kappel G, Kapsammer E, and Retschitzegger W (2000) X-Ray-towards integrating XML and relational database systems. Technical report, July 2000
5. Choi RH, Wong RK (2009) Efficient date structure for XML keyword search. *DASF AA*, pp 549–554
6. Atay M, Chebotko A, Liu D, Shiyong L, Fotouhi F (2007) Efficient schema based XML-to-relational data mapping. *Inf Syst* 32(3):458–476

# A Study on the Location-Based Reservation Management Service Model Using a Smart Phone

**Nam-Jin Bae, Seong Ryoung Park, Tae Hyung Kim, Myeong Bae Lee,  
Hong Gean Kim, Mi Ran Baek, Jang Woo Park, Chang-Sun Shin  
and Yong-Yun Cho**

**Abstract** This paper suggests a location-based reservation management service model that can manage various reservation services and provide related information to users through their smart phones in real-time. The proposed service model is based on a smart phone to relieve inconvenience of existing reservation systems that have limited access effectiveness in space and time or long waiting time. To improve service satisfaction, the proposed service model includes a service server to manage highly qualified and reliable reservation schedules with user's location information from smart phones, and a client to support users to make their application plans through intuitive user interfaces. Therefore, the proposed model can

---

N.-J. Bae · S. R. Park · T. H. Kim · M. B. Lee · H. G. Kim · M. R. Baek · J. W. Park  
C.-S. Shin · Y.-Y. Cho (✉)

Department of Information and Communication Engineering, Suncheon National University,  
Suncheon, South Korea  
e-mail: yycho@sunchon.ac.kr

N.-J. Bae  
e-mail: bakkepo@sunchon.ac.kr

S. R. Park  
e-mail: ghost214@sunchon.ac.kr

T. H. Kim  
e-mail: mcteng@sunchon.ac.kr

M. B. Lee  
e-mail: lmb@sunchon.ac.kr

H. G. Kim  
e-mail: khg\_david@sunchon.ac.kr

M. R. Baek  
e-mail: tm904@sunchon.ac.kr

J. W. Park  
e-mail: jwpark@sunchon.ac.kr

C.-S. Shin  
e-mail: csshin@sunchon.ac.kr



help users to save a long waiting time in various service places, for example food stores, banks, and government office customer service centers.

**Keywords** LSB · Smart phone · Reservation management service

## 1 Introduction

Recently, a reservation service in a customer service center, restaurant and beauty shop becomes a common occurrence. However, because the reservation service is commonly provided through the Internet and a telephone, it is difficult to manage correctly a subscriber's demands and to predict correctly an awaiter's waiting time. Also, if the subscriber does not comply with reservation time or the awaiter comes out from the waiting line without previous notice, a problem in the reservation process may happen. Due to the higher distribution rate of a smart phone, through which location information can be used in real time, efficient location-based service and various reservation services can be provided.

The way to use real-time location information is by GPS and by base station of Smart Phone, which is provided for easy use in development tool of Smart Phone. This service model constructs database of ServiceProvider's location information, and compares user's location of Smart Phone by real-time judgment. Therefore, it is necessary to inscribe service ServiceProvider's location, and for user to install application to Smart Phone or to use QR-Code provided.

In order to make an efficient management of subscriber or awaiter, it is necessary for algorithm to deviate from service location, or to reflect real-time correction for reassignment its order of priority. The proposed service model makes a real-time management of location and correction to reset its order of priority. Therefore, it is necessary to receive real-time location information. And Server needs reservation algorithm for order of priority.

The proposed service model does not use another terminal. It is using location information and technology of Smart Phone with recent usage and distribution rate explosively increased. For this, it is possible to provide reservation service using ads with QR-Code attached to service location providing user's real reservation service, and to propose user's efficient reservation management service model that adjusts subscriber's priority on the basis of real-time location information.

## 2 Related Work

Recently, prevalence of smart phone has been increasing dramatically over the last few years. As a result, various applications (for example: guide traffic, directions etc.) using smart phone is being provided. Above all, one of the most popular services in provided services is QR-Code-based and location-based service.

### 2.1 QR-Code

Quick Response Code (QR code) is the trademark for a type of matrix barcode (or two-dimensional bar code) first designed for the automotive industry in Japan [1]. Bar codes are optical machine-readable labels attached to items that record information related to the item. Initially patented, its patent holder has chosen not to exercise those rights. Recently, the QR Code system has become popular to people due to its fast readability and greater storage capacity compared to standard UPC barcodes [1]. It recently use various places (for example: Business card, Flyers) printed by to publicity. Also, it use to induce access to web page of mobile [2, 3].

One of services QR-Code used is ‘smart phone learning information code scan program’ [4]. ‘Smart phone learning information code scan program’ is program that provides learning information by use camera of smart phone to recognize learning information code (or QR tag). It transfers automatically to screen of learning information code scan after this program activate learning information code scan program. The next can use learning information provide service by transfer to web service that smart phone learning information provided using link information of learning information code.

In this paper, QR-Code technology use to exchange first data between waiting service required place and user, it uses to receive reservation waiting management service (Fig. 1).

### 2.2 Location-Based Services

Location-based services are a general class of computer program-level services used to include specific controls for location and time data as control features in

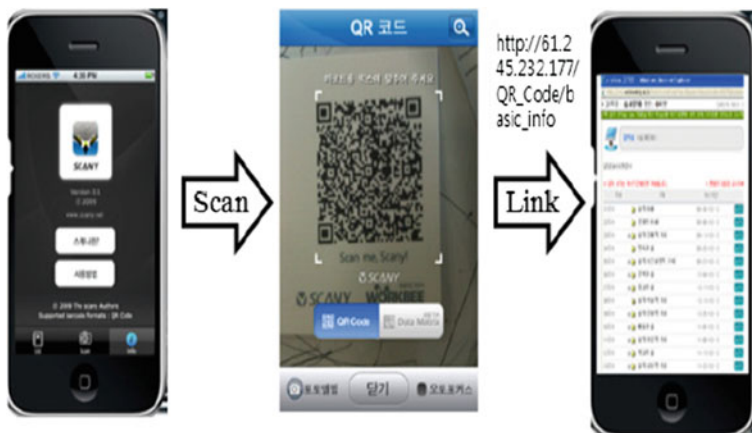


Fig. 1 ‘Daum Map’ service

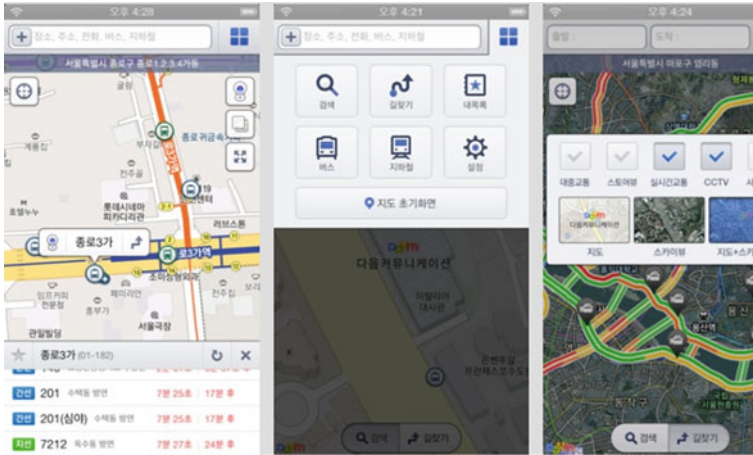


Fig. 2 Moving part of learning information code scan program using smart phone

computer programs. As such (LBS) is an information and has a number of uses in Social Networking today as an entertainment service, which is accessible with mobile devices through the mobile network and which uses information on the geographical position of the mobile device [5]. Recently, many people one of services used is ‘Daum Map’ [6]. When user inputs starting point and destination, ‘Daum Map’ service shows user’s current location, destination and optimal path finding in user’s screen. Also ‘Daum Map’ can choose means of transportation such as walk, subway, taxi, bus. Besides, it gives information such as the fare and the time required for each means of transportation (Fig. 2).

Interworking through geocoding is important, because provided service using location information use a map [7–9]. In this paper, LBS use for calculate required distance and arrival time to reset hard-wired reassignment system in server using location information of smart phone, and use to mark ServiceProvider location or arrival limit area.

### 3 Reservation Management Service Model Based Location Using Smart Phone

#### 3.1 Service Model Construction

This service model is composed by server, service user (Smart Phone), and server has ServiceProvider’s location and waiting management system and priority reassignment system. And service user (Smart Phone) would transfer location information regularly to server collected by geocoding and calculate reaching

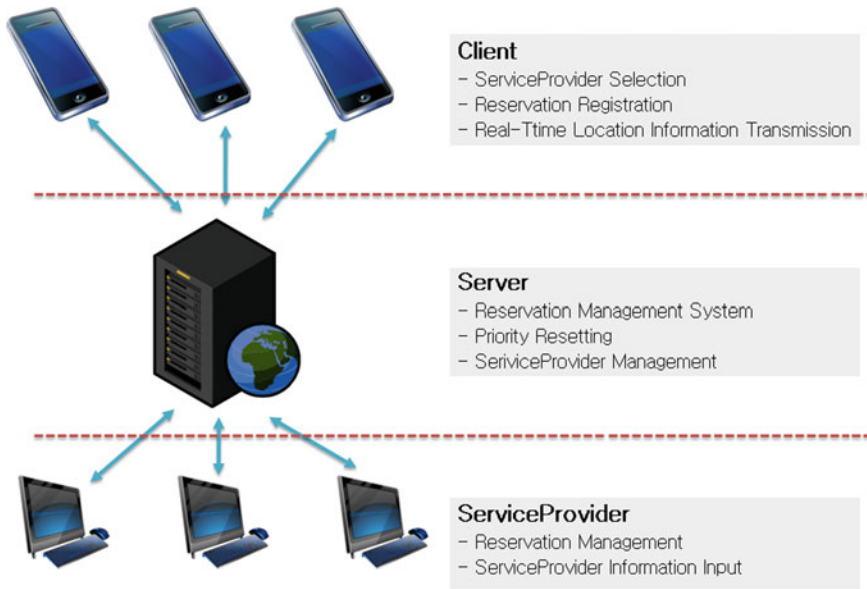


Fig. 3 The overall system configuration diagram

time. Deviation from reaching time range can move back or exclude client of waiting state in waiting lines in reassignment system. Figure 3 represents the whole system block diagram of user reservation management system based on location information using Smart Phone that is proposed by this paper. The proposed system is composed highly by ServiceProvider, server and client.

Client in Fig. 3 represents users available to service with Smart Phone. In this time, user selects his desired ServiceProvider from available ServiceProviders appeared on Smart Phone to user a specific service, and registers it in reservation service. The information of registered user and real-time location information is transferred to server. Server uses user's service register information received from client and real-time location information for reassignment and determining priority of waiting order, and uses the whole service subscribers and service reservation information registered in client to manage state of reservation. Also, Manages the ServiceProvider registration and the basic information. ServiceProvider means restaurant, bank and service center that provides reservation service. And, ServiceProvider can manage subscriber management and reservation service provided by connect to server.

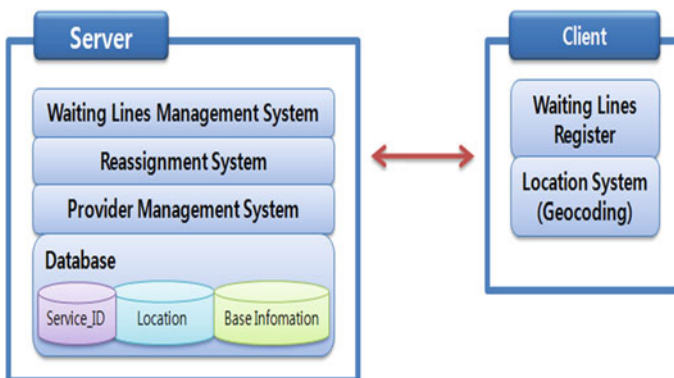
### 3.2 Reservation Management Process Using Location Information

Reservation service model proposed in this paper progresses reservation management between user and a large number of users using real-time location information.

- (1) Using geocoding gather smart phone location information, then transfers to the server.
- (2) Compare entered ServiceProvider location information of database.
- (3) Calculate distance, arrival time etc. in order of priority reset system then prepare to reset the order of priority.
- (4) Transfers order of priority reset to the client.
- (5) Approve User received order of priority reset.
- (6) Reassign the order of priority, then transferred to awaiter management system to adjust the waiting lines.

Figure 4 represents module configuration diagram of each client and location information processing process for reservation management service. This figure includes location information, reservation service and basic information database related system modules.

Server of Fig. 4 is composed highly by waiting lines management system required for the reservation management, assignment and reassignment system for the reservation order of priority, subscriber management system and database to store information of ServiceProvider. Client is composed highly by location-information process system to process location information based geocoding and waiting lines registration module to register each service reservation of user.



**Fig. 4** Server, client configuration diagram

### ***3.3 Client***

The user install application or using separately provided QR-Code by Service-Provider in the client after service start by connect to reservation service. The client provides reservation information including waiting time and a number of awaiter to user. Also, User may receive various benefits including coupon and event information.

### ***3.4 Waiting Management***

User by use information of ServiceProvider and transmitted location in the client judges whether user arrives to the destination or not. If user then arrives to destination, user is bookable. However, if the user is unable to arrive on time and are limited reservation Also, user a high ratio of canceled is expressed to screen of ServiceProvider.

If you need to reset after distinguishing users in reassignment system, then reduce waiting time and waiting lines in reservation management system.

### ***3.5 Order of Priority Assignment and Reassignment***

If User deviates feasible arrival distance on time using location information real-time transmitted from smart phone, then except from awaiter's waiting lines, and then transmit subscriber information to the reservation management system. If user postpones reservation time or cancels reservation, then alike transmit subscriber information to the reservation management system after judge suitable order of priority. Also, If user have a high rate of reservation cancelation or don't input authorization code, then rejudge order of priority.

- Deviated client from arrival limit distance
- Cancel a reservation by client
- Postpone a reservation by client
- Client, a high rate of canceled
- Client Adjusted by ServiceProvider

## **4 Conclusion**

This paper proposed a reservation management service model based on Smart Phone location information technology and can determine the ranking of providing service based on reservation service user's location information and context

information and give an efficient support of waiting management. The proposed service model uses individual Smart Phone and QR-Code technology that has excellent use and access of reservation management service. Also, various applications of reservation service with the proposed service model applied are expected to make an efficient reservation management based on user's location and context information and to raise satisfaction of service user's service quality. The future study will be progressed apt for designing and realizing real reservation management service system according to the proposed service model.

**Acknowledgments** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government. (MEST) (No. 2012-0003026).

## References

1. Wikipedia [http://www.doopedia.co.kr/doopedia/master/master.do?\\_method=view&MAS\\_IDX=101230001172183](http://www.doopedia.co.kr/doopedia/master/master.do?_method=view&MAS_IDX=101230001172183)
2. Lee K (2011) Characteristics of QR code ad and its effects on usage satisfaction and consumers' behavior as a commercial communication tool. *Korean J Advert* 22(3):103–124
3. Park J (2012) A research on expansion of library service by using QR code. *Korean library. Inf Sci Soc* 43(3):1–27
4. Jung W (2010) A design of U-learning study support system. In: 2010 Korea multimedia society autumn conference, vol 13, no 2
5. Wikipedia [http://en.wikipedia.org/wiki/Location-based\\_service](http://en.wikipedia.org/wiki/Location-based_service)
6. 'Daum Map' <http://map.daum.net/>
7. Choi DS (2010) Google API-based expression and utilization plan. The Korean institute of maritime information and communication sciences 2010 autumn conference, pp 672–674
8. Jin K (2008) A study of government policy plans on ubiquitous based lbs services revitalization. Thesis
9. Google, Google Geocoding API. <https://developers.google.com/maps/documentation/geocoding/?hl=ko-KR>
10. Barkhuus L, Dey A (2003) Location-based services for mobile telephony: a study of users' privacy concerns. In: Proceedings of INTERACT 2003, 9th IFIP TC13 international conference on human-computer interaction
11. Gruteser M, Grunwald D (2002) Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the first international conference on MobiSys

# A Real-time Object Detection System Using Selected Principal Components

Jong-Ho Kim, Byoung-Doo Kang, Sang-Ho Ahn, Heung-Shik Kim and Sang-Kyoon Kim

**Abstract** The detection of moving objects is a basic and necessary preprocessing step in many applications such as object recognition, context awareness, and intelligent visual surveillance. Among these applications, object detection for context awareness impacts the efficiency of the entire system and it requires rapid detection of accurate shape information, a challenge specially when a complicated background or a background change occurs. In this paper, we propose a method for detecting a moving object rapidly and accurately in real time when changes in the background and lighting occur. First, training data collected from a background image are linearly transformed using principal component analysis (PCA). Second, an eigen-background is organized from selected principal components with excellent ability to discriminate between object and background. Finally, an object is detected by convoluting the eigenvector organized in the previous step with an input image, the result of which is the input value used on an EM algorithm. An image sequence that includes various moving objects at the same time is organized and used as training data to realize a system that can adapt to changes in

---

J.-H. Kim · H.-S. Kim · S.-K. Kim (✉)  
Department of Computer Engineering, Inje University, Gimhae,  
Gyeongsangnam-do 621-749, Republic of Korea  
e-mail: skkim@inje.ac.kr

J.-H. Kim  
e-mail: luckykjh@daum.net

H.-S. Kim  
e-mail: kimhs@inje.ac.kr

B.-D. Kang  
Researcher, STAR Team, Korea Electronics Technology Institute,  
Bucheon-si, Gyeonggi-do 420-734, Republic of Korea  
e-mail: deweyman@gmail.com

S.-H. Ahn  
Department of Electronic Engineering, Inje University, Gimhae,  
Gyeongsangnam-do 621-749, Republic of Korea  
e-mail: elecash@inje.ac.kr



lighting and background. Test results show that the proposed method is robust to these changes, as well as to the partial movement of objects.

**Keywords** Object detection · Principal Components Analysis (PCA) · Eigen-background · Mixture of Gaussian

## 1 Introduction

Computer vision technology, which was originally developed for human computer interaction, has been applied to a variety of fields such as user interface designs, robot learning, and intelligent surveillance systems. Object detection, which accurately and effectively separates an object from its background, is an essential technology. Without proper foreground/background separation, it would be difficult to detect objects or analyze gestures in the next step of the system such as that required in vision-based robotic manipulation, augmented reality, and gesture recognition.

Therefore, a number of researchers have been studying how to separate an object from its background. Representative methods utilize the difference between a previous frame or a previously saved background and the current frame [1], compressed video information [2], object movement [3], or visual attention [4].

Although methods that utilize difference of images between frames are easy to realize, they are extremely sensitive to illumination changes and are unable to detect objects when there are no moving images or when only a part of an image moves [5]. Methods that use compressed video information do not require a previously saved background and is fast. However, these methods have a downside in that they use different detection methods depending on the technique of compression used [2].

A method based on motion, such as structure from motion, has a fast detection speed, but it also has difficulty in determining the accurate shapes of objects and it is also sensitive to changes in illumination. Finally, methods that use a visual attention model present fast detection speed but they have difficulty in extracting meaningful object contours as well as accurate shape information.

Eigenspace is a technique that can overcome most of these problems. The basic idea behind this method is to analyze the background information obtained from training images and then to construct an eigen-background to recall the expected background. This eigen-background can be used to separate an object from its background, allowing the detection of non-moving objects. The improved eigenspace models presented in [6, 7] adapt the background after the formation of the initial eigen-background. That is, these models continue to learn the eigen-background while inputting new images as training data. However, methods using an eigen-background usually separate a foreground object using the eigenvalues and corresponding eigenvectors with a high explanation rate. Therefore, using these

methods lead to background noise to be detected as an object. In [6], on the other hand, this adaptation is carried out using a synthetic background obtained from the current image and eigen-background after the removal of any foreground object. Also in [6], the dimension of the eigenspace—i.e. the number of eigenvalues and corresponding eigenvectors selected—is kept to an optimum number by applying a clustering algorithm that classifies images based on their illumination conditions. Unfortunately, the criterion to select eigenvalues employed by the method in [6] is still based on the largest eigenvalues.

In this paper, we propose a robust object detection system that addresses most of these problems. In the heart of the proposed method is the construction of an advanced eigenspace that is able to capture more effective information about the background, especially under different lighting conditions and background changes. This step is achieved by employing a clustering algorithm that selects the eigenvalues based on their “power of explanation”, rather than their numerical value.

First, we construct training data using images, including moving people, chairs, etc. We can then use these images to create eigen-backgrounds that are accommodative to background changes. Next, eigen-backgrounds are constructed after the background information has been analyzed through PCA [8]. Next, an eigen-backgrounds that distinguishes an object clearly from the background by using a clustering algorithm is selected. Next, an object is detected by convoluting the eigenvector organized in the previous step with an input image, the result of which is the input value of an EM algorithm.

An EM clustering algorithm, which is an unsupervised learning method (i.e., it does not require a user’s artificial goal value), is used for the detection method. A mixture of Gaussians (MOG) and fuzzy c-means (FCM) [9, 10] are used as an EM algorithm. An object detection system that is robust to changes in illumination is realized using a combination of the EM algorithm and the eigenspace.

## 2 Object Detection System

### 2.1 Overview

Figure 1 shows the main structure of our object detection system. First, the system uses various background images as PCA training data. It analyzes the training-data-set information using PCA and then selects the eigenvalue that properly distinguishes moving objects from their backgrounds. The system then constructs an eigen-background using a selected eigenvalue and uses this background as input data for clustering. The system deducts the resultant value achieved by multiplying the input image and the eigen-background based on the pixel unit. Finally, the system detects moving objects from images clustering based on the resultant value obtained using MOG.

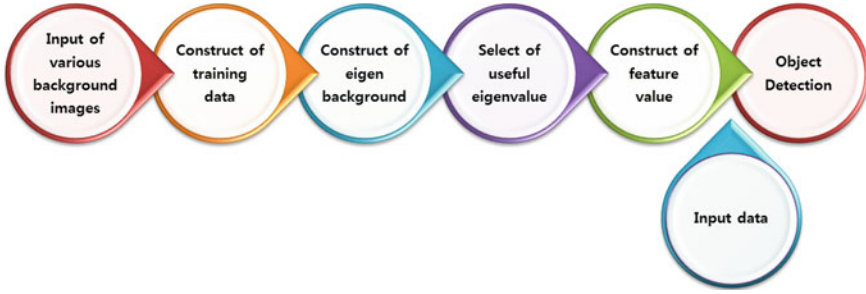


Fig. 1 Main structure for object detection system

### 2.2 Eigen-Background

In this paper, we make an eigen-background using PCA to detect moving objects from the background in an image.

First, the system acquires background samples that are used as training data to construct an eigen-background. Second, the system analyses the training data using PCA and generates an eigen-background after extracting principal components of the background.

The system then changes 2D images into 1D line vector in order to use the training data as PCA input images.

$$S = [I_1, I_2 \dots I_M] = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{bmatrix}, \quad I_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{Nj} \end{bmatrix}, \quad (1 \leq j \leq M) \quad (1)$$

$S$  is an training data images,  $j$  is the index of an image existing within the entire set.  $N$  is the amount of training data, and  $N$  is the number of features.

In this paper, the number of features  $N$  is 76,800 because we use  $240 \times 320$  images. The average of training data is calculated using the following Eq. (2):

$$\Psi_i = \frac{1}{M} \sum_{j=1}^M x_{ij}, \quad (1 \leq i \leq N, 1 \leq j \leq M) \quad (2)$$

$$\Psi = [\Psi_1, \Psi_2, \Psi_3, \dots, \Psi_N]^T$$

where  $x_{ij}$  is an  $i$  order pixel value in  $I_j$ . We use 100 training images in this paper. The deviation is presented as shown in Eq. (4) through determinant (3).

$$\Phi_j = I_j - \Psi \quad (3)$$

$$A = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \cdots & \Phi_{1M} \\ \Phi_{21} & \Phi_{22} & \cdots & \Phi_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{N1} & \Phi_{N2} & \cdots & \Phi_{NM} \end{bmatrix} = [\Phi_1, \Phi_2, \dots, \Phi_M] \quad (4)$$

We create a covariance matrix to ascertain the eigenvalue and eigenvector. Covariance matrix  $C$  is obtained using the following equation:

$$C = \frac{1}{M} \sum_{j=1}^M \Phi_j \Phi_j^T \quad (5)$$

$$= AA^T$$

$C$  is an  $N \times N$  dimension matrix. The number of particular dimensions of this matrix  $N$  is 76,800 for images. It is difficult to calculate 76,800 eigenvalue  $S$  and their corresponding eigenvectors uses a  $76,800 \times 76,800$  dimension covariance matrix. Therefore, the system calculates the eigenvalue and eigenvector in  $M \times M$  dimensions considering an eigenvector on  $A^T A$ .

$$AA^T v_j = \lambda_j v_j \quad (6)$$

When both sides are multiplied by  $A^T$

$$A^T AA^T v_j = \lambda_j (A^T v_j) \quad (7)$$

$A^T v_j$  becomes eigenvector  $v'_j$  for  $A^T A$ . This matrix has  $M \times M$  dimensions.

$$A^T v_j = v'_j, v_j = Av'_j \quad (8)$$

The eigenvector  $v_j$ , of an  $M \times N$  matrix is calculated using Eq. (8).

An  $M$  numbered eigenvalue  $\lambda_j$  and  $N$  numbered eigenvector corresponding to each eigenvalue are determined using the a covariance matrix  $C$ .

$$B_j = \frac{1}{M} \sum_{k=1}^M v_j \Phi_k^T \quad (j = 1, 2, \dots, M) \quad (9)$$

$B_j$  is an eigen-background that has eigenvalue order  $j$ , and  $v_j$  is an eigenvector that has a  $j$  order eigenvalue.  $B_j$  is determined as eigenvector  $v_j$  that has an eigenvalue order  $j$  and a deviation of  $\Phi^T$ .

A general eigen-background method uses an eigenvalue that is arranged in descending order and satisfies Eq. (10).

$$\frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^T \lambda_j} \geq th (=0.9) \quad (10)$$

$T$  is the number of eigenvalues, creates an eigen-background using  $m$ -numbered eigenvalues that set the threshold to  $th$  (0.9), and reflects an explanation rate of up

to 90 %. The eigenvalue that has the highest account rate is not suitable for separating an object from its background and has many noisy elements. Although  $th$  (threshold) is set above 0.9, the separation of the object from its background is not clear because the first eigenvalue is included.

$$B = \frac{1}{m} \sum_{j=1}^m B_{lj}, \quad (1 \leq l \leq m, 1 \leq l \leq N) \quad (11)$$

$$D_j = |I_j \times B| \succ t \quad (12)$$

If an Input image has a higher value than threshold  $th$ , it is an object; if not, it is the background. To construct an eigen-background, the system analyzes the components of the training data.

### 2.3 Object Detection Using EM Algorithm

In this paper, we use a clustering algorithm to determine an eigenvalue that is useful for detecting objects from their backgrounds. We use mixture of Gaussians among clustering algorithms.

#### 2.3.1 MOG (Mixture of Gaussian)

An MOG is a density estimation method used for improving the method of modeling the distribution density of a sample data set as one probability density function. It models the distribution of data using a number of Gaussian probability density functions.

A complete Gaussian probability density function defined as a linear combination of  $K$  number of Gaussians is represented as Eq. (13).

$$f(D_t = u) = \sum_{i=1}^K \omega_{i,t} \cdot \eta(u; \mu_{i,t}, \sigma_{i,t}) \quad (13)$$

where  $\eta(u; \mu_{i,t}, \sigma_{i,t})$  is the  $i$ th Gaussian component with intensity mean  $\mu_{i,t}$  and standard deviation  $\sigma_{i,t}$  and  $\omega_{i,t}$  is the weight based on the  $i$ th component. Typically,  $K$  ranges from three to five depending on the available storage, and here, we use three as the value for  $K$ .

Equation (14) calculates the weight ( $\omega_{i,t}$ ), average ( $\mu_{i,t}$ ), and standard deviation ( $\sigma_{i,t}$ ) initiated when an image is input.

$$\left| D_{i,t} - u_{i,t-1} \right| \leq T \cdot \sigma_{i,t-1} \quad (14)$$

In the above equation, if a deviation in the absolute value is lower than or the same as the standard deviation,  $T$ , which controls the threshold standard deviation rate, becomes 2.5. The next values are renewed if they are lower than threshold  $T$  times the standard deviation. In this paper, 2.5 is used as the threshold  $T$  value.

$$\begin{aligned}\omega_{i,t} &= (1 - \alpha) \omega_{i,t-1} + \alpha \quad \mu_{i,t} = (1 - \rho) \mu_{i,t-1} + \rho D_t \\ \sigma_{i,t}^2 &= (1 - \rho) \sigma_{i,t-1}^2 + \rho (D_t - \mu_{i,t})^2\end{aligned}\quad (15)$$

In the above equation,  $\rho$  is the training rate and is calculated using Eq. (15).

$$\rho \approx \frac{\alpha}{\omega_{i,t}}\quad (16)$$

Here,  $0 \leq \alpha \leq 1$  the user's defining training rate.

If the value in Eq. (14) is higher than the standard deviation which is the modulus deviation  $T$  times the threshold, the next expression is renewed.

$$\omega_{i,t} = (1 - \alpha) \omega_{i,t-1}\quad (17)$$

Finally, an object is separated from its background using the following expression.

$$\sum_{k=i_1}^{i_M} \omega_{k,t} \geq \Gamma\quad (18)$$

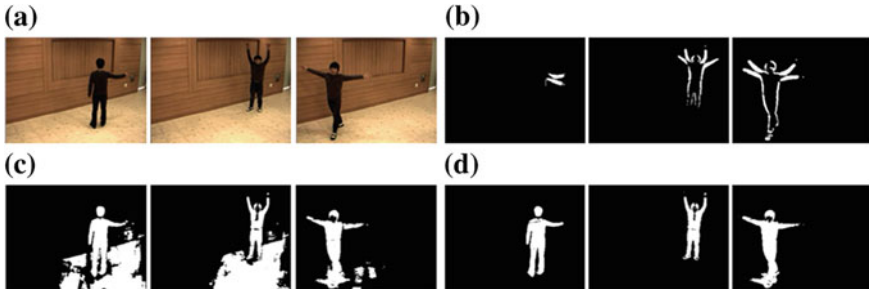
When the total weight of a pixel is more than  $\Gamma$ , it is identified not as an object but an background.

### 3 Experimental Results

The environment for our experiment was Visual C++ on a 3.4 GHz Intel Pentium with Dual CPU, 2 GB RAM and Windows operating system. The proposed method is experimental to image sequences that were captured from IJUData (indoor and outdoor environment with various changes in constituents and illumination, using an HVR-2030 webcam).

#### 3.1 Method Based on Difference Images and Comparison Test

Figure 2 shows the results of a comparison test between a method using difference images and a general eigen-background method and the method proposed in this paper when frames such as those in Fig. 2a are input. Figure 2b show the result of



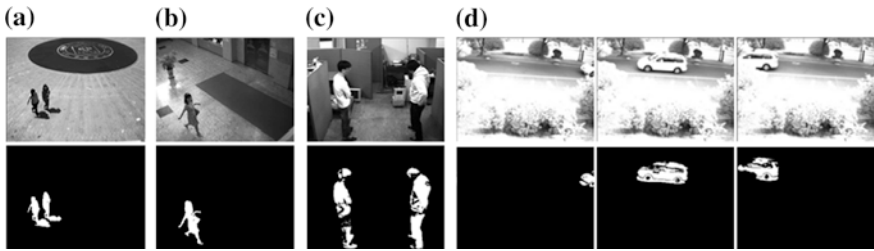
**Fig. 2** **a** Input images, **b** difference image, **c** result of general eigen-background methods and **d** images obtained using proposed method

using a difference image that is different with the previous frame. As you can see in the figure, a body part with no movement is not detected, and only a moving hand is detected. Figure 2c are the results of detection using a general eigen-background. As you can see in the figure, because the principal components have a high explanation rate with lots of noise and shadows, the noise and shadows are detected from the background. Figure 2d show a method using an eigen-background that is improved through the use of the clustering proposed in this paper. No moving body parts are detected accurately, as can be seen in the figure. Because the proposed method in this paper selects a useful eigenvalue that separates a background and object, it does not leave after images. It also removes noise and shadows, and detects an object robustly.

### ***3.2 Experiment on Change in Number of Objects Depending on Various Backgrounds and Lighting Conditions***

To analyze the efficiency of the proposed system, we experimented on changing the number of objects according to various backgrounds and lighting conditions. Figure 3a show the results of detecting two passing pedestrians passing under natural outdoor lighting. Figure 3b are the results of an experiment in a hallway under mixed natural and indoor lighting conditions. Figure 3c show the results of an experiment done under indoor lighting with a complicated background. Figure 3a shows that an object is detected well under natural light but is recognized incorrectly as a shadow under robust light. Figure 3b shows that under a mixture of natural and indoor lighting, the system detects objects without shadows. Figure 3c shows the result of detecting an object from a large number of other objects with a complicated background. Although many parts of the correct object in Fig. 3a are similar to the background region, the object region is correctly detected.

Figure 3d shows the object detection results for a situation in which the colors of the cars and roads are not distinguishable due to a bright natural light. According to our results, the improved eigen-background proposed in this paper



**Fig. 3** Experimental results obtained using various backgrounds and natural lighting

separates light and noise from a background efficiently based on the environment, shadow conditions under different lighting, and a non-moving object. The method proposed in this paper solves the problem of an after image, and the inability to detect non-moving objects. It also does not require artificial background initialization and detects objects from noise and shadows more strongly than a method using a general eigen-background.

## 4 Conclusion

In this paper, we constructed an improved eigen-background to detect an object from its background. First, we constructed various background images as training data to detect an object adaptively under changes in the environment. We then analyzed the training data using PCA and selected the principal components used in analyzing an object from its background using a clustering algorithm.

We solved the problems in which a complicated background is recognized incorrectly as an object using a principal component, which analyzes a background and an object well without depending on the explanation rate and is the traditional method used to analyze the main elements in object detection. The main selected component is used as an eigen-background and is input to the MOG. As a result of convoluting an eigenvector and an input image used as an input value, the existing object detection system using MOG responds sensitively to changes in light. This reduces errors such as noise generated in an image or when an object is recognized as a background when it has little or no movement.

## References

1. Yyilmaz A, Javed O, Shah M (2006) Object tracking: a survey. *ACM J Comput Surv* 38:1–45
2. Stein AN, Herbert M (2009) Local detection of occlusion boundaries in video. *Image Vis Comput* 27:514–522
3. Horn BKP, Schunck NG (1981) Determining optical flow. *Artif Intell* 17:185–203



4. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans PAMI* 20:1254–1259
5. MCHugh JM, Konrad J, Saligrama V, Jodoin PM (2009) Foreground-adaptive background subtraction. *IEEE Signal Process Lett* 16:390–393
6. Rymel J, Renno J, Greenhill D, Orwell J, Jones GA (2004) Adaptive eigen-backgrounds for object detection. In: *International conference on image processing*, vol 3, pp 1847–1850
7. Zhang J, Zhuang Y (2007) Adaptive weight selection for incremental eigen-background modeling. In: *International conference on multimedia and expo*, pp 851–854
8. Turk MA, Pentland AP (1991) Face recognition using eigenfaces. In: *IEEE conference on computer vision and pattern recognition*. pp 586–591
9. Cheung SCS, Kamath C (2005) Robust techniques for background subtraction in urban traffic video. *J Appl Signal Process* 1:2330–340
10. Yang MS, Wu KL, Hsieh JN, Yu J (1999) Alpha-cut implemented fuzzy clustering algorithms and switching regressions. *IEEE Trans Syst* 18:1117–1128

# Trajectory Calculation Based on Position and Speed for Effective Air Traffic Flow Management

Yong-Kyun Kim, Deok Gyu Lee and Jong Wook Han

**Abstract** Trajectory modeling is basic work for 4D-Route modeling, conflict detection and air traffic flow management. This paper proposes a novel algorithm based on coordinate prediction for trajectory calculation. We demonstrated through simulations with flight position and ground speed, and experimental results show that our-trajectory calculation exhibits much better performance in accuracy.

## 1 Introduction

The air traffic management (ATM) system improves safety and efficiency of air traffic by preventing collisions against other aircraft, obstacles and managing aircraft's navigation status. To achieve these purpose the air traffic control system identifies aircraft and displays its location, displays and distributes flight plan data, provides flight safety alerts, and processes controller's requests.

Despite technological advances in air navigation, communication, computation and control, the ATM system is still, to a large extent, built around a rigidly structured airspace and centralized, mostly human-operated system architecture.

The accuracy route calculation in En-route airspace impacts ATM route predictions and Estimated Times of Arrival (ETA) to control fix points. For the airspace controllers, inaccurate trajectory calculation may results in less-than-optimal maneuver advisors in response to a given traffic management problem [1].

---

Y.-K. Kim (✉) · D. G. Lee · J. W. Han

Electronics and Telecommunications Research Institute, 161 Gajeong-dong,  
Yuseong-gu, Daejeon, Korea  
e-mail: ykkim1@etri.re.kr

D. G. Lee

e-mail: deokgyulee@etri.re.kr

There has been significant research in the fields of air traffic flow management and basic trajectory modeling. With the rapid development of air traffic management technology, more and more aircraft would fly in the sky simultaneously [2].

One key factor in route modeling is the uncertainty in present and future estimations of the velocity and position vectors of aircraft.

Many route calculation algorithm account for this uncertainty [3]. On the other hand, only limited research has been done on stochastic route calculation. Given the limited literature on route calculation under uncertainty, some studies concerned with developing probabilistic route calculation model s conclude by stating that there is a need to better understand and utilize route probability estimations in route calculation algorithm [4, 5].

In this paper we propose efficient En-route trajectory calculation algorithm. The remainder of this paper is structured as follows. In the next section, trajectory calculation techniques and theoretical background about trajectory calculation is presented. We present some experimental results of our proposed scheme in Sect. 3, and finally conclude with the conclusion in Sect. 4.

These uncertainties may be due to sensor noise or due to unpredictable disturbances such as wind.

## 2 Trajectory Calculation Techniques

### 2.1 Trajectory Calculation Theory

First, consider a fairly simplified model for trajectory calculation problem. A flow is defined as a set of flights between a departure airdrome and an arrival airdrome. The following simplifications are made (Fig. 1).

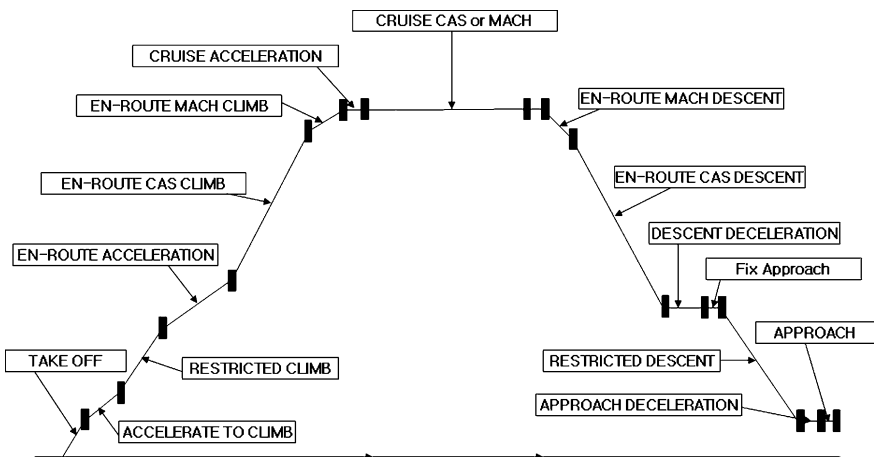


Fig. 1 Typical model of a trajectory calculation modeling

The airspace is considered as an Euclidean space, where all airdrome are at altitude 0. Latitudes and longitudes on the ellipsoid earth surface are converted into (x, y) coordinates by a stereographic projection, and the altitude in feet shall be our z coordinate [4].

### 2.2 Parameters for Trajectory Calculation

For trajectory calculation, we must consider about concept of speed and conversion between other speeds, speed variation by altitude changing and wind parameter.

Fir, airspeed is the speed of an trajectory calculation relative to the air. Among the common conventions for qualifying are: Indicated Airspeed (IAS), Calibrated Airspeed (CAS), True Airspeed (TAS) and Ground speed (GS).

IAS is the airspeed indicator reading uncorrected for instrument, position, and other errors. From current European Aviation Safety Agency (EASA) definitions: Indicated airspeed means the speed of an aircraft as shown on its pitot static airspeed indicator calibrated to reflect standard atmosphere adiabatic compressible flow at sea level uncorrected for airspeed system errors.

Most airspeed indicators show the speed in knots (i.e. nautical miles per hour). Some light aircraft have airspeed indicators showing speed in miles per hour.

CAS is indicated airspeed corrected for instrument errors, position error and installation errors.

CAS is CAS values less than the speed of sound at standard sea level (661.4788 kn) are calculated as follows:

$$V_C = A_0 \sqrt{5 \left[ \left( \frac{Q_C}{P_0} + 1 \right)^{\frac{2}{\gamma}} - 1 \right]} \tag{1}$$

where

$V_C$  is the CAS.

$Q_C$  is the impact pressure sensed by the pitot tube.

$p_0$  is 29.92126 inches Hg at standard sea level.

$A_0$  is 661.4788 kn; speed of sound speed at standard sea level.

This expression is based on the form of Bernoulli’s equation applicable to a perfect, compressible gas. The values  $P_0$  and  $A_0$  are consistent with the International Standard Atmosphere (ISA).

TAS is the physical speed of the aircraft relative to the air surrounding the aircraft. The true airspeed is a vector quantity. The relationship between the true airspeed ( $V_t$ ) and the speed with respect to the ground ( $V_g$ ) is

$$V_t = V_g - V_w \tag{2}$$

where

$V_w$  is wind speed vector.

Aircraft flight instruments, however, don't compute true airspeed as a function of groundspeed and wind speed. They use impact and static pressures as well as a temperature input. Basically, true airspeed is calibrated airspeed that is corrected for pressure altitude and temperature. The result is the true physical speed of the aircraft plus or minus the wind component. True Airspeed is equal to calibrated airspeed at standard sea level conditions.

The simplest way to compute true airspeed is using a function of Mach number

$$V_t = A_0 \cdot M \sqrt{\frac{\tau}{\tau_0}} \quad (3)$$

where  $M$  is Mach number,  $\tau$  is Temperature (kelvins) and  $\tau_0$  is Standard sea level temperature (288.15 K)

Second, speed variation by altitude changing means that when aircraft are climb or descent.

The rate of climb (RoC) is the speed at which an aircraft increases its altitude. This is most often expressed in feet per minute and can be abbreviated as ft/min. Else where, it is commonly expressed in meters per second, abbreviated as m/s. The rate of climb in an aircraft is measured with a vertical speed indicator (VSI) or instantaneous vertical speed indicator (IVSI). The rate of decrease in altitude is referred to as the rate of descent or sink rate. A decrease in altitude corresponds with a negative rate of climb.

There are two airspeeds relating to optimum rates of ascent, referred to as  $V_x$  and  $V_y$ .

$V_x$  is the indicated airspeed for best angle of climb.  $V_y$  is the indicated airspeed for best rate of climb.  $V_x$  is slower than  $V_y$ .

Climbing at  $V_x$  allows pilots to maximize the altitude gain per unit ground distance. That is,  $V_x$  allows pilots to maximize their climb while sacrificing the least amount of ground distance. This occurs at the speed for which the difference between thrust and drag is the greatest (maximum excess thrust). In a jet airplane, this is approximately minimum drag speed, or the bottom of the drag vs. speed curve. Climb angle is proportional to excess thrust.

Climbing at  $V_y$  allows pilots to maximize the altitude gain per unit time. That is,  $V_y$ , allows pilots to maximize their climb while sacrificing the least amount of time. This occurs at the speed for which the difference between engine power and the power required to overcome the aircraft's drag is the greatest (maximum excess power). Climb rate is proportional to excess power.

$V_x$  increases with altitude and  $V_y$  decreases with altitude.  $V_x = V_y$  at the airplane's absolute ceiling, the altitude above which it cannot climb using just its own lift.

Last, we consider about wind parameters. Wind parameter can divide two components (weather fronts and thermal wind) on a large scale.

Weather fronts are boundaries between two masses of air of different densities, or different temperature and moisture properties, which normally are convergence zones in the wind field and are the principal cause of significant weather. Within surface weather analyses, they are depicted using various colored lines and symbols.

The air masses usually differ in temperature and may also differ in humidity. Wind shear in the horizontal occurs near these boundaries. Cold fronts feature narrow bands of thunderstorms and severe weather, and may be preceded by squall lines and dry lines.

Cold fronts are sharper surface boundaries with more significant horizontal wind shear than warm fronts. When a front becomes stationary, it can degenerate into a line which separates regions of differing wind speed, known as a shear line, though the wind direction across the feature normally remains constant. Directional and speed shear can occur across the axis of stronger tropical waves, as northerly winds precede the wave axis and southeast winds are seen behind the wave axis.

Horizontal wind shear can also occur along local land breeze and sea breeze boundaries.

Thermal wind is a meteorological term not referring to an actual wind, but a difference in the geostrophic wind between two pressure levels  $p_1$  and  $p_0$ , with  $p_1 < p_0$ ; in essence, wind shear. It is only present in an atmosphere with horizontal changes in temperature.

In a barotropic atmosphere, where temperature is uniform, the geostrophic wind is independent of height. The name stems from the fact that this wind flows around areas of low (and high) temperature in the same manner as the geostrophic wind flows around areas of low (and high) pressure.

$$f_{VT} = K \times \nabla(\phi_1 - \phi_0) \quad (4)$$

where the  $\phi_x$  are geopotential height fields with ( $\phi_1 > \phi_0$ ),  $f$  is the Coriolis parameter, and  $K$  is the upward-pointing unit vector in the vertical direction. The thermal wind equation does not determine the wind in the tropics. Since  $f$  is small or zero, such as near the equator, the equation reduces to stating that  $\nabla(\phi_1 - \phi_0)$  is small. This equation basically describes the existence of the jet stream, a westerly current of air with maximum wind speeds close to the tropopause which is (even though other factors are also important) the result of the temperature contrast between equator and pole.

### 3 Experimental Results of Proposed Scheme

This section describes the method for computing the various parameters used to compute the position, speed and our route calculation algorithm.

First of all, we need aircraft's position and its speed. Aircraft's position consists of latitude and longitude.

For calculating aircraft's position at specified time, we need airspeed (TAS or GS), wind speed and wind direction.

```

D:\Route\Debug>Route
Input Start Fix Name : SEL

Fix : SEL
1265542E 372449N

Input Stop Fix Name : BELMI

Fix : BELMI
1265929E 371249N

WindSpeed(Knot) : 15
Wind Direction(WindFrom, Degree) : 37
True Air Speed(Knot) : 600

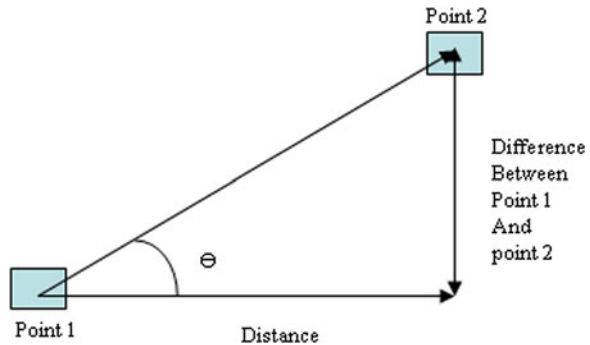
Two Fix Point Distance is 22.886766 Km < 75087.814865 ft >
Tracking Angle is 2.895073 Degree

Ground Speed is 991.723011 < ft/sec >
Estimate Time is 75.714503 < sec >

    0 Sec    372447N 1265540E
   15 Sec    372225N 1265626E
   30 Sec    372003N 1265711E
   45 Sec    371740N 1265755E
   60 Sec    371518N 1265840E
   75 Sec    371254N 1265925E
   75 Sec    371247N 1265927E
    
```

Fig. 2 Result of position prediction algorithm simulation

Fig. 3 Method for calculating different altitude between point 1 and point 2



Aircraft's position can calculate using vincenty's formula at specified time. For using position prediction algorithm, predict aircraft's position is shown in the Fig. 2.

And now we consider about altitude for increase accuracy (Fig. 3).

By computing climb rate, we can calculate aircraft's position in terms of other altitude.

```

D:\#Route\#Route>Climb
Input Start Fix Name : SEL
SEL 's Altitude <ft>: 4000

Fix : SEL
1265542E 372449N
Altitude : 4000.000000 <ft>

Input Stop Fix Name : BELMI
BELMI 's Altitude <ft>: 6000

Fix : BELMI
1265929E 371249N
Altitude : 6000.000000 <ft>

WindSpeed<Knot> : 37
Wind Direction<WindFrom, Degree> : 157
True Air Speed<Knot> : 700

Two Fix Point Distance is 22.886766 Km < 75087.814865 ft >
Tracking Angle is 2.895073 Degree

Ground Speed is 1237.226827 < ft/sec >
Estimate Time is 60.690419 < sec >

    0 Sec      372447N 1265540E
   15 Sec      372149N 1265637E
   30 Sec      371851N 1265733E
   45 Sec      371553N 1265829E
   60 Sec      371255N 1265925E
   60 Sec      371247N 1265927E
    
```

Fig. 4 Result of climb late compute algorithm simulation

Fig. 5 Result of trajectory calculation using position and speed

```

Route Information...
BIGOB -> GOTLO -> BULLS -> KAKSO -> SEL
Time : 250

Aircraft Position 250.000000 Sec
BIGOB -> GOTLO -> BULLS -> 351640N 1283610E
    
```

Climb rate is calculated as below:

$$\text{Climb rate} = \text{TAS} \times \sin \theta \tag{5}$$

Using climb rate, trajectory calculation results in shown in Fig. 4.



With the position and speed, the trajectory of the aircraft can be derived, which is shown in Fig. 5.

## 4 Conclusion

In this paper, we propose trajectory prediction to accurately predict and calculate trajectory a maneuvering aircraft. For increase estimation, we consider wind speed, wind direction and altitude.

From now on, it is further suggested that the proposed algorithm may be extended to the trajectory modeling, which may further improve 4-D trajectory prediction.

**Acknowledgments** This research was supported by a grant (07항공-항행-03) from Air Transportation Advancement Program funded by Ministry of Land, Transport and Maritime affairs of Korean government.

## References

1. Banavar S, Grabbe SR, Mukherjee A (2008) Modeling, optimization in traffic flow management. *Proc IEEE* 96(12):2060–2080
2. James KK, Lee CY (2000) A review of conflict detection and resolution modeling methods. *IEEE Trans Intell Transp Syst* 1(4):179–189
3. Adan EV, Erwan S, Senay S (2009) A two-stage stochastic optimization model for air traffic conflict resolution under wind certainty. In: 28th digital avionics systems conference 2009, pp 2.E.5-1–2.E.5-13
4. Lin X, Zhang J, Zhu Y, Liu W (2008) Simulation study of algorithm for aircraft trajectory prediction based on ADS-B technology. In: Asia simulation conference—7th international conference on systems simulation and scientific computing, pp 322–327
5. Terence SA (2007) A trajectory algorithm to support en route and terminal area self-spacing concepts, NASA/CR-2007-214899

**Part VII**  
**Multimedia Entertainment**

# Design and Implementation of a Geometric Origami Edutainment Application

ByeongSu Kim, TaeHun Kim and JongHoon Kim

**Abstract** Edutainment lies at the intersection of games and learning and includes the benefits of both. To increase students' interest in mathematics and improve their problem-solving abilities, we have designed and implemented a geometric origami edutainment application. This application is based on three heuristic axioms and simple data structures. Students can create any angle or polygon by folding/unfolding the 2D paper displayed on the screen of the Android based system. This application has the benefit of an intuitive interface, and requires a higher level of thinking from the students than that required by traditional origami. We created the application to entertain the students, but also to assist in improving their mathematical problem-solving abilities.

**Keywords** Origami · Geometry · Problem-solving · Edutainment · Android

## 1 Introduction

The digital age in which we currently reside is changing rapidly, and such change will only get faster and more unpredictable as time goes on. In the education field, digital tools are being used to a greater extent in learning activities. In particular, games have had such a large influence on students that they have changed the way these students learn today. Teachers utilize games to enhance the motivation and engagement of students' learning [1].

---

B. Kim · T. Kim · J. Kim (✉)  
Department of Computer Education, Teachers College,  
Jeju National University, Jeju, Korea  
e-mail: jkim0858@jejunu.ac.kr

B. Kim  
e-mail: pigpotato79@naver.com

T. Kim  
e-mail: gtranu@naver.com

Recently, smartphones and tablet PCs have been identified as emerging educational materials. Students carry their own mobile devices, so that they can access a variety of different multimedia sources anywhere and anytime. For example, if a digital game that a student enjoys playing is related to learning English and it can improve his/her English speech, it is not only a game, but also useful learning material. We refer to this kind of game as edutainment.

Edutainment is a hybrid genre that relies heavily on visuals and narratives or game-like formats, but also incorporates some type of learning objective [2]. In this study, we developed a geometric origami edutainment application to enhance students' interest in learning geometry and to improve their problem-solving abilities.

Origami is the Japanese name for the centuries-old art of folding paper into representations of birds, insects, animals, plants, human figures, inanimate objects, and abstract shapes [3]. Most origami designs are attempted by a combination of trial and error and/or heuristic techniques based on the folder's intuition. However, as the designs include various patterns, they could be presented as algorithms based on a set of mathematical conditions [4]. There are many other options for describing shapes other than origami, such as a set of polygon layers. The most attractive feature of origami, however, is that a wide variety of complex shapes can be constructed using a few axioms, simple fixed initial conditions, and one mechanical operation (folding) [5].

## 2 Application Design and Development

### 2.1 Learning Content

Having extracted learning content related to geometric features from the school curriculum of the Republic of Korea [6], we created problems based on the geometric content and reformatted these for implementation on the Android system. The problems are displayed on the screen at each level, from easiest to most difficult, in the same order as the learning content:

- Plane figures (right angle, right triangle, square,  $45^\circ$  angle);
- Triangles (isosceles triangle, equilateral triangle, acute triangle, obtuse triangle);
- Quadrilaterals and polygons (trapezoid, parallelogram, rhombus, rectangle, equilateral polygon).

### 2.2 Design of Android Application

An activity is the basic unit from which the user interface of the Android application is constructed. In other words, an activity is considered to be the screen that a user observes and deals with when responding to system events. In this

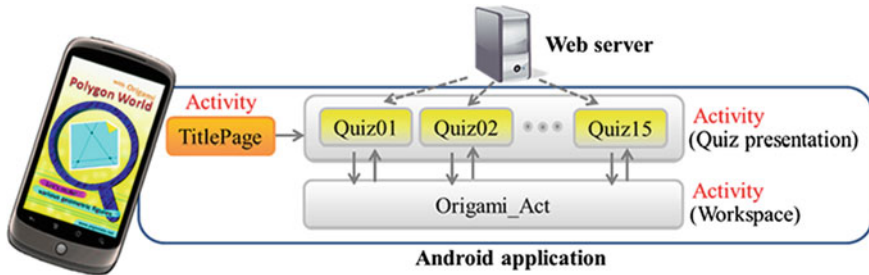


Fig. 1 Structure of activities and interactions

application, there are three main activities: the title page, quiz presentation, and workspace (see Fig. 1).

When a user enters the title page and then starts a quiz presentation, the application attempts to download the data for the first quiz from the Web server. The user can then read the requirements of the quiz and work out a solution in the workspace by manipulating the origami to create a geometric figure for submission as the answer of the current quiz. For example, if the quiz requires a right angle to be constructed, the user would make several folds, thereby creating the angle and then submit the answer by touching three intersection points made by the folded lines and pressing the “submit” button. Figure 2 depicts an activity diagram for Origami\_Act, which shows the flow and logic of the computations and presentations by the system when responding to a user action.

**Implementation of application.** Figure 3 shows screen capture images of a quiz presentation and the workspace. Although many people can fold origami in a variety of different ways to create animals, flowers, insects, or other shapes, it is difficult and complex for programmers to design a 3D application that can be manipulated in the same way as traditional origami. In this study, we imposed certain conditions when running the origami application to allow intuitive manipulation on Android devices.

- The paper represents a 2D environment in which six operations can be performed: restart, pick, fold, unfold, cancel, and submit.
- The user can fold the paper inwards and unfold it outwards.
- The user can fold the paper from ‘point to point’ (see Fig. 4). However, the user cannot fold the one side over to the other side.
- The user cannot make two consecutive folds on the paper; in other words, the folded paper must be unfolded first, before the user can fold it again.

When a user folds the paper, the shape of the folded paper is shown as one of three different types (see Fig. 5).

When the user presses the ‘PICK’ button, four open vertexes appear at the four corners of the paper, and the user can then touch one of these. Let the first vertex selected by the user be  $V1$ . The user then selects a second point,  $V2$ , within or on the edge of the paper. Three different types of folded paper are formed according to

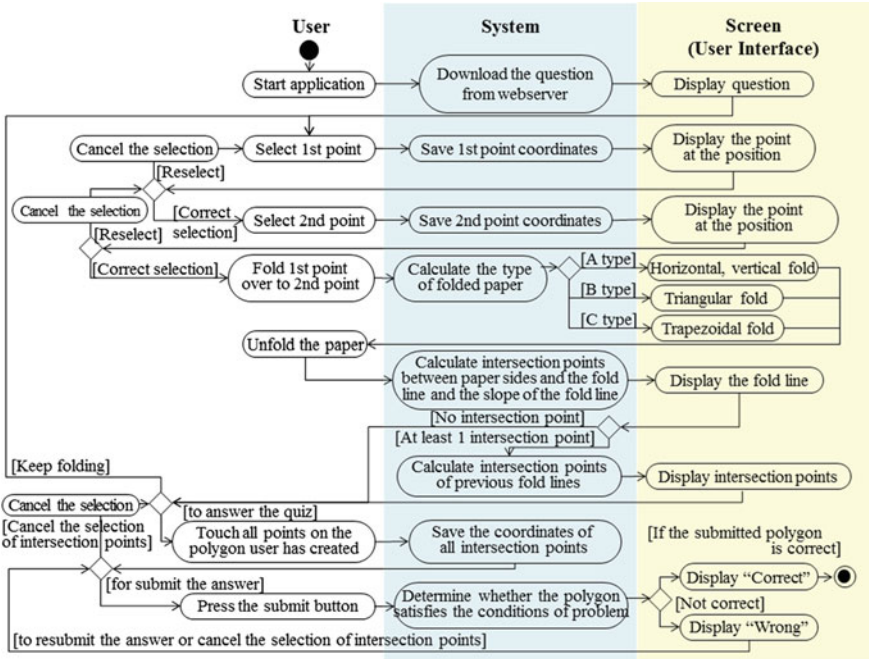
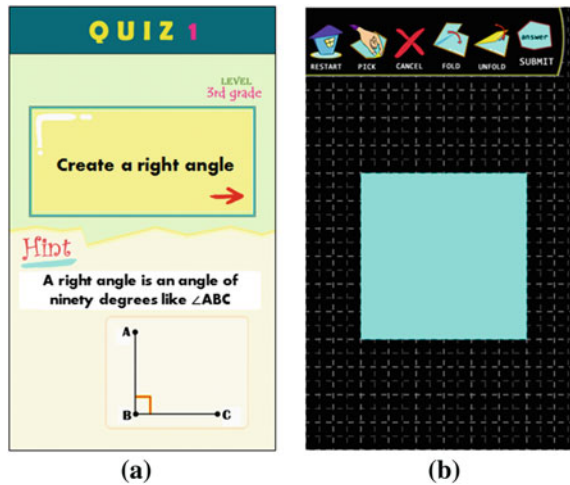
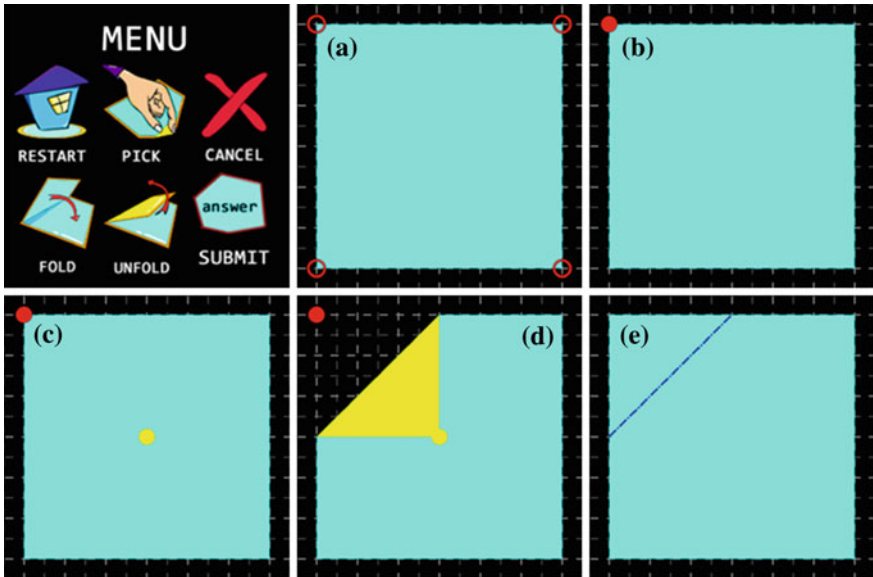


Fig. 2 Activity diagram for Origami\_Act

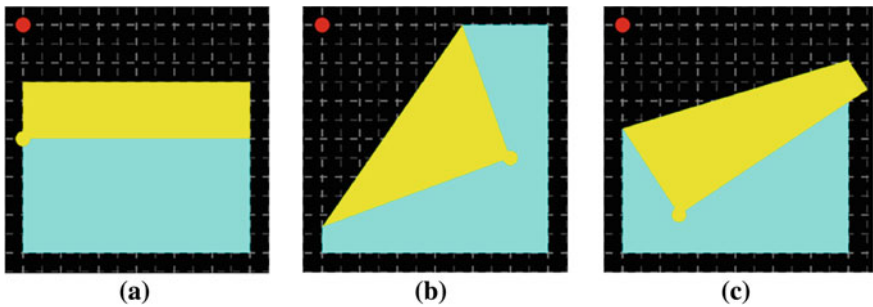
Fig. 3 Screen capture images of a quiz presentation and the user’s workspace. a Quiz presentation. b Workspace



the positions of  $V1$  and  $V2$ . Let the neighboring vertex of  $V1$  in a clockwise direction be  $R$ , and that in an anticlockwise direction be  $L$ . Further, let  $Q(R)$  be the quadrant of the circle with center  $R$  and  $Q(L)$  be the quadrant of the circle with



**Fig. 4** Operation menu (*top left*) and the shape of the origami based on the user’s action. **a** State when the user presses the ‘PICK’ button. **b** State when the user touches an open point. **c** State when the user touches an inner point. **d** State when the user presses the ‘FOLD’ button. **e** State when the user presses the ‘UNFOLD’ button



**Fig. 5** Three variations of folded paper. **a** Type A: Rectangular type. **b** Type B: Triangular type. **c** Type C: Trapezoid type

center  $L$ . We defined the following simple axioms with respect to the patterns of the folded paper.

$$\text{Type A: } V2 \in \overline{V1R} \cup \overline{V1L} \tag{1}$$

$$\text{Type B: } V2 \in Q(R) \cap Q(L) \tag{2}$$

$$\text{Type C: } V2 \in (Q(R) \cup Q(L)) - (Q(R) \cap Q(L)) - (\overline{V1R} \cup \overline{V1L}) \tag{3}$$

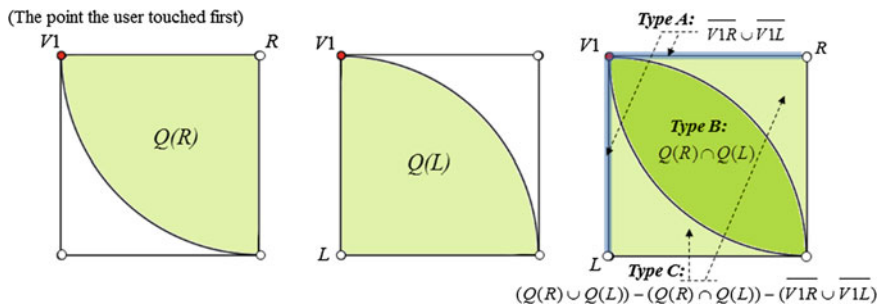


Fig. 6 Origami axioms with respect to the type of folded paper

If point  $V2$  is on line  $V1R$  or line  $V1L$  described by Axiom (1), the paper is folded as Type A (rectangle type). If point  $V2$  is on the intersection of quadrant  $Q(R)$  and  $Q(L)$  described by Axiom (2), then the paper is folded as Type B (triangle type). If point  $V2$  does not lie on line  $V1R$  or line  $V1L$ , and is on the outside of the intersection of quadrant  $Q(R)$  and  $Q(L)$  as described by Axiom (3), the paper is folded as Type C (trapezoid type). Figure 6 illustrates these axioms.

The axioms defined above can be proved mathematically [7]. However, it was more convenient and simple to design and implement the application.

**Example of playing the game.** When users attempt to find the answer to a quiz, they should consider the following: How many times should the paper be folded? How can a fold line be made at a particular position? How can a line be created with the same length or angle as a previously created line? Using this process during the learning process stimulates the user’s thinking. If the quiz requires a parallelogram to be constructed, the user can find a solution by carrying out several fold/unfold steps and then touching the vertexes to outline the parallelogram (see Fig. 7). While creating this shape, the user can also calculate the lengths of line segments using the grid in the background.

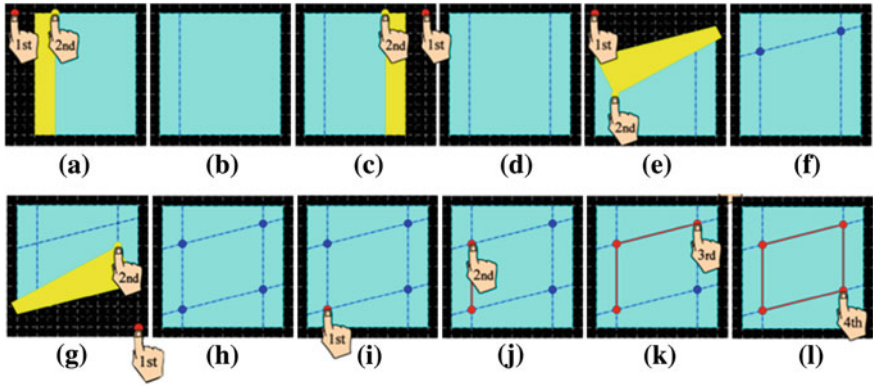
Once the user has created the parallelogram as the solution, he/she can submit it by pressing the ‘SUBMIT’ button. When the system receives the coordinates of the polygon vertexes, it calculates the distances and angles and checks that the submitted polygon meets the requirements of the quiz.

**Data structure.** When the user selects a point on the paper in the application, key information about the point is saved in a data structure. Assume a tuple  $P$ , with four different variables representing the point; that is, the key index, x-coordinate, y-coordinate, and array  $L$  containing the index numbers of lines on which  $P$  is located. This information is used to create the path of the polygon (see Fig. 8).

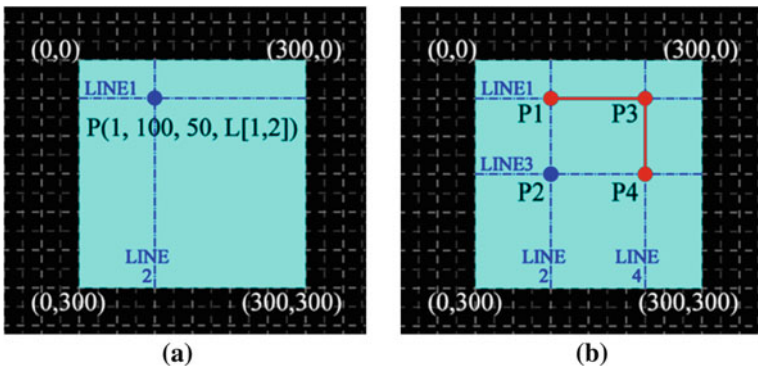
If the user folds the paper four times as in Fig. 8b, four points and four lines are created. The points are defined by the data depicted below:

P1 (1, 100, 50, L[1, 2])	P2 (2, 100, 150, L[2, 3])
P3 (3, 225, 50, L[1, 4])	P4 (4, 225, 150, L[3, 4])

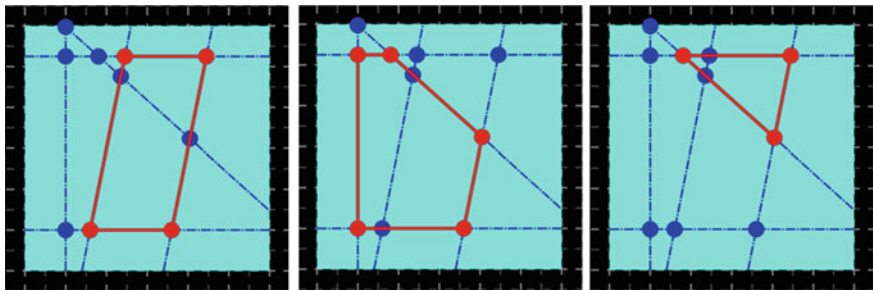




**Fig. 7** Procedure for creating a parallelogram (PICK, FOLD, and UNFOLD mean pressing the corresponding menu buttons, while ‘touch’ implies the user’s action). **a, c, e, g** PICK-touch-FOLD. **b, d, f, h** UNFOLD. **i** Touch 1st point. **j** Touch 2nd point. **k** Touch 3rd point. **l** Touch 4th point



**Fig. 8** Data structure and use of points. **a** Data saved for a point. **b** Creating a path when the user touches P1, P3 and P4 in order



**Fig. 9** Some *polygons* on the paper using the same points

When the user touches P1 and P3, the application checks array L of each point. If both arrays have the same index numbers of lines, they are connected with a red line. When the user touches another point, the application operates in the same way. In this algorithm, users can design any polygon using the points they have created by folding the paper and connecting the points (see Fig. 9).

### 3 Conclusion

In the digital age, the boundaries between games and learning applications have disappeared. In the geometric origami edutainment application we developed, users can have extraordinary experiences creating angles and polygons with digital paper. From an education perspective, drawing a polygon on real paper requires only one dimension of thinking, whereas creating a polygon using this application requires a higher dimension of thinking. We expect that this application can assist students by improving their geometric problem-solving abilities. In addition, it could be used as educational material in ICT based mathematics learning.

### References

1. Garris R, Ahlers R, Driskell JE (2002) Games, motivation, and learning: a research and practice model. *Simul Gaming* 33(4):441–467
2. Okan Z (2003) Edutainment: is learning at risk? *Brit J Educ Technol* 24(3):255–264
3. Kasahara K, Misaki I (1967) *Creative origami*. Japan Publications, Tokyo
4. Lang RJ (1996) A computational algorithm for origami design. In: 12th annual ACM symposium on computational geometry. SCG, Pennsylvania, pp 98–105
5. Nagpal R (2001) *Programmable self-assembly: constructing global shape using biologically-inspired local interactions and origami mathematics*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology
6. Ministry of Education, Science, and Technology (2008) *The school curriculum of the Republic of Korea*, Ministry of Education, Science, and Technology, Seoul
7. Alperin RC (2000) A mathematical theory of origami constructions and numbers. *N Y J Math* 6:119–133

# Gamification Literacy: Emerging Needs for Identifying Bad Gamification

Toshihiko Yamakami

**Abstract** Gamification is a collection of game-design-origin know-how that facilitates service engagement. It is used in a wide range of applications, marketing, enterprise management, and education. The positive side of gamification provides useful tools for improving engagement. At the same time, there are negative sides of gamification when it is used excessively or maliciously. The author proposes the concept of gamification literacy.

## 1 Introduction

The market size of mobile social games in Japan has demonstrated a radical growth in the past couple of years. The investigation of best practices in mobile social games shows that the design is based on micro-management of goal-achievement cycles with intensive feedback systems. This intensive feedback system can be utilized as gamification in order to improve engagement in a wide range of services. However, excessive game techniques of mobile social games lead to some social problems.

This provides an important lesson teaching us that we have to master a new kind of literacy in order to manage gamification in the virtual world of consumer services and enterprise systems. The author proposes the concept of gamification literacy. The author examines some anomalies of mobile social games and discusses the components of gamification literacy.

The aim of this research is to identify a new type of literacy in the emergence of gamification in the virtual world.

---

T. Yamakami (✉)

ACCESS, Software Solution, 1-10-2 Nakase, Mihama-ku, Chiba-shi 261-0023, Japan

e-mail: Toshihiko.Yamakami@access-company.com

URL: <http://www.access-company.com/>

Literacy refers to the ability to read for knowledge, write coherently, and think critically about the written word (Wikipedia).

Game design has attracted a wider scope of audience because of its non-game applications. McGonigal discussed how game design techniques and mechanisms can solve real-world problems [1]. She explained how the combination of an artificial goal, a set of artificial rules, and feedback with voluntary participation creates challenges and fun for the user.

Literacy that deals with gamification has not been covered well in the past literature.

The originality of this paper lies in its examination of a new kind of literacy in service design using gamification.

## 2 Observation

### 2.1 Industry Landscape of Mobile Social Games

The market size of mobile social games in Japan grew on the order of 100 billion Japanese yen since 2010. The amount in 2012 is estimated by Mitsubishi-UFJ-Morgan-Stanley securities to be 400 billion Japanese yen. Early mobile social games that were designed for feature-phones were relatively simple. However, the rapidity of market growth provided rich cash flows and boosted the profits of early winners. The rich cash flows enabled rich graphic representations and intensive data mining that produced a completely different landscape with an intensive selection process.

### 2.2 Social Problems

The overheating market situation has brought about several social problems depicted in Table 1. The industry started to depend on heavy users that pay tens of thousands Japanese yen per month. In the case of GREE, 3.5 % of heavy users provide approximately 70 % of its revenue.

**Table 1** Social problems

Aspect	Summary
Addicted users	Combination of higher goals and routine simple mini-games promote addiction with a daily routine repetition
Speculative systems	Gambling-like achievement using a very small probability of success with extreme, rare rewards promote speculative attitudes in users
Unfair probability adjustment	Completion gatcha, a gambling system, uses arbitrary probability adjustments with misleading expectation of completion

The top-ranking Japanese mobile social game vendors have enjoyed high profitability, approximately 50 %. This high profitability is driven by the heavy users that pays more than five thousand Japanese yen per month. The focus on charging heavy users is an easy way to create efficient revenue-generating engines. Therefore, the top-ranking game vendors have increased their dependence on charging heavy users in the past couple of years. The most popular technique is the completion gatcha. Completion gatcha is a method of awarding rare in-game items in mobile games only when the player has bought a full set of other in-game items. It is a kind of lottery game played on a mobile phone.

In May 2012, the Consumer Affairs Agency of Japan announced that completion gatcha had led legal issues. The Consumer Affairs Agency decided to ban “kompu gatcha” (completion gatcha) online games played on mobile phones starting from July 2012 under the Law against Unjustifiable Premiums and Misleading Representations.

### 3 Gamification Literacy

#### 3.1 Gamification

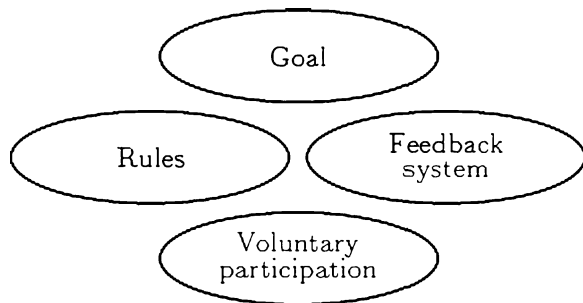
The framework of a game defined by J. McGonigal for a game is depicted in Fig. 1.

It should be noted that this framework can be applied to a wide range of services including social services. When an appropriate goal and a set of rules can be designed, many services can be leveraged by the applications of game-origin techniques to increase user engagement.

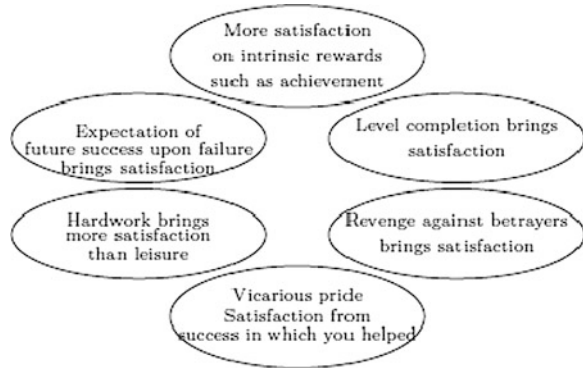
The positive attitudes of human beings explored by positive psychology are depicted in Fig. 2.

This indicates that a human beings are an extraordinarily positive creatures with emphasis on internal rewards, hard work, an eagerness to help others, high engagement of promises, and so on.

**Fig. 1** Framework of a game (by J. McGonigal)



**Fig. 2** Positive attitudes of human beings



All these attributes contribute to constructing positive societies. The engineering of hope with the expectation of achieving rewards drives people in the virtual world.

These positive attitudes represent the possibility that artificial goals and rules can accommodate great enthusiasm from users when managed appropriately. No fantasy is required when the appropriate feedback mechanism is ensured, because internal achievement provides high-level user motivation. This is a core factor that feature-phone based mobile social games in Japan demonstrated through radical user-base growth and revenue growth in the past couple of years, even though they provide simple multimedia representations that are inferior to those of smartphones.

### ***3.2 Definition of Gamification Literacy***

The components of gamification literacy are depicted in Table 2.

This shows how people have awareness of virtual-world persuasion. First, people love to compete. The virtual world amplifies this tendency through a wide range of visualizations. Also, the virtual world amplifies this through 24 h a day, 365 days a year of competition. Competition can be enhanced with globalization and anonymity.

Second, the virtual world can manipulate probability in selection and rewards. People easily build up false expectations in the artificial setting of the virtual world.

Third, people are susceptible to time. Time is something we have to measure in order to identify. People have weak points in the management of asynchronous events. This attribute can be easily exploited by virtual-world service providers. For example, we are sensitive to time-limited offers, one of the classic persuasion techniques.

Fourth, people are susceptible to a speculative mindset and gambling. The real-world has many constraints on gambling. The virtual world has very relaxed

**Table 2** Components of gamification literacy

Component	Description
Competition literacy	Awareness of control techniques using social competition
Probability literacy	Awareness of control techniques using probability, including misleading probability impression
Time management-based persuasion literacy	Awareness of in-game persuasion using time limitation and clock-based visualization techniques
Speculation literacy	Tolerance of speculative game mechanisms
Addiction literacy	Awareness of combination of low-level and high level game machismo that drives addiction to a game
Rareness-based persuasion literacy	Awareness of rare item-based persuasion
Social invitation literacy	Awareness of background persuasion techniques of friend invitation
Completion literacy	Awareness of in-game persuasion techniques using completing a set

regulations on gambling and speculation. Therefore, we have to raise our own awareness of gambling and the speculative mindset in the virtual world.

Fifth, people are susceptible to a routine-task-based stimulus. This can be easily constructed when a repetitive routine task is combined with a higher goal. Many mobile social games use routine-task techniques to leverage the addiction to a game. We need to be aware of this addiction-promoting process.

Sixth, the virtual world can coin a wide range of rarity. Rareness is another classic example of persuasion. In the virtual world, rareness is easily manipulated compared to in the real world. People need to increase their awareness of this type of persuasion in the virtual world.

Seventh, many service providers use the technique of social invitation. As Facebook reached one billion active users, social invitation became recognized as a powerful marketing tool. As people increase their encounters with social invitation in the virtual world, it increases the necessity for the awareness of social invitation. In the technique of social invitation, there are two different roles, inviters and invitees. Both sides need a new kind of awareness of the consequences of social invitation. Also, people need to master a kind of ability to refuse or ignore social invitations for a healthy social life in the virtual world. Social invitation can invoke social problems such as Ponzi schemes even in the real world. In the era of the socially connected Internet, it is more important to leverage social invitation literacy for a healthy Internet life.

Eighth, there is the mechanism of completion in the virtual world. As a user approaches the completion of a task, there comes a stronger desire to complete that task. This psychological process was exploited in mobile social games in Japan in 2011, and that led to the ban of completion gatcha in July 2012. Completion gatcha is considered to be a driving factor that enabled a multi-billion-dollar business for mobile social games in Japan. Generally, completion can be used to lure users to elicit actions to complete a task or collection. In order to protect users from paying

extremely high amounts like hundreds of dollars per month, people need to become aware of this kind of persuasion technique in the virtual world.

The lessons that we learn from the rise of the mobile social game business in Japan include the power of data-mining capabilities for game-tuning. The today's data mining infrastructure provides a real-time analysis capability to collect behavior logs from millions of users and parse them for clues in order to tune games by the hour. In order to cope with this massive data-empowered infrastructure, end users need to have knowledge of behind-the-scenes engineering principles and to have the will-power to resist the temptations leveraged by these design principles. Gamification literacy provides them with the basic skills.

The list shown above is not exhaustive. As people's lives are merged with the virtual world, service providers in the virtual world continue to invent different patterns of persuasion. It is necessary to raise awareness of this increasing persuasion in the virtual world. It is also necessary to constantly maintain the set of gamification literacy in order to protect end users and stability of the virtual-world economy.

Not only consumers, but also workers in the workplace may experience more time staying in the virtual world. This raises an awareness of virtual-world workplace literacy. Gamification can leverage the performance of the virtual-world workplace. However, it is beyond the scope of this paper.

### ***3.3 Gamification Classifier Framework***

Examples of bad gamification in the mobile social game industry are depicted in Table 3.

In order to systematically deal with gamification literacy, the author provides a view model of the gamification classifier framework, as depicted in Fig. 3.

This multi-stage framework provides a skeleton that accommodates the flow-chart of gamification identification.

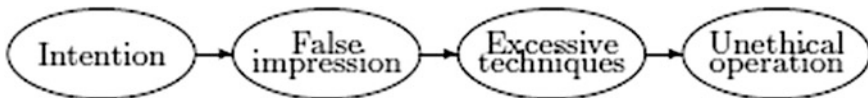
### ***3.4 Education of Gamification Literacy***

When the technology to impact end users on a massive scale emerged, we had to coin a new type of literacy in order to deal with it. Communication technology created radio and television technologies that provide real-time massive distribution of information. Such a type of real-time information broadcasting was impossible before the invention of that technology. The convenience of technologies also brought concerns such as brain-washing and biased broadcasting. These concerns built up the so-called media literacy, the literacy to read the behind-the-scene intention and provide unbiased judgments of broadcast information.



**Table 3** Examples of bad gamification

Item	Description
Untrustworthy manipulation of gambling	Setting up rewards as a gambling system. As a direct award, the user is entitled with the right to enter a lottery. The probability of winning the lottery may be controlled in an arbitrary way, such as controlling the winning probability to be extremely small when a user gets closer to completion of a target set
Untrustworthy tuning of a game	Changing game parameters in an untrustworthy way without notices
Intensive competition	Putting an extremely rare item for daily top ranking, weekly top ranking, or monthly top ranking users
Intensive time-limited offer	Setting a bonus time such as gaining double points or triple points for a limited time span, such as five minutes or ten minutes to heat-up the competition
Arbitrary management	Creating arbitrary parameters, levels, enemies. Arbitrarily raising the upper limit to lure consumers such as by giving the impression of a final limit of 100, but then raising it to over 100 after a certain time frame
False impression	Setting up non-human characters to be listed in the top rank. Arbitrary management of the “rareness” of items
Extravagant advertisement	Misleading advertisement of the rareness of gift items, time-limited offers, and so on

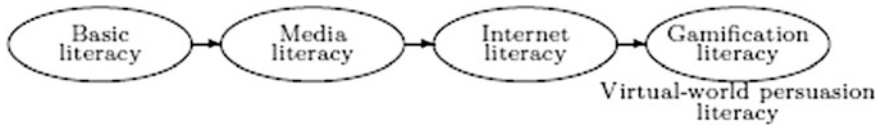


**Fig. 3** Gamification classifier framework

When Internet technology started to flourish, many people experienced new types of conflicts such as flaming on bulletin boards. In addition, new areas of concern about privacy protection on the Internet emerged. These concerns built up the so-called Internet literacy, the literacy to detect privacy issues and avoid unnecessary interpersonal conflicts in the social interactions leveraged by Internet technology.

As social service engineering emerges and the stay time on social services becomes longer, it is necessary to deal with behind-the-scenes social service techniques. The author calls this gamification literacy, with emphasis on the back-end technology that enhances user engagement in social services.

A detailed discussion of gamification literacy is beyond the scope of this paper. The author hopes that coining this new terminology will raise awareness of the impacts, and risks of social service engineering that has come to prevail on this planet in the past couple of years. The author feels that this is a starting point to build up a gamification literacy that will lead to part of modern literacy education in the coming years.



**Fig. 4** Transition of literacy

The transition of literacy as the methods of massive real-time proliferation of information raised is depicted in Fig. 4.

## 4 Discussion

### 4.1 Advantages of the Proposed Approach

The mass media industry has been challenged by emerging Internet services. Not only for information retrieval, but also social interaction is now empowered by Internet services including social network services such as Facebook, Twitter, and so on.

The stay time in such virtual world social lives continues to increase. There is no foreseeable obstacle to prevent its growth.

Additionally, many teenagers are becoming digital citizens who begin their primary social lives in the virtual world.

These changes have brought about concerns about how we can develop appropriate manners and techniques to deal with a new style of social life.

The increase stay time in social lives in the virtual world leverages a strong demand for social service engineering. New growth in social service engineering has both positive and negative side effects.

For the positive side, the services utilizes advanced computer and communication technologies with anytime and anyplace capabilities to improve the social quality of life. For the negative side, the immaturity of social service engineering may damage social lives and the virtual world economy with inappropriate use of technology. Some negative aspects are already visible in the heating-up of the mobile social game industry in Japan.

The author coins the concept, gamification literacy, in order to highlight the importance of awareness of new types of literacy we need for the new type of social life that exists today.

Gamification literacy helps people:

- increase their willpower to resist temptations in online services with awareness of backend mechanisms,
- prevent themselves from paying extremely high bills through awareness of operational manipulation,
- prevent themselves from becoming network service addicts through awareness of backend persuasion methodologies.

## **4.2 Limitations**

This research is a descriptive study. The quantitative measures for verifying multiple aspects of gamification literacy discussed in this paper remain for further study.

Concrete education design methodology of gamification literacy is beyond the scope of this paper.

## **5 Conclusion**

The game industry needs to cope with this problem through their self-regulation and ethics. At the same time, considering the huge social graph and rapid propagation of word-of-mouth mechanisms on social network services, it is important to raise awareness on the consumer side. Not only for games, but also a wide range of social services can utilize the general principles of gamification.

The author coins the concept, gamification literacy, to highlight the needs for such a new awareness and techniques for daily life.

## **Reference**

1. McGonigal J (2011) Reality is broken: why games make us better and how they can change the world. The Penguin Group, New York

# Automatic Fixing of Foot Skating of Human Motions from Depth Sensor

Mankyu Sung

**Abstract** This paper presents a real time algorithm for solving foot skating problem of a character whose motions are controlled in on-line puppetry manner through motion depth sensor such as Kinect. The IR (Infrared Light) projection-based motion sensor is very sensitive to occlusion and light condition. One significant artifact from these characteristics is foot skating, which means that feet are floating over the ground even when the character is standing still. Our algorithm is working as a post processing that applies the orientation data from sensors to a character first, and then keeps monitoring the current character status and applying foot skating solver next whenever the problem occurs. For this, we propose an adaptive threshold method that make positional and velocity threshold change automatically. Inter-frame smoothing is then followed to make sure there is no discontinuity on the motions.

## 1 Introduction

The emergence of new motion sensor technology provides users a new experience of a natural user interface for interactive contents. A primal example is Microsoft's *Kinect* [1]. Ever since it was introduced to video game community, it has drawn a lot of attention for its effectiveness in capturing movement. Natural interaction with contents using full body increases immersive feeling to the contents, which is a critical point for attracting people. The current skeleton tracking of Kinect sensor is based on the *depth image* [2, 3]. The depth image represents the distance map from the sensor to the environment including users. Through the machine learning algorithm or by solving numerical kinematic configuration of users, most of the sensors are able to obtain the joint positions and orientation at real time rate.

---

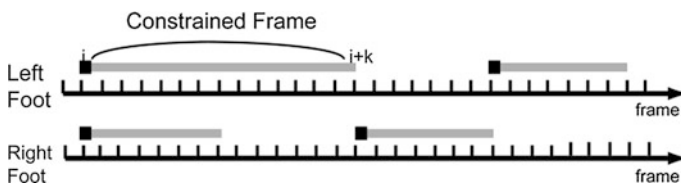
M. Sung (✉)

Keimyung University, 1895 Dalgubeol-Daero, Daegu 704-701, Korea  
e-mail: mksung@kmu.ac.kr

Current motion sensors, however, do have some limitations. Because they are not a motion capture device—which use more than eight optical cameras to compensate the occlusion problem—the data from sensor has a jittering and significant artifacts such as foot skating [4]. The foot skating means that one or both feet are floating over the ground even when the character is standing still. This is the most disturbing artifact in character animation. This artifact comes from the several reasons. First, estimated depth data from the sensor might have some errors. This is inevitable because the sensor, which uses IR (Infrared sensor) projection, is very sensitive to light condition and occlusions. Second, the skeleton tracking algorithm might have some bugs. Although current depth sensors provide SDK for building applications that need skeleton tracking data, the orientation and position data often have low confidence values [5].

In this paper, we address this foot skating problem. Our approach is not targeting on improving the existing depth camera or accuracy of tracking algorithm. Instead, we try to solve this problem as a post processing job. That is, our algorithm is working right after applying original data onto a character. We then keep checking the character status after applying original data. If any artifact such as foot skating is happening on the character, we trigger the foot skating solver to fix it. Figure 1 shows the result of our algorithm. Although jerkiness of motion data can be smoothed out through a filtering technique, the foot skating cannot be solved with ease. First, we need to figure out when foot skating happens. Second, if foot skating occurs, then we need to determine the exact foot position. Third, we have to change the leg configuration in order to locate the foot to the position exactly while keeping original leg configuration as much as possible. This might have a lot of answers, but we need to choose right one based on our criteria. At the final step, we need to minimize the pop-out effect that might be caused due to the change of configuration.

In our algorithm, we define the *constrained frames* as frames that one or both feet touch the ground. We recognize the constrained frame by analyzing the foot velocity and distance from the ground. Under the condition that in general the foot plants continue for multiple frames, once the constrained frames start, then we use the very first foot position as the target foot position. Then, we move the root joint to the target when the target is located too far away. In this process, we apply simple moving averaging (SMA) technique to minimize discontinuity. Next, we apply the special IK solver to locate the foot to the goal position. The IK solver



**Fig. 1** In general, the constrained frames continue for multiple frames. We choose the foot position of first frame of each constrained frame blob as a target

should consider the skin mesh collapsing during computation so that the skin mesh does not twist. Finally, we apply rotation interpolation with neighboring frames for cleanup any pop-out.

## 2 Related Work

For obtaining depth image, several different approaches have been used for the camera. Stereo, Time-of-flight (TOF) and IR-structured-light method are three major technologies for implementing the depth camera today [6, 7]. In the case of Kinect, it uses the continuously projected infrared structured light to estimate 3D structure of objects [8, 9]. Currently, it is a unique consumer hardware that has interactive rate with full range of body types and sizes.

Foot skating has been a nagging problem in character animation community for a long time, especially for animators using motion capture data. Kovar et al. proposed an explicit solver for the problem [10]. They use an analytic IK solver which adjusts leg length to keep the foot to an exact position on the ground for the motion capture data. In our approach, we use the similar fast analytic IK solver to obtain the hinge orientation of leg. However, one fundamental difference is that our foot skating is working in on-line manner. Therefore, our algorithm does not have any information about next frames whereas motion capture data have full information about previous and future data.

## 3 Algorithm

### 3.1 Adaptive Constrained Frame Detection

Our foot skating cleanup algorithm consists of five steps. First step is to detect constrained frame. Essentially, this step checks whether the current character's feet are contacting the ground or not. We have two different ways for doing it. First, we can do all checking on the sensor domain. From the user's foot position received from sensor, we can figure out if its foot is touching the ground or not. Second, we can do that on the character domain. After applying data on the character first, we check the current character's foot position in the virtual environment. The sensor domain computation is highly risky because it is very sensitive to the sensor location and user position. So, we choose the character domain approach in our algorithm.

In particular, we determine the foot's status through two parameters. These parameters are foot position and velocity. Let's say that current foot position of  $P_f$  and the foot position at previous time frame is  $\hat{P}_f$ , then the velocity  $V_f$  is computed as following:

$$V_f = r\left(\frac{P_f - \dot{P}_f}{\Delta t}, n\right)$$

$$r(v, n) = \text{roundXL}(v, n)$$

where  $\Delta t$  is the inter-frame duration time and  $r(v, n)$  returns the rounded value at  $n$ th digit of  $v$ .

Given  $y$  coordinate of  $P_f$ , which is the up vector coordinate that represents the distance from the ground, and its velocity  $V_f$ , we check whether these two values are under the threshold value  $P_{threshold}$  and  $V_{threshold}$  at the same time. For relaxing the sensitivity, we apply rounded values instead of the original value. If two conditions are satisfied, we mark that frame is constrained. One point to note is that the whole character's global position is closely related with the root (pelvis) joint position. Therefore, initial root position of the character is very important. In our algorithm, we calculate the length between foot (end-effector) and root joint, and lift up the character as far away from the ground as possible so that the character stands up above the ground. One downside is that setting constant threshold values might be too risky because these values should be changed depending on applications. Finding threshold value is a time consuming job that requires a lot of try-and-errors. In our approach, we take *adaptive* approach for finding these values. Basically, the user set the rough initial threshold value at the beginning, then the algorithm incrementally adjusting these values automatically. One important cue that makes this technique possible is that for human motion, one of the feet should be on the ground all the time. Therefore, while we applying the initial threshold values to detect constrained frames, if both feet are turned out to be not-constrained, then the threshold values need to be changed to make one or both feet be constrained. Let's say  $P_f(r)$  is the right foot position,  $P_f(l)$  be the left foot position,  $V_f(r)$  and  $V_f(l)$  are their velocity respectively. If both feet are non-constrained, then we check which values are above the threshold values on both feet at the same time. For example, if  $P_f(r)$  and  $P_f(l)$  are above the threshold value  $P_{threshold}$ , then new  $P_{threshold}$  are set as  $MIN(P_f(r), P_f(l)) + \delta$ . Likewise, if  $V_f(r)$  and  $V_f(l)$  are above the threshold value  $V_{threshold}$ , then the new  $V_{threshold}$  is set as  $V_{threshold} = MIN(V_f(r), V_f(l))$ . In our approach, we set the  $\delta = 0.1$ .

### 3.2 Determining Foot Position

Once we find out that the current frame is constrained, then we need to fix the foot on the ground during the time being constrained. The most important job for this is to decide the exact foot position. The new foot position,  $P_f$ , is then fed into IK solver as an argument for obtaining proper leg configuration. In general, the constrained frames continue for a fair of amount of time. During that time, we need only a single foot position for fixing the foot. If we are working on off-line processing, where we have all previous and future data, we can easily find a single

foot position that minimizes deviation from original data over constrained frames [12]. However, because we are working on on-line processing, we don't know how long this constrained frame period continues. Therefore, we use the foot position of the first occurred constrained frame as target foot position.

### 3.3 Adjusting Root Joint Position

The target foot position may be too far to be reached even when the character is fully stretching its leg. This problem can be solved by adjusting the root joint position to the target foot position. Let  $o$  be the offset vector from the hip to the root joint and  $P_r$  be the root joint position. Also, let  $P_t$  the target foot position and  $l$  be the full limb length. Then, the target foot joint,  $P_t$ , is reachable only if  $\|P_t - (P_r - o)\| \leq l$ . In other words, when we think of it as a sphere that centered at  $P_r - o$  with radius  $l$ , the  $P_r$  should be inside the sphere in order to be reachable. Because both feet can be constrained at the same time, we have to consider maximum two spheres at the same time. In this case, the root should be projected onto surface of two spheres' intersection region [10, 11]. Because of relocation of root joint, there might be small discontinuity on the character. To minimize the discontinuity, we apply the simple moving average (SMA) on the root joint position stream. Suppose that newly computed root position from this projection technique is  $\hat{P}_r$  and original position is  $P_r$ , then SMA performs  $P_r = \frac{(\hat{P}_r + P_r)}{2}$  to get the final position. This computation is iterating during the constrained frames.

### 3.4 Applying IK Solver

Once we fix the root position, next step is to adjust the leg to place the foot in the target position. Specifically, we need to find the hip and knee angle. This is accomplished by two-link analytic IK algorithm. We use the similar method that Lucas et al. used for their paper [12, 13]. The IK solver is composed of two phases. The first phase is to rotate the knee angle so that the length between hip and target foot position ( $\overrightarrow{P_h P_t}$ ) matches the length between hip and current foot position ( $\overrightarrow{P_h P_f}$ ). The knee angle  $\theta_k$  can be obtained as follows:

$$\theta_k = \arccos \left( \frac{l_1^2 + l_2^2 + 2\sqrt{l_1^2 - \hat{l}_1^2}\sqrt{l_2^2 - \hat{l}_2^2} - \|P_h - P_f\|^2}{2\hat{l}_1\hat{l}_2} \right)$$

where  $l_1$  and  $l_2$  denote the length of thigh and knee joint respectively and  $\hat{l}_1$  and  $\hat{l}_2$  are the length of projected vector onto rotation axis. The rotation axis is obtained by cross product between  $\overrightarrow{P_h P_k}$  and  $\overrightarrow{P_f P_k}$  where  $P_k$  is the knee joint position.



One significant problem happens when the  $\theta_k$  is close to 180 degrees, which corresponds to a fully straight leg. As addressed in<sup>4</sup>, this causes an unnatural extension and contraction of knee angle for small change of target position. To prevent it, we limit the knee rotation to the maximum knee angle, say  $\theta_{\max}$  (170°, in our algorithm), and if the angle is bigger than  $\theta_{\max}$ , we put a damping on the angle. Suppose that original knee angle before phase 1 is  $\hat{\theta}_k$ , and the difference between  $\hat{\theta}_k$  and  $\theta_k$  becomes  $\Delta\theta_k$ . Then, the damped knee angle  $\bar{\theta}_k$  is computed as follows:

$$\bar{\theta}_k = \hat{\theta}_k + \int_{\hat{\theta}_k}^{\theta_k} I\left(\frac{\theta - \theta_{\max}}{\pi - \theta_{\max}}\right) d\theta$$

$$I(\theta) = 2\theta^3 - 3\theta^2 + 1$$

$\frac{\theta - \theta_{\max}}{\pi - \theta_{\max}}$  is the normalization operation with  $\theta_{\max}$ .  $I(\theta)$  is the cubic polynomial guaranteeing C1 continuity because  $I(1) = 0$ ,  $I(0) = 1$  and  $\frac{dI}{d\theta}(1) = \frac{dI}{d\theta}(0) = 0$ .

One problem is that all additional hip rotations  $q$  about the axis  $\frac{P_i P_j}{P_i P_i}$  satisfy the target constraint. Among all possible rotations, we choose the orientation that makes the hip and knee angle closer to the original orientations relative to their parent coordinate systems.

When we adjust the joint angle, we have to consider the skin mesh twisting problem as well. Most of skinning is built with linear blending skinning technique where each vertex of skin mesh is assigned weight values for small set of joints. the left and the right character have an exactly same skeleton pose but the left character's skin is twisted. To prevent this problem, given old joint orientation  $q$  and newly determined joint angle  $\hat{q}$ , we first computes angle  $\phi$  about the twisted axis, say X axis, then multiplied  $\hat{q}$  with the corresponding rotation  $\Delta\theta_\phi$  to compensate the twist.

## 4 Experiments

We perform an experiment to validate our algorithm with Microsoft Kinect sensor. The NITE middle ware from PrimeSense is used for retrieving position and orientation data from the sensor. Also, Open source game engine (OGRE) is used for rendering characters. All experiments were carried out on an Intel Xeon 3.20 GHz processor PC with a graphics acceleration card. Figure 2 shows our experiment setup. We use the Kinect sensor as the motion sensor and perform actions 2–3 m in front of the sensor. The skeleton has 15 joints. But, currently the NITE middle ware supports only 11 of them.

Figure 2 is screen shots that compare two cases when we do not use our algorithm and when we use the foot skating clean up algorithm. Note that yellow balls on the heel appear when the frame is identified constrained. Red balls indicate the foot positions when foot IK solving finished. When the yellow balls



**Fig. 2** *Top left:* Kinect sensor for experiments. *Top right:* An 3D character is controlled by the user in on-line puppetry manner. *Bottom left:* Screen shots before applying foot skating cleanup. *Bottom right:* Screen shots after applying foot skating cleanup. *Yellow ball* indicates whether this frame is constrained or not

match red balls totally, it means that our algorithm fixed foot on the ground. For evaluating performance, we display the frame rate information on the top left of the screen at run time, which shows more than 30 frame/s.

## 5 Discussion

In this paper, we have introduced a simple and practical foot skating cleanup algorithm for a character whose motions are controlled in the on-line puppetry manner through the depth sensor. One advantage of our algorithm is that even when character's pose is collapsed because of some errors in the sensor, our

algorithm can keep reasonable pose all the time as long as the frame is identified constrained. Also, not only for legs, but we can also easily apply our algorithm for arms. One limitation of our approach is that our algorithm depends on the middle ware providing joint angle data through skeleton pose estimation algorithm from depth image. As a result of this limitation, we are not currently supporting multi-user foot skating cleanup, although the sensor itself has the capability for identifying multiple users.

**Acknowledgments** This research was supported by the Bisa Research Grant of Keimyung University in 2012.

## References

1. Tsunoda K, Fitzgibbon A (2010) Kinect for Xbox 360—the innovation journey (June 2010). Microsoft Research Faculty Summit
2. Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A (2011) Real-time human pose recognition in parts from single depth images. In: CVPR'2011: Proceedings of the 2011 IEEE Computer Society conference on computer vision and pattern recognition. IEEE Computer Society
3. Zhu Y, Fujimura K (2010) A bayesian framework for human body pose tracking from depth image sequences. *Sensors* 10(5):5280–5293
4. PrimeSense: PrimeSense™ NITE 1.3 Algorithms notes (2010) PrimeSense
5. PrimeSense: PrimeSense™ NITE 1.3 Control programmer's guide (2010) PrimeSense
6. Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. *Comput Vis Image Underst* 104:90–126
7. Ganapathi V, Plagemann C, Koller D, Thrun S (2010) Real time motion capture using a single time-of-flight camera. In: CVPR'2010: Proceedings of the 2010 IEEE Computer Society conference on computer vision and pattern recognition. IEEE Computer Society
8. Microsoft Kinect Teardown: Microsoft Kinect Teardown, <http://www.ifixit.com/Teardown/Microsoft-Kinect-Teardown/4066/>
9. Carmody T, How motion detection works in Xbox kinect, <http://www.wired.com/gadgetlab/2010/11/tonights-release-xbox-kinect-how-does-it-work/>
10. Kovar L, Schreiner J, Gleicher M (2002) Footskate cleanup for motion capture editing. In: SCA'2002: Proceedings of the 2002 ACM SIGGRAPH/eurographics symposium on computer animation, New York, NY, USA. ACM, pp 97–104
11. Shin HJ, Lee J, Shin SY, Gleicher M (2001) Computer puppetry: an importance-based approach. *ACM Trans Graph* 20:67–94

**Part VIII**  
**IT and Multimedia Applications**

# A Study on the Development and Application of Programming Language Education for Creativity Enhancement: Based on LOGO and Scratch

YoungHoon Yang, DongLim Hyun, EunGil Kim, JongJin Kim and JongHoon Kim

**Abstract** Students meet with many problems during programming education in computer education. Programming education have been several research that are published that helps develop students' creativity in the process of exploring solutions through flexible. Especially in elementary school students, receive training in programming for the first time feel constraint to lean programming language. But the use of EPL reduces difficulties. So Students can focus on the thinking. Therefore, in this paper, a draft was produced and applied to observe the creative elements increasing of elementary school students. And modify and supplement the draft through improvements to a draft obtained by applying.

**Keywords** Educational programming language · Computer education · Creativity

---

Y. Yang · D. Hyun · E. Kim · J. Kim (✉)  
Department of Computer Education, Teachers College, Jeju National University,  
Jeju, Korea  
e-mail: jkim0858@jejunu.ac.kr

Y. Yang  
e-mail: atriple1981@naver.com

D. Hyun  
e-mail: gody5@naver.com

E. Kim  
e-mail: eunjg@mail@naver.com

J. Kim  
Polytechnic Seoul Kangseo College, Seoul, Korea  
e-mail: jjkim70@kopo.ac.kr

## 1 Introduction

Programming education is very suitable education for the direction of education. Program planning, designing, implementing and revising procedure is very similar to solving social problems in the reality. Students will, therefore, experience problem solving indirectly and enhance their creative problem-solving skills through programming in various methods on given problems [1].

EPL ensures to design one's thoughts easily through the use of the intuitive command and block-script based coding. This will reduce repulsion of programming education and initiate interests. Students may also visually identify their program movement process thereby check and modify errors.

The study prepared draft teaching materials to examine the effect of EPL education, and executed TORRANCE TTCT (Drawing) Type-A inspection, a creative thinking test, before and after implementing the education, respectively [2]. Draft teaching materials are crafted based on 40 min classes in the amount of eight sessions suitable for circumstances of schools; education is implemented on elementary students in fourth grade by an elementary school teacher who pursues the master and doctor's courses.

Teaching materials are revised and supplemented considering education details, such as the cognitive level of students and conditions of schools based on the inspection results of before and after education and experience of using draft teaching materials.

## 2 Preceding Research

There are many researches reporting that programming education via LOGO enhances creativity. With regard to thinking skills, Clement surveyed the affect of LOGO programming through numerous researches and published the results that LOGO affects mathematical knowledge and enhances thinking skills (problem-solving ability and creativity) [3]. In addition Clement (1991) divided 73 students in the 8 years into three groups; a LOGO programming group, a group using creativity enhancement programs, not LOGO, and a comparison group, and examined the affect of LOGO experience in creativity. The experiment result provided affirmation that there was Figural Creativity learning transfer after learning LOGO and also uncovered that LOGO can contribute in promoting verbal creativity [1].

The preceding research shows a positive role of LOGO in the enhancement of creativity. However, as there are no comparative analyses with various new EPLs, it is challenging to describe the pros and cons of LOGO only when compared with EPL. Hence, there is a limitation in selecting ELP that considered the characteristics of students.

One of the most popular ELP that has introduced recently is Scratch. Today, vibrant research is being conducted on Scratch, and here are a few examples as follows.

As a result of applying the learning content configured base on creative problem solving model (CPS) to improve complex cognitive skills and strategy to stimulate learners' intrinsic motivation during programming in sixth grades discretionary activity time, the study confirmed that Scratch programming learning is effective in improving learners' intrinsic motivation and problem solving skills [4].

The programming education using Scratch is a positive influence in effectiveness and satisfaction with respect to the cognitive area of learning. Especially, the performance of Scratch use is statically higher to visual-inclined learners. Therefore, we can presume that the programming process of elementary students has a great impact in learning effect.

However, Scratch also demonstrated independent influence only and there is a lack of research on differentiating the benefit of Scratch only through comparison with other ELPs.

Hence, the study carried out comparative analysis between Scratch and LOGO, which are known to provide help in improving creativity through preceding research and examined the enhancement of detailed elements of creativity by developing draft teaching materials, and revised problems suggested from field application of draft teaching materials in order to provide materials that can be utilize in education and EPL selection suitable for education purpose as well as student's characteristics.

### **3 Development of Draft Materials**

#### ***3.1 Education Details and Configuration System***

Most of all, in order to teach programming, somewhat esoteric topic to elementary students, we need to continuously induce their interests. If we suggest visual stimulus which lead to a result, students will feel a sense of accomplishment alongside improved interest on learning as well as reduce repulsion on programming. In addition, the content should be configured in education details that could improve their creativity. The study developed teaching materials that ensure programming and demonstrating creativity as well as expanding the scope of their thinking, rather than a mere suggestion and analysis of new programs.

While LOGO and Scratch have features that they can easily approach as educational programs, they have a difference in programming environment. Suppose LOGO writes program when command is input, Scratch uses command that uses a method to drag and stack blocks expressed in graphic.

Hence, we come to determined that the approach methods should be different in teaching the two educational languages. We used LOGO to advance outputs

through stepwise education of command while Scratch is used to increase the difficulty of making content by subject.

### Details of Teaching Material Development

It is a transitional stage towards the formal operational period from concrete operational period, and the study selected as follows considering the programming language features and the level of elementary school students in fourth grade who encounter programming for the first time [4].

#### LOGO

We let students to introduce various diagrams and draw them using each different command by session, and introduced variables as well as control command and incorporated a variety of features of LOGO more in detail.

#### Scratch

The study abstracted educational elements which can be used in Scratch in the elementary school level based on programming concept. We abstracted educational elements ensuring to learn and practice the basic functions and operating methods of Scratch whereby configured in education data development.

### Configuration System of Teaching Materials

The steps to derive creative outputs were initially introduced in the creative research of Wallas in 1926 [5]. Wallas divided the courses of obtaining creative outputs in four steps as Fig. 1.

Learners will become familiar with the knowledge of details relevant to individuals in the Preparation Step, analyze and understand acquired information in the Incubation Step, a solution will be provided in the Illumination Step and the solution will be verified in the Verification Step.

Base on these, the study first configured steps to apply LOGO and Scratch programs and developed educational materials accordingly. The study configured Fig. 2 by considering the features of each program language for program configuration of single session amount, and divided into four steps to ensure manifestation of self-directive creative thinking and encourage divergent thinking by associating with learning objectives.

#### Step 1 (Concept understanding and finding principles).

Understand the details of today's study as well as related concept and principles and execute simple programs firsthand. Concept and principles should be provided with figures and also include details studied in the previous sessions.

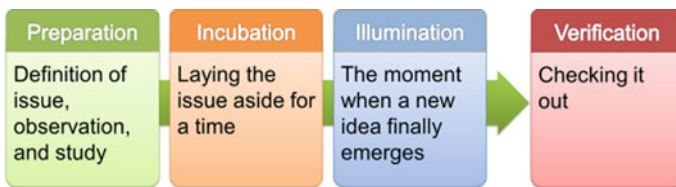


Fig. 1 The Wallas model for the process of creativity



**Fig. 2** Material configuration system

	LOGO	Scratch
Step 1	Concept Understanding	Understanding Principles
Step 2	Familiarizing Command	Command Examining
Step 3	Programming	Command Executing
Step 4	Thinking	Problem Solving

*Step 2 (Familiarizing and examining command).*

In the steps, learns examine a concept related to problems to study today and explore rules and instructions based on completed output. The discovered rules and command will be used at the programming.

*Step 3 (Programming and command executing).*

This step will ensure programming using the rules and command discovered in the previous step.

*Step 4 (Contemplating and problem solving).*

We designed this step to make new programs using a concept or express differently by modifying programs related to details already learned.

## 4 Improvement Measures Study Through Field Application

### 4.1 Field Application and Analysis

Based on the developed draft worksheet, we organized four groups, which divided into men and women, respectively, targeting each class of fourth grade of Shinjeju Elementary School (Scratch) and Gangjung Elementary School LOGO) Jeju Special Self-governing Island, and implemented pre-post inspections through TORRANCE TTCT (diagram) creativity test A-type [2]. We also configured the homogeneous group by abstracting samples from each group as well as comparative analysis on detailed elements creativity including pre-post analyses.

As a result of comparative analysis on the impact of each language on the creativity area of students, both languages as shown in Table 1 were helpful in improving creativity. Especially, whereas LOGO causes positive influence in the fluency area among creativity area, Scratch affects on resistance and abstractness. These results implies that the use of various functions via the command of LOGO are suitable to the features of fluency, and that the process of program implementation using the graphic of Scratch can be checked visually suits the characteristics of abstractness and resistance, which result in improving the area of each creativity.

**Table 1** Analysis of pre-post creativity enhancement

EPL	Gender	Creative element	(Average)	t	Freedom degree	Significance probability (both side)
LOGO	Male	Fluency	-17.38	-4.950	7	0.002**
		Originality	-9.63	-1.925	7	0.096
		Abstractness	-8.38	-0.706	7	0.503
		Sophistication	-5.38	-2.641	7	0.033*
	Female	Resistance	-5.63	-0.802	7	0.449
		Fluency	-25.10	-6.138	9	0.000**
		Originality	1.80	0.596	9	0.566
		Abstractness	-7.30	-0.959	9	0.362
Scratch	Male	Sophistication	-1.70	-0.749	9	0.473
		Resistance	-1.50	-0.271	9	0.793
		Fluency	-6.67	-2.082	8	0.071
		Originality	-5.44	-1.135	8	0.289
	Female	Abstractness	-31.56	-3.914	8	0.004**
		Sophistication	-2.33	-1.228	8	0.254
		Resistance	-14.11	-2.615	8	0.031*
		Fluency	-11.78	-2.075	8	0.072
Female	Originality	-4.22	-0.907	8	0.391	
	Abstractness	-25.22	-2.889	8	0.020*	
	Sophistication	-3.22	-1.442	8	0.187	
	Resistance	-13.11	-5.063	8	0.001**	

\*  $p < 0.05$ , \*\*  $p < 0.01$

It may seem that the features of each ELP played positive roles in the enhancement of creativity, but a programming process using a command in the case of LOGO suffers a downside in that it is difficult to detect errors—if they occur—in the execution of the program process, while a programming process using Scratch block entails a restriction and unnecessary long process of programming in using advanced function. Hence, we should solve a challenge to establish detailed and systematic education process suitable for the development steps of students when teaching the programs considering these limitations of the programs.

## 4.2 Improvement Measures for Draft Teaching Materials

### Education Details and Configuration

In a programming process, it is necessary to configure a process that emphasizes problem solving based on learning simple programming techniques. Therefore, it is needed to provide additional details of teaching and learning plans as well as learning materials produced in the existing eight sessions. In particular, trails should be partitioned to ensure students to fully understand the concept in the step that introduce (variables and logical operation) the upper mathematical concept.

## **Education Details**

Programming courses that require systematic design, namely the function call and recursive call, need adequate thinking and the time to practice by adjusting difficulties, and various methods should be introduced to students to ensure to solve problems by adding the basic algorithm details. The focus should be on the use of effective combination, not mere basic function learning. And the length of code should be reduced as well. Even if an overall length of the code becomes lengthier, it should be bundled up by feature in order to reduce fatigability of students. This problem was displayed especially with Scratch's code but the lengthier of a code that uses a block, the more the visual fatigability.

Therefore, functions should be properly divided by feature whereby approach using a calling method by the need of the function.

## **5 Improved Teaching Material Development**

### ***5.1 Teaching Material Improvement Details***

The study believed that there is no problem for elementary school students to use EPL. It was, however, challenging for fourth graders to understand the mathematical concept. Therefore, we extended the learning materials that are prepared in eight sessions of the draft teaching materials into 15 sessions and add more sessions in the process of concept introduction to ensure seamless application for students to apply in class. With respect to the basis of the 15 sessions, our suggestion is based on the circumstance of schools in which compiles about 16 h of one semester among 64 h of the total annual training hours of discretionary activity as information education for students in fourth, fifth, and sixth grades. In addition, in order to improve thinking skills using the acquainted features, the study adjusted the configuration of sessions and provided additional relation/logical operation as well as algorithm topics.

In the case of LOGO, we stressed the process that solve problems through the recursive call of functions, and as for Scratch, we divided code preparation by feature so as to call when needed in order to reduce visual fatigability resulting from lengthy code and emphasized by displaying the flow of thinking in a diagram form.

## **6 Conclusion**

The study produced and input draft teaching materials on LOGO and Scratch, the most used EPL and conducted cross-tab analyses on the influence of the detailed areas of creativity between the two languages, which are unprecedented in

preceding research in addition to the improvement of creativity often observed in the existing research. Moreover, we identified problems in design and the course of draft teaching development through on-site inspection and confirmed the matters that required attention when the materials are used in the actual field whereby developed suitable teaching materials for the characteristics of each language in addition to increase the vast use of the materials.

To foster creative citizens who meet the demand of the age, it is vital to continue stepwise education through the establishment of elaborated education courses, instead of mere one-off education. With respect to computer education, we believe it is possible to establish efficient education courses if the instructional designed used in the study is fragmented and incorporated to various circumstances.

There are many educational languages besides LOGO and Scratch used in the study and a variety of languages are used abroad in education courses. Hence, Korea should also develop research on a variety of ELPs as soon as possible.

## References

1. Clements DH (1991) Enhancement of creativity in computer environments. *Am Educ Res J* 28(1):173–187
2. Torrance EP (1999) Torrance test of creative thinking: norms and technical manual. Scholastic Testing Services, Bensenville
3. Clements DH, Battista MT (1989) Learning of geometric concepts in a logo environment. *J Res Math Educ* 20(5):450–467
4. Jeongbeom S, Soenghwan C, Taewuk L (2008) The effect of learning scratch programming on students' motivation and problem solving ability. *J Korean Assoc Inf Educ* 12(3):323–332
5. Wallas G (1926) *The art of thought*. Harcourt, San Diego

# Design and Implementation of Learning Content Authoring Framework for Android-Based Three-Dimensional Shape

EunGil Kim, DongLim Hyun and JongHoon Kim

**Abstract** This study was conducted to create a more tangible educational environment by allowing learners to directly control three-dimensional learning contents through the touch interface of smart devices. Furthermore, because there are limitations to the acquisition and provision of three-dimensional learning contents due to difficulties in producing them, the proposed framework was designed to allow teachers and learners to directly produce and share contents. The proposed framework is based on intuitive XML language and the application was built to enable playback and authoring in Android-installed devices. In addition, a server environment was constructed for contents sharing. The feasibility of the proposed framework was verified through expert evaluation and its potential for utilization of new learning contents was positively evaluated.

**Keywords** Android · 3D learning contents · m-Learning

## 1 Introduction

Today's smart devices have expanded the scope of educational contents because have much more advanced hardware compared to PDAs in the past and are equipped with various sensors such as GPS, acceleration, and compass.

---

E. Kim · D. Hyun · J. Kim (✉)

Department of Computer Education, Teachers College, Jeju National University,  
Jeju, Korea

e-mail: jkim0858@jejunu.ac.kr

E. Kim

e-mail: computing@korea.kr

D. Hyun

e-mail: gody5@naver.com

**Fig. 1** Framework for bidirectional contents authoring and playing



Educational media using smart devices can provide infinite learning contents through the Internet and enable the provision of three-dimensional contents with visual, sensory expressions.

Based on these advantages of smart devices, this study investigated a method of creating learning contents that can be directly used by learners using touch sensors. Furthermore, three-dimensional learning contents that can express even the sense of texture, etc. to attract learners' interest were produced. However, for effective education, measures to widely supply 3D learning contents were required. To create 3D learning contents, programming abilities as well much time, effort, and costs are required to implement the development environment.

Thus, in order to achieve quantity and quality of 3D learning contents, this study proposed a contents production framework in which anyone can easily produce and use learning contents. This implies that from the educational perspective, teachers and students with different educational levels can produce a wide variety of learning contents as shown in Fig. 1.

Android provides frequent updates for continuous improvement and many new hardware control APIs [1]. This study defined XML format for 3D learning contents based on Android and designed a framework for analyzing and showing them to learners in a controllable form.

## 2 Theoretical Background

### 2.1 Android

The core of the Android platform is the Linux kernel which plays the role of an operating system that manages device drivers, resources, and so on. Above the kernel, there are OpenGL for 3D graphic, etc. [2]. As part of the Android project, Google developed the Dalvik virtual machine for optimal design of low-power mobile devices after a lot of research. It can reduce the capacity of applications by combining various Java class files into .dex and reusing duplicate information. Most applications are developed in Java and access the kernel and libraries through

the Dalvik virtual machine [3]. The application framework contains telephone communication, position tracking, and content provider, and applications are developed using APIs [4].

## 2.2 OpenGL

OpenGL is managed by the Khronos Group Consortium founded by many global companies.

In the OpenGL pipeline as shown in Fig. 2, geometric data such as vertices, lines, and polygons go through Evaluator and Per-Vertex Operation, but pixel data (pixels, textures, etc.) go through different paths, gather at Rasterization and are recorded in the Frame Buffer.

Display List stores OpenGL commands, geometric and pixel information for later execution and it can improve performance because cache can be applied. Evaluator transforms the information inputted through a polynomial function into coordinates to obtain vertices used in curves and surface shapes. One point in a 3D space is projected into a vertex on a screen through Per-Vertex Operation. Furthermore, if advanced functions such as lighting are activated, lighting effect-related operation is performed for the color information of vertices. The main feature of Primitive Assembly is clipping. After this step, a geometric primitive with a perfect shape consisting of transformed or clipped vertices is created for color, depth, texture coordinates, and rasterization steps.

Pixel data are processed along a different path from the geometric data in the OpenGL rendering pipeline, and the pixels saved in an arrangement of a specific format in the system memory are read first. To these data, operations such as scaling and basis application are applied. This process is called Pixel Transfer Operation, and the result can be saved in the texture memory or directly rasterized.

In the Rasterization step, all the geometric data and pixel data are transformed into fragments. Each fragment is an array that includes colors, depth, line thickness, dot thickness, and anti-aliasing. Each fragment corresponds to a pixel in the

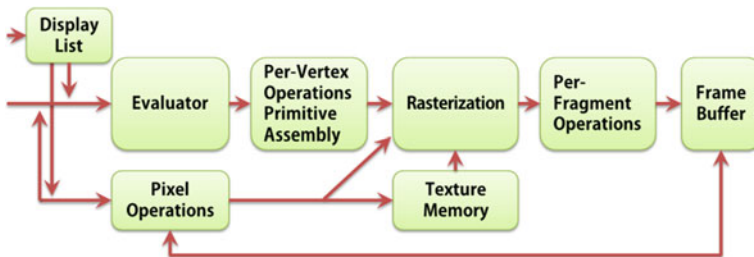


Fig. 2 OpenGL pipeline [5]

Frame Buffer. These fragments are changed or removed through Per-Fragment Operations before they are stored in the Frame Buffer.

### 3 Construction of Framework

The overall flow of the framework proposed in this study for producing and playing learning contents for 3D object is shown in Fig. 3.

When the application and learning contents are started, the Main Activity communicates in XML for update information with the server through Thread at the background. Thus, learners can receive continuously updated learning contents. Not only the update information, but also the learning contents are defined in XML which allows communication in reduced volumes. The XML information of the learning contents is analyzed by XML Parser in the DOM method. The analysis results are created in a tree form which allows easy interaction with learners for insertion and deletion of 3D objects [6]. The space coordinates of the required vertices and the texture space coordinates for expression of texture are calculated through Vertex Operation. The calculation results are processed by OpenGL through the Renderer interface of GLSurfaceView and displayed as 3D object on the screen [7].

#### 3.1 Contents Authoring and XML Script

The contents authoring part also works in the Android-embedded device. Users input 3D object by touching them to the 3D projection guideline and pile them up to gradually create the total shape. For the positions of the inputted shapes, the row, column and layer data are created in XML which is parsed during playback and the space coordinates of each 3D object are calculated. User can select a surface quality of the 3D object and the resource ID of the image representing the selected quality is saved in XML. Then it is applied to the surface when the contents are played back together with the space coordinates of the 3D object.

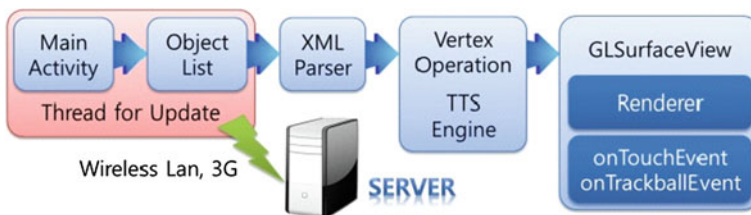


Fig. 3 Total flowchart for learning contents



A learning scenario consists of text in XML. The scenario text includes explanation about learning method and is converted to speech in multilanguages and told to learners using the TTS engine of Android [8]. Learning scenarios output the viewpoint of 3D objects in line with the sequence through interaction with users.

### ***3.2 Contents Playback and Control***

The contents playback part largely consists of XML Parser that analyzes XML information and saves it by element, Renderer that shows the 3D object on the screen, and the TTS engine that transforms text to speech. The interaction between contents and user is based on touch and trackball control.

The XML information is analyzed in the DOM method, and the resource elements related to the screen display of 3D object are divided by delimiters and sent to Vertex Operator. The Vertex Operator calculates the space vertex coordinates of each face to express the location of the 3D object together with their surface texture. The length of one side of the 3D object is defined by calculating the value of the longest axis so that the prospect of the OpenGL virtual camera is not exceeded, and the center of the entire 3D object is specified as the origin of the virtual space so that it will not go out of the screen during control [9, 10]. The space coordinates of the vertex calculated by the Vertex Operator and the surface texture information are sent to the Renderer. In the renderer, the 3D object is rasterized through the OpenGL ES API provided by the Android album, saved in the Frame Buffer and displayed on the screen.

The scenario elements of XML are analyzed by Parser and saved in the multi-dimensional array for each angle element. When user requests the saved information by entering the sequence by touch input, the text element values are sent to the TTS engine and outputted as speech and the external resource information is displayed as Android View.

## **4 Implementation of Authoring and Playback Framework**

To verify the feasibility of the proposed framework for authoring and playing 3D object learning contents, it was implemented in the development environment based on Android 2.3 Gingerbread.

Users can freely rotate the 3D object to the desired direction via the Android touch and trackball interface and the figures were instantly drawn by the Frame Buffer. The rotation information, text saved in the XML scenario can be provided as buttons for user operation as needed. In addition to the default learning contents, users can also produce and play scenarios.

The learning contents that are created by user through the authoring part can be provided to other users through sharing. When shared, the XML file of the learning contents is sent to the server and the server updates the learning contents list and version. Other users who run the application can compare their version with the server version and download the XML file of the updated contents. Then the updated list of contents is displayed through the Android ListView.

This framework application was installed and tested in a smart phone with the hardware specifications of CPU QSD8250 1 GHz, RAM 512 MB, and WVGA display. It took less than 1 s for parsing the XML script, preparing the TTS engine, and displaying the content on the screen. However, considering the various factors of the mobile environment, external resource loading was not added to the XML.

## 5 Expert Evaluation

The usability and possible improvements of the prototype application implementing the framework were evaluated through experts who were 20 regular licensed teachers of class 1 who had at least 10 years of experience as elementary school teacher.

The expert evaluation method was to answer a questionnaire after demonstration and use of the application. The experts checked one of the answers according to 5-level Likert scale in 2.5 point intervals, or selected an answer and stated their opinions depending on the questions.

The prototype application was largely divided into playback and control part and authoring part for this survey. The answers to the questions about the playback and control part were quantified by the 5-level Likert scale. The prototype application received 8.875 points for the reflection of learner controls in the expression of learning contents, and 9 points for realistic expressions. Furthermore, the touch interface of smart devices was positively evaluated at 8.25 for user interaction with contents. Thus, the use of touch and trackball interface of smart devices for control of learning contents seems to be effective. The authoring part using touch interface received 8.375 points for easiness of creating learning contents and 7.75 points for provision of features required for creating learning contents.

The demand for provision of learning contents about various basic 3D object in addition to the cube shape that was developed for the prototype was the highest at 30 %, followed by the demand for easier and simpler authoring part, the demand for manual for authoring and use of contents at 20 %, and the demand for the development of applications that can be run on other platforms at 15 %.

The analysis results for the requirements for activation of 3D learning contents using smart devices, the requirement for a sufficient volume of default learning contents was the highest at 60 %, followed by the requirement for interaction among learners using various sensors which is the advantage of smart devices.

This result suggests that users demand learning methods using various sensors (gravity, acceleration, GPS, etc.).

## 6 Conclusions

This study implemented a prototype of learning contents for 3D object that can be used in education on the basis of highly portable smart devices equipped with various sensors. The learning contents of our study allow teachers and learners to directly produce learning contents through touch interface and to share contents with one another as they are written in XML. Furthermore, scenarios can be defined to control the flow of learning contents and it is also easy to interact with the contents through the touch interface. From the aspect of communication environment of mobile devices, they are also positive for the new learning environment using smart devices considering that small size XML learning contents can be quickly shared.

The feasibility of the application produced on the proposed framework was verified through expert evaluation by teachers and the future improvements were discussed. The evaluation result was positive for the new educational environment, and the use of various sensors of smart devices for interaction with learning contents was required. Future studies will investigate learning methods using various sensors such as gravity, acceleration, and GPS in addition to touch, and develop learning contents that can be directly produced and controlled by learners based on these methods.

## References

1. Murphy ML (2010) *Beginning Android2*. Apress, New York
2. Hashimi S, Komatineni S, MacLean D (2010) *Pro Android2*. Apress, New York
3. Dalvik virtual machine internals. <http://developer.android.com/videos/index.html#v=ptjedOZEXPM>
4. Google. <http://developer.android.com/guide/basics/what-is-android.html>
5. Munshi A, Ginsburg D, Shreiner D (2008) *OpenGL ES 2.0 programming guide*. Addison-Wesley Professional, Boston
6. w3schools. [http://www.w3schools.com/Dom/dom\\_parser.asp](http://www.w3schools.com/Dom/dom_parser.asp)
7. Google. <http://developer.android.com/guide/topics/graphics/opengl.html>
8. Text-To-Speech and Eyes-Free Project. <http://developer.android.com/videos/index.html#v=xS-ju61vOQw>
9. Pulli K, Aarnio T, Miettinen V, Roimela K, Vaarala J (2007) *Mobile 3D graphics*. Morgan Kaufmann, San Francisco
10. Khronos. <http://www.khronos.org/opengles/sdk/docs/man>

# A Study on GUI Development of Memo Function for the E-Book: A Comparative Study Using iBooks

Jeong Ah Kim and Jun Kyo Kim

**Abstract** Currently used electronic books (hereafter referred as “e-book”) do not reflect people’s memo-taking behavior patterns, as an intuitive Graphic User Interface (hereafter referred as “GUI”) is not used in e-books. A study was carried out to suggest a GUI prototype that applies users’ memo-taking behavior patterns to the memo function of iBooks to enable greater usability. The prototype herein suggested is a memo GUI prototype that can apply people’s real-life memo-taking behavior to the memo function based on the iBooks interface. For the next step, a usability test was conducted on the suggested prototype and the iBook’s memo interface through four environmental factors and eighteen evaluation factors. Five research subjects participated in the usability test on two types of interface, and a questionnaire was analyzed using a paired t-test (T-test). Analysis of the questionnaire showed that users were highly satisfied with the usability of the newly suggested prototype, compared to iBooks.

**Keywords** E-book application · GUI · Usability test

## 1 Introduction

### 1.1 Background and Purpose

Technology development today brings out new media such as smartphones, tablet PCs, and e-books providing a viewer function. These remove the inconvenience of carrying books around and offer the convenience of selecting books through the

---

J. A. Kim

Design Department, Chung-Ang University, Seoul, Korea

e-mail: sam2496kr@hotmail.com

J. K. Kim (✉)

Visual Design Department, Chung-Ang University, Seoul, Korea

e-mail: kjk3134@korea.com

Internet and applications instead of visiting a bookstore in person. Although e-book applications have diversified in accordance with this rapidly changing market, little research has been performed on the GUI environment of e-book applications from the perspective of usability. Specifically, research is now required on a memo function for users who would like to 'take notes', rather than on information provided in e-books. To achieve this, the study aims to research individuals' memo-taking behavior in an analogue environment to suggest a memo GUI prototype that can apply their memo-taking behavior to e-books.

## ***1.2 Range and Methods of Research***

For the study, two types of memo-taking behavior patterns were derived from 10 study participants' memo-taking behavior, and they were applied to the iBooks memo interface to suggest a GUI prototype. Then a usability test was conducted on the iBook's current memo interface and the newly suggested interface. Usability test factors were derived through 5 experts' verification, and a survey questionnaire was prepared for the usability test. It was conducted on five research subjects for three days, and analyzed using a T test. Furthermore, the study range was restricted to the iBooks interface environment, and the survey was conducted on a prototype, not a developed version of the suggested interface.

## **2 Memo**

### ***2.1 Questionnaire Survey of Memo-Taking Behavior***

A questionnaire survey of 150 people was conducted from March to April, 2012. The questionnaire was conducted on individuals who used analogue books to identify what their behavior was when taking memos, and on the usability of the currently-used iBook memo function. The result showed that auxiliary tools (e.g. post-it) used for taking memos account for 38.8 % of memo-taking when reading books, followed by hardly taking memos, taking memos on the book, and others, which accounted for 30.1 %, 20.8 %, and 10.3 % respectively. Furthermore, when using auxiliary tools (e.g. Post-It notes) to take memos, the results showed that underlining parts related to Post-It memos accounts for 12 % while randomly sticking those memos accounts for 10 %; taking memos in a blank space does not decrease the readability of a book. According to an analysis of the results, memo-taking behavior can be classified into the two following patterns: first, people take memos in blank spaces in the book, and secondly, people use auxiliary tools to take memos.

## 2.2 Memo Functions in E-Book Reader Applications

### 2.2.1 A Comparison of Memo Functions in e-book Reader Applications

Currently, around the world, the Amazon Kindle and iBooks are the most well-known e-book readers and applications. The Jungle Kindle is an e-book reader, and iBooks and similar e-book applications are being developed around the world for the Android and iOS platforms. These e-book platforms offer readers different things in terms of size, GUI, and so on. When comparing the Amazon Kindle and iBooks’ memo function, the following becomes apparent: the Amazon Kindle was developed for use on the iPad and Android, and the Amazon Kindle application and Aldiko book reader application can be used on the Android, but not all platforms offer the memo function (Fig. 1).

### 2.2.2 iBooks and Usability

The iBook’s memo function is simple to use. Drag and select a part to take a memo, and then press a memo selection button. Then a memo pop-up window appears, and users can take memos. A memo can then be typed. After the memo is done, tap any space other than the memo pop-up window, and the pop-up window disappears and a memo icon appears on the end at the right. A simple questionnaire survey was conducted concerning the usability of the iBook’s memo function. The results showed low satisfaction in that only 25 % of the respondents were satisfied with the usability of the iBook’s memo function. On the question of whether or not the function is convenient to use, just 33 % of the respondents responded positively, which indicates poor usability for users. A short essay question about what inconveniences related to the current memo function occur produced many different opinions, including: difficult typing, hard to retrace



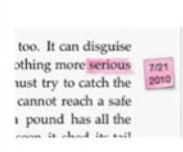

type	iBooks	Amazon Kindle	type	iBooks	Amazon Kindle
Memo GUI			Memo Icon		
	<ul style="list-style-type: none"> <li>- Large form for entering notes</li> <li>- Similar design with Post-it</li> </ul>	<ul style="list-style-type: none"> <li>- Context shouldn't change</li> <li>- User enters Note on 6 lines of text disappears.</li> </ul>		<ul style="list-style-type: none"> <li>- Shows date note entered</li> <li>- Tapping note brings up form on left</li> </ul>	<ul style="list-style-type: none"> <li>- Very small icon indicate note is present</li> <li>- Tapping note brings up form on left</li> </ul>

Fig. 1 A comparison of memo functions

memos, hard to take several memos in one line, the memo pop-up window blocks one's view of one's writing, memos are non-interactive and hard to use, use of writing or pictures is highly recommended, etc.

### ***2.3 Prototype Memo GUI***

A new prototype memo GUI was suggested, adjusted for GUI composition factors, four environmental factors, and inconveniences derived from the survey: difficult typing, hard to retrace memos, hard to take several memos in one line, memo pop-up windows blocking one's view of one's writing, memos are non-interactive and hard to use, use of writing or pictures was highly recommended, etc. The characteristics of the prototype are five-fold, as follows: First, many memo icons can be put into one line. Second, a sketch can be offered. Third, location of memo icons can be changed by users. Fourth, windows are slightly transparent to show the text below. And, fifth, font size and color can be modified. This prototype memo GUI is described in Fig. 2 in detail.

## **3 Usability Test of iBook Memo and Prototype Memo**

### ***3.1 Usability Test Method***

A comparative study was conducted between the prototype memo GUI, to which new functions have been added, and conventional iBook memo functions. Prerequisites for testing usability were the extraction of evaluation factors and the composition of four environmental factors into a matrix.

### ***3.2 Collected Factor Data for Usability Test***

This study conducted a usability evaluation using eighteen usability evaluation factors of e-book applications. By creating a matrix with the collected factors, five experts distinguished overlapping parts and colors of unnecessary parts to show the importance status as a number, and extracted factors from collected data. After this, factors were derived from the experts' evaluation. Table 1 shows these derived factors to test usability, which can be used as a basis for designing and deriving a questionnaire for a usability test. The following eighteen factors can be suggested as a guideline for an evaluation framework to be a principle for usability tests of tablet PCs. A usability test can be conducted using the following factors:



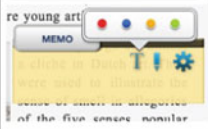
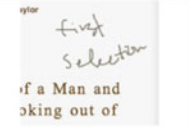

New prototype GUI		New prototype GUI	
	<ul style="list-style-type: none"> <li>-New design for memo, and new functions added.</li> <li>-Sketch function is available</li> <li>-Transparency is added</li> </ul>		<ul style="list-style-type: none"> <li>-Memo can be edited its size by user by drag and drop</li> </ul>
	<ul style="list-style-type: none"> <li>-Memo can use many colors of Fonts</li> </ul>		<ul style="list-style-type: none"> <li>-It is available to write without background</li> </ul>
	<ul style="list-style-type: none"> <li>- Memo icons can be moved &amp; edited by users freely</li> </ul>		

Fig. 2 Prototype for iBook memo function

### 3.3 Selection of Research Subjects

The matrix was composed by using the above usability evaluation factors and the four environmental elements of e-book applications extracted above, and an evaluation was performed to give scores based on each evaluation factor. Prior examination of each five users showed that all of them had used an iPad for more than six months, and two of them had previously used an e-book while the other three had not. Therefore, before the experiment, a time of 30 min was set in which users with no e-book experience could use an e-book; this was to reduce the difference between those who had used an e-book and those who had not.

### 3.4 Results and Analysis on Usability Test

The usability test questionnaire was analyzed using T test. In other words, an average score for each question item showed the degree of research subjects’

**Table 1** Evaluation of e-book application usability

Usability test factors of e-book applications
Consistency, Efficiency, Readability, Simplicity, Aesthetics, Alert, Operability, Accessibility, Effort, Intuition, Frequency of Mistakes, Clarity, Learning Ability, Satisfaction, Personalization, Help, Memorization, Understanding of Users, Feedback



satisfaction. If the significant probability (p value) is  $0.05 \leq \alpha$ , a null hypothesis is rejected and the analysis shows a significant difference. If the significant probability (p value) is  $0.05 > \alpha$ , a null hypothesis is not rejected and the analysis shows no significant difference. Only evaluation results showing a significant difference were evaluated, and usability was differentiated on items showing 3.0 (normal), a significant difference. In other words, a difference of 3.0 or above is normal or above and can be manipulated while a difference below 3.0 causes inconvenience to manipulate. Figure 3 shows is a questionnaire to compare usability tests.

**4-5-1 Questionnaire Analysis**

Items showing a significant probability (p-value) of less than 0.05 were classified. The criteria for this were the eighteen types of usability test items, and GUI visual factors of e-books such as color, layout, typography, multimedia, graphics, and navigation. Multimedia was deleted in that it is not applicable to the criteria. According to analysis results, a new prototype showed a significant probability of 0.05 or below in the following twelve evaluation items of color: Consistency, Efficiency, Readability, Simplicity, Aesthetics, Accessibility, Effort, Clarity, Learning Ability, Satisfaction, Personalization, and Memorization. The new

	Usability Factors	N	iBooks		New Prototype		F value	P value
			average	standard deviation	average	standard deviation		
Consistency	Graphic	10	3,60	,966	3,80	,789	1,353	,257
	Navigation	10	3,60	,516				
	layout	10	3,70	,949				
	Multimedia	10	3,20	,632				
	Color	10	3,70	,675				
	Typographic	10	3,00	,943				
	Total	60	3,47	,812	3,10	,994		
Efficiency	(Graphic, Navigation, layout, Multimedia, Color, Typographic) Total	60	3,35	,899	4,20	,632	1,482	,211
Readability	Total	60	2,93	1,056	3,70	,823	,327	,894
Simplicity	Total	60	3,23	1,047	3,00	,816	2,147	,074
Aesthetics	Total	60	3,28	,922	3,60	,516	,421	,832
Accessibility	Total	60	2,83	,924	3,60	1,075	9,636	,000
Effort	Total	60	2,75	,866	3,70	,675	2,284	,059
Intuition	Total	60	3,23	,851	3,90	,738	1,949	,101
Frequency of mistake	Total	60	2,63	,843	3,70	,675	1,507	,203
Clarity	Total	60	3,07	,756	3,60	1,075	1,763	,136
Learning ability	Total	60	3,32	,676	3,50	1,179	1,292	,281
Satisfaction	Total	60	3,00	,759	3,37	,920	,894	,492
Personalization	Total	60	2,33	1,115	3,40	,699	1,200	,322
help	Total	60	2,05	,928	3,40	,516	,860	,514
Memorization	Total	60	1,77	,945	4,20	,675	2,262	,061
Feedback	Total	60	1,67	,729	2,70	,949	,869	,508

**Fig. 3** T-test results

prototype showed a significant probability of 0.05 in ten evaluation items in layout, in color, in typography, in graphic, and in navigation. As such, compared to the conventional iBook's memo function, the newly suggested prototype's memo function resulted in higher satisfaction in that it showed a significant probability of 0.05 or below, and it showed a very significant difference in evaluation items. Also, the new prototype showed a significant probability of 0.05 in ten evaluation items in accessibility as well.

## 4 Conclusion

The study herein was carried out to suggest a GUI prototype where usability based on memo-taking behavior patterns can be applied to the iBook's memo function as an e-book application. A memo GUI prototype that can apply individuals' memo-taking behavior to memo functions based on the iBook's interface environment was suggested. For the next step, a usability test was conducted on the suggested prototype and the iBook's memo interface using four environmental factors and eighteen heuristic evaluation factors verified by five UI experts. Five research subjects participated in the usability test on two types of interface, and a questionnaire was analyzed using T test. Through the questionnaire analysis, users were more satisfied with the usability of the newly suggested prototype interface, compared to the iBook interface. As a result, a new prototype contains typographic improvements to the current inconvenient environment, the addition of a memo icon and a new editing interface, various GUI improvements including the ability to adjust size automatically, and improvements to the overall design. A test comparing the usefulness of this newly improved environment using the most commonly used e-book platform, iBooks, was done and proved that usefulness had greatly increased.

**Acknowledgments** This work has been supported by the Chung-Ang University research fund.

## References

1. Gong J, Tarasewich P (2004) Guidelines for handheld mobile device interface design. In: Proceedings of decision sciences, Institute annual meeting
2. Bahr GS, Nelson MM (2007) Development of a multiple heuristics evaluation table (MHET) to support software development and usability analysis. In: Universal access in human computer interaction: coping with diversity. Springer, Berlin
3. Nielsen J (2000) [www.useit.com/alertbox/20000319.html](http://www.useit.com/alertbox/20000319.html)
4. Gardiner E, Ronald GM (2010) The electronic book. In: Suarez MF, Woudhuysen HR (eds) The Oxford companion to the book. Oxford University Press, Oxford, pp 164
5. Kim JA, Kim JK (2012) Methods of portable PC GUI usability evaluation. J Digit Des 12(1):289–298, Korea Digital Design Society

6. Jeong Ah Kim, Jun Kyo Kim, Study on investigation and analysis of UI design trend of e-book applications, *J Korea Soc Des Trend* 36:253–264
7. Kim SH (2011) A study on the graphic user interface design for improving usability. Seoul National University of Science and Technology, Seoul
8. Chang W, Ji YG (2011) Usability evaluation for smart phone augmented reality application user interface. *Soc E-bus Stud* 35–47
9. e-book Oxford Dictionaries (2010) Oxford university press experimental study on usability evaluation of e-Book terminal of Seungjin Kwak and Kyoungjin Bae
10. Mehrabian A (1968) Communication without words. *Psychol Today* 56(4):53–56
11. Conati C, Gertner A, VanLehn K (2002) Using Bayesian networks to manage uncertainty in student modeling. *User Model User-Adap Inter* 12:371–417
12. Dadgostar F, Ryu H, Sarrafzadeh A, Overmyer S (2005) Making sense of student use of nonverbal cues for intelligent tutoring systems. In: *Proceedings international conference of ACM SIGCHI*, vol 122, pp 1–4
13. Wentzel K (1997) Student motivation in middle school: the role of perceived pedagogical caring. *J Educ Psychol* 89(3):411–419
14. Lehman B, Matthews M, D’Mello S, Person N (2008) What are you feeling investigating student affective states during expert human tutoring sessions. In: *ITS*
15. Ekman P (1989) *The argument and evidence about universals in facial expressions of emotions*. Wiley, New York
16. Meijer M (1989) The contribution of general features of body movement to the attribution of emotions. *J Nonverbal Behav* 13(4):247–268
17. Pavlovic V, Sharma R, Huang T (1997) Interpretation of hand gestures for human-computer interaction: a review. *IEEE Trans Pattern Anal Mach Intell* 19(7):677–695

# Relaxed Stability Technology Approach in Organization Management: Implications from Configured-Control Vehicle Technology

Toshihiko Yamakami

**Abstract** The requirement for quickly solving complicated problems poses a fundamental challenge to the modern organization. In order to cope with this challenge, an organization needs to tune its management methodology to reduce time and costs and to increase its organizational efficiency. The author proposes a theory called a Relaxed Stability Organization (RSO) framework from implications learned from Relaxed Stability Technology (RST) in Configured-Control Vehicle (CCV) technology. The author discusses the overall organizational challenges in general. Then, the author discusses how the viewpoint of RST technology can be applied to organizational management. The author presents the framework of RSO Theory.

## 1 Introduction

Gamification is a paradigm that utilizes techniques originating from game theory to improve user engagement with services. This framework can be applied to services, marketing, and education. This technique also promises to be useful for improving enterprise management. The rapidly changing industrial landscape increases demands for agility in the decision making process and execution process in the enterprises. In order to cope with these demands, the author proposes the Relaxed Stability Organization (RSO) framework as an analogy of Configured-Control Vehicle concept in military aviation technology. In this paper, the author outlines the concept of an RSO framework and its implications for computer-supported cooperative work.

---

T. Yamakami (✉)

ACCESS, Software Solution, 1-10-2 Nakase, Mihama-ku, Chiba-shi 261-0023, Japan

e-mail: Toshihiko.Yamakami@access-company.com

URL: [www.access-company.com](http://www.access-company.com)

## 2 Background

The aim of this research is to identify a framework that can cope with the agility of decision and execution in the fast-changing industrial landscape.

Organization is an open system. Grudin presented eight challenges for groupware from social dynamics [1]. Past research presented a range of different approaches, business process reengineering, open innovation, agile process, and so on.

The originality of this paper lies in its examination of relaxed stability in the context of organizational management.

## 3 Lessons from Relaxed Stability Technology

### 3.1 *What is the Aim of RST?*

Fly-by-wire (FBW) is a system that replaces the conventional manual flight controls of an aircraft with an electronic interface (Wikipedia). The movements of flight controls are converted to electronic signals that are transmitted by wires. The fly-by-wire system also allows automatic signals sent by the aircraft's computers to perform functions without the pilot's input, as in systems that automatically help stabilize the aircraft.

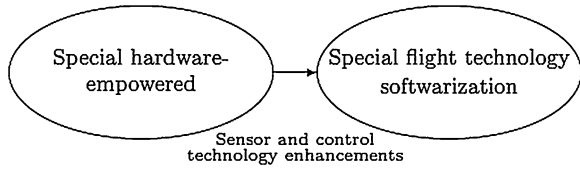
In aviation, relaxed stability is the tendency of an aircraft to change its attitude and angle of bank of its own accord (Wikipedia).

Fly-by-wire technology facilitated the Control-Configured Vehicle (CCV), which is an aircraft that utilizes fly-by-wire flight controls. In military and naval applications, it is now possible to fly military aircraft that have relaxed stability. Relaxed Stability Technology escapes from the restrictions induced by aviation design that focuses on stability. Although this increases the risks of aircraft crashes, it also increases flexibility in air-combat aviation. The increased flexibility in aircraft attitude control leads to an increased kill ratio.

### 3.2 *Why Don't CCVs Have Extra Hardware?*

Increased flexibility in aviation was the target of research in Control-Configured Vehicle technology in the 1970s. The basic design for aircraft was based on stability. The demand for stability restricted the flexibility of aviation, which led to a drawback in air-combat capabilities. Relaxed Stability Technology was introduced in order to increase the flexibility of air-combat capabilities. In order to bring relaxed stability, the initial test aircraft was equipped with special hardware (e.g. special wings). This special hardware facilitated an increased flexibility in aircraft attitude. During development, the sensor and control technology was

**Fig. 1** Softwarization of special hardware in advancement of RST technology



**Table 1** Softwarization of special hardware

Aspect	Description
Advantages	Improved indefectibility. Flexibility of added features
Disadvantages	Increased risk of instability. Increased demands of control processing power

enhanced. The final CCV achieved relaxed stability without the initial special hardware. The transition is depicted in Fig. 1.

The implications of this softwarization of special hardware are summarized in Table 1.

The most significant advantage is the improved indefectibility. The lack of physical characteristics makes the radar-detection of the relaxed stability-empowered aircraft difficult. This brings a critical advantage in the air-combat situations.

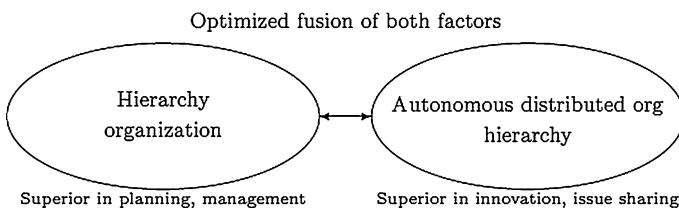
### 3.3 Implications for Organizational Management from RST

Advances in information technology increase the rapid speed of industry changes.

The needs to cope with these changes have increased on a global scale. Market changes are fast and radical, and today’s organization have to cope with these challenges using fast decision making and execution capabilities.

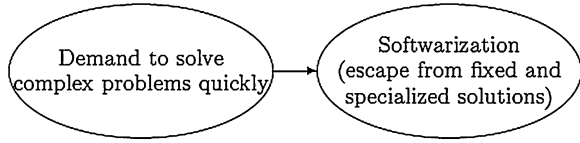
In order to cope with the challenges of market dynamism, organizations need to use a fusion use of different aspects of organization structures, as depicted in Fig. 2.

Market demands require increased flexibility of organizational capabilities, with softwarization of organization management, as depicted in Fig. 3.

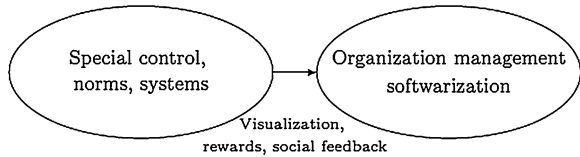


**Fig. 2** Utilization of multiple aspects of organization structures

**Fig. 3** Softwarization of organization management



**Fig. 4** Examples of the softwarization of organization management



In order to facilitate the improvement of agility in organizations, it is necessary to relax the existing constraints. These constraints were created to maintain the stability of organization. Relaxation requires additional spontaneous stabilization mechanisms to deal with decreased stability.

Examples of softwarization of organization management are depicted in Fig. 4.

There are several approaches to relaxed organizations. For example, it is feasible to construct a special team with specially-trained members. Another example is software control of an organization using visualization, rewards, and social feedback. The other is the softwarization of organization management, using software-controlled feedback based on ubiquitous real-time monitoring systems.

## 4 RSO Theory

### 4.1 Definition

RSO is defined as an organization that is managed by a highly coordinated visualization and feedback system in order to improve agility and flexibility with the intentional elimination of norms and regulations in existing organizations. In this context, “existing organizations” refers to organizations with legacy norms, regulations and fixed structures, such as hierarchy organizations.

### 4.2 What are Sensors and Controllers in RSO?

The lessons from mobile social game design for gamification of organization management are depicted in Table 2.

It is interesting to note that these techniques are deployed by skillful managers with conscious or unconscious manners in a real world landscape. Frankly speaking, these techniques are universal metrics that deal with the creation of

**Table 2** Lessons from mobile social game design

Aspect	Summary
Visualization	Visualization of achievement and the next target
Real-time human rewards	Rewards from human beings in a real-time manner
Sense of honor	Constant awareness of honor

human hopes. Hope represents a certain type of expectation that combines rewards with a confidence of achievement. Engineering that deals with the creation of human hopes has been widely recognized in the game design of mobile social games.

### 4.3 RSO Framework

The transition to a flexible organization is depicted in Fig. 5.

In order to pursue a flexible organization, it is necessary to remove fixed norms. The fixed norms are replaced with dynamic management with monitoring and rewards. Monitoring and rewards contribute to building a positive expectation with a flexible goal-achievement cycle.

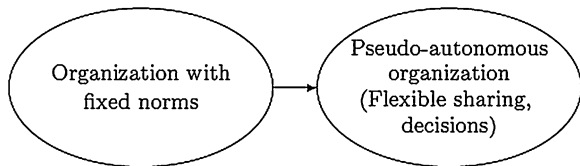
A framework that facilitates RSO is depicted in Fig. 6.

Flexible organizations are open to external resources. Spontaneous collaboration is facilitated through various sharing and socialization mechanisms such as lunch socials. Teams are virtual, therefore, the human capital is sought on demand. Sometimes, massive external resources are utilized using cloud-sourcing where a large-scale problem is split into small pieces. Each piece can be solved in a distributed and global manner. Millions of people can contribute to the problem solving through cloud sourcing. Wikipedia is one of such examples.

The transition to SRO is depicted in Fig. 7.

Softwarization in RSO is depicted in Fig. 8.

**Fig. 5** Transition to flexible organization



**Fig. 6** Framework that facilitates RSO

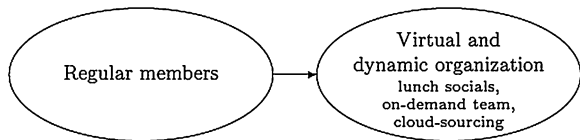




Fig. 7 Transition to RSO

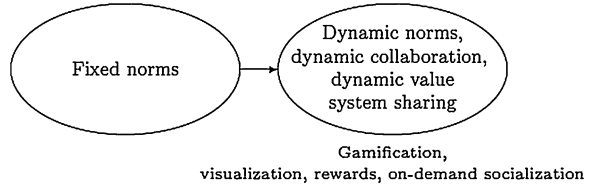
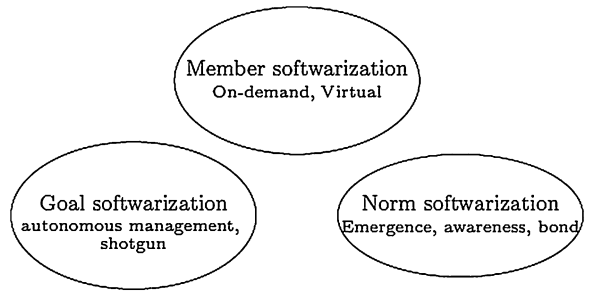


Fig. 8 Softwarization in RSO



There are three types of softwarization in RSO. One is member softwarization, where on-demand teams and virtual teams are built using an open innovation scheme. Another is goal softwarization. Relaxation of goal management is another contributing factor for RSO. The other is norm softwarization. Relaxed stability facilitates the management of flexible norms. This is a short-term instability factor. However, it contributes to the agility of the organization.

The design principles of RSO are depicted in Fig. 9.

Detailed analysis of RSO design principle is beyond the scope of this paper.

The author proposes three principles of RSO to serve as the basis for further research. One is minimum organization design where the minimum set of organizational components is established for relaxed stability. Another is tool-based dynamic team management. Relaxed stability requires constant monitoring and feedback to compensate instability of an organization. Tools are required for this purpose. The other is virtual team management, where the members of the team are flexible.

RSO-based organization management is depicted in Fig. 10.

Stability is required to maintain an organization. The main purpose of RSO is to decrease stability in order to increase organizational performance. For this

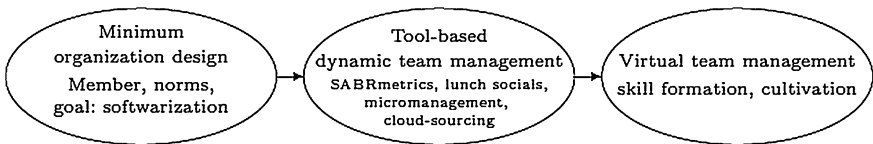
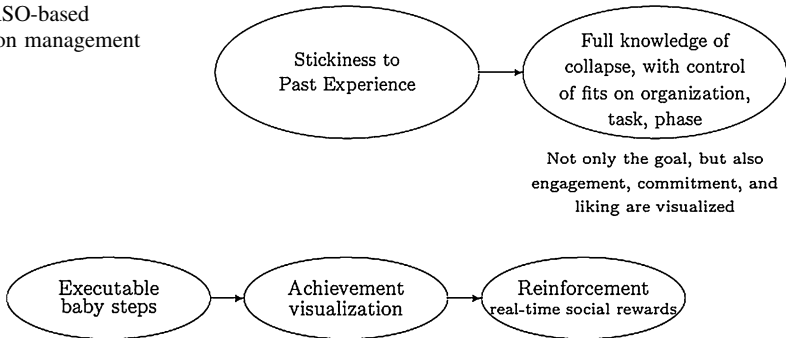


Fig. 9 Design principles of RSO

**Fig. 10** RSO-based organization management



**Fig. 11** Basic cycle in the micro-management

purpose, the large-scale know-how of collapses is required, just as the relaxed stability technology in military aviation required massive knowledge of aviation collapses.

Engagement, commitment, liking and culture are studied for this purpose. Dynamism of organization and organizational performance requires detailed analysis.

Gamification in enterprise environments follows the basic cycle of micro-management as depicted in Fig. 11.

For monitoring and feedback mechanisms, gamification is a promising candidate technology to be deployed in RSO. The first step is executable baby steps to facilitate achievements and changes. The second step is achievement visualization that increases engagement and emotional rewards for hard work. The third step is reinforcement using real-time social rewards. A detailed discussion of component design in this cycle remains for further research.

## 5 Discussion

### 5.1 The Advantages of the Proposed Approach

The author discusses a new umbrella concept, RSO, in analogy to RST in military aviation. There are two merits to this approach:

The factors that drive RSO are depicted in Table 3.

The concept that intentionally introduces decreased stability.

The concept that management can be maintained with decreased stability without any additional special mechanisms in office systems.

For the first point, stability is an unwritten rule of common organizations. It is difficult to manage an organization without stability. Stability serves a purpose for work metrics, workplace norms, interpersonal relationships, and so on. It is unique to place importance on the intentional decrease of stability.

**Table 3** Factors that drive RSO

Factor	Summary
Open innovation	Today's organization needs to cope with the demands for open innovation, leveraging capabilities to accommodate open innovation with cross-boundary collaboration
Demands for agile operation	It is necessary to facilitate agility to cope with a changing industrial landscapes. It is also necessary to take initiative for changes
Demands for flexible management for productivity	In order to leverage workplace productivity and satisfaction, it is necessary to cope with flexible management through which detailed work contexts are considered and honored

With regard to the second point, it is analogous to the use of visualization and feedback systems in gamification. For example, the original RST utilizes real-time sensor and control systems.

Today's organization has to cope with challenges and increased demands for work efficiency, improvement of workplace satisfaction, and flexibility that deals with a changing external landscapes.

The accommodation of intentional decrease of stability (in analogy to RST in military aviation) in order to reengineer organizational management is a revolutionary idea.

## 5.2 Limitations

This research is a qualitative study. Quantitative measures for verifying multiple aspects of RSO discussed in this paper remain for further study.

Acceptance of RSO in a real world environment is beyond the scope of this paper. The concrete design methodology of an RSO-oriented office systems is beyond the scope of this paper.

## 6 Conclusion

The author discusses a new umbrella concept for corporate management, RSO, in analogy to the RST used in military aviation. In military aviation, the demands for increased flexibility in air-combat capabilities led to RST, the intentional decrease of stability.

The author proposes RSO concept based on an inspiration gained from RST in the organizational management. The emergence of gamification provides another insight into the fact that available new technologies can provide new fits in the organizational management.

**Acknowledgments** The author expresses thanks to Toshiaki Fujii, NTT Comware, for his insightful suggestions on RST.

## Reference

1. Grudin J (1994) Groupware and social dynamics: eight challenges for developers. *CACM* 37(1):92–105

# Mapping and Optimizing 2-D Scientific Applications on a Stream Processor

Ying Zhang, Gen Li, Hongwei Zhou, Pingjing Lu, Caixia Sun and Qiang Dou

**Abstract** Stream processors, with the stream programming model, have demonstrated significant performance advantages in the domains signal processing, multimedia and graphics applications, and are covering scientific applications. In this paper we examine the applicability of a stream processor to 2-D stencil scientific applications, an important and widely used class of scientific applications, which compute values using neighboring array elements in a fixed stencil pattern. We first map 2-D stencil scientific applications in FORTRAN version to the stream processor in a straightforward way. In a stream processor system, the management of system resources is the programmers' responsibility. We then present several optimizations, which avail the stream program for 2-D stencil scientific applications, of various aspects of the stream processor architecture. Finally, we analyze the performance of optimized 2-D stencil scientific stream applications, with the presented optimizations. The final stream scientific programs gain from 2.56 to 7.62 times faster than the corresponding FORTRAN programs on a Xeon processor, with the optimizations playing an important role in realizing the performance improvement.

---

Y. Zhang (✉) · G. Li · H. Zhou · P. Lu · C. Sun · Q. Dou  
School of Computer, National University of Defense Technology, Changsha 410073, China  
e-mail: zhangying@nudt.edu.cn

G. Li  
e-mail: genli@nudt.edu.cn

H. Zhou  
e-mail: hwzh@nudt.edu.cn

P. Lu  
e-mail: pjl@nudt.edu.cn

C. Sun  
e-mail: cxsun@nudt.edu.cn

Q. Dou  
e-mail: qd@nudt.edu.cn

# 1 Introduction

Stream processors [1, 2] have demonstrated significant performance advantages in media applications [3]. Many researchers are interested in the applicability of stream processors to scientific computing applications [4].

The stream processor architecture, which has many differences from the architecture of a conventional system, is designed to implement the stream programming model [5]. Although language implementation exploit the model's features well, they do so at such a comparatively low-level; it is mainly the programmer's responsibility to manage system resources. Moreover, compared to other stream applications, such as media applications, scientific computing applications have more complex data traces and stronger data dependence. Therefore, writing a high-performance scientific stream program is rather hard and important to get right and high performance.

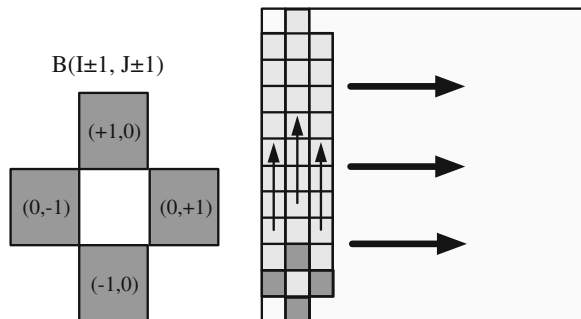
2-D stencil scientific applications, an important and widely used class of scientific applications, have special data access trace and compute values using neighboring array elements in a fixed stencil pattern. This stencil pattern of data accesses is then repeated for each element of the array. This paper first take 2D Jacobi iteration as an example to illuminate our mapping and optimization, and our presented methods can be used to any 2D stencil scientific application. Figure 1 presents the code for 2-D Jacobi iteration. The Jacobi iteration kernel consists of a simple 4-point stencil in two dimensions, shown in the first part of Fig. 2. On each loop iteration, four elements of the array are accessed in the 4-point diamond stencil pattern shown on the left. As the computation progresses, the stencil pattern is repeatedly applied to array elements in the column, sweeping through the array, as shown in the second part of Fig. 2.

**Fig. 1** Code for 2-D Jacobi iteration

```

A(N,N), B(N,N)
do J=2,N-1, I=2,N-1
  A(I,J) =
  C*(B(I-1,J) +
  B(I+1,J) +
  B(I,J-1) +
  B(I,J+1))
    
```

**Fig. 2** Data traces for 2D Jacobi



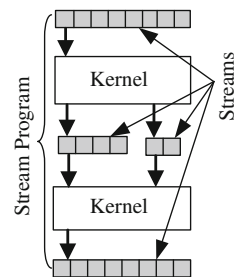
In this paper, we use the language streamC/kernelC [1] to map FORTRAN version of 2-D stencil scientific application to the stream processor. A straightforward mapping method is first given to map 2-D Jacobi iteration to the stream processor; optimizations are then proposed to improve the overall performance of the mapped stream program. Finally, the performance of the stream programs for some typical-D stencil scientific applications, and the effectiveness of our optimizations are measured through a number of experiments. Compared with FORTRAN program on a Xeon, our stream programs finally achieve from 2.56 to 7.62 times speedup.

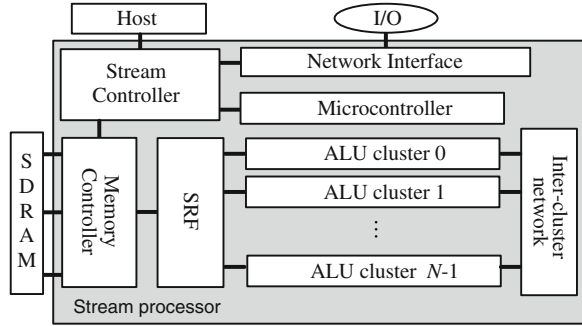
## 2 Background

The stream processor architecture is developed to speed up stream applications with intensive computations. The stream programming model divides an application into a stream-level program that specifies the high-level structure of the application and one or more kernels that define each processing step [6]. Each kernel is a function that operates on streams, sequences of records (Fig. 3).

Popular languages implementing the stream programming model include StreamC/KernelC [1] for Imagine and Merrimac processor, Brook [7] for GPU and SF95 for FT64 processor. These languages can also be used to develop stream programs for Cell Processors [8]. All these stream architectures have the characteristic of SIMD stream coprocessors with a large local memory for stream buffering. Figure 4 shows a simplified diagram of such a stream processor. A stream-level program is run on the host while kernels are run on the stream processor. A single kernel that operates sequentially on records of streams is executed on clusters of ALUs, in a SIMD fashion. Only data in the local register files (LRFs), immediately adjacent to the arithmetic units, can be used by the clusters. Data passed to the LRFs is from the Stream Register File (SRF) that directly access memory. On-chip memory is used for application inputs, outputs and for intermediate streams that cannot fit in the SRF.

**Fig. 3** Stream programming model



**Fig. 4** A stream processor

### 3 Straightforward Map and Optimization

#### 3.1 Straightforward Map

The implementation of mapping applications to the stream programming model can be thought of as a code transformation on programs that consist of a series of loops that process arrays of records. The data traces of different references in the innermost loop are extracted into different streams, and the computations performed by each loop are encapsulated inside a kernel. The remaining code composes the stream program.

When 2-D Jacobi iteration is mapped, the corresponding stream program declares five streams: four that correspond to the data accessed by four references to array B in I-loop and a fifth,  $a_0$ , that corresponds to the data accessed by the array  $a$  in I-loop. After declaring the streams, the stream program then calls a kernel that processes the streams  $b_0, b_1, b_2$  and  $b_3$  to produce the stream  $a_0$ . The kernel is declared, taking four input streams (“istreams”), one output stream (“ostream”) and one microcontroller variables as arguments. It first reads the values of the microcontroller variables, then loops over the records in the input streams computing records in the output stream.

#### 3.2 Exploiting the Reuse of Record

The SRF is banked into lanes such that each lane supplies data only for its connected cluster. Records of a stream are interleaved among lanes. If a cluster needs data residing in another cluster, it gets the data by inter-cluster communication.

In the FORTRAN code of 2-D Jacobi iteration, two neighboring elements, i.e.  $B(I + 1, J)$  and  $B(I - 1, J)$ , of array B are involved in each iteration of I-loop. When mapping such loops, we have two choices as follow:

- Organize data as different streams to start and to end at different offsets into the original data arrays. The data covered by the two references,  $B(I - 1, J)$  and



$B(I + 1, J)$ , in I-loop are organized into the streams  $b_0$  and  $b_1$  in referred order. Although the data in the records of every stream are almost the same, just displaced, all the streams must be loaded from off-chip memory.

- Organizing data covered by the two references in I-loop as a single stream. In this way, the streams  $b_0$  and  $b_1$  in *StreamJacobi* are merged into one stream  $b((j + 1) * N, (j + 2) * N - 1)$ . However, during the kernel example execution,  $cluster_i$  must communicate with  $cluster_{i+2}$  to gather the needed record by inter-cluster communication.

We reorganize the data distribution, with adjacent records distributed on the same lane. Thus, the optimized stream program has the same number of memory transfers with the second choice, but does not require any inter-cluster communication. The steps of exploiting the reuse of records for *StreamJacobi* are given below.

**Step 1** Organize all records covered by the two array references  $B(I + 1, J)$  and  $B(I - 1, J)$  as a new stream, with the same order as the array  $B$ .

**Step 2** Set the stride of the new stream be  $Length_{stream}/N_{cluster}$  and the record length be  $Length_{stream}/N_{cluster} + 2$ , where  $N_{cluster}$  is the number of clusters in the stream processor and  $Length_{stream}$  is the length of the new stream. This means the original records from  $i \times Length_{stream}/N_{cluster}$  to  $(i + 1) \times Length_{stream}/N_{cluster} + 2$  in the new stream become the  $i$ th record of the new derived stream,  $b_0'$ , residing in the  $i$ th cluster. Data distribution is transposed as shown, with neighboring records distributed on the same lane, such that  $cluster_i$  gets neighboring records from the lane of itself without any inter-lane communication.

**Step 3** Divide all other stream references into  $N_{cluster}$  parts by setting the stride be  $Length_{stream}/N_{cluster}$  and the record length be  $Length_{stream}/N_{cluster}$ .

**Step 4** Update original kernel to process the records in the corresponding new order.

After the optimization, the final stream-level program, has little inter-cluster communication and less memory transfers.

### 3.3 Exploiting the Reuse of Streams

The stream length, record length and stride of the streams  $b_2'$  and  $b_3'$  are changed as those of the stream  $b_0'$ , with the changed streams named  $b_2''$  and  $b_3''$ ; correspondingly, the kernel is updated to process the correct records, with the changed code. The relationship among the locations, accessed by the streams  $b_0'$ ,  $b_2''$  and  $b_3''$  on neighboring iterations. The stream  $b_3''$  on iteration  $i$ , the stream  $b_2''$  on iteration  $i + 1$  and the stream  $b_0'$  on iteration  $i + 2$  access the same locations. Since the values of the basic stream  $b$  are unchanged, the stream  $b_2''$  does not require accessing off-chip memory but accesses the SRF to get the values that are used by the stream  $b_3''$  in a previous iteration. Similarly,  $b_0'$  can access the SRF to get the values that are used by the stream  $b_3''$  in two previous iterations.

However, stream compilers cannot recognize and utilize the reuse supplied by the streams  $b_0'$ ,  $b_2''$  and  $b_3''$ . This is because the start and end bound of these streams are variables, which means they are unknown when stream compilers allocate the SRF for streams.

We optimize it by introducing four basic streams,  $b_{00}$ ,  $b_{01}$ ,  $b_{02}$  and  $a_{00}$ , initializing  $b_{02}$  and  $b_{01}$  before the loop, defining  $b_3''$  and  $a_0'$ , loading the values referred to by  $b_3''$  to  $b_{00}$ , replacing the references  $b_0'$ ,  $b_2''$ ,  $b_3''$  and  $a_0'$  with basic streams  $b_{02}$ ,  $b_{01}$ ,  $b_{02}$  and  $a_{00}$  respectively, saving the output  $a_{00}$  to the locations defined by  $a_0'$  and moving the values of  $b_{01}$  and  $b_{00}$  to  $b_{02}$  and  $b_{01}$  at the end of the loop body. The function *streamCopy*( $s$ ,  $t$ ) copies records of  $s$  to  $t$ . An SRF-to-memory copy generates a save of  $s$  to memory; a memory-to-SRF copy generates a load of  $s$  to the SRF; an SRF-to-SRF copy generates a save of  $s$  to memory and a load of  $s$  to the SRF buffer that holds  $t$ . We can effect the reuse by replacing streams that have unknown starts and ends with streams that have constant starts and ends, i.e. basic streams, and explicitly transferring original reuse to the reuse among streams with constant starts and ends. The stream compiler will recognize and utilize the reuse in the transformed code. However, it does so at the expense of introducing two expensive SRF-to-SRF data moves. Since these moves implement a permutation of values in the SRF, we can eliminate the need for moves by unrolling to the cycle length of the permutation, i.e. 3 times and permuting the stream references in each unrolled loop bodies.

## 4 Experimental Setup and Performance Evaluation

In our experiments, we use Isim [2], a cycle-accurate stream processor simulator supplied by Stanford University, to get the performance of the stream applications with different versions. Table 1 summarizes the applications we used. The baseline configuration of the simulated stream processor and its memory system is detailed in Table 2, and is used for all experiments unless noted otherwise. For comparison, 2-D stencil scientific applications in FORTRAN is compiled by Intel's IA32 compiler (with max speed optimization option), and run on a Xeon processor, one class of the most popular machines used for scientific computing applications now. Table 3 shows the configuration of the Xeon processor.

**Table 1** Application programs

	Problem size
2-D Jacobi	256*256
2-D Laplace	1 K*1 K
MG	128*128*128
QMRCGSTAB	800*800
MVM	832*832

**Table 2** Baseline parameter of Isim

Parameter	Value
Cluster number	8
LRF	38.4 KB
SRF	512 KB
Off-chip DRAM	4 GB
Frequency	2 GHz

**Table 3** Configuration of the Xeon processor

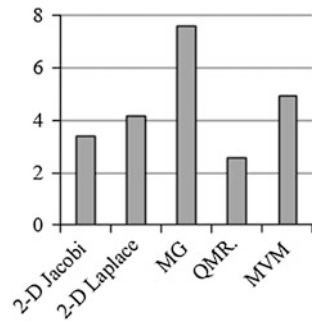
Parameter	Value
Core number	8
Frequency	2.7 GHz
L1 Cache	8*32 K*2
L2 Cache	8*256 K
L3 Cache	20 M

### 4.1 Overall Performance

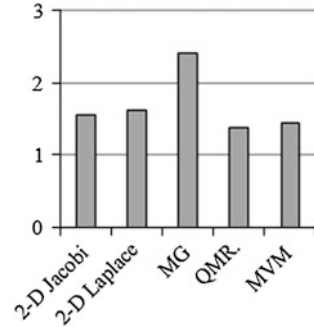
The performance of 2-D stencil scientific stream applications with all optimizations is first presented to evaluate the stream processor’s ability to process 2-D stencil scientific applications. Figure 5 shows the speedup yielded by final stream programs, over FORTRAN programs. Optimized stream programs yield from 2.56 to 7.62 times speedup, which indicates the stream processor can successfully process such class of scientific applications. This is because plenty of ALUs process computations in 2-D stencil scientific applications; data reuse is all exploited; memory transfers are overlapped with kernel execution perfectly.

Compared with other applications, MG has much record reuse and stream reuse, and thus gets highest speedup, 7.62. But for QMRCGSTAB, only part of its computation is accelerated, and thus only gets a speedup of 2.56.

**Fig. 5** Speedup yielded by Isim over Xeon



**Fig. 6** Speedup from exploiting the reuse of records



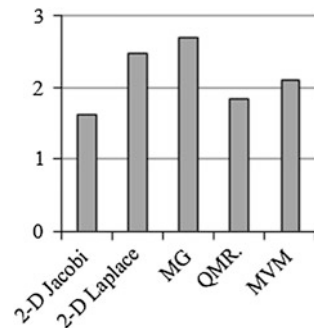
### 4.2 Exploiting the Reuse of Records

We now demonstrate the effectiveness of exploiting the reuse of records that reorganizes streams to reduce off-chip memory transfers. Figure 6 demonstrates marginal speedup, due to the reduction of memory transfers. One of two input records of Jacobi iteration reuses previous generated data; one of two input records of Laplace reuses previous generated data; two of three input records of MG reuse previous generated data; one of two input records of QMR reuses previous generated data; one of two input records of MVM reuses previous generated data. Correspondingly, they yield speedup of similar trend.

### 4.3 Exploiting the Reuse of Streams

As a stream processor reduces memory transfers only by capturing the reuse among streams in the SRF, this optimization is important. We evaluate the impact of exploiting the reuse of streams on program performance. Our stream programs benefit greatly from this optimization. Figure 7 demonstrates the speedup attained with this optimization. One out of two input streams of Jacobi iteration reuse the

**Fig. 7** Speedup from exploiting the reuse of streams



data generated on the previous iteration; two out of three input streams of Laplace reuse the data generated on the previous iteration; two out of three input streams of MG reuse the data generated on the previous iteration; two out of three input streams of QMRCGSTAB reuse the data generated on the previous iteration; two out of three input streams of MVM reuse the data generated on the previous iteration. Thus, a lot of memory transfers are reduced for each stream program. Without the optimization, the stream compiler cannot identify the reuse, all input streams must be loaded from off-chip memory and each kernel must wait until its input streams are loaded from off-chip memory. Stream reuse removes the appearance of streams with unknown starts and ends, thus making the stream compiler able to identify reuse.

**Acknowledgments** This work was supported by NSFC (61003075, 61103193,61103011, 61103014).

## References

1. Rixner S (2001) Stream processor architecture. Kluwer Academic Publishers, Boston
2. Kapasi U, Dally W, Rixner S, Owens J, Khailany B (2002) The imagine stream processor. In: Proceedings of 2002 IEEE international conference on computer design, pp 282–288
3. Gordon M, Maze D, Amarasinghe S, Thies W, Karczmarek M, Lin J, Meli A, Lamb A, Leger C, Wong J et al (2002) A stream compiler for communication-exposed architectures. ACM SIGARCH Comput Archit News 30(5):291–303
4. Fatica M, Jameson A, Alonso J STREAMFLO: an Euler solver for streaming architectures, submitted to AIAA conference
5. Kapasi U, Rixner S, Dally W, Khailany B, Ahn J, Mattson P, Owens J (2003) Programmable stream processors. Computer 36(8):54–62
6. Das A, Dally WJ, Mattson P (2006) Compiling for stream processing. In: proceedings of the 15th international conference on parallel architectures and compilation techniques PACT '06. ACM Press, New York, pp 33–42
7. Buck I, Foley T, Horn D, Sugerman J, Fatahalian K, Houston M, Hanrahan P (2004) Brook for gpus: stream computing on graphics hardware. ACM Trans Graph 23(3):777–786
8. Kahle JA, Day MN, Hofstee HP, Johns CR, Maeurer TR, Shippy D (2005) Introduction to the cellmultiprocessor. IBM J Res Dev 49(4/5):589–604

# Development of an Android Field Trip Support Application Using Augmented Reality and Google Maps

DongLim Hyun, EunGil Kim and JongHoon Kim

**Abstract** In this study, an application to support field learning was developed to apply to education Location Based Service (LBS) which is expanding with the spread of smartphones. To select functions in tune with the purpose, a requirements survey was conducted for current elementary school teachers, and the application was developed by reflecting the requirements based on an analysis of the survey results. Through the developed application, teachers can provide information to students and students can acquire information (location, pictures, description) required for field learning and carry out autonomous activities. The developed application was explained and demonstrated to current elementary school teachers who were given an experience of using it. A survey to verify its effectiveness was conducted and it was found that the application had a high potential for being utilized in field learning.

**Keywords** Field learning · Augmented reality · Android

## 1 Introduction

Ubiquitous Internet access is changing the users of smartphones. These social changes are leading to changes in curriculums and educational methods, and particularly, field trips are becoming more important. This study developed a field

---

D. Hyun · E. Kim · J. Kim (✉)

Department of Computer Education, Teachers College, Jeju National University,  
Jeju, South Korea

e-mail: jkim0858@jejunu.ac.kr

D. Hyun

e-mail: gody5@naver.com

E. Kim

e-mail: computing@korea.kr

trip support application that can be used in the future, following the changes in society and the educational environment, using augmented reality and Google Maps in the Android platform. Such application was developed by experienced teachers for use in schools. It will promote cooperative learning by providing information to teachers and students and enabling information sharing among students using the features of smartphones.

## **2 Theoretical Background**

### ***2.1 Augmented Reality***

There are mainly three methods of implementing augmented reality: the layer method, the marker recognition method, and the markerless method. The layer method identifies the location and position of smartphones using the phone camera, GPS, and sensors, and adds information to the images captured by the camera. The marker recognition method positions specific markers at the locations where information must be displayed, and the markers captured by the camera are recognized and the corresponding information, displayed. The markerless method extracts the feature points from the images captured by the camera that replaced the markers. The application in this study was implemented with the augmented reality of the layer method, considering the limited computation capacity of smartphones and users' need for easy addition of information.

### ***2.2 Related Studies***

So-Hee Kim (2004) and Yoon-Kyung Min (2005) designed and developed a field learning support system [1, 2]. In these two studies, however, the actual activities of students are carried out through the Web and wireless terminals only send information to the Web.

Seung-Ah Lee (2010) designed and developed a field learning support application for performing missions of a scenario using Android-based smartphones [3]. However, it simply provides students with missions and students carry out the missions using the functions of smartphones.

Many studies presented applications for supporting field learning. However, they just use MMS service or Web pages and are not optimized to the widespread smartphones and other smart devices or they just designed and proposed the systems.

A difference between our study and previous studies is the use of LBS applications through smartphones. In this study, the information supply and sharing and missions which were implemented in previous studies were combined with the status

and location of users, and they were applied to field learning in a more diverse and effective ways such as augmented reality and maps. Furthermore, user convenience was improved by providing all these functions through a single application.

### 3 Scope and Functions of the Application

To define the scope and functions of the field learning support application, a survey on the required functions of the field learning application was conducted with 15 current elementary school teachers who were using smartphones or had high interest in them. The key questions and results of the survey are shown in Table 1.

This survey found that elementary schools were performing field learning once or twice per semester and up to once a month depending on the circumstances of the school and for the percentage of places for field learning, the percentage of outdoor places was a little higher than that of indoor places. Furthermore, for the functions of field learning support applications, they required information sharing between teachers and students and between students, destination search through maps and augmented reality, and notification about approaching the destination.

## 4 Implementation of Field Learning Support Application

### 4.1 Implementation of Augmented Reality

To synchronize the camera image with the image of the destination, the coordinates to be indicated on the screen were calculated by combining the values of the sensors.

The azimuth of the destination was computed by substituting the GPS coordinates of the destination with those of the smartphone using the computation algorithm issued by the National Geographic Information Institute [4].

**Table 1** Content and results of requirements survey

Question 1. How often do you perform field learning a year on average?			
• 3–4 times	47 %	• 5–6 times	40 %
• 7–8 times	7 %	• 9 times or more	7 %
Question 2. What are the percentages of indoor and outdoor places for field learning?			
• Indoor	40 %	• Outdoor	60 %
Question 3. Which functions of the application would assist field learning? (multiple answers are allowed)			
• Information input and sharing through maps			80 %
• Search for destinations through augmented reality			60 %
• Display of destinations and current locations on maps			53 %
• Notification of arrival at destinations and confirmation of inputted information			33 %



**Fig. 1** Augmented reality screen



Compass and land prices were added to the developed application for user convenience. The application screen is shown in Fig. 1.

## ***4.2 Use of Google Maps***

A destination can also be added to the map by touching such destination on the map and getting the GPS values. The destination objects are managed as a list and can be easily added or removed. Users can add the name, description, and images of destinations in the Alert Dialog Window to add a destination. The name and description can be entered through the smartphone keyboard, and the image can be added by taking a picture with the smartphone camera or selecting one of the existing images. The application screen is shown in Fig. 2.

## ***4.3 Sharing Through the Web***

The application allowed the sharing through the Web of information created by teachers. Students can also share their data and use them for small group learning and cooperative learning activities. The application screen is shown in Fig. 3.

## ***4.4 Development of the Destination Approach Notification Feature***

The user must be notified when the destination set in AR Activity or Map Activity is near. This feature is provided as an Android API, so the developer only needs to set the destination coordinates, approach distance, action on notification, etc.



Fig. 2 Screens created using Google Maps

Fig. 3 Screen for sharing through the Web



In the developed application, the approach distance was set at 10 m and the observation time, at always; and the action on notification was set at viewing the destination information on a Alert Dialog Window. The application screen is shown in Fig. 4.

**Fig. 4** Screen for destination approach notification



## 5 Expert Assessment

The feasibility and improvements of the developed field learning application were diagnosed through expert assessment. 15 current class 1 elementary school teachers were selected for the experts.

For this expert assessment, the application was explained to the experts and they practiced it before answering a questionnaire by checking one of the answers in five-step Likert scale with 1 point intervals. The expert assessment areas and results are shown in Table 2.

The expert assessment results showed generally high satisfaction levels on the application developed in this study, and the utilization potential of the application

**Table 2** Expert assessment areas and results

Area 1. Satisfaction on screen layout	
• Satisfaction on the augmented reality screen layout and menu	4.5
• Satisfaction on the map screen layout and menu	4.5
• Satisfaction on the displayed information and buttons in the notification window	4.6
Area 2. Satisfaction on the implemented functions	
• Sharing and expression of information (location, pictures, description) on the destinations through a server	4.8
• Search for destinations through augmented reality	4.4
• Display of destinations and current location on the map screen	4.6
• Notification for approaching the destination	4.7
Area 3. Intuitiveness and convenience of operation	
• Intuitiveness and convenience of operation on the augmented reality screen	4.6
• Intuitiveness and convenience of operation on the map screen	4.6
• Intuitiveness and convenience of operation for switching between augmented reality screen and map screen	4.7

in field learning was found to be high. In particular, information sharing and notification of approaching a destination were highly evaluated as they could be applied to outdoor learning as well as field learning.

## 6 Conclusion

Society is changing fast, and schools are seeking changes in their educational methods in line with the social changes. This study paid attention to the rapid propagation of smartphones and the emphasis on field trips in the changed curriculums. It is expected that learning handsets similar to smartphones will be supplied to students and that the importance of field trips in schools will increase.

The developed application can be useful for the future educational environment. More active studies in this area are needed to prepare for the future educational environment.

## References

1. Google Public Interface Map. <http://developer.android.com/reference/java/util/Map.html>
2. Kim SH (2003) A wireless/wired field-experience learning support system using mobile handsets. Master's Thesis, Ewha Womans University
3. Min YK (2005) Design and implementation of support system for field education learning based ubiquitous. Master's Thesis, Ewha Womans University
4. National Geographic Information Institute. <http://www.ngii.go.kr/kor/board/download.do?rbsIdx=31&idx=278&fidx=1>

# Implementation of Automotive Media Streaming Service Adapted to Vehicular Environment

Sang Yub Lee, Sang Hyun Park and Hyo Sub Choi

**Abstract** Among the variety of vehicle technology trend issues, the biggest one is focus on automotive network system. Especially, optical network system is preferred. The aim in optical network system including vehicular environment information is that realization of media streaming service beyond the current car audio system. This paper is introduced the implementation of media streaming service which is consist of optical network as called MOST (Media Oriented System Transport) and realization of car audio system adapted to vehicular environment information collected from (On-Board Diagnosis) OBD via (Controller Area Network) CAN.

**Keywords** In-vehicle network systems · MOST · OBD · CAN · Car sound systems · Media streaming service

## 1 Introduction

With the passage of time, vehicle technology in modern times has been developed rapidly. Recent of today's automobile research area has been gradually changed from mechanics to electronics to the way of offering entertainment service to customs that can serve the connection to smart device easily and conveniently.

---

S. Y. Lee (✉) · S. H. Park · H. S. Choi  
Jeonbuk Embedded System Research Centre, Korea Electronics Technology  
Institute, Dunsan-ri, Bongdong-eup, Wanju-gun, Jeolabuk-do, Pyeongtaek-si 565-902,  
Republic of Korea

e-mail: syublee@keti.re.kr

S. H. Park

e-mail: shpark@keti.re.kr

H. S. Choi

e-mail: hschoi@keti.re.kr

In particular, most of people want to be experienced in high quality audio streaming service while they drive. According to customer's demands, MOST system is developed to provide an efficient and cost effective fabric to transmit audio data between any devices attached to the harsh environment of automobile. Compared to conventional car audio system which has been used in many of audio electrical line, bundle of lines per an audio channel, dissimilarly being with the one optical line in order to transit the audio data source, the advanced streaming service adopted optical network system makes light harness system. As the simply network construction, it can be obtained the high fuel efficiency from the usage of plastic optical fiber and for optical signal characteristics, it makes be free on electro-magnetic problems. Considered on driving environment, car audio system as the streaming service gathering automobile information via CAN enables to tune the volume automatically for their vehicular conditions. To develop the self-contained audio system, the complex information having vehicular status is needed to be accessed more comfortable.

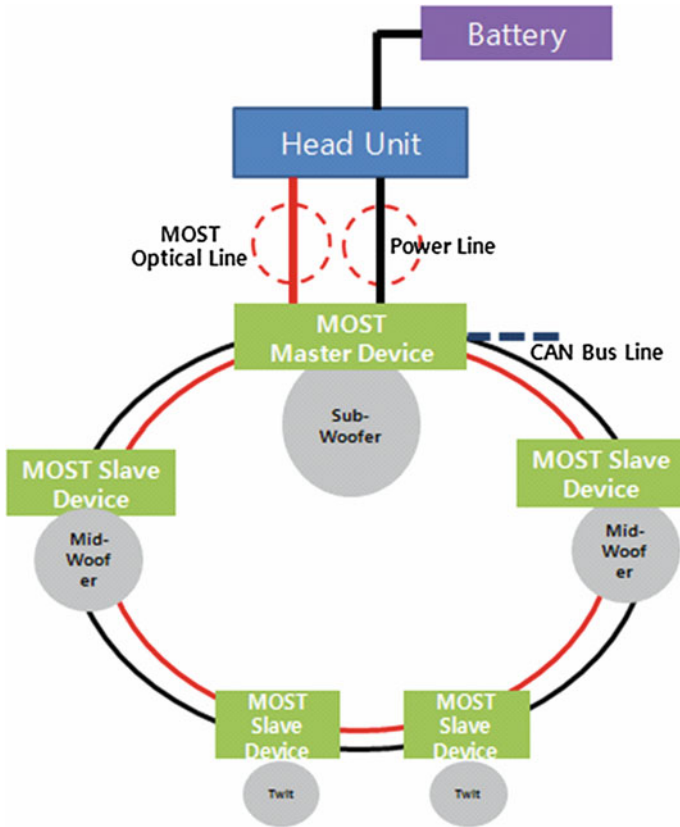
In Sect. 2, MOST network system and MOST data processing are introduced for vehicle environment scenario. The vehicle information processing module which classified into vehicular status information to be collected from CAN network is described in Sect. 3. In Sect. 4, platform and demonstration of streaming service are shown. This paper is concluded in Sect. 5.

## 2 MOST Network System

MOST is the de-factor standard for efficient and cost effective networking of automotive multimedia and infotainment system [1, 2]. The current MOST standards released MOST150, 150 means that 150 Mbps network bandwidth with quality of services is available. To meet the demands from various automotive applications, MOST network system provides three different message channels: control, synchronous used in streaming service and asynchronous channel only for packet data transfer. In describe in Fig. 1, proposed network system is consist of MOST devices with CAN bus line. In conception of the network topology, MOST devices are divided into master and slave mode. Master mode device with CAN bus can be realized the optimum sound effect depending on vehicular environment information gathered by OBD interface in car.

### 2.1 MOST Steaming Data Processing

For the data processing, it is explained in streaming data part and network service one. Being presented the streaming data part; the bandwidth of the streaming data channel can be calculated using the following formula:



**Fig. 1** MOST Network System with CAN bus line

$$\text{Bandwidth} = 93 \times 8 \text{ bit} \times 48 \text{ kHz} \tag{1}$$

Synchronous data is used for the real-time transmission of audio data. Before data transmission, the data transmitted for the synchronous connection must be established by the connection master in network system. For this method, one socket has to be created and connected at the interfaces to the frame and to the local resource. Thus, up to 93 stereo connections can be established simultaneously on a MOST 150 frame. The content of the frame remains unmodified until the frame arrives back at the sending node. The quasi-static establishment of connections on a channel is denominated as (Time Division Multiplexing) TDM. The data are transmitted cyclically in a specified time pattern at the same frame position. There is no repetition in the case of communication errors. A valid value is then available the next cycle.

On the network service side, MOST devices have the unique interface which transfers the data between the physical layer network and processor as an external host controller [3]. Working in conjunction with the clock manager, the network

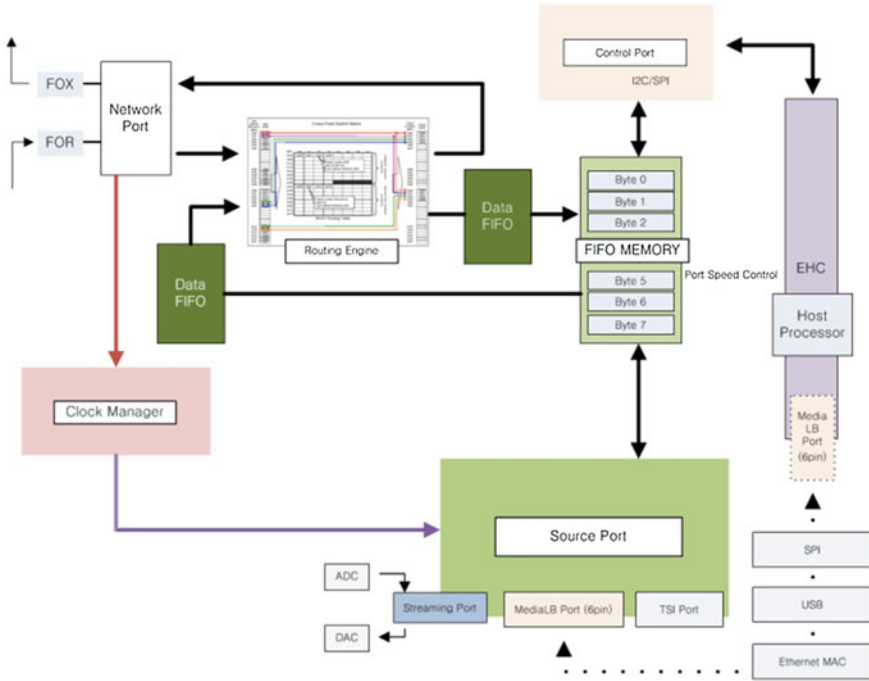


Fig. 2 The Data Process for MOST Frames

port recovers the network clock for time synchronization. And then, receiving data is decoded and delivered to the microprocessor via (Media Local Bus) MLB. Transferred data is routed into appropriate memory destination on and off in platform (Fig. 2).

### 3 Vehicular Status Information

#### 3.1 Processing Module for Vehicle Information

As described in this paper, the specific module collecting automobile data is called as the vehicle information processing module. In the processing module, vehicle data transmission and channel connection is served in communication module and vehicle communicator executes the classification of information which factors on streaming service can be affected by while driving. Through the vehicle data analyzer and container, required vehicle status data is transferred to the platform using internal bus line without delay.



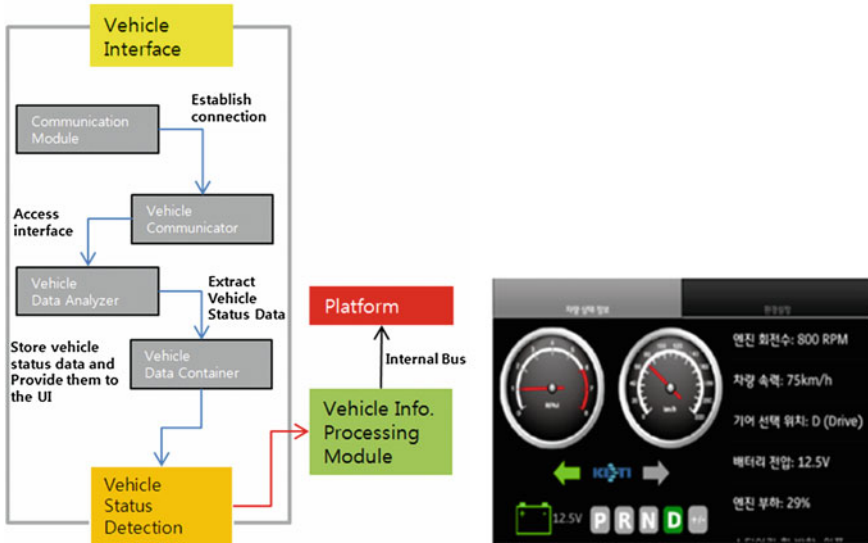


Fig. 3 The process of CAN Transceiver and Interpreter

Vehicular sensing module as depicted in Fig. 3 is coded into vehicular status information and status transit method. For applications, expressed black box on right side in Fig. 3, vehicle processing module is displayed engine rpm, vehicle speed, gear information, battery status and engine load to panel of platform.

## 4 Implementation of Streaming Service

### 4.1 System Architecture

The external host controller communicates with network interface controller via I2C bus, MLB and connected with OBD via CAN bus [4, 5]. As shown below, the streaming port included in network interface controller can be used for stereo audio exchange between the network and physical audio port. The external host controller may support for managing the streaming audio exchange remotely.

As described in Fig. 6, in order to make the data connection path, SRAM interfacing method is needed generally. Data transferred from the host bus accesses into SRAM memory space and linked in network interface controller through the MLB embedded in FPGA area. MLB is an on-PCB or inter-chip communication bus, designed to a common hardware interface. Especially, media local bus supports the MOST streaming data [6] Fig. 4.

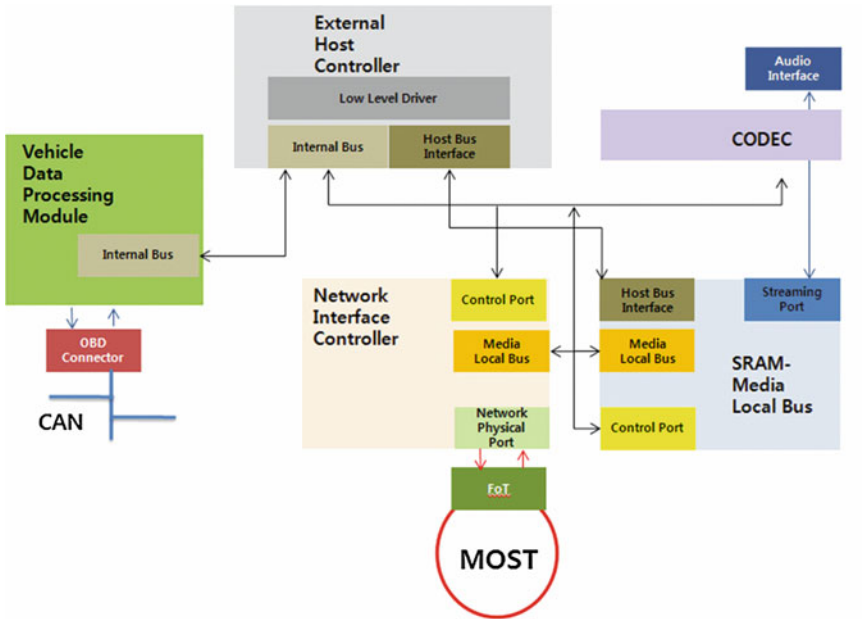


Fig. 4 System Architecture of Streaming Service

### 4.2 Streaming Service Platform

As mentioned that media local bus having SRAM memory interface is enable to exchange data frame. Designed logic block is mapped to Xilinx Vertex chipset and applied to streaming service platform. Table 1 provides a summary of the developed field programmable gate array utilization which can be shown that small memory usage can be expected the area effective size when this is made into system on chip level.

Table 1 The summary of device utilization

Target device	XC5VLX110	
Slice logic utilization	Number of slice registers	6,986
	Number of slice LUTs	15,347
	Number of route-thrus	897
	Number of occupied slices	6,103
	Number of LUT filp flop pairs used	17,233
	Number of BlockRAM/FIFO	18
	Number of BUFG/BUFGCTRLs	8
	Number of DCM_ADVs	2
	Total memory used	630 KB

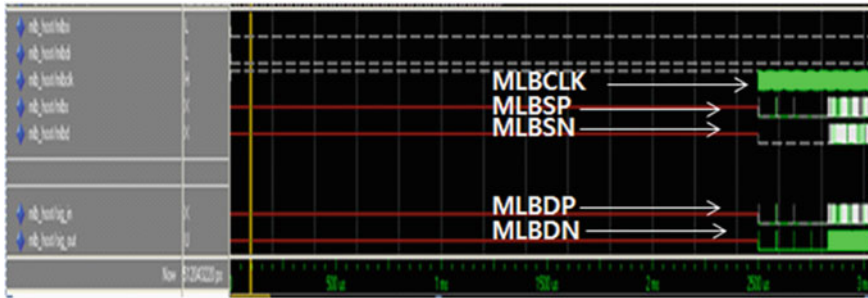


Fig. 5 The timing analysis of MLB

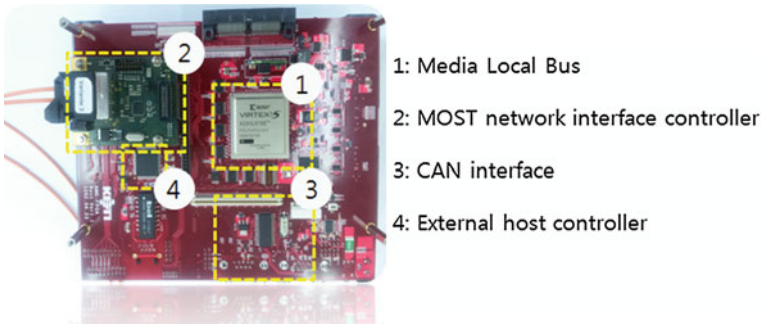


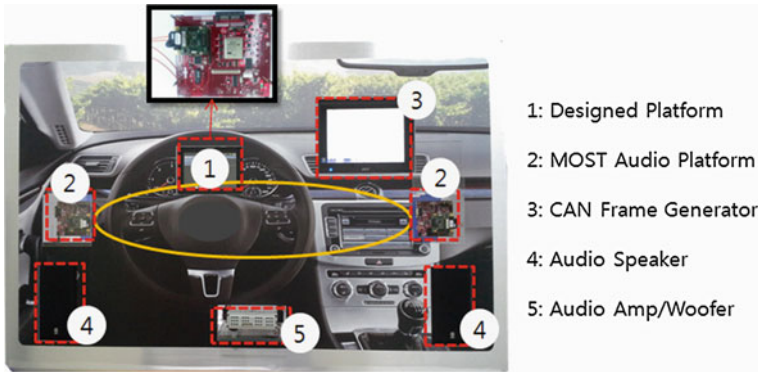
Fig. 6 Streaming Service Platform

All designed logic functions are showing the data and transmission commands when streaming service begins (Figs. 5, 6).

### 4.3 Streaming Service based on User Demanded Scenario

A proposed streaming service is reached to scenario as types of outside and inside for drivers and passengers. The environment of car audio system to be applied to proposed service are defined by below

- Outside Scenario1: Automotive audio sound can be affected by vehicle speed when moving car goes into high speed level, for being windy outside, the streaming sounds seems to be low with noise and it makes same situation when being rainy outside. Thus volume of sound system has to be up for itself.
- Outside Scenario2: Because driver has to be attention without sound disturbance, Automotive audio sound has to be down when driver want to move back, that is the gear shift paddle is located in rear mode.



**Fig. 7** System Demonstration

- Inside Scenario: Existed car audio system cannot control the woofers and speakers independently, however proposed streaming system which is assigned the network id for each device supports as a self-contained equipment. Especially, it can be the realization of optimum sounds effect according to the user's position and need in car.

As shown in Fig. 7, system demonstration is built with designed platform, audio platform connected with audio speaker and amplifier, CAN frame generator which makes the virtual environment via OBD interface applied to car. With running this set, it realizes that driver and passenger experience the car audio sound depending on vehicular status automatically without voluntary control.

## 5 Conclusion

For the trend of car infotainment system is moving to the high quality audio sound system, this paper is introduced the development of the audio streaming service based on optical network system. Particularly, optical network system, MOST, to be optimized sound level depending on vehicle environment is satisfied with reducing the weight and ensuring the reliability for the free of electro-magnetic problems. With vehicle status information, designed network platform realized and demonstrated that self-contained MOST speaker and woofer can be controlled and tuned their volume level automatically to be served to passengers more conveniently. Especially, outstanding point for implementation of adaptive audio streaming service is represented to hardware architecture and software frame depending on demanded vehicle environment scenario.

## References

1. Grzemba A (2012) MOST Book from MOST25 to MOST150. MOST Cooperation, FRANZIS
2. Strobel O, Rejeb R, Lubkoo J (2007) Communication in automotive system principles, limits and new trends for vehicles, airplanes and vessels. In: IEEE ICTON, pp 1–6
3. Wegmuller M, von der Weid JP, Oberson P, Gisin N (2000) High resolution fiber distributed measurements with coherent OFDR. In: Proceedings of ECOC'00, paper 11.3.4, p 109
4. Lee SY (2010) A method of vehicular data collection and processing using wireless communications interface. In: KICS winter conference
5. Godavarty S, Broyles S, Parten M (2000) Interfacing to the on-board diagnostic system. In: Proceedings of IEEE vehicular technology conference 52nd-VTC, vol 4, pp 24–28
6. Lee SY, Park SH, Choi HS, Lee CD (2012) Most network system supporting full-duplexing communication. In: IEEE ICACT, Korea, pp 1272–1275

# The Evaluation of the Transmission Power Consumption Laxity-Based (TPCLB) Algorithm

Tomoya Enokido, Ailixier Aikebaier and Makoto Takizawa

**Abstract** In order to realize energy-aware information systems, it is critical to discuss how to reduce the total electric power consumption of information systems. We consider applications of a communication type where a server transmits a large volume of data to clients. A client first selects a server in a cluster of servers and issues a file transmission request to the server. In this paper, we newly propose the transmission power consumption laxity-based (TPCLB) algorithm to select a server for a cluster of servers so that the total power consumption in a cluster of servers can be reduced. We evaluate the TPCLB algorithm in terms of the total power consumption and elapse time compared with the round-robin (RR) algorithm.

**Keywords** Green IT technology · Energy-aware information systems · TPC model · ETPC model · TPCLB algorithm

## 1 Introduction

In information systems, a client issues a request to a server and the server mainly consumes the power to perform the request as a process. We discuss software-oriented aspects of information systems to reduce the electric power consumption

---

T. Enokido (✉)  
Rissho University, Shinagawa, Japan  
e-mail: eno@ris.ac.jp

A. Aikebaier  
National Institute of Information and Communications Technology (NICT),  
Koganei, Japan  
e-mail: alisher@nict.go.jp

M. Takizawa  
Seikei University, Musashino, Japan  
e-mail: makoto.takizawa@computer.org

in a cluster of servers to realize energy-aware information systems [1–3]. In this paper, we consider communication type applications [2, 3] where a server transmits a large volume of data to clients.

The *transmission power consumption (TPC)* model of a server to transmit files is discussed in the paper [2]. In the TPC model, the electric power consumption of a server to transmit files to clients depends on the total transmission rate of the server. The approximated linear function to show how much a server consumes the electric power for transmission rates is derived from experimental studies of file transfer between servers and clients [2]. In the TPC model, the rotation speed of each fan is assumed to be fixed, i.e. the power consumption of each fan is constant. In current servers, the rotation speed of each fan can be changed to keep the temperature of each device. Thus, the total power consumption of a server depends on not only the power consumption of computation and communication devices but also cooling devices. The *extended TPC (ETPC)* model of a server is proposed to perform processes of communication type, where the power consumption of cooling devices are considered [3]. In this paper, we newly propose the *transmission power consumption laxity-based (TPCLB)* algorithm based on the TPC and ETPC models to select one of servers for communication type applications so that the total power consumption of servers can be reduced. We evaluate the TPCLB algorithm compared with the basic round-robin (RR) algorithm [4].

In Sect. 2, we present the file transmission model of a server. In Sect. 3, we present the power consumption models of a server. In Sect. 4, we discuss the TPCLB algorithm based on the TPC and ETPC models. In Sect. 5, we evaluate the TPCLB algorithm compared with the RR algorithm.

## 2 Transmission Model

Let  $S$  be a cluster of multiple data transmission servers  $s_1, \dots, s_n$  ( $n \geq 1$ ), each of which provides clients with the same data  $d$ . Each server  $s_t$  holds a full replica of the data  $d$ . Let  $C$  be a set of clients  $c_1, \dots, c_m$  ( $m \geq 1$ ). A client  $c_s$  issues a data transmission request to a load balancer  $K$ . The load balancer  $K$  selects one server  $s_t$  in the server cluster  $S$  and forwards the request to the server  $s_t$ . On receipt of a request, the server  $s_t$  transmits a reply file  $f_s$  to the requesting client  $c_s$ . Each request from a client  $c_s$  is performed as a process  $p_{ts}$ . Here, a notation  $p_{ts}$  shows a process  $p_s$  performed on a server  $s_t$  for a client  $c_s$ . A term *process* means an *application process* created for a request in this paper.

Let  $CT_t(\tau)$  be a set of current transmission processes on a server  $s_t$  at time  $\tau$ .  $NT_t(\tau)$  shows the number of current processes,  $NT_t(\tau) = |CT_t(\tau)|$ . Suppose a server  $s_t$  concurrently transmits files  $f_1, \dots, f_m$  to a set  $C_t (\subseteq C)$  of clients  $c_1, \dots, c_m$  at rates  $tr_{t1}(\tau), \dots, tr_{tm}(\tau)$  ( $m \geq 1$ ), respectively, at time  $\tau$ . Let  $b_{ts}$  be the maximum network bandwidth [bps] between a server  $s_t$  and a client  $c_s$ . Let  $Maxtr_t$  be the maximum transmission rate [bps] of the server  $s_t$  ( $\leq b_{ts}$ ). The total transmission rate  $tr_t(\tau)$  of the server  $s_t$  at time  $\tau$  is given as  $tr_t(\tau) = tr_{t1}(\tau) + \dots + tr_{tm}(\tau)$ . Here,

$0 \leq tr_i(\tau) \leq Maxtr_i$  for each server  $s_i$ . Each client  $c_s$  receives a file  $f_s$  at receipt rate  $rr_s(\tau)$  at time  $\tau$ . Let  $Maxrr_s$  indicate the maximum receipt rate of the client  $c_s$ .  $rr_s(\tau) \leq Maxrr_s$ .

Let  $TR_{ts}$  be the total transmission time [s] of a file  $f_s$  from a server  $s_i$  to a client  $c_s$ . Let  $minTR_{ts}$  show the minimum transmission time  $|f_s|/\min(Maxrr_s, Maxtr_i)$  [s] of a file  $f_s$  from a server  $s_i$  to a client  $c_s$  where  $|f_s|$  indicates the size [bit] of the file  $f_s$ .  $TR_{ts} \geq minTR_{ts}$ . Let  $tr_{ts}(\tau)$  be the transmission rate [bps] of a file  $f_s$  from a server  $s_i$  to a client  $c_s$  at time  $\tau$ . Suppose a server  $s_i$  starts and ends transmitting a file  $f_s$  to a client  $c_s$  at time  $st_{ts}$  and  $et_{ts}$ , respectively. Here,  $\int_{st_{ts}}^{et_{ts}} tr_{ts}(\tau)d\tau = |f_s|$  and the transmission time  $TR_{ts}$  of the server  $s_i$  to the client  $c_s$  is  $(et_{ts} - st_{ts})$ . The transmission laxity  $lt_{ts}(\tau)$  [bit] of transmission time is  $lt_{ts}(\tau) = |f_s| - \int_{\tau}^{et_{ts}} tr_{ts}(\tau)d\tau$  at time  $\tau$  ( $st_{ts} \leq \tau \leq et_{ts}$ ), i.e. how many bits of the file  $f_s$  the server  $s_i$  still has to transmit to the client  $c_s$  at time  $\tau$ .

First, we consider a model where a server  $s_i$  satisfies the following properties:

**[Server-bound model]** If  $Maxrr_1 + \dots + Maxrr_m \geq Maxtr_i$ ,  $\sum_{c_s \in CT_i(\tau)} tr_{ts}(\tau) = \sigma_i(\tau) \cdot Maxtr_i$  at every time  $\tau$ .

Here,  $\sigma_i(\tau) (\leq 1)$  is the transmission degradation ratio of a server  $s_i$ . In this paper, we assume  $\sigma_i(\tau) = \gamma_i^{NT_i(\tau) - 1}$  ( $0 < \gamma_i \leq 1$ ) at time  $\tau$ . Here, the effective transmission rate  $maxtr_i(\tau)$  of the server  $s_i$  is  $\sigma_i(\tau) \cdot Maxtr_i$  at time  $\tau$ .

Suppose a client  $c_s$  cannot receive a file  $f_s$  from a server  $s_i$  at the maximum transmission rate  $Maxtr_{ts}$ , i.e.  $Maxrr_s < Maxtr_i$ . Here,  $tr_{ts}(\tau) = Maxrr_s$ .

**[Client-bound model]** If  $Maxrr_1 + \dots + Maxrr_m \leq Maxtr_i$ ,  $\sum_{c_s \in CT_i(\tau)} tr_{ts}(\tau) = Maxtr_i \cdot (Maxrr_1 + \dots + Maxrr_m)/Maxtr_i$  at time  $\tau$ .

Even if every client  $c_s$  receives a file  $f_s$  at the maximum rate  $Maxrr_s$ , the effective transmission rate is not degraded, i.e.  $\sigma_i(\tau) = 1$ .

In a fair allocation algorithm, the transmission rate  $tr_{ts}(\tau)$  for each transmission process  $p_{ts}$  in the set  $CT_i(\tau)$  is the same, i.e.  $tr_{ts}(\tau) = maxtr_i(\tau)/NT_i(\tau)$ . However, the maximum receipt rate  $Maxrr_s$  of the client  $c_s$  might be smaller than  $maxtr_i(\tau)/NT_i(\tau)$ . Here, the rate  $(maxtr_i(\tau)/NT_i(\tau) - Maxrr_s)$  is not used. In order to more efficiently use the total transmission rate  $maxtr_i(\tau)$ , at the higher receipt rate a client  $c_s$  would like to receive, at the higher transmission rate a server  $s_i$  allocates to the client  $c_s$ . In this paper, the transmission rate  $tr_{ts}(\tau)$  for each client  $c_s$  at time  $\tau$  is allocated by the following algorithm:

1.  $V = 0; R = 0; TS = maxtr_i/NT_i(\tau);$
2. For each client  $c_s$ ,  $tr_i(\tau) = TS$  and  $R = R + (TS - Maxrr_s)$  if  $Maxrr_s \leq TS$ .  
Otherwise,  $tr_i(\tau) = Maxrr_s$  and  $V = V + (Maxrr_s - TS)$ .
3. For each client  $c_s$ ,  $tr_i(\tau) = tr_i(\tau) + V \cdot (Maxrr_s - tr_i(\tau))/R$  if  $tr_i(\tau) < Maxrr_s$ .

### 3 Power Consumption Model

We would like to discuss how much electric power a server  $s_i$  consumes to transmit files to clients.  $maxE_i$  is the maximum electric power consumption rate [W] of the server  $s_i$  to transmit files.  $minE_i$  is the minimum electric power



consumption rate [W], i.e. the server  $s_t$  is in idle state.  $minETR_t(\tau)$  is the minimum electric power consumption rate [W] of the server  $s_t$  to transmit a file. If  $tr_t(\tau) > 0$ ,  $minETR_t$  is constant. Otherwise,  $minETR_t = 0$ .  $E_t(\tau)$  is the electric power consumption rate [W] of the server  $s_t$  at time  $\tau$  where  $minE_t \leq E_t(\tau) \leq maxE_t$ .

In our previous studies [2], the *transmission power consumption (TPC)* model for a server  $s_t$  is proposed. Let  $PC_t(tr_t(\tau))$  show the electric power consumption rate [W] of a server  $s_t$  at time  $\tau$  where the server  $s_t$  transmits files to clients at the total transmission rate  $tr_t(\tau)$ . Here, the TPC model for a server  $s_t$  is given as follows:

**[Transmission power consumption (TPC) model]**

$$E_t(\tau) = PC_t(tr_t(\tau)) = \beta_t(m) \cdot \delta_t \cdot tr_t(\tau) + (minETR_t + minE_t). \quad (1)$$

Here,  $\delta_t$  is the power consumption rate of a server  $s_t$  to transmit one Mbits [W/Mb].  $m$  is the number  $NT_t(\tau)$  of transmission processes on the server  $s_t$ .  $\beta_t(m)$  shows how much power consumption rate is increased for the number  $m$  of transmission processes,  $\beta_t(m) \geq 1$  and  $\beta_t(m) > \beta_t(m - 1)$ . There is a fixed point  $maxm_t$  such that  $\beta_t(maxm_t - 1) \leq \beta_t(maxm_t) = \beta_t(maxm_t + h)$  for  $h > 0$ . The number  $maxm_t$  shows the maximum number of processes which can be performed on a server  $s_t$ .  $maxPC_t = \beta_t(maxm_t) \cdot \delta_t \cdot Maxtr_t + (minETR_t + minE_t)$  gives the maximum power consumption rate  $maxE_t$  of the server  $s_t$  to transmit files.

In the TPC model, the rotation speed of each fan is assumed to be fixed. The *extended transmission power consumption (ETPC)* model for a server  $s_t$  where the rotation speed of each fan can be changed is given as follows:

**[Extended transmission power consumption (ETPC) model]**

$$E_t(\tau) = \begin{cases} maxE_t & \text{if } tr_t(\tau) > 0. \\ minE_t & \text{otherwise.} \end{cases} \quad (2)$$

The rotation speed of each fan is the maximum to transmit a file  $f_s$  if at least one process  $p_{ts}$  is performed on a server  $s_t$ . The amount of the power consumption rate to rev up to the maximum rotation speed of fans for transmitting files is so large that the power consumption rate to transmit files can be neglected. Hence, the server  $s_t$  consumes the power at the maximum power consumption rate  $maxE_t$  if at least one process is performed on the server  $s_t$ .

## 4 TPCLB Algorithm

We discuss how to estimate the total power consumption laxity  $lpc_t(\tau)$  [Ws] of a server  $s_t$  at time  $\tau$ . Suppose a new transmission process  $p_{ts}$  is started on a server  $s_t$  at time  $\tau$ . Here, the transmission laxity  $lt_{ts}(\tau)$  of each current process  $p_{ts}$  in the set  $CT_t(\tau)$  is decremented by the transmission rate  $tr_{ts}(\tau)$  at time  $\tau$ . If the transmission laxity  $lt_{ts}(\tau)$  gets 0 at time  $\tau$ , the process  $p_{ts}$  terminates.

Given a process set  $CT_t(\tau)$  at time  $\tau$ , we can estimate time when each process in  $CT_t(\tau)$  to terminate and the power consumption laxity  $lpc_t(\tau)$  of a server  $s_t$ . The power consumption laxity  $lpc_t(\tau)$  shows how much power a server  $s_t$  consumes to perform every current processes in  $C_t(\tau)$  at time  $\tau$ , which is given by the following procedure **CommLaxity**( $s_t, \tau$ ):

```

CommLaxity ( $s_t, \tau$ ) {
  if  $CT_t(\tau) = \phi$ , return (0);
   $tr_t(\tau) = \sum_{c_s \in CT_t(\tau)} tr_{ts}(\tau)$ ; /* total transmission rate.*/
   $lpc_t = E_t(\tau)$ ; /* formulas (1) or (2). */
  for each current process  $p_{ts}$  in  $CT_t(\tau)$ , {
     $lt_{ts}(\tau + 1) = lt_{ts}(\tau) - tr_{ts}(\tau)$ ;
    if  $lt_{ts}(\tau + 1) = 0$ ,  $CT_t(\tau + 1) = CT_t(\tau) - \{p_{ts}\}$ ;
  }
  return ( $lpc_t + \mathbf{CommLaxity}(s_t, \tau) + 1$  [unit time]);
}

```

In the *transmission power consumption laxity-based (TPCLB)* algorithm, a load balancer  $K$  selects a server  $s_t$  where **CommLaxity**( $s_t, \tau$ ) is the minimum in the server cluster  $S$  at time  $\tau$ . The load balancer  $K$  issues the request to the server  $s_t$ .

## 5 Evaluation

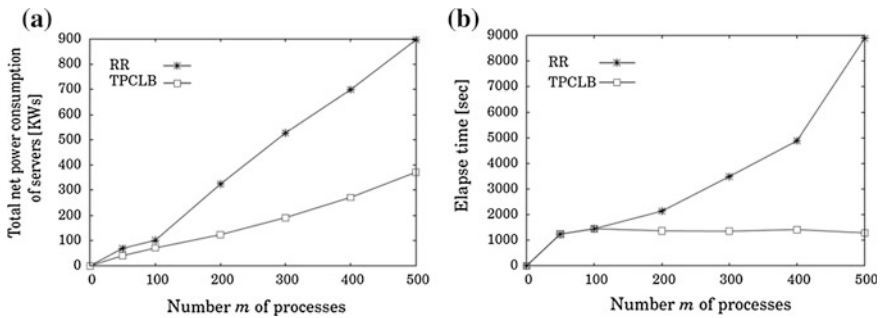
We evaluate the TPCLB algorithm in terms of the total power consumption and elapse time compared with the round-robin (RR) algorithm [4]. There are eight servers  $s_1, \dots, s_8$  each of which holds full replicas of files  $f_1, f_2, f_3$ , and  $f_4$  of 1024, 512, 256, and 103 [MB], respectively, as shown in Table 1,  $S = \{s_1, \dots, s_8\}$ . Parameters of each server are defined based on our previous experimentations. The ETPC model holds for the servers  $s_3$  and  $s_7$ . On the other hand, the TPC model holds for the other servers.

A number  $m$  of clients randomly download files from one server  $s_t$  in the server cluster  $S$ . The maximum receipt rate  $Maxrr_s$  of each client  $c_s$  is randomly selected between 10 and 100 [Mbps]. Each client issues a transfer request of a file  $f_h$  ( $h \in \{1, 2, 3, 4\}$ ) to a load balancer  $K$  at time  $st_s$ . The file  $f_h$  is randomly selected in  $f_1, \dots, f_4$ . The starting time  $st_s$  of each client is also randomly selected between 1 and 1000 [s] at the simulation time. Each client  $c_s$  issues one file transfer request at time  $st_s$  in the simulation. In the evaluation, the TPCLB and RR algorithms are performed on the same traffic pattern.

Figure 1a shows the total net power consumption [KWs] which is obtained by subtracting the power consumption  $minE_t$  of each server  $s_t$  in the idle state from the total power consumption during the simulation, i.e. the power consumption for

**Table 1** Servers

Servers	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$	$s_6$	$s_7$	$s_8$
<i>Model</i>	TPC	TPC	ETPC	TPC	TPC	TPC	ETPC	TPC
$Maxr_t$	160	447	778	802	160	447	778	802
$\delta_t$	0.11	0.02	–	0.012	0.11	0.02	–	0.012
$\gamma_t$	1	1	1	1	1	1	1	1
$minE_t$	105	149	97	96	105	149	97	96
$maxE_t$	282	273	141	230	282	273	141	230
$minETR_t$	4	3	–	2	4	3	–	2
$maxm_t$	10	10	–	10	10	10	–	10
$\beta_t(m)$	1.09	1.5	–	1.42	1.09	1.5	–	1.42



**Fig. 1** Total power consumption and elapse time. **a** Total power consumption [KWs]; **b** Elapse time [s]

number  $m$  of processes. In the TPCLB algorithm, the total net power consumption can be more reduced than the RR algorithm.

Figure 1b shows the elapse time [s] of the TPCLB and RR algorithms for number  $m$  of processes. For  $m \leq 100$ , the elapse time of TPCLB and RR algorithms are almost the same since servers hold enough transmission rates. For  $m > 100$ , the number of transmission overloaded servers in the RR algorithm increases. However, the elapse time of TPCLB algorithm does not increase. This means the transmission rate of servers is efficiently used in the TPCLB algorithm than the RR algorithm.

From the evaluation results, the TPCLB algorithm is more useful than the RR algorithm.

## 6 Concluding Remarks

In this paper, we newly proposed the transmission power consumption laxity-based (TPCLB) algorithm for a cluster of heterogeneous servers which follow the TPC and ETPC models to reduce the total power consumption of a cluster of

servers for communication type applications. In the TPCLB algorithm, a server  $s_r$  whose power consumption laxity  $lpc_r(\tau)$  is the minimum in a cluster of servers is selected for a new request from a client  $c_s$  at time  $\tau$ . We evaluated the TPCLB algorithm in terms of the total power consumption of a cluster of servers and the elapse time compared with the RR algorithm. From the evaluation, the TPCLB algorithm is more useful than the RR algorithm.

## References

1. Enokido T, Aikebaier A, Takizawa M (2011) Process allocation algorithms for saving power consumption in peer-to-peer systems. *IEEE Trans Ind Electron* 58(6):2097–2105
2. Enokido T, Aikebaier A, Takizawa M (2011) An extended power consumption-based algorithm for communication-based applications. *J Ambient Intell Humanized Comput* 2(4):263–270
3. Enokido T, Takizawa M (2012) The extended transmission power consumption model for communication-based applications. In: *Proceedings of the 15th international conference on network-based information systems (NBIS-2012)*, pp 112–119
4. Job scheduling algorithms in linux virtual server (2010). <http://www.linuxvirtualserver.org/docs/scheduling.html>

# The Methodology for Hardening SCADA Security Using Countermeasure Ordering

Sung-Hwan Kim, Min-Woo Park, Jung-Ho Eom  
and Tai-Myoung Chung

**Abstract** In this paper, we considered that SCADA system has few authorized users and access control is one of the most important values for cyber security. We propose the method which reducing the success probability of attacker's penetration using ordered countermeasures. We assume that any system has two or more safety countermeasures for authentication. It follows that setting multiple countermeasures in chain and making a causal relationship before and after action. And then, we making an access procedure matrix for it and sharing them among authorized users. As doing so, we can prevent attacker's penetration and reduce risk level by hacking.

**Keywords** Security hardening · Penetration success probability · Ordered countermeasure

---

Jung-Ho Eom is co-author of this paper.

---

S.-H. Kim · M.-W. Park · T.-M. Chung  
Department of Computer Engineering, School of Information and Communication  
Engineering, Sungkyunkwan University, Suwon-si, Republic of Korea  
e-mail: shkim47@imtl.skku.ac.kr

M.-W. Park  
e-mail: mwpark@imtl.skku.ac.kr

T.-M. Chung  
e-mail: tmchung@ece.skku.ac.kr

J.-H. Eom (✉)  
Military Studies, Daejeon University, 62 Daehakro, Dong-Gu, Daejeon, Republic of Korea  
e-mail: eomhun@gmail.com

## 1 Introduction

In recent years, there has been an increasing interest in cyber attacks on SCADA such as ‘Stuxnet’ in 2010, ‘Duqu’ in 2011 and ‘Flame’ in 2012.

The latest cyber-attacks have a variety of target, Attacker’s penetration is becoming more precise. According to this trend, there is a great deal of research on the cyber defense method for SCADA.

In case of cyber attacks taking control of the critical infrastructure, the damage caused by the malicious use will be very critical. Because of this risk, The United States government has been paying attention to cyber security since the early 2000s.

In this paper, we look at the major features and security-related issues of SCADA system. And we propose a method to reduce the attacker’s penetration success probability.

This paper is organized as follows. In Sect. 2, we present the related works on SCADA security. In Sect. 3, we deal with security issues on SCADA and major features of SCADA related to cyber defense. In Sect. 4, we explain the method to reduce attacker’s penetration success probability. In Sect. 5, we demonstrate a more detailed process with a case study. In Sect. 6, we summarize this paper and provide a conclusion and suggestions for future work.

## 2 Related Work

One of the most significant current discussions in cyber security is SCADA cyber defense. Many kinds of studies including key management, multiple password, and attack/defense tree have been conducted to strengthen SCADA security.

Beaver et al. [1], Dawson et al. [2], Pietre-Cambacedes and Sitbon [3] conducted a study on key management for SCADA network and Ni et al. [4], Adar and Wuchner [5], Taylor et al. [6] and Haimes and Chittester [7] conducted a study on cryptographic algorithms from the point of risk management. Chiasson et al. [8], Topkara et al. [9] studied multiple password and proposed the graphical passwords and passwords separation. We present a new approach to SCADA security to build up this idea.

## 3 SCADA System Security and Access Control

### 3.1 Security Issues in SCADA System

SCADA is a supervisory and control system for large-scale facilities such as power plants and steel mills. The initial SCADA system has no concept on network, but nowadays most SCADA systems are operating in a network environment [10].

As already mentioned, Cyber attacks on SCADA systems continue to increase due to the generalization of network and evolution of cyber attacks. Types of cyber attacks on SCADA systems are as follows.

First, look at the proportion of internal and external attacks, prior to 2000 insider attack by a disgruntled insider occupied 70 % of total. But outsider attacks through the network have occupied 70 % of the total since 2000 [11]. In recent years, the SCADA system can be accessed through the Internet. For this reason, it is becoming increasingly important issue that network protection for SCADA system [12].

Second, the major form of SCADA attack aims to intercept the system administrative authority. Malicious codes such as ‘Stuxnet’ attacks on the control authority for PLC (Programmable Logic Controllers) or RTU (Remote Terminal Units). PLC and RTU control the field devices directly by converting electrical signals into physical signals [13]. The final goal of the SCADA attack would be a great confusion rather than financial gain. So, it can be seen that countermeasures of access control are very important for both attack and defense.

### ***3.2 Internet SCADA Architecture***

Undoubtedly, some security consultants didn’t know the characteristics of the system and could not provide effective advice about the countermeasures for hacking attacks. In the view of cyber defense, the features of SCADA system are as follows.

First, the scale of SCADA system users is smaller than other common systems. SCADA systems communicated not with users but with devices mainly. So, there are few users in the SCADA system. Thus, key management is simple and access load is low.

## **4 SCADA System Security Hardening Methodology**

### ***4.1 Background***

Let us consider the following. The act of opening the door is access and the locking mechanism is the control method for access. Only authorized users can open the lock in normal conditions. There exists a possibility that abnormal access using an illegal key or breaking the lock can open. The locking device may be only one. Also depending on the level of security, it would be set two or more.

The following figures depict the door lock that has two block devices. One is a number device another is key.

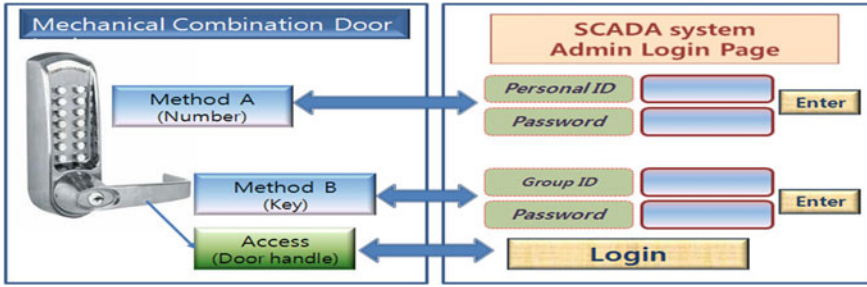


Fig. 1 Countermeasure combination

The open door (access) process in Fig. 1 can be represented by the following formula.

- Common access procedure

- ① *Open(KeyLock)* → ② *Open(NumberLock)* → ③ *Pull(handle)or*
- ① *Open(NumberLock)* → ② *Open(KeyLock)* → ③ *Pull(handle)*

Let  $i$ th Countermeasure in system  $X = CM(x_i)$ ,  
 Attacker’s Penetration Success Probability on  $CM(x_i) = PSP(CM(x_i))$ ,  
 Authorized user’s unlocking success probability is 1,  
 $n =$  total number of countermeasures ( $n \geq 2$ )  
 Unauthorized user penetration success probability. Then

$$\begin{aligned}
 PSP(SystemX) &= PSP(CM(x_1)) \times \dots \times PSP(CM(x_n)) \\
 &= \prod_{i=1}^n PSP(CM(x_n))(n \geq 2)
 \end{aligned}
 \tag{1}$$

### 4.2 Ordered Countermeasures

Service this section will explain the security hardening method step by step based on the previous section. The core of the proposed method is setting a causal and procedural relationship between the countermeasures. Generating the dependency condition between the countermeasures is the first step. In case that ‘a’ is a prerequisite of ‘b’, it is expressed as follows.

$$a \Rightarrow b \neq b \Rightarrow a$$

Any system X has two countermeasure; countermeasure ‘a’ and ‘b’. The countermeasure ‘a’ must be unlocked before countermeasure ‘b’. That case can be presented as follow formula:

$$CM(x_a) \Rightarrow CM(x_b)$$



It means that CM(a) should be open (unlock) before the execution of CM(b). Let's suppose some system has  $n(n \geq 2)$  countermeasures. And the all countermeasure of that system have to execute only once. That system can be created procedure number of  $n!$ .

$$nPr = \frac{n!}{(n-r)!}, \quad nPr = n!(n=r)$$

Also, let  $X'$  is the security hardening system on  $X$ , attacker's penetration success probability is as follows:

$$PSP(X') = \frac{\prod_{i=1}^n PSP(CM(x_n))}{n!} \tag{2}$$

Since  $n$  is 2 or more, the hardening method could reduce the attacker's success probability below 50 %.

### 4.3 Access Procedure Matrix and Main Process

The second is the process of creating a matrix of all possible procedures as shown below table. After creating the procedure matrix, it is being distributed to all authorized users. As we have mentioned previously, SCADA systems have a small and limited user group. Thus, the system shares the access matrix easier than other common system (Table 1).

### 4.4 Access Procedure Using Ordered Countermeasures

Where AC = Authorized Client, SCS = Control Server of SCADA  
 AP(number) = Access Procedure number  
 CM(SCS<sub>a</sub>) = Countermeasure A of SCADA control server system  
 AP(1) = CM(SCS<sub>a</sub>) ⇒ CM(SCS<sub>b</sub>)  
 P(CM(SCS<sub>a</sub>)) = Password of Countermeasure A in system SCS  
 The main procedure described so far can be expressed as follows.

- (1) Authorized user request access procedure matrix number for SCS  
 AC → SCS : AC request number of AP for SCS
- (2) SCS provide the procedure number (Let #1 : CM(SCS<sub>a</sub>) ⇒ CM(SCS<sub>b</sub>))  
 SCS → AC : AP(1)

**Table 1** Access procedure matrix

Procedure number	First	Second	...	n th
1	CM 1	CM 2	...	CM n
2	CM 2	CM 1	...	...
...	...	...	...	...
n!	CM n	...	...	CM 1

(3) Authorized user access countermeasure A with password of A

$$AC \rightarrow CM(SCS_a) : AC \text{ unlock } CM(SCS_a) \text{ by } P(CM(SCS_a))$$

(4) Authorized user access countermeasure B with password of B

$$AC \rightarrow CM(SCS_b) : AC \text{ unlock } CM(SCS_b) \text{ by } P(CM(SCS_b)).$$

(5) AC access the SCS : Finish.

## 5 Demonstration

In this chapter, we explain the security hardening method using the example case. We assume a virtual power generation SCADA system as below. And we compare the penetration success probability before and after using the proposed hardening method (Table 2).

Firstly, we apply the example table data to the formula 1.

$$\begin{aligned} PSP(PP) &= \prod_{i=1}^n PSP(CM(PP_n)) \\ &= PSP(CM(PP_1)) \times PSP(CM(PP_2)) \times PSP(CM(PP_3)) = 0.3 \times 0.2 \times 0.1 = 0.006 \end{aligned}$$

For comparison, we substitute the formula 2.

$$\begin{aligned} PSP(\text{Hardening PP}) &= \frac{\prod_{i=1}^n PSP(CM(PP_n))}{n!} = 0.3 \times 0.2 \times 0.1 / 6 = 0.001 \\ &= 0.1(\%) \end{aligned}$$

**Table 2** Power plant system specification

Item	Description
<i>Purpose of SCADA</i>	Power plant (PP) management
<i>Total Countermeasure number</i>	3[(CM(PP <sub>1</sub> ), CM(PP <sub>2</sub> ), CM(PP <sub>3</sub> )]
<i>Total number of procedure</i>	3! = 6
<i>Penetration Success Probability</i>	CM(PP <sub>1</sub> ) = 0.3, CM(PP <sub>2</sub> ) = 0.2, CM(PP <sub>3</sub> ) = 0.1

As a result, we can see that the penetration success probability was reduced by a factor of  $n!$ .

## 6 Conclusion and Future Work

In this paper, we explain the security hardening method by setting the causal relationship between the access control countermeasures. The primary target of SCADA cyber attack is obtaining an access control. For this reason, we focused on the reducing the attacker's access control penetration probability. It is difficult to say that the proposed method is absolute method. But it is clear that proposed method is a useful methodology for SCADA system security. In the future, we will refine algorithm for ordered countermeasures and evaluating the validity. We will also conduct a further study to apply this in risk management.

**Acknowledgments** This work was supported by the IT R&D program of MKE/KEIT. [10041244, Smart TV 2.0 Software Platform].

## References

1. Beaver C, Gallup D, Neumann W et al (2002) Key management for SCADA. Cryptog information systems security dept, Sandia Nat. Labs, Technical Report SAND 2001-3252
2. Dawson R, Boyd C, Dawson E et al (2006) SKMA: a key management architecture for SCADA systems. In: Proceedings of the 2006 Australasian workshops on grid computing and e-research ACSW Frontiers '06, vol 54, pp 183-192
3. Pietre-Cambacedes L, Sitbon P (2008) Cryptographic key management for SCADA systems-issues and perspectives. International conference on information security and assurance ISA 2008. pp 156-161
4. Ni M, McCalley JD, Vittal V et al (2003) Online risk-based security assessment. IEEE Trans Power Syst 18:258-265
5. Adar E, Wuchner A (2005) Risk management for critical infrastructure protection (CIP) challenges, best practices and tools. First IEEE international workshop on critical infrastructure protection
6. Taylor C, Krings A, Alves-Foss J (2002) Risk analysis and probabilistic survivability assessment (RAPSA) an assessment approach for power substation hardening
7. Haimes YY, Chittester CG (2005) A Roadmap for quantifying the efficacy of risk management of information security and interdependent SCADA systems. J Homel Secur Emerg Manage 2:1-21
8. Chiasson S, Forget A, Stobert E et al (2009) Multiple password interference in text passwords and click-based graphical passwords. In: Proceedings of the 16th ACM conference on computer and communications security CCS '09. pp 500-511
9. Topkara U, Atallah MJ, Topkara M (2006) Passwords decay, words endure: secure and reusable multiple password mnemonics. In: Proceedings of the 2007 ACM symposium on applied computing SAC '07. pp 292-299

10. Cai N, Wang J, Yu X (2008) SCADA System security: complexity, history and new developments, industrial informatics. INDIN 2008. 6th IEEE international conference on 2008. pp 569–574
11. Ijure VM, Laughter SA, Williams RD (2006) Security issues in SCADA networks. Computer and security 2006. pp 498–506
12. Qiu B, Gooi HB (2000) Web-based SCADA display systems (WSDS) for access via internet. IEEE transactions on power systems, vol 15. pp 681–686
13. Chunlei W, Lan F, Yiqi D (2010) A simulation environment for SCADA security analysis and assessment. International conference on measuring technology and mechatronics automation (ICMTMA) 2010, vol 1. pp 342–347

# Development and Application of STEAM Based Education Program Using Scratch: Focus on 6th Graders' Science in Elementary School

JungCheol Oh, JiHwon Lee and JongHoon Kim

**Abstract** For this study, we reviewed theoretical background of STEAM education and domestic and international case studies in STEAM education. By doing so, we developed and applied the STEAM Education Program through the use of Scratch. This program is designed for the 3rd (“Energy and Tools”) and 4th (“Combustion and Extinguishing”) lessons of 6th graders’ science in elementary school. As a result, the creativity index and positive attitude about science of the students who went through the researched program increased with meaningful difference compared to that of the sample population. The result of this study shows that ‘The STEAM Education Program,’ using Scratch, can improve creativity. And it is sure that it brings positive changes for the Science Related Affective Domains.

**Keywords** STEAM · Scratch · Creativity · Scientific attitude · Fluency

## 1 Introduction

In the present century, science and technology has combined with human life in a more humane and artistic way than ever before. The iPhone, introduced by Steve Jobs, an artistic engineer, is a good example of how science and technology combine with human life in a human-friendly and artistic way. This societal

---

J. Oh · J. Lee · J. Kim (✉)  
Department of Computer Education, Teachers College,  
Jeju National University, Jeju, Korea  
e-mail: jkim0858@jejunu.ac.kr

J. Oh  
e-mail: lov0502@naver.com

J. Lee  
e-mail: torchere@naver.com

demand has been reflected in the educational community, and the Ministry of Education and Science of South Korea has established various educational strategies for educating students to create fused talents [1].

In 2009, Partnership for Twenty-first-Century Skills, an organization in the U.S., suggested essential skills that students need to learn and master in order to succeed in the twenty first century [2]. The organization suggested that students should learn such skills as creativity, critical thinking, problem solving, communication, and collaboration through art, mathematics, science, economics, and history. That is, learners need to develop the ability to fuse diverse skills holistically based on creativity to be able to adapt to the rapidly changing society of the twenty first century and to get ahead of the times.

The government stressed the importance of STEAM education to train people with such holistic talents and is now preparing various strategies for educating them. This study aimed to develop a STEAM education program that can be applied in the field, using technology and engineering Scratch, which can be easily accessed by learners in the aforementioned context.

## 2 Theoretical Background

### 2.1 *Need for and Definition of STEAM*

Smart STEAM stands for Science, Technology, Engineering, Arts, Mathematics and means learning the fused knowledge of various fields. The effort to find the cause of the economic crisis in the U.S. led to the identification of the decrease in the academic performance of mathematics and science learners as the cause [3, 4]. To address this problem, STEAM aims to promote the learners' motivation for learning and to educate people so as to help them become capable of solving multidisciplinary problems.

Regarding STEAM fusion education, Yakman (2008) presented a pyramid model consisting of several levels, from continuing education to the classification of the detailed study contents and stated that the interdisciplinary integrative level was appropriate for elementary school education [5].

Although in most of the current STEAM education programs computers serve only as auxiliary tools, in this study, it was used as a main activity in the fusion education for applying and utilizing the science lessons. That is, the students can clearly understand the lesson contents and can see the process and result of making programs real-time. Thus, the possible errors in the real-life experiment can be reduced, and the students can have opportunities to apply and express the scientific principle in various ways using divergent and creative methods, without temporal and spatial restrictions.

### 3 Design and Making of the Steam Education Program

#### 3.1 Design of the Stages of Steam Teaching and Learning

Study stages were established as follows to develop the STEAM education program using Scratch, and to apply the developed program. The study was conducted for 6 months, according to the study stages. The teaching and learning stages for science and technology presented by Miaoulis (2009) and the creative comprehensive design stages for people with inter-disciplinary talents presented by Korea Foundation for the Advancement of Science and Creativity (2011) were reviewed to design the stages for STEAM teaching and learning that will be applied to the education field.

Table 1 shows the stages for creative design presented by Korea Foundation for Advancement of Science and Creativity when it introduced the concept of creative comprehensive design education. The foundation explained that the stages are the characteristics of the new STEAM education.

Based on this, six stages of STEAM teaching and learning Table 2 were determined, and the program was conducted. In particular, during the “making or synthesizing” and “testing” stages, creative work was accomplished by frequently making and modifying the activities. Thus, continuous testing and feedback are required.

**Table 1** Stages for creative design presented by Korea foundation for advancement of science and creativity

1. Setting of objective	2. Planning and designing	3. Analysis of design	4. Making	5. Test	6. Evaluation
-------------------------	---------------------------	-----------------------	-----------	---------	---------------

**Table 2** Stages of STEAM teaching and learning

1. Experiencing priming water for an idea	Define the issues and experience priming water for an idea related with the issue
2. Coming up with an idea	Create various ideas and share them with collaborators
3. Planning and design fusion	Establish a plan for materializing the idea, and make a design by fusing related studies
4. Making or synthesizing	Make or synthesize works based on the idea using scientific, technological, engineering, and artistic methods
5. Testing	Test, and Feedback or modifying
6. Evaluation	Inter-collaborators’ evaluation, and refinement of the idea through evaluation

### 3.2 Formulation of the Steam Education Program

The units were constructed, using the stages of the STEAM teaching and learning activities and the categorized science experiment themes. Activities for making games using Scratch were included considering that the learners are elementary school students, who are highly interested in games, and that various artistic activities, such as plotting a story, selecting background music, and drawing characters, are complexly performed for game creation.

At the “experiencing priming water for an idea” stage of each unit, the teachers pre-sent Scratch games to help the learners come up with an idea. At the “coming up with an idea” stage, the learners are allowed to present and discuss the games that they want to make, and by doing so, to share the ideas with one another.

At the “planning and integrative design” stage, the learners are allowed to make a storyboard based on the various science experiments presented in the textbook. Many storyboards can be made, according to the experiment, or only one storyboard can be made in detail. The learners are allowed to make storyboards freely, without restrictions in theme and expression. At the “making and synthesizing” stage, the learners are allowed to formulate “instructions on how to play the game made” based on the storyboard made, and to make games using Scratch.

At the “testing” stage, the learners can modify the game by playing it with their collaborators, and by reviewing it. There was no boundary between the “making or synthesizing” and “testing” stages, allowing the learners to make, synthesize, and review activities whenever they needed to do so, and by doing so, to complete creative works like those shown in Fig. 1.

In addition, the learners can write what they felt through these two stages in “Instructions on how to play the game made.” The learners who made a maze game through several stages added mazes to the game by reviewing the game together with their collaborators (Fig. 2) and by completing the final game.

At the “evaluation” stage, the learners can display the games that they made, and can play these, so as to evaluate one another’s games. The learners can also refine their ideas by sharing their completed ideas with others.

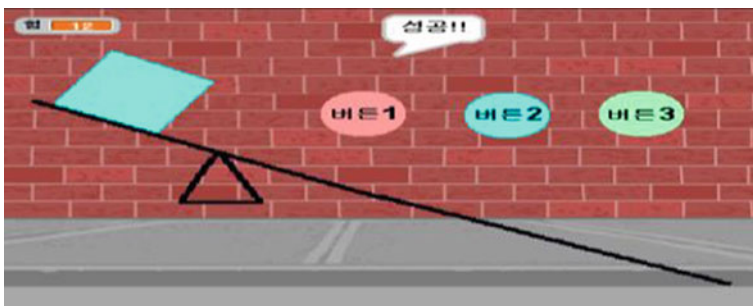


Fig. 1 The game completed after the review with the collaborators, and after modification



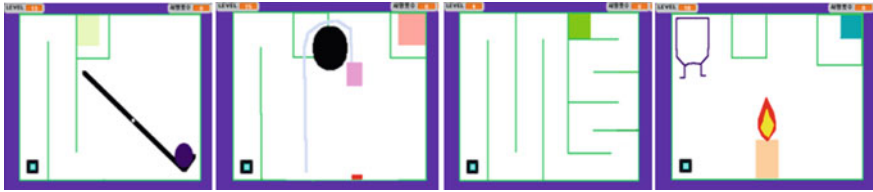


Fig. 2 Games made by other learners

## 4 Application of the Program and Analysis of the Results

### 4.1 Study Design and Control of Variables

Two classes in the 6th grade of Elementary School in Jeju City, Jeju Island, South Korea were included in the study (Table 3). The classes were assigned as the experimental and control groups, respectively. For 10 weeks, from the fourth week of September 2011 to the first week of December 2011, the Scratch-based STAEM education program was conducted in the experimental group, on the two units of the science textbook.

As shown in Table 3, a test of baseline creativity and an evaluation of the affective characteristics associated with science were performed in the two groups, and the homogeneity of the two groups was validated. For the control group, a normal science class was conducted as planned.

### 4.2 Test Tools

The area that this study intended to identify through objective validation after the application of the Scratch-based STEAM education program was the creativity and affective characteristics associated with science. To determine if the creativity had been enhanced, Torrance’s TTCT (diagram) Creativity Test Type A was performed before and after the experiment. In addition, “Evaluation System for Affective Characteristics (Attitude) Related with National Science” developed by Hyonam et al. [6] at the Science Education Institute of Korea National University of Education, based on the theory on science-related attitude presented by Klopfer

Table 3 Experimental group

Division	Number of students		
	Male	Female	Total
The experimental group	12	13	25
Comparison group	12	13	25
Total	24	26	50

and the criteria for evaluation items presented by Edward, was used for the test of affective characteristics related with science. Type A consisted of items evaluating the awareness of and interest in science, and type B consisted of items evaluating scientific attitude. For the reliability of the items, the Chronbach alpha coefficients were 0.83 and 0.86 for types A and B, respectively. As both exceeded 0.8, they were deemed reliable.

### 4.3 Results of the Creativity Test and Interpretation of the Results

To verify the homogeneity of the factors of creativity between the experimental and control groups, the mean of each area of creativity in each group was tested through a *t* test, using SPSS 12.0 for Windows, before the experimental treatment ( $p = 0.05$ ). As shown in Table 4, the significance probability for the creativity index was  $p = 0.929$ ; thus, there was no significant difference in creativity between the experimental and control groups. The other areas of creativity, such as fluency, creativity, abstracting ability, and delicacy, were not significantly different between the two groups as the significance probability was higher than 0.05.

Ten weeks later, the creativity test was performed again on the experimental and control groups. As shown in Table 5, the creativity and originality were significantly different between the groups, with  $p = 0.036$  and  $p = 0.039$ , respectively. In particular, the post-experiment creativity increased by 10.17, and the significance level was remarkably different, with  $p = 0.000$  ( $p < 0.05$ ) in the experimental group.

Then, the difference in creativity before and after the experiment was compared within the experimental group and was analyzed. As shown in Table 6, in the

**Table 4** Test of baseline creativity

Domain	Class	N	Average	The standard deviation	T	Note that the probability
Fluency	Pre	25	115.32	24.81	-1.619	N.S.
	Post	25	122.72	18.57		0.118
Originality	Pre	25	104.80	25.51	-0.202	N.S.
	Post	25	106.04	23.09		0.842
Abstractness of title	Pre	25	95.44	29.87	0.507	N.S.
	Post	25	90.36	37.30		0.617
Elaboration	Pre	25	94.20	18.45	0.676	N.S.
	Post	25	91.36	14.17		0.505
Resistance to premature closure	Pre	25	99.60	17.60	-0.663	N.S.
	Post	25	101.56	11.88		0.513
Creativity index	Pre	25	105.99	19.25	-0.090	N.S.
	Post	25	106.37	16.15		0.929

\*  $p < 0.05$ , N.S. not significant, N number of cases

**Table 5** Post results comparing

Domain	Class	N	Average	The standard deviation	T	Note that the probability
Fluency	Pre	25	125.52	19.12	-2.225	0.039*
	Post	25	132.96	13.89		
Originality	Pre	25	117.00	25.82	-2.187	0.039*
	Post	25	129.24	22.81		
Abstractness of title	Pre	25	100.64	22.93	0.724	N.S.
	Post	25	95.28	29.65		
Elaboration	Pre	25	91.64	18.45	-0.593	N.S.
	Post	25	93.96	14.88		
Resistance to premature closure	Pre	25	104.28	15.62	1.212	N.S.
	Post	25	99.92	15.19		
Creativity Index	Pre	25	114.29	14.67	-4.104	0.000*
	Post	25	116.54	16.29		

\* $p < 0.05$ , *N.S.* not significant, *N* number of cases

comparison of the fluency, originality, and creativity index before and after the experiment within the experimental group, the significance probability was less than 0.05; thus, the difference was significant. It was found that Scratch-based STEAM education had a positive effect on the improvement of the fluency, originality, and creativity index.

#### ***4.4 Results of the Test of Affective Characteristics Related with Science***

The evaluation of affective characteristics developed by Hyonam et al. [6] consists of three categories (awareness, interest, and scientific attitude), and each item is graded based on a 5-point Likert scale. The results of the pre-experiment test showed that compared with the control group, the awareness of science was low (0.548 points), the interest in science was high (0.008), and the scientific attitude was high (0.284) in the experimental group (Table 7).

The post-experiment test showed that the mean of awareness (C) of science increased by 0.168, the interest in science (I) by 0.231, and the scientific attitude (A) by 0.281. In particular, compared with the control group, the awareness and interest of the experimental group considerably increased. The results showed that the Scratch-based STEAM education program had a positive effect on the affective characteristics related with science in the experimental group.

**Table 6** Results of the test of creativity by time point

Domain	Class	N	Average	The standard deviation	T	Note that the probability
Fluency	Pre	25	122.72	18.57	-3.556	0.002*
	Post	25	132.96	13.89		
Originality	Pre	25	106.04	23.09	-5.705	0.000*
	Post	25	129.24	22.81		
Abstractness of title	Pre	25	90.36	37.30	-0.571	N.S.
	Post	25	95.28	29.65		
Elaboration	Pre	25	91.36	14.17	-0.732	N.S.
	Post	25	93.96	14.88		
Resistance to premature closure	Pre	25	101.56	11.88	-0.543	N.S.
	Post	25	99.92	15.19		
Creativity index	Pre	25	106.37	16.15	-3.323	0.003*
	Post	25	116.54	16.29		

\* $p < 0.05$ , *N.S.* not significant, *N* number of cases

**Table 7** Results of the test of affective characteristics related with science

Domain	Class	N	Average	The standard deviation	T	Note that the probability
Cognition	Comparison	Pre	25	3.599	0.419	+0.014
		Post	25	3.613	0.351	
	Experiment	Pre	25	3.051	0.366	+0.168
		Post	25	3.219	0.440	
Interest	Comparison	Pre	25	3.293	0.585	+0.003
		Post	25	3.296	0.586	
	Experiment	Pre	25	3.301	0.579	+0.231
		Post	25	3.532	0.615	
Attitude	Comparison	Pre	25	3.081	0.424	+0.231
		Post	25	3.312	0.460	
	Experiment	Pre	25	3.365	0.443	+0.281
		Post	25	3.646	0.661	

## 5 Conclusions

This study aimed to develop and apply a STEAM education program that can be applied to elementary school education, and to show the effect of the program according to the education for people with integrated talents stressed by the Ministry of Education and Science. Towards these ends, foreign and local studies on STEAM were reviewed, and appropriate subjects and contents that could be applied to the education were selected and combined with Scratch, an educational programming language. In addition, to increase the appropriateness of the program, discussion with and inter-views of experts in schools and colleges were

performed during the selection of education contents and the designing of the stages of teaching and learning. Then, the developed STEAM education program was applied to the students for 10 weeks, under the condition where all the possible variables were controlled. The fluency, originality, and creativity index significantly increased in the experimental group, which used the STEAM education program, as opposed to the control group, and the positive answers in the area of awareness of and interest in the affective area considerably increased in the experimental group. These results indicate that the Scratch-based STEAM education program has a positive effect on the creativity and affective characteristics related with science. In addition, the STEAM education program developed in this study has the two following meanings: (1) STEAM education was utilized without interrupting the flow of the units of the science textbook, to allow the smooth progression of the current curriculum; and (2) using the educational programming language, methods of fusing science, technology, engineering, arts, and mathematics to attain what STEAM intends to achieve were presented.

## References

1. Ministry of Education Science and Technology (2010) The future Republic of Korea to open using creative talent and advanced science technology
2. P21 Framework Definitions (2009) The partnership for 21st century skills
3. Tarnoff J (2010) STEM to STEAM—recognizing the value of creative skills in the competitiveness debate
4. Puffenberger A (2010) The STEAM movement: it's about more than hot air
5. Yakman G (2008) STEAM education: an overview of creating a model of integrative education. In: Proceeding of PATT on 19th ITEEA conference, pp 335–358
6. Hyonam K, WanHo C, JinWoo J (1998) National assessment system development of science-related affective domain. *Korean J Sci Edu* 18(3)357–369

**Part IX**  
**Advanced Technologies and Applications**  
**for Cloud Computing and Sensor Networks**

# Performance Evaluation of Zigbee Sensor Network for Smart Grid AMI

Yong-Hee Jeon

**Abstract** Smart Grid is a convergent system of the existing power grid and Information Technology (IT). In Smart Grid system, AMI (Advanced Metering Infrastructure) is a system with various sensors to evaluate data related with the usage of various utility resources such as electricity, gas, and water. In this paper, IEEE 802.15.4 MAC protocol for the AMI sensor networks is analyzed. Based on the analysis, OPNET simulator was implemented for the performance evaluation. Particularly, this paper focuses on the effect of protocol parameters to the performance. For these parameters, the simulation results are presented to achieve the most efficient usage of network under different traffic loads. Based on the performance evaluation results, it is revealed that real-time data transmission is possible if the total offered load is restricted under 50 % with 100 nodes and the offered load period from 20 to 50 % makes the best trade-off in terms of network throughput, average delay, etc.

**Keywords** Smart grid · AMI sensor network · Performance · Simulation · OPNET

## 1 Introduction

In Smart Grid system, it is required to monitor and control in real-time the electric power grid for the efficient operation. Therefore the Smart Grid communication networks have a strict latency requirement such as the maximum message transmission time. As an important component of the Smart Grid system, AMI refers to the collection of systems to evaluate data related with the usage of various utility resources such as electricity, gas, and water.

---

Y.-H. Jeon (✉)

Department of IT Engineering, Catholic University of Daegu, Gyeongsan,  
Republic of Korea  
e-mail: yhjeon@cu.ac.kr

Among the communication types of Smart Grid AMI, IEEE 802.15.4 Zigbee is one of the mostly considered types due to the superior performance characteristics. Zigbee uses the IEEE 802.15.4 as its MAC layer protocol. In the MAC protocol, there are numerous protocol parameters that affect the performance of Zigbee communication, such as Beacon Order (BO), Superframe Order (SO), Number of Backoffs (NB), Backoff Exponent (BE), etc. This paper intends to examine the effect of those parameters to the delay performance of AMI sensor networks, including security aspects [1].

## 2 Background

In the star topology of IEEE 802.15.4, communication types may be categorized into two types such as beacon-enabled mode and non beacon-enabled mode based on data communication types between coordinator and end device. In this paper, beacon-enabled mode with single hop star-topology is assumed. According to the IEEE 802.15.4 superframe structure, it consists of beacon interval for the synchronization of the superframe, active period for the actual data transmission, and inactive (sleep) period for low power consumption. Beacon interval and active period are determined by Beacon Order (BO) and Superframe Order (SO) values respectively. Active period is further divided into Contention Access Period (CAP) and Contention Free Period (CFP) depending on whether each node trying to transmit data contends with other nodes or not. In the CAP, Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) protocol is used for data transmission. In order to get Quality of Service (QoS) guarantee and for the successive data transmission, end device may request CFP to its coordinator [2, 3].

The slotted CSMA/CA algorithm starts by setting the appropriate initial values. Number of Backoffs (NB) refers to how many times the device may backoff due to the unavailability of the medium and is set to zero. Contention Window (CW) refers to the number of backoff periods that need to be clear of channel activity before the packet is allowed to transmit and is set to 2. Backoff Exponent (BE) determines the number of backoff periods for a device to wait prior to attempting the channel access. The default value of  $\text{macMinBE}$  is set to 3. Then, the boundary of the next backoff period is determined and a random number is generated in the range of 0 to  $2^{\text{BE}} - 1$ .

The procedure then counts down for this number of backoff periods, which is so called Random Backoff Countdown (RBC). Then the algorithm performs Clear Channel Assessment (CCA) and checks whether the channel is busy or not. If the channel is idle, the value of CW is decremented by one. After the CCA is performed one more time and if the channel is idle, the corresponding node occupies the medium. If the channel is busy, the value of CW goes back to 2, the number of NB is incremented by one, and the value of BE is chosen again. When the value of NB is greater than the limit of  $\text{macMaxCSMABackoffs}$ , the transmission is assumed to fail and the algorithm is terminated [3].



**Table 1** Simulation parameters

Parameter	Value or types
Traffic model	CBR or Poisson
Number of nodes	[10 ... 100]
BO = SO	[0 ... 14]
CW	2
NB	5
macMinBE	[1 ... 5]
Packet size	103 byte
Offered loads	[20 % ... 300 %]

### 3 OPNET Simulation and Results

There are various protocol parameters in IEEE 802.15.4 such as BO, SO, NB, BE, etc. The number of combination for these parameters may be more than 9,000, considering 16 values of the SO parameter, 16 values of the BO parameter, 6 values of the NB parameter, and 6 values of the BE parameter. Thus parameter tuning is required for the efficient usage of Smart Grid AMI sensor networks. In this paper, OPNET simulator was implemented for the performance evaluation of Zigbee sensor networks. The performance test bed of Zigbee sensor networks was used with 100 end devices and 1 coordinator [4].

Table 1 shows simulation parameters. In general, Poisson traffic model is used due to the convenience of mathematical analysis. In the simulation, there was no significant difference between Poisson and Constant Bit Rate (CBR) models. Since AMI data is sampled periodically, CBR traffic model is assumed in this paper. In the simulation, the performance of uplink traffic from AMI node to coordinator is only considered in CSMA/CA beacon-enable mode [5, 6].

The maximum data rate of physical layer is assumed as 250 kbps and acknowledgement is not considered. OPNET model is broadly categorized into Node model and Process model. Node model is used to model each network device and process model is used to simulate a practical operational behavior of network device.

Figure 1 shows the values of throughput under different traffic loads by SO and BO parameters. For low SO (BO) values, especially when the SO value equals to zero, slotted CSMA/CA mechanism results in more than 50 % of overhead under all traffic loads. This is due to the fact that many backoff periods are wasted by CCA deference and beacon frames are generated more frequently. CCA deference occurs when the remaining backoff period in the superframe is smaller than the required number for the transmission of total frame. It is shown that the effect of SO to the network throughput is decreasing when the value of SO is larger than 4. It is analyzed that the probability of CCA deference becomes lower for higher SO values and thus the throughput may be increased by the reduction of collision. The figure shows that the corresponding throughput is 80 % for SO values greater than or equal to 4 and falls in the range of 40–76 % for lower than 4, both for lower than 50 % of traffic loads.

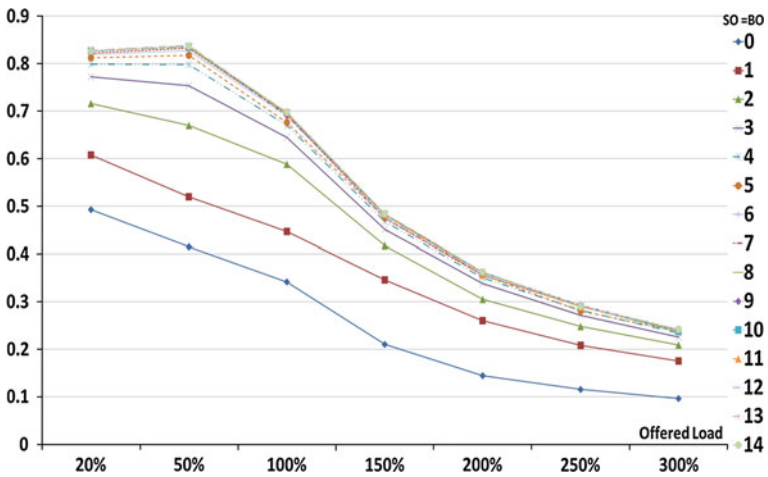


Fig. 1 Throughput under different traffic loads by SO, BO parameter values

Figure 2 shows average delay under different traffic loads by SO, BO parameter values. Similar to the effects on the throughput and link utilization, the average delay becomes saturated for high traffic loads and the effect is decreasing for the SO values greater than or equal to 4. However the delay is increasing as the SO values becomes larger. This is because the backoff delay is not increasing for lower SO values while the delay is increasing for higher SO values [5, 6].

Figure 3 shows that how the delay varies as times go by. In this figure, the large delay appears periodically with a constant interval. This phenomenon is analyzed to occur by the reason that many nodes try to access the medium right after beacon signal and thus increased collision.

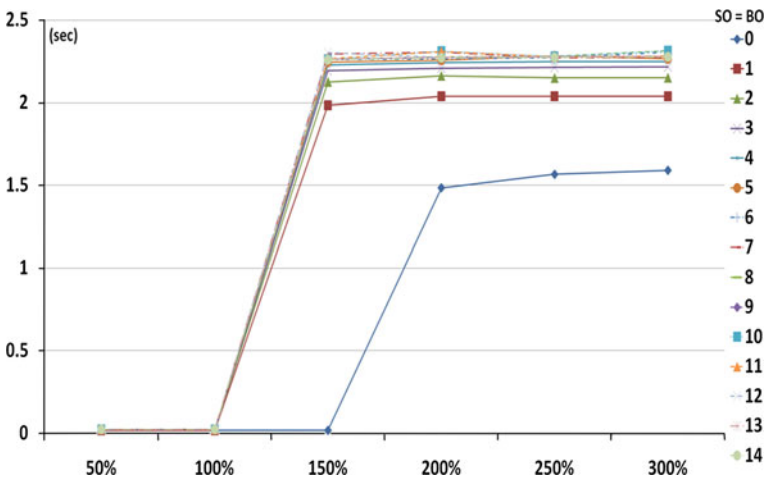


Fig. 2 Average delay under different traffic loads by SO, BO parameter values

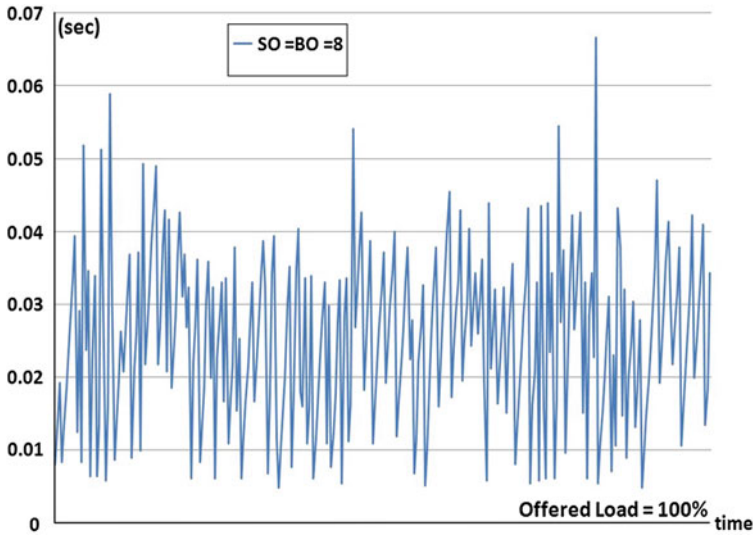
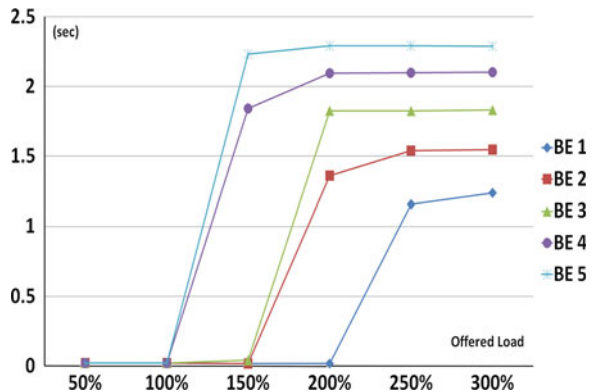


Fig. 3 Delay variations by time when SO = BO = 8 and offered traffic load = 100 %

Figure 4 shows the average delay by BE values.

BE is an important parameter in the backoff algorithm of slotted CSMA/CA. This parameter affects to the backoff delay prior to the channel access trial. In IEEE 802.15.4, this parameter may have the values in the range of 0–5. However, the range of 1–5 is used in the simulation because the value 0 disables the collision avoidance in the first iteration. As expected, if the macMinBE increases, the average delay increases as well. This is due to the fact that the probability of success becomes higher for the increased backoff delay by the increased macBinBE values.

Fig. 4 Average delay under different traffic loads by BE values



The utility may be defined as  $\{(\text{throughput} \times \text{reference delay})/\text{delay}\}$ . It can be used to estimate the level of performance by the measured delay. The value was maximized in the range of 20–50 % period of offered traffic load. In the range of 50–100 % of offered traffic load, the utility value is decreasing slowly. When the offered traffic load is  $\geq 150$  %, the network performance becomes severely deteriorated, resulting in the deadlock. Based on the utility graph, it was revealed that the offered traffic load to achieve an optimal efficient usage of the Zigbee sensor network falls between the range of 20 and 50 %.

In addition to the delay performance of Zigbee sensor network as discussed so far, the security function for the network is also important. In the AMI system design and implementation, the confidentiality and integrity of sensor network data must be guaranteed. However, due to the page limitation, simulation results with regard to the security function are not described in this paper.

## 4 Conclusion

This paper deals with the performance issues of Smart Grid AMI Zigbee sensor network. In particular, the effect of protocol parameters in IEEE 802.15.4 to the network performance was examined. Based on the simulation results, it is revealed that the network may have an optimal operation with regards to network throughput, average delay, etc. in the range from 20 to 50 % of offered traffic loads. To deliver AMI sensor network data in real-time, it is analyzed that the offered traffic load is required to be restricted below 50 %.

## References

1. Zigbee Alliance (2009) Zigbee smart energy profile 2.0 technical requirements document
2. IEEE 802.15.4 Standard-2003, Part 15.4 (2003) Wireless medium access control (MAC) and physical (PHY) layer specifications for low-rate wireless personal area networks (LR-WPANS), IEEE-SA standards board
3. Misić J, Misić VB (2008) Wireless personal area networks—performance, interconnections and security with IEEE 802.15.4. Wiley Series on Wireless Communications and Mobile Computing, Wiley, Chichester
4. OPNET Technologies Inc. OPNET Modeler Wireless Suite—ver. 11.5A. <http://www.opnet.com>
5. Rohm D, Goyal M, Hosseini H, Divjak A, Bashir Y (2009) A simulation based analysis of the impact of IEEE 802.15.4 MAC parameters on the performance under different traffic loads. *Mob Inf Syst* 5:81–99
6. Koubaa A, Alves M, Nefzi B, Tovar E (2006) A comprehensive simulation study of slotted CSMA/CA for IEEE 802.15.4 wireless sensor networks. Paper presented at IEEE International Workshop on Factory Communication Systems, pp 183–192

# P2P-Based Home Monitoring System Architecture Using a Vacuum Robot with an IP Camera

KwangHee Choi, Ki-Sik Kong and Joon-Min Gil

**Abstract** We propose Peer-to-Peer-based (P2P-based) home monitoring system architecture to exploit a vacuum robot with an IP camera, a movable IP camera, without requiring a number of cameras for the whole monitoring at home. The key implementation issues for the proposed home monitoring system are (1) the easy configuration of a vacuum robot to connect to Wi-Fi networks, (2) the session management between a vacuum robot and a home monitoring server, and (3) the support of the Network Address Translator (NAT) traversal between a vacuum robot and a user terminal. In order to solve these issues, we use and extend the Wi-Fi Protected Setup (WPS), the Session Initiation Protocol (SIP), and the UDP hole punching. For easy configuration of a vacuum robot, we also propose the GENERATE method which is an extension to the SIP that allows for the generation of a SIP URI. In order to verify the feasibility of the proposed system, we have implemented the prototype and conducted the performance test using the authoritative call generator.

**Keywords** Home monitoring · Vacuum robot · Home networks · SIP

---

K. Choi

Service Development Unit, LG Uplus, 34 Gajeong-Dong, Yuseong-Gu,  
Daejeon 305-350, Korea  
e-mail: theidea@lguplus.co.kr

K.-S. Kong

Department of Multimedia, Namseoul University, 21 Maeju-Ri, Seonghwan-Eup,  
Seobuk-Gu, Cheonan, Chungnam 331-707, Korea  
e-mail: kskong@nsu.ac.kr

J.-M. Gil (✉)

School of Information Technology Engineering, Catholic University of Daegu, 13-13  
Hayang-Ro, Hayang-Eup, Gyeongsan, Gyeongbuk 712-702, Korea  
e-mail: jmgil@cu.ac.kr

## 1 Introduction

According to the recent trend of ubiquitous computing, we can access information and services anywhere and at any time via any device. A home network is used for communication between the digital devices typically deployed in the home, usually a small number of personal computers (PCs) and consumer electronics such as a vacuum robot or an IP camera. An important function of the home network is the sharing of Internet access, that is, often a broadband service through a fiber-to-the-home (FTTH), cable TV, Digital Subscriber Line (DSL) or mobile broadband Internet service provider (ISP). If the ISP only provides one IP address, a router including a Network Address Translator (NAT) allows several computers and consumer electronics to share the IP address. Today, the deployment of a dedicated hardware router including a wireless access point (AP) is common, often providing Wi-Fi access. Recently, there are many products in the monitoring system [1, 2]. However, most of them only considered immovable cameras such as an IP camera or a web camera which is connected to the wired Internet. In this paper, in order to overcome such the limitation, we design and implement the home monitoring system exploiting a movable camera of a vacuum robot, which is the first work that has never been reported for the home monitoring system. In our home monitoring system, for easy configuration of a vacuum robot, the Wi-Fi Protected Setup (WPS) [3] is adopted, and the GENERATE method, which is an extension to the Session Initiation Protocol (SIP) [4], is proposed. Also, for signaling and P2P-based NAT traversal, the SIP and the UDP hole punching [5] are mainly used, respectively. In order to verify the capacity of the proposed system, the performance test is conducted by the authoritative call generator.

## 2 P2P-Based Home Monitoring Service

### 2.1 Overall Architecture

We describe the overall architecture for P2P-based home monitoring service, as depicted in Fig. 1. In the home network, one or more home routers, working as NATs, such as an AP or a home gateway, are connected to the main Internet.

The home device (HD) such as a vacuum robot is connected to the AP through Wi-Fi. Therefore, a vacuum robot is always located behind one or more home NATs such as an AP or a home gateway. The user terminal (UT) such as a smart phone or a PC is connected to the Internet via wired or wireless networks. Smart phones connected to mobile networks such as 3G or 4G are often located behind the NAT since mobile network operators have widely deployed the NATs to cope with IPv4 address pool exhaustion [5]. The home monitoring server (HMS) and the media relay server (MRS) are connected to the Internet, respectively.

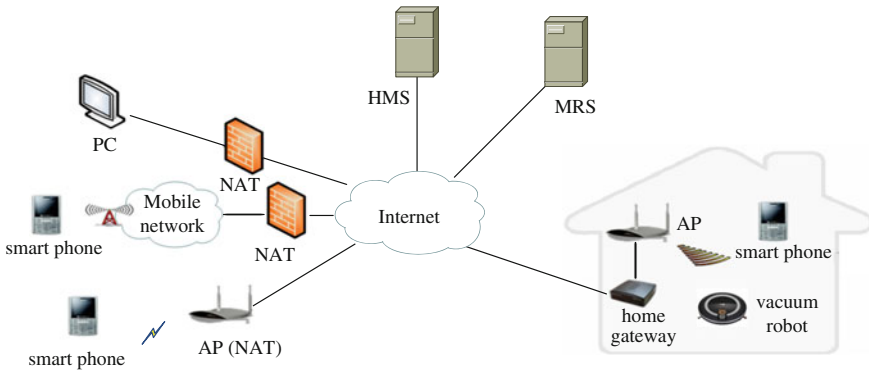


Fig. 1 System architecture for P2P-based home monitoring service

### 2.2 Components of the Proposed Architecture

In Fig. 2, the architecture for P2P-based home monitoring service consists of four main components: a HMS which controls session establishment based on SIP, a MRS which supports the media relay in the case that a HD or a UT is behind a symmetric NAT, a HD agent which is the embedded software in a HD such as a vacuum robot, and a UT agent which is the embedded software in a UT such as a smart phone.

The HMS contains the functionalities of SIP, UDP hole punching, and NAT type check. Firstly, the HMS acts as a SIP registrar, a back-to-back user agent (B2BUA), and a location server which provide user registration, user authentication, session routing, user location, and information management for HDs and UTs. Secondly, the HMS acts as a rendezvous server which helps two clients to set up direct P2P UDP session using UDP hole punching. Finally, the HMS has the NAT type check server which differentiates between cone NATs and symmetric NATs. According to

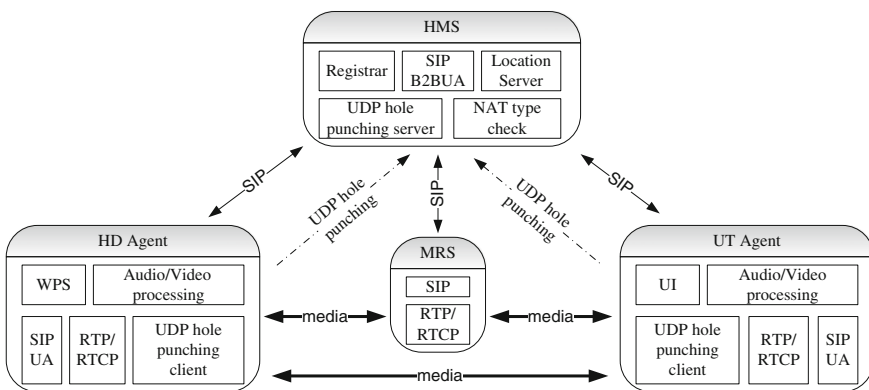


Fig. 2 Components for P2P-based home monitoring service

the Simple Traversal of UDP over NATs (STUN) protocol, NAT implementation is classified as a full cone NAT, an address restricted cone NAT, a port restricted cone NAT, or a symmetric NAT [6]. The UDP hole punching does not work with symmetric NAT even if this mechanism enables two clients to set up a direct P2P UDP session. In order to overcome this drawback, the MRS is activated to relay all the traffics between the peers in the case of symmetric NAT. The HD agent consists of WPS pairing, SIP user agent, UDP hole punching client, audio/video processing, and RTP/RTCP stack. The HD agent supports PBC-based WPS pairing which is familiar to most consumers to configure a network and enable security. It also contains SIP UA functionality for SIP session control, and performs UDP hole punching. Considering the low bandwidth of mobile networks, Speex and motion JPEG are used for audio codec and video codec, respectively. The UT agent consists of SIP user agent, UDP hole punching client, audio/video processing, RTP/RTCP stack, and user interface control for the user.

### 2.3 Service Flow

In this subsection, we describe the signaling flows of the home monitoring service. First of all, a user should configure a HD such as a vacuum robot to use home monitoring service. Figure 3 shows the message flow for the configuration of a HD. The flow consists of the WPS pairing step and the SIP Uniform Resource Identifier (SIP URI) generation step.

We assume that a user notifies his or her HD of the unique identify (e.g., MAC address) to the HMS through the on/off line subscription. In order to enable for a HD to access Wi-Fi networks easily and securely, the PBC-based WPS, as already mentioned, is used.

The AP and the HD have a physical or software-based button. During the setup period (e.g., 2 min) which follows the push of the AP's button, the HD can join the network by pushing its button if it is in range (**Step 1**). Due to the individual privacy issue, it is not desirable that the MAC address of the HD is exposed. In order to more easily provide a SIP URI for the HD, we propose the GENERATE method which is an extension to the SIP that allows for the generation of a SIP URI.

Figure 4 shows the example of the GENERATE method. In the GENERATE method, the hashing value of the MAC address of the HD is inserted using 'key' parameter, newly defined, into the 'from' header field. The user name values in the 'from' and 'to' fields are assigned to 'generate' since the HD has no SIP user name until it receives 200 OK. The HMS responds 200 OK including the SIP URI for the HD using 'val' parameter, newly defined, in the 'from' header field (**Step 2**).

The HD should maintain a UDP session with the HMS to receive a home monitoring request. Figure 5 shows the message flow for the establishment of a UDP session with the HMS. For the SIP registration, the HD agent sends REGISTER message to the SIP registrar of the HMS (S:5060). The AP (NAT) receives the REGISTER message and allocates a temporary IP address and a port for



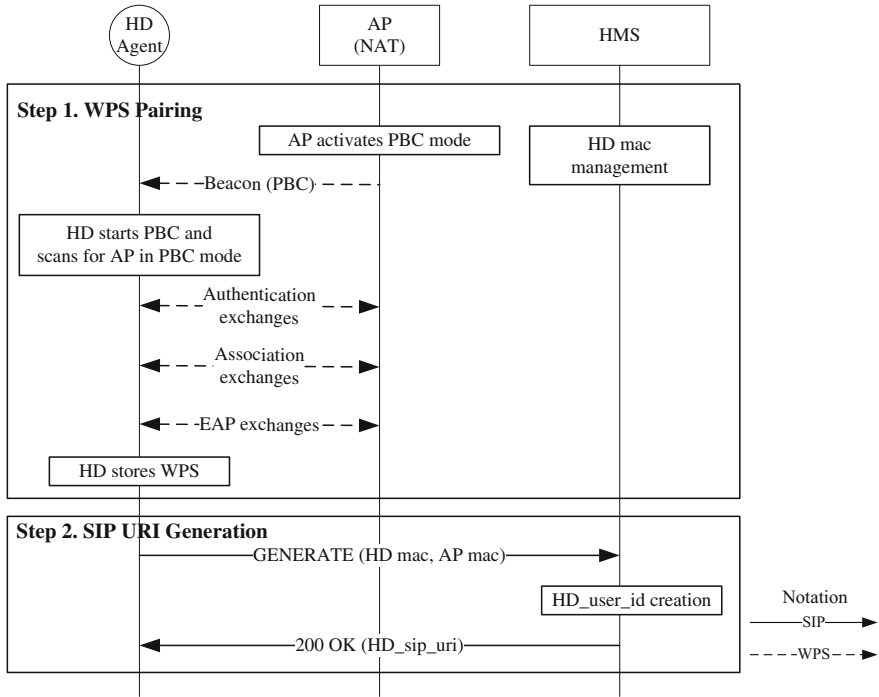


Fig. 3 Message flow for configuration of a HD

Fig. 4 The example of the GENERATE method

```

GENERATE sip:domain.com SIP/2.0
From:sip:generate@domain.com;tag=aaa50ab3d;key=3
8e4fcbe049be2fc4aba6537e78664f2
To: sip:generate@domain.com
...

SIP/2.0 200 OK
From:sip:generate@domain.com;tag=aaa50ab3d;val=h
omedevice1@domain.com
To: sip:generate@domain.com;tag=s694272
...
    
```

outgoing connection. Then, it translates the source IP address and the port for the message from A:a to A':a', and sends the message to the HMS. The HMS compares the IP address and the port in 'via' filed of the REGISTER message with the IP address and the port received from the network and the transport layers. If they are different, the HMS sends 200 OK message with received = A' and rport = a' parameters in the topmost 'via' header field to the UA (Step 3). After receiving 200 OK, the HD agent sends the UDP message, defined as CHECKNAT, to the NAT type check server of the HMS (S:5061). The HMS can obtain the port, a'', received from the transport layer. The HMS can know if the HD is behind the symmetric NAT by comparing a' with a''. This NAT type status is used by the home

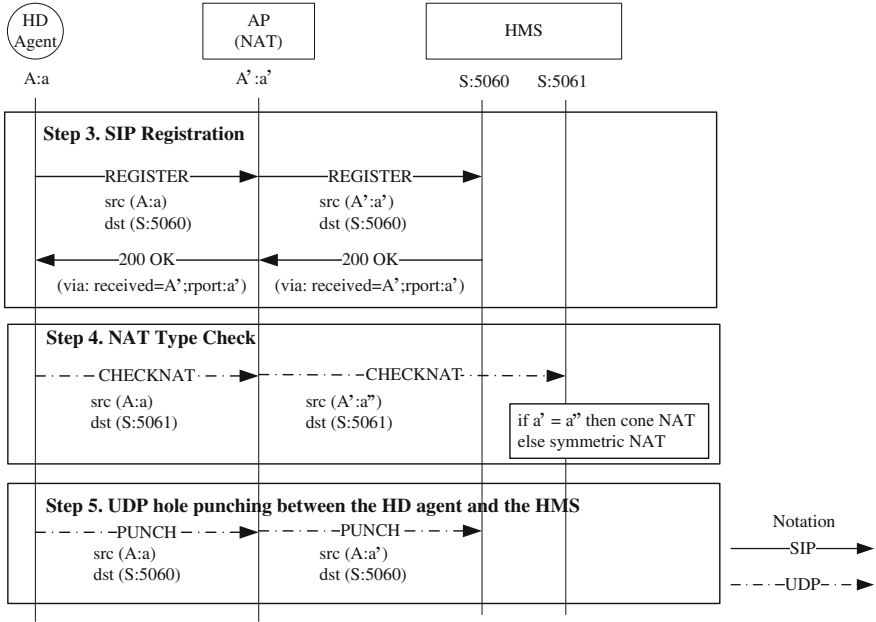


Fig. 5 Message flow for establishment of a UDP session between the HD agent and the HMS

monitoring flow, mentioning later (Step 4). The HD agent uses the UDP hole punching mechanism for NAT traversal, and usually transmits UDP message, defined as PUNCH, to the HMS periodically within the hole punching timer value of the NAT (Step 5). Similarly, the UT agent can establish a UDP session with the HMS.

Suppose the UT (e.g., smart phone) wants to establish a UDP session directly with the HD (e.g., vacuum robot). The procedure for the home monitoring service is as follows:

- The UT agent does not know how to reach the HD agent, so the UT agent sends the INVITE message, including its private IP address and port, to the HMS.
- The HMS checks if both the UT and the HD are not behind symmetric NATs.
- If both are not behind symmetric NATs, the procedure shown in Fig. 6a is performed. Otherwise, the procedure in Fig. 6b is performed.

Figure 6a shows the message flow for the home monitoring service in the P2P-based communication case. After receiving the INVITE from the UT agent, the HMS sends the INVITE message, including the UT's private and public IP addresses and ports, to the HD agent. The HD agent responds 200 OK, including its private IP address and port, to the HMS. Similarly, the HMS sends the response message, which is including the HD's private and public IP addresses and ports, to the UT agent (Step 6). The HD agent and the UT agent start sending the UDP

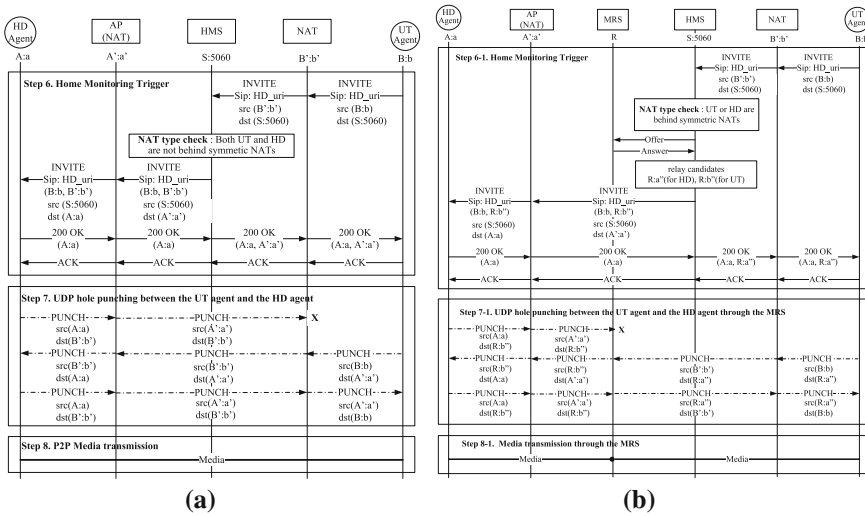


Fig. 6 Message flow for home monitoring service. a P2P case. b Relay case

messages, PUNCH, to establish a P2P session (Step 7). Once the session is established, each agent can send not only video media but also control data (Step 8).

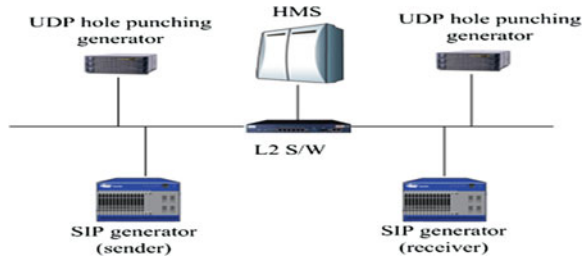
Figure 6b shows the message flow for the home monitoring service in the relay-based communication case. The NAT traversal with the UDP hole punching does not work with symmetric NAT. To overcome this drawback, the HMS uses the MRS in the case of symmetric NAT. The HMS obtains the relay pairs (e.g., R:a'' and R:b'') for the HD and the UT from the MRS. The HMS works in the same way as the Step 6, except replacing each agent's public IP address and port with relaying IP address and port (Step 6-1). The HD agent and the UT agent start sending the UDP messages, PUNCH, to establish a session through the MRS (Step 7-1). Once the session is established, each agent can send not only video media but also control data through the MRS (Step 8-1).

Suppose that a user controls the vacuum robot using his or her smart phone in home. The UT and the HD are behind the common NAT. Since 78 % of the commercial NATs do not support hairpin translation, each agent starts to establish a UDP session directly using opponent's private IP address and port gained from Step 6 or Step 6-1 [5]. Each agent can send not only video media but also control data directly.

### 3 Performance Test

In order to verify the capacity and the stability of the proposed system, the performance test is essential. In this section, we focus on conducting the performance test of the HMS to find out tolerable load, aspects of SIP signaling traffic in the

**Fig. 7** Experimental test environment



proposed system. Figure 7 shows the experimental test environment. We use a Spirent's Abacus-5000 [7], which is one of the most popular SIP traffic generators, and the UDP hole punching generators, which can send simple UDP packets, respectively.

For performance evaluation, test conditions are listed in Table 1. The number of SIP subscribers, which consist of UTs and HDs, are set to 16,000. The SIP registration interval is set to 3,600 s, which is the default expiration for SIP registrations [8]. The UDP hole punching interval is set to 20 s [5]. Call holding time is set to 180 s, which is commonly used in the telecommunication simulation [9].

The performance test of the HMS proceeds as follows:

- The SIP subscribers, which represent UTs and HDs, register with the HMS by sending SIP REGISTER method every 3,600 s.
- Currently, they send the UDP packet, PUNCH, to the HMS every 20 s.
- The 30 Call attempts Per Second (CPS) are generated by senders which represent UTs. These calls are accepted by receivers which represent HDs.

We evaluated the performance of the HMS with increasing Call attempts Per Second (CPS) values by 5, starting at 10 CPS. The measurements and statistics reported by Abacus-5000 in Fig. 8 have shown that the performance of the HMS is 54.40 CPS with tolerable delay values in the test conditions. The test elapsed time is 2 h 12 min 21 s. The total call completion ratio and the total registration success ratio are 99.99 % (431,918 completions of 431,930 call attempts) and 100 % (210,644 successes of 210,644 registration attempts), respectively. This means that the HMS outperforms 95 % which is the call completion rate criteria for Internet telephony in Korea [10]. In this test, the performance factor is only SIP signaling traffic since media packets are transmitted by the P2P way, as mentioned in

**Table 1** Test conditions

Item	Value	Remarks
SIP subscriber	16,000	UTs (8,000), HDs (8,000)
Registration interval	3,600	Second
Hole punching interval	20	Second
Call holding time	180	Second

**Measurements Summary**

Delay Type	Count	Minimum	Average	Maximum
Call length terminate (s)	215959	179.919	180.118	180.395
Call length originate (s)	215959	180.014	180.126	180.407
Response Time (msec)	215959	2	9	227
Call Setup (msec)	215959	64	80	387
Tear Down (msec)	215959	2	9	242
Post dial delay (msec)	215959	77	108	458
Reg 4XX response time	210644	2	10	243
Reg 200 response time	210644	2	9	222
Reg success time	210644	6	19	354

**Test Status**

Test status: DONE  
 Start time: 07:59:12 PM 08/01/2011  
 End time: 10:20:38 PM 08/01/2011  
 Elapsed time: 02:12:21

**Statistics**

SIP Subscriber	Script Attempts	Script Compl.	% Script Compl.	Call Attempts	Call Compl.	% Call Compl.	Call Attempts per Sec	Reg. Successes	Reg. Failures	Reg. Retry Attempts	Reg. In Progress	Errors
Total	431918	431918	100.00	431930	431918	99.99	54.40	210644	0	210644	0	12
Orig	215959	215959	100.00	215971	215959	99.99	27.20	106644	0	106644	0	12
Term	215959	215959	100.00	215959	215959	100.00	27.20	104000	0	104000	0	0

Fig. 8 The performance test result for the HMS reported by Abacus-5000

Fig. 6a. From the performance test of the HMS shown above, we can find that the HMS can support more than 50 CPS signaling load stably in given conditions.

### 4 Conclusion

In this paper, we proposed the P2P-based home monitoring service exploiting a vacuum robot with an IP camera, without requiring a number of cameras for the whole monitoring at home, which is the first trial in the literature. In order to provide the easy configuration of a vacuum robot, the session management, and the NAT traversal, we used the WPS, the SIP, and the UDP hole punching, respectively. Also, in order to configure a vacuum robot easily, we proposed the GENERATE method which is an extension to the SIP that allows for the generation of a SIP URI. In addition, we implemented the HMS, the MRS, the HD agent, and the UT agent to validate the effectiveness and feasibility of the proposed system. The performance test has shown that the HMS supports 99.99 % call completion ratio. These results indicate that the HMS outperforms 95 % which is the call completion rate criteria for Internet telephony in Korea.

**Acknowledgments** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A4A01015777).

### References

1. Zhuang H, Wang Z (2006) IP-based real time video monitoring system with controllable platform. In: Proceedings of 2nd IEEE/ASME international conference on mechatronic and embedded systems and applications, pp 1–4
2. Lam K, Chiu C (2003) Mobile video stream monitoring system. In: Proceedings of 11th ACM international conference on multimedia, pp 96–97

3. Alliance W (2007) Wi-Fi protected setup specification. WiFi Alliance Document
4. Rosenberg J, Schulzrinne H, Camarillo G, Johnston A, Peterson J, Sparks R, Handley M, Schooler E (2002) SIP: Session Initiation Protocol. IETF RFC 3261
5. Ford B, Srisuresh P (2005) Peer-to-peer communication across network address translators. In: Proceedings of USENIX annual technical conference, pp 139–140
6. Rosenberg J, Weinberger J, Huitema C, Mahy R (2003): STUN—Simple traversal of user datagram protocol (UDP) through network address translators (NATs). IETF RFC 3489
7. Ji L, Yin X, Wang X (2007) Conversational model based VoIP traffic generation. In: Proceedings of 3rd international conference on networking and services, pp 14–19
8. Rosenberg J (2004) A session initiation protocol (SIP) event package for registrations. IETF RFC 3680
9. Carothers C, Fujimoto R, Lin Y, England P (1994) Distributed simulation of large-scale PCS networks. In: Proceedings of 2nd international workshop on modeling, analysis, and simulation of computer and telecommunication systems, pp 2–6
10. Telecommunications Technology Associations (2005) TTAS.KO-01.0077. Voice quality criteria of internet telephony

# Design and Simulation of Access Router Discovery Process in Mobile Environments

DaeWon Lee, James J. Park and Joon-Min Gil

**Abstract** With the development of mobile communications and Internet technologies, smart phones have become a necessity of life. To have better connection status, power consumption, and faster transmission speed, the most of mobile users want to access 802.11 wireless networks that are well known as a Wi-Fi. When entering a new area, a mobile host (MH) decides to use one of access routers (ARs) on available networks. However, since previous works are focused on layer 2 handoff for faster connection at 802.11, only the subsystem identification (SSID) and signal strength are considered to choose its new connection. This can fail to provide the MH with a suitable AR. Therefore, more information needs to be used to determine the suitable AR and seamless connectivity. In this paper, we extend a prefix information option in the router advertisement message to include the status information of an AR, such as status of the MH, capacity of the router, current load of the router, and depth of the network hierarchy. Also, we propose a decision engine by which the MH can analyze the status information of ARs and determine a suitable AR automatically based on the information. By analyzing our simulation results, we found that our AR discovery process has several advantages. For the MH, the packet loss can be reduced with the increase of wireless connection period. Additionally, load balancing was achieved for the AR and router, and the network topology was also able to become more efficient.

---

D. Lee

Division of General Education, SeoKyeong University, Jeongneung 4-dong, Sungbuk-gu, Seoul 136-704, Korea  
e-mail: daelee@skuniv.ac.kr

J. J. Park

Department of Computer Science and Engineering, SeoulTech, 172 Gongreung 2-dong, Nowon-gu, Seoul 139-743, Korea  
e-mail: parkjonghyuk1@hotmail.com

J.-M. Gil (✉)

School of Information Technology Engineering, Catholic University of Daegu, 13-13 Hayang-ro, Hayang-eup, Gyeongsan-si, Gyeongbuk 712-702, Korea  
e-mail: jmgil@cu.ac.kr

**Keywords** Fast scanning · Seamless connectivity · Neighbor discovery protocol · 802.11 networks

## 1 Introduction

Today, 802.11 wireless networks are the most popular access networks as demand for mobile access continues to increase. To have better connection status, power consumption, and faster transmission speed, most of mobile users want to access the 802.11 wireless networks that are well known as a Wi-Fi. The Mobile Hosts (MHs) access to 802.11 networks in one of two different modes. An MH can form spontaneous networks as a Mobile Router (MR; ad hoc mode) or it can get connected to an Access Router (AR), which is directly connected to a backbone network (infrastructure mode) [1]. When the MH changes its point of attachment, it should quickly discover and attach to a new point of attachment to reconnect to the network. This is known as a handoff. Most previous studies are focused on layer 2 handoffs in 802.11 networks [2–6]. When an MH starts up or enters a new cell, it needs to discover its environment including radio frequencies, neighbor points of attachment, and available services. Generally, the MH may have several available ARs. The user of the MH decides to use one of the ARs, based only on the SubSystem IDentification (SSID) and signal strength of the user's Internet connection [1]. In our previous work [7], we presented AR selection algorithm and compared its performance with that of pure 802.11 scanning algorithm.

In this paper, we focus on utilizing the layer 3 information to determine a suitable AR of available ones. The information includes status, capacity, current load, and depth of network hierarchy. In order to correctly describe the status information of ARs, we extend a prefix information option in the router advertisement message [8]. We also design a Decision Engine (DE) to analyze reachable ARs and determine a suitable AR. In the proposed protocol, the handoff MH selects the suitable AR automatically, and thus the proposed protocol provides load balance of ARs and efficiency of network topology.

This paper is organized as follows. In [Sect. 2](#), we briefly describe related work. [Section 3](#) describes the need for AR status information and presents the extended prefix information option. AR discovery algorithm is given in [Sect. 4](#). [Section 5](#) presents the performance evaluation with simulation results. Finally, [Sect. 6](#) concludes the paper.

## 2 Related Work

Most related studies on the 802.11 discovery process have focused on reducing the scanning latency during a layer 2 handover, when an MH roams from one AP to another. One simple method to reduce the full scanning latency is to use selective



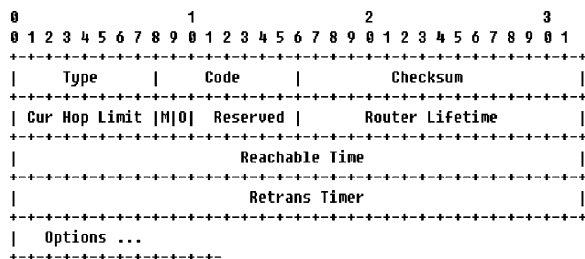
scanning, which allows the scanning of a subnet of channels instead of probing each of them. The other methods have focused on reducing the value of the scanning timers (MinCT and MaxCT) [2, 3] by fixing the potential best values for both timers, presenting theoretical considerations and simulation results. The smooth handover [4] and the periodic scanning [5] are based on splitting the discovery phase into multiple sub phases. The objective of this division is to allow an MH to alternate between data packet exchange and the scanning process. An MH builds a list of target APs, maintaining information only on channel and SSID. These methods focus only on reducing latency to minimize the disconnected time of the MH. However, they do not ensure that the MH connects to a suitable AP. Instead, more information needs to be used to connect to the suitable AR. This information cannot be provided by a probe signal at layer 2. To collect this information, we instead focus on a neighbor discovery protocol at layer 3. The router advertisement message is one message format from the neighbor discovery protocol in IPv6 [8].

The neighbor discovery protocol corresponds to a combination of the IPv4 protocol ARP, ICMP, RDISC, and ICMPv4. The router advertisement messages contain prefixes that are used to determine whether another address shares the same link and/or address configuration, a suggested hop limit value, and so on. To collect the information to find a suitable AP, we focus on the router advertisement message [8], as shown in Fig. 1. However, the router advertisement message has not enough reserved fields to provide more information. Thus, we focus on the prefix information that must be used in the hierarchy architecture in IPv6. Figure 2 shows the prefix information option [8].

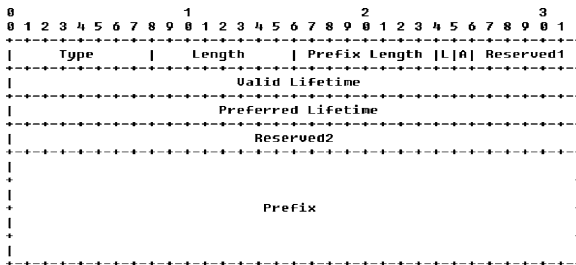
### 3 Extended Prefix Information Option

To access a new AR in an L2 handoff, the AR provides only SSID and authentication. However, the information provided by L2 is insufficient to determine a suitable AR. In this paper, we focus on an L3 handoff for AR selection. There are four information elements needed to determine the suitable AR. First, the status of the AR is necessary. Second, the maximum capacity of the AR is necessary. Third, the expected AR load is necessary; because several MHs use the same AR in a cell,

**Fig. 1** Router advertisement message format



**Fig. 2** Prefix information option format



the network bandwidth is limited. The MH should find a free AR. The last is the depth of the network hierarchy, which represents the logical location with respect to a border gateway in a subnet. To prevent frequent handoffs and provide fast transmission in a subnet, the MH should connect with a higher-level AP in a subnet hierarchy. The additional attributes in the extended router advertisement message are as follows: (1) Status: status of AR (stationary/portable), (2) Cap: maximum capacity of AR, (3) Load: current load of AR, (4) Depth: depth of network hierarchy.

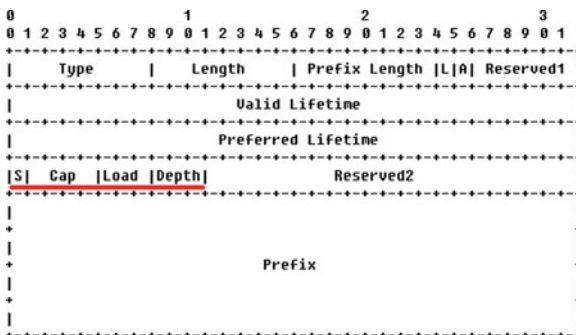
Figure 3 shows a format of the extended prefix information option. However, we use the minimum bits in the reserved field because the signaling overhead on wired/wireless links is an important issue.

### 4 AR Discovery Algorithm

In this section, we propose a decision engine (DE) that analyzes the router advertisement messages and determines a suitable AR. Table 1 shows our AR discovery algorithm with the DE.

The proposed AR discovery algorithm consists of three parts. First, information is received from the router advertisement message. Second, it is used for decision-making, which is divided into two parts: an active state for an MH that moves frequently and an idle state for an MH that moves rarely. The last part of the

**Fig. 3** Extended prefix information option format



**Table 1** AR discovery algorithm

---

```

1 If Power up or entering new subnet
  1.1 Send router solicitation message to ARs
  1.2 Wait for router advertisement messages of ARs
2 For all elements in each router advertisement messages
  2.1 Compare status
  2.2 Compare bandwidth
  2.3 Compare signaling strength
  2.4 Compare hierarchy
  2.5 Compare loadratio
  2.6 Decide candidate AR_list
/* the priority of active MH: status > bandwidth > signaling
  strength > hierarchy > loadratio */
/* the priority of idle MH : status > hierarchy > signaling
  strength > loadratio > bandwidth */
3 If Connection is needed
  3.1 Scanning candidate ARs
  3.2 Verify AR states
  3.3 Select state best AR
  3.4 Send binding update to new AR

```

---

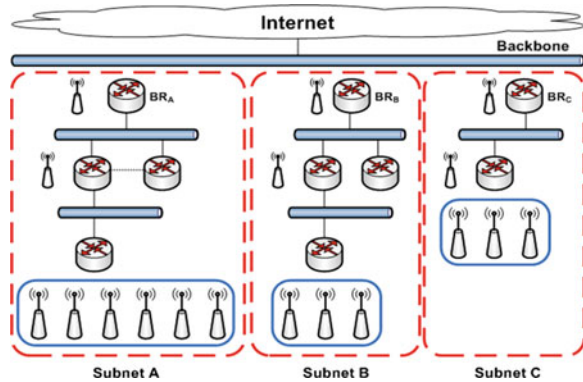
algorithm covers what happens when the MH loses its connection. The MH broadcasts a neighbor-solicitation message and then connects to a new AR, following decision-making.

## 5 Performance Evaluation

### 5.1 Simulation Environment

We tested the performance of the proposed discovery process with AR discovery algorithm and compared it with that of 802.11 discovery process. To this end, we developed a simulator in JAVA and incorporated the DE into it. Figure 4 shows the simulation environment used to evaluate the proposed discovery process. Our simulations are conducted with three subnets, each of which has a border router (BR) to connect to the Internet. Subnet A is composed of three hierarchies. BR<sub>A</sub> consists of an AR and six routers. Each router on subnet A has six ARs. Also, subnet B is composed of three hierarchies. BR<sub>B</sub> consists of an AR and two routers. Each router on subnet B has three ARs. Subnet C is composed of two hierarchies. BR<sub>C</sub> consists of an AR and a router. A router on subnet C has three ARs. Two kinds of mobile devices are generated: one is an MR that changes its point of attachment by random movement and the other is an MH. And, 70 MRs and 150 MHs were generated. Both of them were randomly located in the initial state and each of them had random mobility.

**Fig. 4** Simulation environment

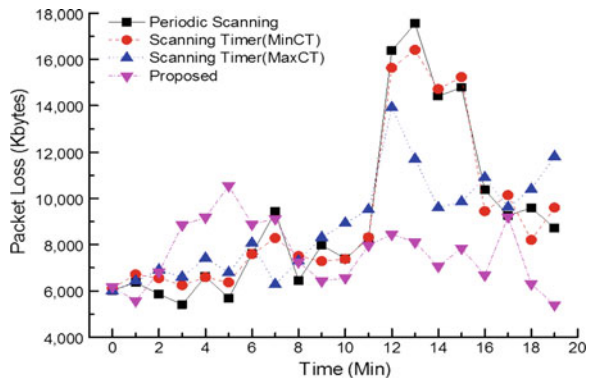


### 5.2 Simulation Results and Analysis

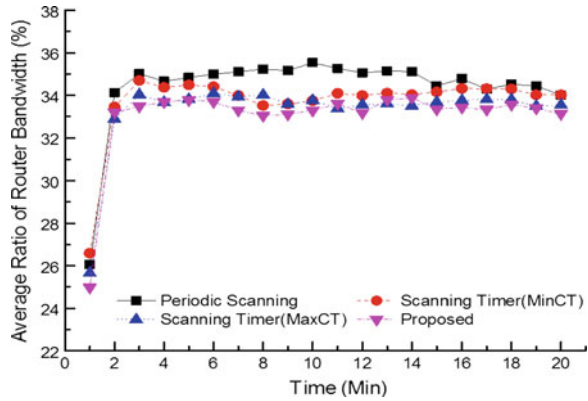
In our first experiment, we compared the packet loss of the proposed discovery process with that of previous works (periodic scanning, scanning timer with MinCT, and scanning timer with MaxCT). We measured the packet loss of each method in 20 min. Figure 5 shows the comparison of the packet loss. The results of this figure show that the proposed discovery process has an improvement of 20 (periodic scanning), 20 (MinCT), and 15 % (MaxCT) in the reduction of the packet loss.

Figure 6 shows the average ratio of router bandwidth in 20 min. The average ratio of router bandwidth means the average utilization of all routers in a domain. The proposed discovery process outperformed the other discovery processes and evenly distributed except in the initial state. The average ratio of the proposed discovery process is 32.9 %. The periodic scanning, MinCT, and MaxCT have the average ratio of 34.3, 33.7 and 33.2 %, respectively. The proposed discovery process showed an improvement of 1.0–1.5 %, as compared to the other three methods. From the results of Fig. 6, we can see that the proposed discovery process can achieve the load balancing of ARs and topological stability.

**Fig. 5** Comparison of packet loss



**Fig. 6** Comparison of average ratio of router bandwidth



## 6 Conclusion

In this paper, we addressed the decision making process for determining a suitable AR when an MH enters a new cell. We proposed a DE to find the suitable AR on the 802.11 based MANET. An MH may have several available networks when powered up or moving into a new area. The user of the MH decides to use one of the ARs on an AR list. However, a decision based only on the AR’s name and signal strength for the user’s Internet connection has a limitation to provide the suitable AR for the MH. To determine the suitable AR, we focused on the neighbor discovery protocol in L3. We extended the prefix information option in the neighbor discovery protocol to include the AR’s status information, such as status, capacity, current load, and depth of network hierarchy. The simulation results showed that the proposed AR discovery process has the following advantages: for the MH, the packet loss can be reduced with the increase of wireless connection period. Additionally, load balancing was achieved for the AR and router, and the network topology was also able to become more efficient.

**Acknowledgments** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A4A01015777).

## References

1. RFC 5416: Control and provisioning of wireless access points (CAPWAP) protocol binding for IEEE 802.11. <http://www.rfc-editor.org/rfc/rfc5416.txt>
2. Shin S, Forte AS, Rawat AS, Schulzrinne H (2004) Reducing MAC layer handoff latency in IEEE 802.11 wireless LANs. In: Proceedings of 2nd international workshop on mobile management and wireless access protocols, pp 19–26
3. Velayos H, Karlsson G (2004) Techniques to reduce the IEEE 802.11b handoff time. In: Proceedings of 2004 IEEE international conference on communications, pp 3844–3846

4. Liao Y, Gao L (2006) Practical schemes for smooth MAC layer handoff in 802.11 wireless networks. In: Proceedings of 2006 international symposium on world of wireless, mobile and multimedia Networks, pp 181–199
5. Montavont J, Montavont N, Noel T (2005) Enhanced schemes for L2 handover in IEEE 802.11 networks and their evaluations. In: Proceedings of IEEE 16th international symposium on personal, indoor and mobile radio communications, pp 1429–1434
6. Koutsopoulos I, Tassiulas L (2007) Joint optimal access point selection and channel assignment in wireless networks. *IEEE/ACM Trans Network* 15(3):521–532
7. Lee D, Kim Y, Lee H (2012) A study of an optimal discovery process in mobile ad hoc network. In: Proceeding of The 2012 FTRA international conference on advanced IT, engineering, and management
8. RFC 4861: Neighbor discovery for IP version 6 (IPv6). <http://www.rfc-editor.org/rfc/rfc4861.txt>

# Integrated SDN and Non-SDN Network Management Approaches for Future Internet Environment

Dongkyun Kim, Joon-Min Gil, Gicheol Wang and Seung-Hae Kim

**Abstract** For years, computer scientists have been dreaming of innovating Internet in terms of performance, reliability, energy efficiency, security, and so on. However, it is nearly impossible to carry out practical large-scale experiments and verification, since new software and programs are hardly evaluated on the current Internet environment where routers and switches are totally closed. In this context, Software Defined Networking (SDN) concept has been introduced to deploy software-oriented and open programmable network coupled tightly with traditional Internet (non-SDN) environment. This tight integration results in easy and efficient deployment of SDN domains into non-SDN infrastructure, but it is also required that integrated management methods for SDN and non-SDN be developed. This paper proposes how SDN and non-SDN can be managed in an integrated manner in terms of two aspects: (1) non-SDN driven management approach and (2) SDN oriented management approach, in order to achieve reliable Future Internet environment.

**Keywords** SDN · Non-SDN · Network management · Future internet

---

D. Kim · G. Wang · S.-H. Kim  
Korea Institute of Science and Technology Information, 52-11 Eoeun-dong,  
Yuseong-gu, Daejeon, South Korea  
e-mail: mirr@kisti.re.kr

G. Wang  
e-mail: gcwang@kisti.re.kr

S.-H. Kim  
e-mail: shkim@kisti.re.kr

J.-M. Gil (✉)  
School of IT Engineering, Catholic University of Daegu, 13-13 Hayang-ro, Hayang-eup,  
Gyeongsan-si, Gyeongbuk, South Korea  
e-mail: jmgil@cu.ac.kr

## 1 Introduction

There have been drastic demands over the past years in the Internet, compared to the earlier Internet designs [1]. One of the challenges is the requirement of open networking that makes it possible for users to innovate Internet by experimenting and applying newly developed technologies over practical large-scale networks. Internet innovation known as Future Internet, currently, is deeply involved with network performance, reliability, energy efficiency, security, etc., while Internet is entirely closed not to allow any new user-oriented software to be installed and tested on. Only can network vendors access and control the network devices on Internet regarding programmability, reconfiguration, and any innovative efforts.

Therefore, various researches have been performed to cope with the new user demands, and one of them is Software Defined Networking (SDN) with OpenFlow protocols [2] that has invoked significant interest in reconsidering traditional aspects of Internet architecture and design. Two important features of SDN are (1) implementation of network control plane decoupled from data plane (network forwarding hardware), and (2) relocation of control plane from hardware switch equipped with a typical low performance CPU.

Another interesting feature of SDN is that it can easily be integrated into the current Internet environment generally at layer 2 local networks. For example, OpenFlow devices are mostly Ethernet switches to support packet forwarding, which indicates that SDN switches are basically to be incorporated with other non-SDN Ethernet switches for layer-2 communications. Based on SDN's inherent integration with classical local networks (i.e., non-SDN), remote SDN domains communicate with other SDN domains via Internet (i.e., non-SDN) [3]. So, in order to keep the overall SDN and non-SDN environment reliable, it is inevitable that SDN and non-SDN need to interoperate in a tightly coupled way, and should be managed constantly and reliably through well-designed network management model.

In this context, this paper proposes combined network management models for both SDN and non-SDN environments adopting Distributed virtual Network Operations Center (DvNOC) [4] and several other related works, which will be introduced more in detail in [Sect. 2](#). The proposed models are induced based on two approaches: (1) non-SDN adaptation into SDN in an SDN-oriented way and (2) SDN adjustment into non-SDN in a non-SDN-oriented way. The main difference of SDN and non-SDN management is that SDN management is interactive while non-SDN management is not two-way. SDN controller(s) and devices communicate with each other by exchanging (and analyzing) a variety of management information (e.g., status, topology), which results in "super-active" network management. Non-SDN management is generally not interactive though: there is no centralized controller, but one or more management servers which contact network devices and gather management information from them, whereas network devices cannot do the same jobs. Therefore, the suggested management models in



this paper principally deal with how contrastive management methods can be integrated as the combined and efficient network management model.

The remainder of this paper is organized as follows. In Sect. 2, we introduce related works for the proposed network management approaches. Section 3 describes DvNOC architecture as a non-SDN management environment. The SDN and non-SDN integrated management models based on Future Internet are explained in Sect. 4. Finally, Sect. 5 concludes the paper.

## 2 Related Work

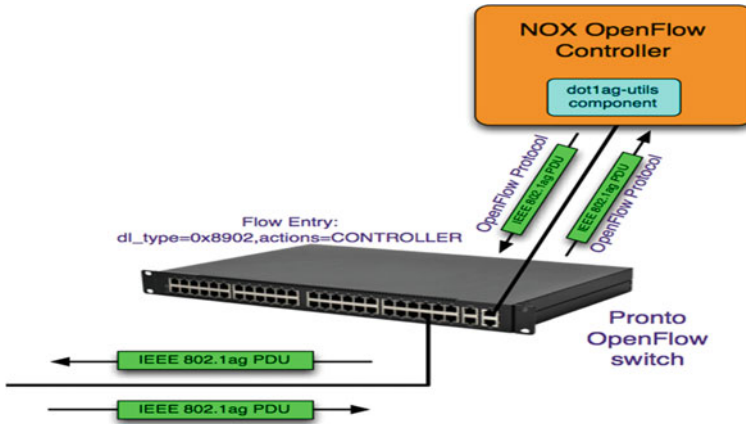
There are several open source SDN controllers that have been developed and being improved so far (in Table 1). Many experimental SDN applications exploit the open source controllers in the field of network virtualization, energy efficiency, mobility, traffic engineering, wireless mobile video streaming, and so on. Each of the controllers can be used for the proposed network management model as well in terms of Disaster Manager (DM) implementation and East/Westbound Interface/API designs with non-SDN manager.

Slice Around the World (SATW) initiative [5] is influenced by various Future Internet testbed activities in Europe, Asia, and elsewhere outside USA, and it consists of three components such as application/service, programmable SDN infrastructure, and international experimental facility. K-GENI [6] has participated in SATW initiative since 2011. Both K-GENI and SATW initiative are playgrounds on which SDN and non-SDN integrated network management model can be applicable.

IEEE 802.1ag [7] is a standard specifying Ethernet OAM for connectivity fault management (CFM) of paths through 802.1 bridges and local area networks. OpenFlow/SDN-based Ethernet OAM [8] adopts IEEE 802.1ag standard to exchange 802.1ag Ethernet frame based on NOX in Table 1, and to manage Open

**Table 1** Open source SDN controllers in public domain

Name	Language	Platform(s)	License	Developer(s)	Notes
OpenFlow Ref. [11]	C	Linux	OpenFlow License	Stanford U., Nicira	Extensibility necessary
NOX [12]	Python, C++	Linux	GPL	Nicira	Actively developed
Beacon [13]	Java	Win, Mac, Linux, Android	GPL(core), FOSS L.	Stanford U.	Web UI framework
Maestro [14]	Java	Win, Mac, Linux	LGPL	Rice U.	–
Trema [15]	Ruby, C	Linux	GPL	NEC	Emulator included
RouteFlow [16]	Python, C++	Linux	Apache L.	CPqD	Virtual IP routing



**Fig. 1** IEEE 802.1ag frame monitoring using NOX controller

vSwitch based on CCM (Continuity Check Message). This research is very relevant to the integration of SDN and non-SDN management in the Ethernet layer as shown in Fig. 1, which describes how to operate and manage SDN-based Ethernet environment using IEEE 802.1ag PDU exchanges, interacting with NOX OpenFlow Controller. In turn, this SDN management can be extended to other non-SDN Ethernet devices using the IEEE 802.1ag standard.

Disaster Manager (DM) [9] is a more active management system than IEEE 802.1ag based OAM, devised to store and analyze the disaster-related information of network nodes by embedding a DM firmware into OpenFlow devices. An embedded DM actively acquires more than thousands of monitoring data from SysLog, for example, generated everyday in OpenFlow switches, and the DM analyzes the numerous datasets so that network devices can perform self-directed network disaster isolation and autonomous intelligent network management. The proposed management model incorporates DM as a premier building block to achieve “super-active” network management over combined SDN and non-SDN ecology.

### 3 Distributed Virtual Network Operations Center

DvNOC [4] was designed as a virtual network management framework for hybrid research network (that is a combination of circuit-oriented network and packet-switching network for providing both end-to-end dedicated lightpath and layer-3 routed path [10]). Virtual network resources can be managed by researchers and network engineers through DvNOC framework. Moreover, since recent advanced applications require global end-to-end network environment for collaborative researches over many individual network domains and countries, Network

Operations Center (NOC) to NOC cooperation is getting more and more important between multi-domain networks. In this regard, DvNOC provides the following functionalities to support collaborative efforts on multi-domain NOCs: *Multi-domain Network Awareness, Efficient NOC-to-NOC Cooperation, and User-oriented Virtual Network Management.*

DvNOC incorporates new features (e.g., resource repository) for special end-users (researchers and experimenters) as well as conventional users (network operators and engineers) in addition to the traditional NOC facilities so that the DvNOC manages resource information collected from local networks inside a network domain, ultimately in order to share the resource information with other NOCs on DvNOC domain. Each associated dNOC (distributed NOC) exchanges the operational dataset with others, interfacing with virtual NOC (vNOC).

Since DvNOC supports several specific functionalities that classical Internet management system hardly provides, we consider DvNOC a non-SDN management framework to be combined with SDN. Non-SDN oriented DvNOC architecture basically includes data acquisition, federation engine, and user interface, all of which are interacting with resource repository. Among those capabilities, federation engine comprises data ownership and policy management as well to coordinate multi-domain data exchanges over Autonomous Distributed Networks (ADNs) correctly. Users and applications can acquire the datasets stored in resource repository (by federation engine and data acquisition modules) in a very similar way as northbound interface/API in SDN provides network resource information for applications and users. In order to achieve northbound interface in DvNOC, it is capable of following four OpenAPIs (equivalent with Northbound Interface/APIs in SDN): *getListDataSet()*, *getSpecificDataSet()*, *getWholeDataSet()*, *setDataSet()*.

## **4 Proposed SDN and Non-SDN Integrated Management Approaches Based on DvNOC Framework**

SDN is designed, by nature, to embrace classical non-SDN networks including local and wide area networks. In addition, targeted users of both SDN and non-SDN are anything but dissimilar regarding services, applications, and users, such as end-to-end virtual network services, high performance data transmission, and experimental and research uses, etc. Therefore, it is inferred that integrated management of SDN and non-SDN infrastructure may lead to considerable synergistic effect for diverse future Internet demands.

In this paper, two network management models are proposed for integrated SDN and non-SDN environment based on DvNOC and related works introduced in [Sect. 2](#). Basically, these two models aim to the same direction, but the detailed configurations are revised. Regarding SDN, a controller communicates with network hardware using OpenFlow protocol, while it interacts with applications and

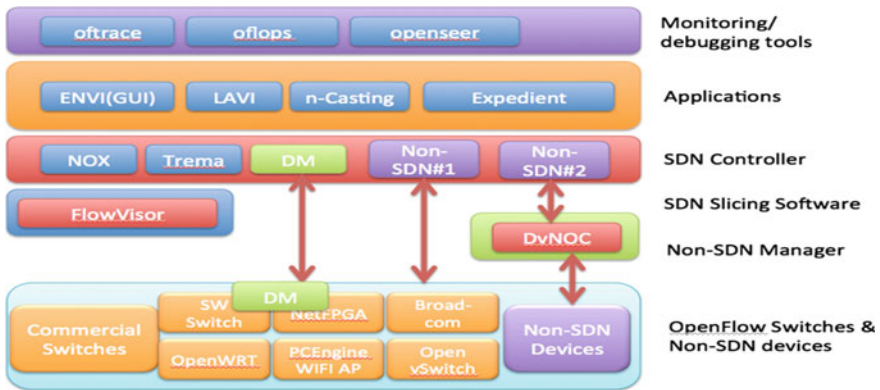


Fig. 2 The first model: SDN-oriented non-SDN management integration

users through northbound interface and APIs. Considering this structure, the simple way of managing combined SDN and non-SDN infrastructure is that an SDN controller acquires resource information from non-SDN hardware. That is, an open source SDN controller like NOX needs to be designed and implemented including some modules to interface with non-SDN devices. The first model is proposed under this consideration using DvNOC as a non-SDN management framework. In this sense, an SDN controller communicates with DvNOC in different layers, SDN in upper layer and DvNOC in lower layer as shown in Fig. 4, where an SDN controller can carry out supervisory management for integrated SDN and non-SDN network environment efficiently.

In Fig. 2, three main functionalities are included in the controller layer. First, DM is equipped with OpenFlow protocol in SDN controllers for more dynamic and interactive network management as well as network disaster isolation. Second, controllers embrace non-SDN communication module (non-SDN#1), adopting SNMP, Netconf, TL1, CLI, etc., which are classical network management protocols on Internet. Third, DvNOC is combined with controllers (non-SDN#2), achieving multi-domain network management based on hybrid research networks in order to meet some application and user demands in Table 2. Nonetheless, there is a trade-off between non-SDN#1 and non-SDN#2. Non-SDN#1 has high development overhead but good performance due to dense coupling, while non-SDN#2 costs less overhead, but it may have relatively low performance.

Figure 3 indicates non-SDN based network management model requiring quite less development and verification overheads compared to the first model in that both controller and DvNOC only need to accommodate standard east/westbound interface and APIs of SDN scheme. The other standard interfaces such as north/southbound interface/APIs in Fig. 4 are also supposed to be equipped with controller in general SDN architecture. Therefore, only does DM need to be additionally implemented for controller and DvNOC. The inadequacy of the second model is that it doesn't conduct as much supervisory management as the first

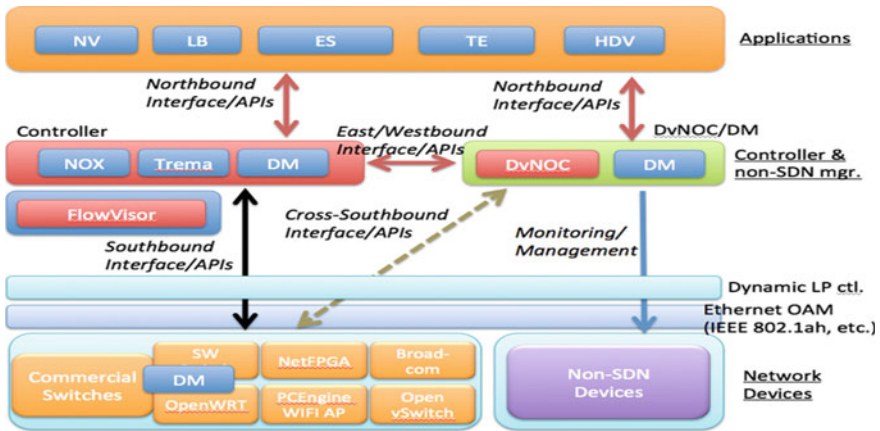


Fig. 3 The second model: non-SDN-oriented SDN management integration

model does, because DvNOC communicates parallel with controller in the same layer using east/westbound interfaces.

In the meanwhile, for the bright side, the DM implemented in DvNOC may interact with SDN hardware when it comes to fault and failure management using cross-southbound interface and APIs. In this way, DvNOC is acting like a second-controller regarding breakdown management, which enhances overall integrated network reliability by providing an alternative to solve single point of failure problem of one centralized controller. In addition, both SDN and non-SDN environment are closely coupled with Ethernet OAM layer functionalities (IEEE 802.1ag, IEEE802.ah, OpenFlow based CCM exchanges) and dynamic circuit control systems. This coupling makes it possible for users to achieve the implementation and deployment of end-to-end virtual networks, high-end application networks, dedicated QoS networks, etc., from SDN to non-SDN and vice versa.

### 5 Conclusion and Future Works

This paper proposed two network management approaches for integrated SDN and non-SDN environment, since there are user, application, and service demands to incorporate two different network infrastructure, mostly for the purpose of advanced researches and experiments. The proposed approaches mainly describe how SDN controllers and DvNOC, a specialized non-SDN network management framework, can interact with each other in order to maintain reliable end-to-end user oriented Future Internet environment. Our future works will evaluate implementation aspects of prototypes based on proposed network management models, considering each model's tradeoff in terms of performance, development overheads, and reliability.

## References

1. Clark D (1988) The design philosophy of the DARPA internet protocols. *ACM SIGCOMM Comput Commun Rev* 18(4):106–114
2. McKeown N, Anderson T, Balakrishnan H, Parulkar G, Peterson L, Rexford J, Shenker S, Turner S (2008) OpenFlow: enabling innovation in campus networks. *ACM SIGCOMM Comput Commun Rev* 38(2):69–74
3. Levin D, Wundsam A, Heller B, Handigol N, Feldmann A (2012) Logically centralized? State distribution trade-offs in software defined networks. In: 1st workshop on hot topics in software defined networks, pp 1–6
4. Kim D (2008) User oriented virtual network management based on dvNOC environment. *Int J Comput Sci Netw Secur* 8(10):59–65
5. SATW Initiative. <http://groups.geni.net/geni/ticket/913>
6. K-GENI Initiative. <http://groups.geni.net/geni/wiki/K-GENI>
7. IEEE 802.1ag Standard. <http://standards.ieee.org/getieee802/download/802.1ag-2007.pdf>
8. Pol R (2012) Ethernet OAM integration in OpenFlow. In: 27th NORDUnet conference
9. Song S, Hong S, Guan X, Choi B-Y, Choi C (2013) NEOD: network embedded on-line disaster management framework for software defined networking. In: IFIP/IEEE international symposium on integrated network management. Accepted for publication
10. Ham J, Dijkstra F, Grosso P, Pol R, Toonk A, Laat C (2008) A distributed topology information system for optical networks based on the semantic Web. *Opt Switch Networking* 5(2–3):85–93
11. OpenFlow Reference. <http://www.openflow.org/wp/tag/reference-implementation/>
12. NOX Controller. <http://www.noxrepo.org/>
13. Beacon Controller. <https://openflow.stanford.edu/display/Beacon/Home>
14. MAESTRO Platform. <http://code.google.com/p/maestro-platform/>
15. Trema Controller. <http://trema.github.com/trema/>
16. RouteFlow Controller. <https://sites.google.com/site/routeflow/>

# Analysis and Design of a Half Hypercube Interconnection Network

Jong-Seok Kim, Mi-Hye Kim and Hyeong-Ok Lee

**Abstract** This paper proposes a new half hypercube interconnection network that has the same number of nodes as a hypercube but reduces the degree by approximately half. To evaluate the effectiveness of the proposed half hypercube, its connectivity, routing, and diameter properties were analyzed. The analysis results demonstrate that the proposed half hypercube is an appropriate interconnection network for implementation in large-scale systems.

**Keywords** Half hypercube · Hypercube variation · Interconnection network

## 1 Introduction

The need for high-performance parallel processing is increasing because modern engineering and science application problems require many computations with real-time processing. A parallel processing system can connect thousands of processors with their own memory, or even more via an interconnection network enabling inter-processor communication by passing messages among processors through the network. An interconnection network can be depicted with an

---

J.-S. Kim  
Department of Computer Science, University of Rochester,  
Rochester 14627, USA  
e-mail: Rockhee7@gmail.com

M.-H. Kim  
Department of Computer Science Education, Catholic University of  
Daegu, Daegu, South Korea  
e-mail: mihyekim@cu.ac.kr

H.-O. Lee (✉)  
Department of Computer Education, Suncheon National University, Suncheon, South Korea  
e-mail: oklee@sunchon.ac.kr

undirected graph. The most common parameters for evaluating the performance of interconnection networks are degree, connectivity, diameter, network cost, and broadcasting [1, 2].

In an interconnection network, degree (relevant to hardware cost) and diameter (relevant to message transmission time) are correlated. In general, the throughput of an interconnection network is improved with a higher degree because the diameter of the network is increased when its degree is increased. However, a parallel computer design increased the hardware costs of an interconnection network, because of the increased number of processor pins. An interconnection network with a lower degree can reduce hardware costs, but its latency and throughput are degraded because the message transmission time is increased. Due to such characteristics of interconnection networks, the network cost (= degree  $\times$  diameter) is a typical parameter used to evaluate interconnection network performance [3].

The hypercube is a typical interconnection network topology and is widely used in both research and commercial fields due to its advantages that can easily provide a communication network structure as required in various application areas. The hypercube is node- and edge-symmetric with a simple routing algorithm, maximum fault tolerance, and simple recursive structure. Additionally, it can be easily embedded in various types of existing interconnection networks [4, 5]. However, it has the drawback of increasing network costs associated with the increased degree when the number of nodes increases. To improve this shortcoming, a number of variations of the hypercube have been proposed, such as multiple reduced hypercube [3], twisted cube [6], folded hypercube [7], connected hypercube network [8], and extended hypercube [9]. This paper proposes a new variation of the hypercube that reduces the hypercube degree by approximately half with the same number of nodes: the Half Hypercube (HH). We denote an  $n$ -dimensional half hypercube as  $HH_n$ . To validate the effectiveness of the proposed HH, performance measurement parameters were analyzed, such as connectivity, routing, and diameter.

This paper is organized as follows. Section 2 presents the definition of the proposed HH and discusses its properties, including connectivity. Section 3 proposes and analyzes a simple routing algorithm and the diameter of the HH. Section 4 summarizes and concludes the paper.

## 2 Definition and Properties of the Proposed Half Hypercube

The hypercube  $Q_n$  ( $n \geq 2$ ) is defined as an  $n$ -dimensional binary cube where the nodes of  $Q_n$  are all binary  $n$ -tuples. Two nodes of  $Q_n$  are adjacent to each other if and only if their corresponding  $n$ -tuples differ in one bit at exactly one position [10].  $Q_n$  is an  $n$ -regular graph with  $2^n$  nodes and its diameter is  $n$ . In this paper,  $\bar{S}$  indicates the complement of the binary string  $S(= s_n s_{n-1} \dots s_1)$ ; that is, it is

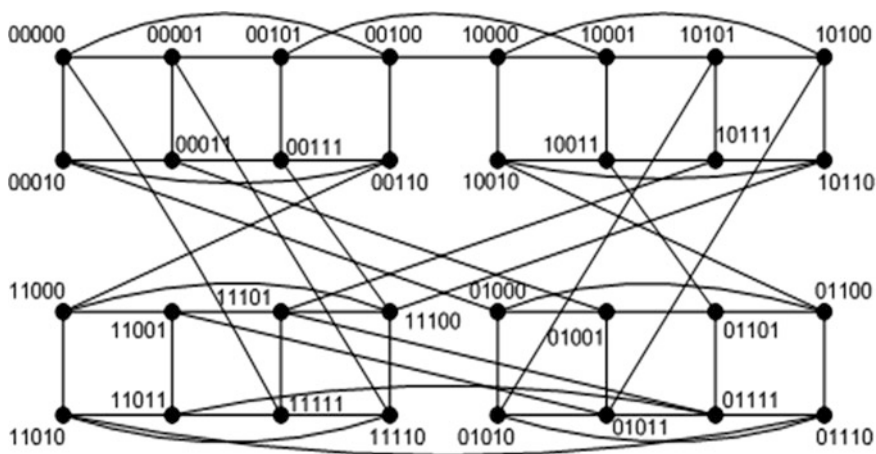


obtained by inverting all the bits in the binary number (inverting 1's for 0's and vice versa).

We denote an  $n$ -dimensional half hypercube as  $HH_n$  and represent its node with  $n$  binary bits. Let the address of node  $S$  in  $HH_n$  be  $s_n s_{n-1} s_{n-2} \dots s_i \dots s_3 s_2 s_1$  ( $n \geq 3$ ). There are two types of edge in  $HH_n$ : the  $h$ -edge, which connects node  $S(= s_n s_{n-1} s_{n-2} \dots s_i \dots s_3 s_2 s_1)$  to a node that has the complement in exactly one  $h$  position of the bit string of node  $S$  ( $1 \leq h \leq \lceil n/2 \rceil$ ), and the  $sw$ -edge (i.e., swap-edge), which connects node  $S(= s_n s_{n-1} s_{n-2} \dots s_h s_{h-1} \dots s_3 s_2 s_1)$  to a node in which the  $\lfloor n/2 \rfloor$  leftmost bits of the bit string of  $S$  and the  $\lfloor n/2 \rfloor$  bits on the right-side of  $S$  starting from the  $\lfloor n/2 \rfloor$  bit are swapped ( $h = \lfloor n/2 \rfloor$ ). For example, when  $n$  is even, node  $S(= s_n s_{n-1} s_{n-2} \dots s_{\lfloor n/2 \rfloor + 1} s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} s_{\lfloor n/2 \rfloor - 2} \dots s_3 s_2 s_1)$  is connected to node  $(s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} s_{\lfloor n/2 \rfloor - 2} \dots s_3 s_2 s_1 s_n s_{n-1} s_{n-2} \dots s_{\lfloor n/2 \rfloor + 1})$  in which  $(s_n s_{n-1} s_{n-2} \dots s_{\lfloor n/2 \rfloor + 1})$  and  $(s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} s_{\lfloor n/2 \rfloor - 2} \dots s_3 s_2 s_1)$  of  $S$  are exchanged. When  $n$  is odd, node  $S$  is connected to node  $(s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} s_{\lfloor n/2 \rfloor - 2} \dots s_3 s_2 s_n s_{n-1} s_{n-2} \dots s_{\lfloor n/2 \rfloor + 1} s_1)$  in which  $(s_n s_{n-1} s_{n-2} \dots s_{\lfloor n/2 \rfloor + 1})$  and  $(s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} s_{\lfloor n/2 \rfloor - 2} \dots s_3 s_2)$  of  $S$  are swapped. However, if two parts of  $\lfloor n/2 \rfloor$  bits to exchange in the bit string of node  $S$  are the same, the node  $S$  connects to node  $s_n s_{n-1} s_{n-2} \dots s_{\lfloor n/2 \rfloor + 1} s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} s_{\lfloor n/2 \rfloor - 2} \dots s_3 s_2 s_1$ , which is the one's complement of the binary number of node  $S$ . Figure 1 shows an example of a 5-dimensional half hypercube ( $HH_5$ ). The degree of  $HH_n$  is  $\lceil n/2 \rceil + 1$ , which adds the number of  $h$ -edges ( $1 \leq h \leq \lceil n/2 \rceil$ ) and one of the  $sw$ -edges. Table 1 presents the degree of  $HH$  according to the dimension of  $HH$  graphs.

**Lemma 1** An  $HH_n$  graph is expanded with recursive structures.

*Proof* An  $HH_n$  graph is constructed with the nodes of two  $(n-1)$ -dimensional  $HH_{n-1}$  graphs by adding one  $sw$ -edge or  $h$ -edge. The address of each node in an  $HH_j$  graph is represented as  $j$  bit strings with binary numbers  $\{0, 1\}$  (i.e., with a



**Fig. 1** Example of a 5-dimensional half hypercube ( $HH_5$ )

**Table 1** Degree of  $HH_n$  according to its dimension

Dimension		3	4	5	6	7	8	...	13	14	15	16	...	$n$
Edge Type	$h$ -edge	2	2	3	3	4	4	...	7	7	8	8	...	$\lceil n/2 \rceil$
	$sw$ -edge	1	1	1	1	1	1	...	1	1	1	1	...	1
Degree		3	3	4	4	5	5	...	8	8	9	9	...	$\lceil n/2 \rceil + 1$

binary number of length  $j$ ). We denote  $HH_j^0$  when the bit at the  $j + 1$  position of the address of a node (i.e., at the  $j + 1$  position of the bit string of a node) is binary 0, and denote  $HH_j^1$  when it is binary 1. Let us examine the expansion of  $HH_n$  by dividing it into two cases: when the dimension  $n$  is even and  $n$  is odd.

Case 1 When expanded from an odd-dimension ( $j$ ) to an even-dimension ( $k$ ),  $k = j + 1$ .

Let us construct an  $HH_k$  graph by connecting a node of  $HH_j^0$  and a node of  $HH_j^1$  in  $HH_j$  ( $k = j + 1$ ). When expanded from an  $HH_j$  graph to an  $HH_k$  graph, the  $sw$ -edges in  $HH_j$  will be replaced with new  $sw$ -edges in the expanded  $HH_k$  graph. A node of  $HH_j^0$  the address bit of which is binary 1 at the  $n/2$  position, will be connected to a node of  $HH_j^1$  through an  $sw$ -edge in  $HH_k$ . When the bit is binary 0, the node will be connected to a node of  $HH_j^0$ .

Case 2 When expanded from an even-dimension ( $k$ ) to an odd-dimension ( $l$ ),  $l = k + 1$ .

We construct an  $HH_l$  graph by connecting a node of  $HH_k^0$  and a node of  $HH_k^1$  in  $HH_k$ . When expanded from an  $HH_k$  graph to  $HH_l$ , the  $sw$ -edges in  $HH_k$  will be replaced with new  $sw$ -edges in the expanded  $HH_l$  graph. A node of  $HH_k^0$ , the address bit of which is binary 1 at the  $\lceil n/2 \rceil$  position, will be connected to a node of  $HH_k^1$  via an  $sw$ -edge in  $HH_l$ . When the bit is binary 0, the node will be connected to a node of  $HH_k^0$ .

**Lemma 2** *There exist  $2^{\lceil n/2 \rceil} \lceil n/2 \rceil$ -dimensional hypercube structures in an  $HH_n$  graph.*

*Proof* An  $h$ -edge of  $HH_n$  connects node  $S (= s_n s_{n-1} s_{n-2} \dots s_h \dots s_3 s_2 s_1)$  to a node, the address bit of which is the complement of the bit string of node  $S$  at exactly one  $h$  position ( $1 \leq h \leq \lceil n/2 \rceil$ ). The  $h$ -edge of  $HH_n$  is equal to the edge of a hypercube. Therefore, the structure of a partial graph constructed from  $HH_n$  via the  $h$ -edge is the same as the structure of a  $\lceil n/2 \rceil$ -dimensional hypercube. Let us assume that a partial graph that has the same structure as an  $\lceil n/2 \rceil$ -dimensional hypercube is  $HH_n^{\lceil n/2 \rceil}$  in  $HH_n$  and refer to it as a cluster. Here, the address of each node in a cluster is  $s_h \dots s_3 s_2 s_1$ . The number of clusters consisting of the  $HH_n$  graph is  $2^{\lceil n/2 \rceil}$  because the number of bit strings that can be configured by the bit string  $s_n s_{n-1} s_{n-2} \dots s_h$  is  $2^{\lceil n/2 \rceil}$ . Node (or edge) connectivity is the minimum number of nodes (or edges) that must be removed to disconnect an interconnection network to

two or more parts without duplicate nodes. If a given interconnection network remains connected with the removal of any arbitrary  $k-1$  or fewer nodes, but the interconnection network becomes disconnected with the removal of any arbitrary  $k$  nodes, then the connectivity of the interconnection network is  $k$ . When the degree and node connectivity of a given interconnection network are the same, we say that the interconnection network has maximum fault-tolerance [3]. It has been proven that  $k(G) \leq \lambda(G) \leq d(G)$  where the node connectivity, edge connectivity, and degree of interconnection network  $G$  are denoted as  $k(G)$ ,  $\lambda(G)$ , and  $d(G)$ , respectively [6, 11]. Through the proving process of  $k(\text{HH}_n) = \lambda(\text{HH}_n) = d(\text{HH}_n)$  in Theorem 1, we will demonstrate that the proposed  $\text{HH}_n$  has maximum fault-tolerance.

**Theorem 1** *The connectivity of  $\text{HH}_{n,k}(\text{HH}_n) = \lceil n/2 \rceil + 1 (n \geq 3)$ .*

*Proof* Let us prove that  $\text{HH}_n$  remains connected even when  $n$  nodes are deleted from  $\text{HH}_n$ . Through Lemma 2, we know that an  $\text{HH}_n$  graph is composed of clusters, and all clusters in  $\text{HH}_n$  are hypercubes and two arbitrary nodes are connected via an  $sw$ -edge. Assuming that  $X$  is a partial graph of  $\text{HH}_n$  where  $|X| = n$ , it will be proven that  $k(\text{HH}_n) \geq \lceil n/2 \rceil + 1$  by demonstrating that  $\text{HH}_n$  remains connected even after the removal of  $X$ . This will be done by dividing two cases in accordance with the location of  $X$ . We denote the  $\text{HH}_n$  in which  $X$  is deleted as  $\text{HH}_n - X$  and a node of  $\text{HH}_n$  as  $S$ .

Case 1 When  $X$  is located in one cluster of  $\text{HH}_n$ :

The degree of each node in a cluster is  $\lceil n/2 \rceil$ . If  $\lceil n/2 \rceil$  nodes adjacent to an arbitrary node  $S$  of the cluster are the same as the nodes to be deleted from  $X$ ,  $\text{HH}_n$  is divided into two components: an interconnection network  $\text{HH}_n - X$  and a node  $S$ . However, all nodes in a cluster are linked to other clusters in  $\text{HH}_n$  via  $sw$ -edges, and  $2^{\lceil n/2 \rceil} - 1$  clusters in which  $X$  is not located are also connected to other clusters via  $sw$ -edges. Therefore,  $\text{HH}_n - X$  is always connected when  $X$  is included in only a cluster of  $\text{HH}_n$ .

Case 2 When  $X$  is located across two or more clusters of  $\text{HH}_n$ :

As the nodes of  $X$  to be deleted are included across two or more clusters of  $\text{HH}_n$ , the number of nodes to be deleted from a cluster is at most  $\lceil n/2 \rceil - 1$ . However, even if  $\lceil n/2 \rceil - 1$  nodes adjacent to an arbitrary node  $S$  of a cluster are deleted, the nodes of the cluster in which node  $S$  is included remain connected because the degree of a node in a cluster is  $\lceil n/2 \rceil$ . Although the other node to be removed is located in a different cluster, and not in the cluster that includes node  $S$ , it is still clear that  $\text{HH}_n$  remains connected. Therefore,  $k(\text{HH}_n) \geq \lceil n/2 \rceil + 1$  because  $\text{HH}_n$  always remains connected after removing  $X$  from any clusters in  $\text{HH}_n$  and  $k(\text{HH}_n) \leq \lceil n/2 \rceil + 1$  because the degree of  $\text{HH}_n$  is  $\lceil n/2 \rceil + 1$ . Consequently, the connectivity of  $\text{HH}_{n,k}(\text{HH}_n) = \lceil n/2 \rceil + 1$ .

### 3 Routing Algorithm and Diameter of $HH_n$

This section analyzes a simple routing algorithm and diameter of  $HH_n$ . We assume that an initial node  $S$  is  $s_n s_{n-1} s_{n-2} \dots s_{n/2} \dots s_3 s_2 s_1$  and a destination node  $T$  is  $t_n t_{n-1} t_{n-2} \dots t_{n/2} \dots t_3 t_2 t_1$ . A simple routing algorithm can be considered in two cases depending on whether  $n$  is an even or an odd number.

Case 1 When  $n$  is an even number

If an initial node  $S(= s_n s_{n-1} s_{n-2} \dots s_{n/2} s_{n/2-1} \dots s_3 s_2 s_1)$  is presented with two  $\lceil n/2 \rceil$  bit strings  $A(= s_n s_{n-1} s_{n-2} \dots s_{n/2})$  and  $B(= s_{n/2-1} \dots s_3 s_2 s_1)$ , node  $S$  can be denoted as  $AB$ . In the same way, a destination node  $T(= t_n t_{n-1} t_{n-2} \dots t_{n/2} t_{n/2-1} \dots t_3 t_2 t_1)$  can be denoted as  $CD$  where  $C(= t_n t_{n-1} t_{n-2} \dots t_{n/2})$  and  $D(= t_{n/2-1} \dots t_3 t_2 t_1)$  (i.e.,  $C$  is the  $\lceil n/2 \rceil$  leftmost bits and  $D$  is the  $\lceil n/2 \rceil$  rightmost bits in the bit string of node  $T$ ).

*Simple routing algorithm-even:*

- (1) Convert the bit string of  $B$  in node  $S(= AB)$  with the bit string  $C$  in node  $T(= CD)$  using  $h$ -edge ( $1 \leq h \leq \lceil n/2 \rceil$ ).
- (2) Exchange the bit string of  $A$  with that of  $C$  in node  $S(= AC)$  using  $sw$ -edge.
- (3) Convert the bit string of  $A$  in node  $S(= CA)$  with  $D$  of node  $T(= CD)$  using  $h$ -edge.

In Phases (1) and (3), the bit string  $B$  is converted with  $C$  and  $A$  is converted with  $D$  using the hypercube routing algorithm.

**Corollary 1** When  $n$  is even, the length of the shortest path is  $2 \times \lceil n/2 \rceil + 1 = n + 1$  by the above *simple routing algorithm-even*.

Case 2 When  $n$  is an odd number

Let the initial node be  $S(= s_n s_{n-1} s_{n-2} \dots s_{\lfloor n/2 \rfloor + 1} s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} \dots s_3 s_2 s_1)$ , the  $\lfloor n/2 \rfloor$  leftmost bits of the bit string of  $S$  be  $A(= s_n s_{n-1} s_{n-2} \dots s_{\lfloor n/2 \rfloor + 1})$  and the  $\lfloor n/2 \rfloor$  bits on the right-side of  $S$  starting from the  $\lfloor n/2 \rfloor$  bit be  $B(= s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} \dots s_3 s_2)$ . Then, node  $S$  can be denoted as  $(ABs_1)$ . In the same way, a destination node  $T(= t_n t_{n-1} t_{n-2} \dots t_{\lfloor n/2 \rfloor + 1} t_{\lfloor n/2 \rfloor} t_{\lfloor n/2 \rfloor - 1} \dots t_3 t_2 t_1)$  can be denoted as  $(CDt_1)$ , where  $C(= t_n t_{n-1} t_{n-2} \dots t_{\lfloor n/2 \rfloor + 1})$  and  $D(= t_{\lfloor n/2 \rfloor} t_{\lfloor n/2 \rfloor - 1} \dots t_3 t_2)$ .

*Simple routing algorithm-odd:*

- (1) Convert the bit string of  $B$  in node  $S(= ABs_1)$  with the bit string  $C$  in node  $T(= CDt_1)$  using  $h$ -edge ( $1 \leq h \leq \lfloor n/2 \rfloor$ ).
- (2) Exchange the bit string of  $A$  with that of  $C$  in node  $S(= ACs_1)$  using  $sw$ -edge.
- (3) Convert the bit string of  $As_1$  in node  $S(= CA s_1)$  with  $Dt_1$  of node  $T(= CDt_1)$  using  $h$ -edge ( $1 \leq h \leq \lfloor n/2 \rfloor$ ).

In Phases (1) and (3), the bit string  $B$  is swapped with  $C$  and  $As_1$  is swapped with  $Dt_1$  using the hypercube routing algorithm.

**Corollary 2** When  $n$  is odd, the length of the shortest path is  $2 \times \lfloor n/2 \rfloor + 2$  by the above *simple routing algorithm-odd*.

Through the proposed simple routing algorithms, we can see an upper bound for the diameter of  $HH_n$ , thus proving Theorem 2.

**Theorem 2** *The upper bound on the diameter of  $HH_n$  is  $n + 1$  when  $n$  is even and  $2 \times \lfloor n/2 \rfloor + 2$  when  $n$  is odd.*

## 4 Conclusion

This paper proposed a half hypercube interconnection network  $HH_n$  (a new variation of the hypercube) that reduced the degree by approximately half,  $n/2$ , even though it has the same number of nodes as a hypercube. To evaluate the effectiveness of the proposed half hypercube, we analyzed its connectivity and diameter properties. We also analyzed a simple routing algorithm of  $HH_n$  and presented an upper bound for the diameter of  $HH_n$ . These results demonstrate that the proposed half hypercube is an appropriate interconnection network for implementation in a large-scale system.

**Acknowledgment** This research was supported by Basic Science research program through the National research Foundation of KOREA (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A4A01014439).

## References

1. Leighton FT (1992) Introduction to parallel algorithms and architectures: arrays, hypercubes. Morgan Kaufmann Publishers, San Francisco
2. Mendia VE, Sarkar D (1992) Optimal broadcasting on the star graph. IEEE Trans Parallel Distrib Syst 3(4):389–396
3. Sim H, Oh JC, Lee HO (2010) Multiple reduced hypercube (MRH): a new interconnection network reducing both diameter and edge of hypercube. Int J Grid Distrib Comput 3(1):19–30
4. Saad Y, Schultz MH (1988) Topological properties of hypercubes. IEEE Trans Comput 37(7):867–872
5. Seitz CL (1985) The cosmic cube. Commun ACM 26:22–33
6. Abraham S (1991) The twisted cube topology for multiprocessor: a study in network asymmetry. J Parallel Distrib Comput 13:104–110
7. El-Amawy A, Latifi S (1991) Properties and performance of folded hypercubes. IEEE Trans Parallel Distrib Syst 2(1):31–42
8. Ghose K, Desai KR (1995) Hierarchical cubic network. IEEE Trans Parallel Distrib Syst 6(4):427–435
9. Kumar JM, Patnaik M (1992) Extended hypercube: a hierarchical interconnection network of hypercubes. IEEE Trans Parallel Distrib Syst 3(1):45–57
10. Livingston M, Stout QF (1988) Embedding in hypercubes. Math Comput Model 11:222–227
11. Akers SB, Harel D, Krishnamurthy B (1987) The star graph: an attractive alternative to the N-Cube. In: Proceedings of the international conference on parallel processing, pp 393–400

# Aperiodic Event Communication Process for Wearable P2P Computing

Tae-Gyu Lee and Gi-Soo Chung

**Abstract** Wearable computing has been proposed as an alternative to the best computing interfaces and devices for the ubiquitous computing. A digital wear can be a main element of wearable computers. This study shall apply digital yarn as a material of data communications for the purpose to take advantage a digital garment. Wearable P2P application communications are consisted of periodic or aperiodic methods. This paper proposes an aperiodic event process for wearable P2P computing. It shows the transmission process that collects from a digital garment at random time. Specially, the process supports the recovery transfer process when the aperiodic event messages are failed.

**Keywords** Aperiodic communication · Wearable computing · P2P communication · Digital garment

## 1 Introduction

Nowadays, as we enter the era of ubiquitous computing, the computing appliances are more gradually closer to human and the using time of information devices has exponentially increased. A wearable computing has been proposed as an alternative to the best mobile computing devices for these ubiquitous computing [1].

A digital garment accounts a principle element of wearable computing. This study shall apply digital yarn as a material of data communications for the purpose to take advantage a digital garment [3]. This conductive micro-wire digital yarn can be applied in a general garment knitting or weaving process as a lightweight weaving textile unlike the existing communication lines. The digital fiber was

---

T.-G. Lee (✉) · G.-S. Chung  
Korea Institute of Industrial Technology (KITECH), Ansan 426-791, Korea  
e-mail: tigerlee88@empal.com

already developed, but still is in incomplete status to be used as a communication standard configuration and transport platforms [2, 3].

Wearable P2P application communications are consisted of periodic or aperiodic methods. Periodical communications support the regular collection of information, and aperiodic communications support the transfer of information when a particular event has occurred. This paper proposes an aperiodic event process for wearable P2P computing. It shows the transmission process that collects from a digital garment at random time. Specially, the process supports the recovery transfer process when the aperiodic event messages are failed [4–6].

An *event message channel* is a temporary storage location for data while the data is being transferred. The event channel is often used for supporting a recovery of loss data frames. There have the problems that if the channel size were too large, it makes the efficient use of channel resources worse and otherwise, if it has the smaller channel capacity, it makes the available channel bandwidth inefficient due to a waste of the network failure of digital yarn. Therefore, it is necessary to set up the optimal size of the event channel capacity.

Typically, an event process consists of First In First Out (FIFO) communication structure. The event message model presented in this paper is based on the sender-receiver communication channel of Ethernet, which is widely used in communication systems.

This paper describes about the overview of aperiodic message communication system in Sect. 2. Section 3 shows an aperiodic P2P event processes. Section 4 shows the application scenarios for wearable P2P commutations. Finally, Sect. 5 concludes this paper.

## 2 Aperiodic Message Communication

Wearable aperiodic P2P communication system supports the infrastructure of information exchange to build the wearable embedded computing services of a mobile wearable user. In order to build these systems, two or more nodes (MSS: Mobile Service Station) of sharing information have been configured and the wired and wireless communication channels should be organized for the information exchange. More particularly, this study focuses on the wear-embedded wired communication system based on digital yarn.

For aperiodic messages, their arrival durations usually are irregular, however it is assumed that there is a minimum inter-arrival time in order to guarantee its temporal constraint. Aperiodic messages can be classified into *urgent* and *normal* states.

The wearable aperiodic P2P communication is a transferring process which sends the information initiated by the sender-peer to the receiver-peer. The aperiodic message transmission system considers the link propagation delay and multiplexing method as a factor which affects to the transmission performance and the channel efficiency.

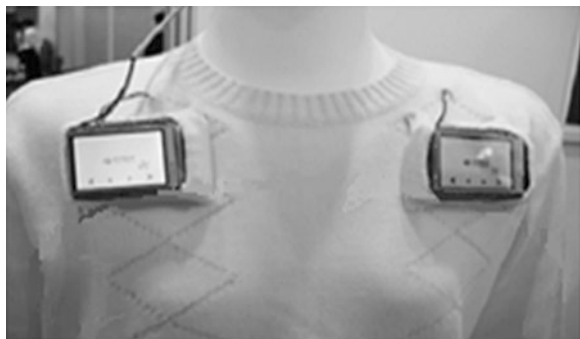
Figure 1 shows the aperiodic P2P computing system on P2P communication channels. Any one peer terminals become an initiator and the other peer terminal become a correspondent.

The P2P system configuration for applications supports a distributed P2P MSS and a dual P2P MSS. First, the distributed MSS applications are corporately performed for providing different computing service on two more terminals. As a distributed P2P system, the cross-work configuration of left-peer and right-peer can be applied to the distributed applications crossover terminals, respectively. Second, the dual P2P MSS supports the same computing services on two more terminals, and one MSS peer performs the active computing as the foreground services and the other MSS peer performs stand-by backup computing. As a dual P2P communication system, the front-peer and rear-peer configuration can be employed to each of application terminal and backup terminal, respectively. The rear-peer can be optionally used as standby terminal.

The wearable aperiodic P2P computing system has the needs to meet the following requirements. First, the P2P computing on two more peers can provide the interactive event message at any time. Second, it can support the mission-critical projects more than single terminal with the same event process image. Third, it can support the load balancing service for transferring the event messages by load-distributing method. There are military wear, police wear, and firefight wear as the mission-critical applications based on the wearable P2P system. These applications are important to consider computing performance as well as safety issues.

As the differentiation with other P2P systems, the existing P2P systems provide the efficient resource sharing and load balancing on wide-area Internet, but the proposed P2P communication system realizes the distributed asynchronous P2P transmission system among the limited specific terminals on wearable local network. Thus, the wearable P2P communication system has the low complexity of transmission links. And it has the high frequency of P2P transfer between the specific terminals. The system should consider the performance aspects and resource-efficient aspects together.

**Fig. 1** Aperiodic P2P communication system in wearable computing





This paper only considers dual P2P communication system for the aperiodic communication performance and system safety. When the dual P2P system organization is considered for experiment and analysis model, the main factors that influence communication performance are asynchronous P2P communication link. The P2P link supports bidirectional communication between front-peer and rear-peer.

To improve the communication performance, we consider the propagation delay of the asynchronous dual P2P link. Also, we will consider the multiple levels of the P2P links to enhance the communication safety.

### 3 Aperiodic P2P Event Process

In the normal event process, the aperiodic P2P event can be defined of a set as event tuple  $E < DN_j, E < SN_i, M_k \gg$ .  $E$  means the event transfer function,  $DN_j$  is a destination peer node,  $SN_i$  is a start peer node, and  $M_k$ . indicates an event message of  $SN_i$ . The following process of Fig. 2 shows these normal event processes. Initialization method of checkpoint cycle uses a time period, or employs a message count period.

In the failure event process, the P2P fault-recovery events can be processed as the following process of Fig. 3. A destination node  $DN_j$  discards the duplicate event messages  $M_k$ . from the same start peer node  $SN_i$ . It determines that the message was lost when the  $k$  index value of the incoming event messages  $M_k$  is not increased sequentially as skipping. In a sender  $SN_i$ , if there is no response message to the outgoing message from a receiver  $DN_j$ , we may decide the transmitted message was lost.

The checkpoint is a snapshot of process control block as resource information for process recovery or synchronization.

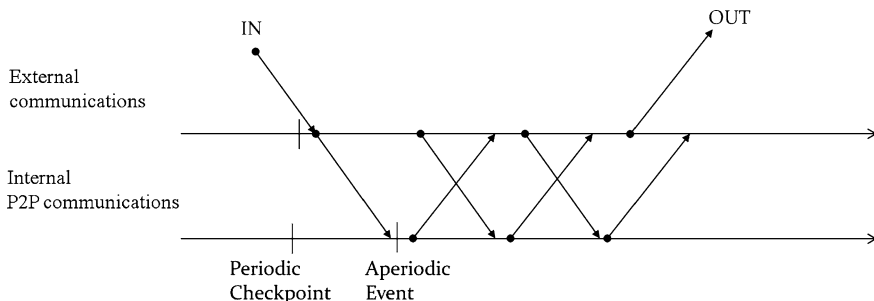
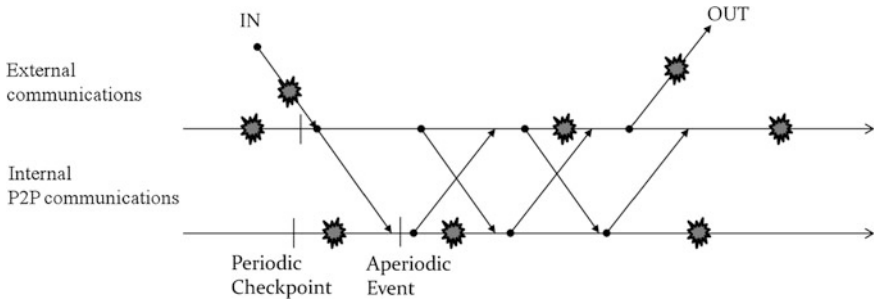


Fig. 2 Normal process in aperiodic P2P communication system



**Fig. 3** Fault-recovery process in aperiodic P2P communication system

The process events are classified into real-time and non-real-time data. First, the real-time data means the continuous information as streaming sensing signals. It can be received from the correspondent and saved to storage devices periodically for easy data management. Second, the non-real-time data means the discrete information as general intermediate messages. It can be received from the correspondent and saved to storage devices a-periodically for easy data management.

The following process sequences describe the sender and receiver communication processes, and present the fault-recovery processes.

Sender process in an starter peer node,  $SN_i$ :

- Start and initialize with its  $SN_i$  and the correspondent's  $DN_j$ ;
- Check a new event and save the process aperiodic checkpoints on any random time;
- Select data resource in collected data;
- Send the event message including data resource to the correspondent,  $DN_j$ ;
- Receive the confirm message from Receiver;
- If (the confirm message = *NAK*) then retry;
- Otherwise, wait for next event;

Receiver process as a correspondent peer,  $DN_j$ :

- Start and initialize with its  $DN_j$  and the correspondent,  $SN_i$ ;
- Receive the event message from Sender,  $SN_i$ ;
- Check a new event and save the process checkpoints on aperiodic time;
- Reply the confirm message with event sequence information to Sender,  $SN_i$ ;

Event fault-recovery:

Each Sender,  $SN_i$ , and Receiver,  $DN_j$ , transfer its process sequences respectively;

Save the event message into the backup memory of the Sender,  $SN_i$ ;

Remove the messages of mismatched sequences;

Remove orphan or duplicate messages;

*Fault detection();*

Detect the crash fault using beacon signals by correspondent;

Detect the temporal fault using the inform message by faulty peer;

*Standby state();*

Wait for the alive message of faulty peer;

Skip or block the P2P messages;

*Alive state();*

Rollback to last checkpoint;

Send alive message including checkpoint position;

Receive the confirm message with synchronization;

Perform the recovery computing and communication using the message backup memory of the sender  $SN_i$ ;

*P2P event synchronization();*

Send event sync message with the event sequence numbers to the correspondents;

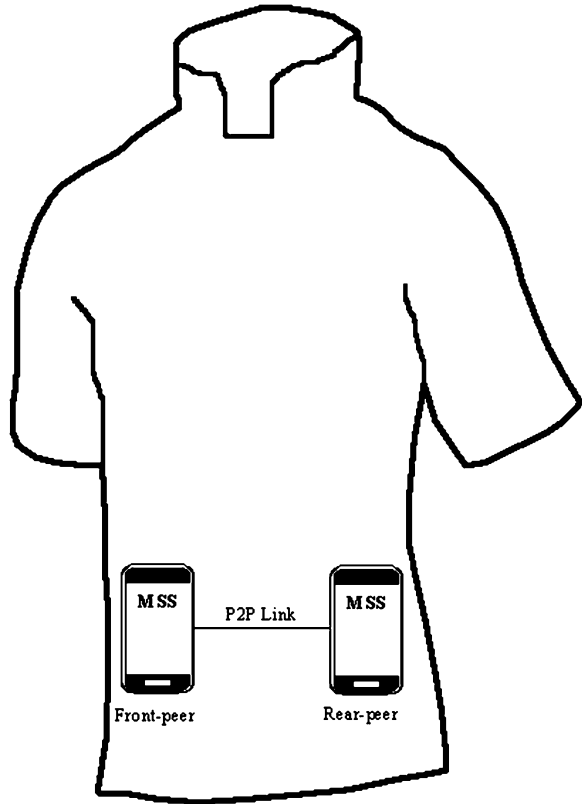
## 4 Wearable Aperiodic P2P Application

In this section, as a wearable P2P application, the mission-critical operations are performed such as police and military missions as shown in Fig. 4. The wearable system collects aperiodic event of non-real-time messages. For example, the military applications with digital garment have the interactive messages between a MSS peer and the other MSS peer as the following scenario sequences.

These are aperiodic P2P application scenarios as the mission-critical services for military digital garment.

- (1) Start top-down commands of orders and mission commitment;
- (2) Determine the mission and communicate internal P2P communication links;
- (3) Read circumstances using embedded sensors and start reporting;

**Fig. 4** P2P applications for wearable communication system



- (4) Execute mission monitoring using internal computing and P2P communication link;
- (5) Real-time message reporting of special situations as aperiodic event;
- (6) Mission can be changed and aborted optionally;
- (7) Mission is reconfirmed and a self-assessment results is performed;
- (8) Perform the expanded mission;
- (9) Mission off commands are delivered between a peer and the correspondent peer by the commander;
- (10) Sleep for waiting the new missions after a current mission is off;

The components of wearable embedded mobile smart station (MSS) are consisted as follows. First, as a dual P2P communication system, the front-peer and rear-peer configuration is employed to each of application terminal and backup terminal, respectively. The rear-peer can be selectively used for standby dual terminal.

Second, as a distributed P2P system, the cross-work configuration of left-peer and right-peer is applied to the distributed applications crossover terminals, respectively.

When the dual P2P system organization is considered for experiment and analysis model, the key factors that influence communication performance are P2P communication link. The P2P link supports P2P bidirectional communication between front-peer and rear-peer.

## 5 Conclusions

This work has presented the aperiodic P2P communication logic in the wearable computing fields using digital yarn. Then it has proposed the multiple links and the fault-recovery process as the methods for enhancing the performance and feasibility of the P2P communication system. Finally, the aperiodic communication application scenarios have been shown the needs of multiple links and the effectiveness of fault-recovery process simultaneously.

## References

1. Lee T (2012) Information life cycle design and considerations for wearable computing. *EMC Technol Serv LNEE* 181:501–508
2. Vassiliadis S, Provatidis C, Prekas K, Rangussi M (2005) Novel fabrics with conductive fibers. In: *Intelligent textile structures—application, production & testing international workshop, GREECE, May 2005*
3. ChungGS (2009) Digital garment for data communication using digital yarn. In: *2009 Korean-German smart textile symposium, pp 57–67, Sept 2009*
4. Lee T, Chung G (2012) Wearable P2P communication system organization on digital yarn. *Ubiquitous information technologies and applications. Lect Notes Electr Eng* 214:601–609
5. Wang Z, Shen X, Chen J, Song Y, Wang T, Sun Y (2005) Real-time performance evaluation of urgent aperiodic messages in FF communication and its improvement. *Elsevier Comput Stand Interfaces* 27:105–115
6. Kato S, Fujita XY, Yamasaki N (2009) Periodic and aperiodic communication techniques for responsive link. In: *15th IEEE international conference on embedded and real-time computing systems and applications, pp 135–142*

# Broadcasting and Embedding Algorithms for a Half Hypercube Interconnection Network

Mi-Hye Kim, Jong-Seok Kim and Hyeong-Ok Lee

**Abstract** The half hypercube interconnection network, has been proposed as a new variation of the hypercube, reducing its degree by approximately half with the same number of nodes as an  $n$ -dimensional hypercube,  $Q_n$ . This paper proposes an algorithm for one-to-many broadcasting in an  $n$ -dimensional half hypercube,  $HH_n$ , and examines the embedding between hypercube and half hypercube graphs. The results show that the one-to-many broadcasting time of the  $HH_n$  can be accomplished in  $n + 1$  when  $n$  is an even number and in  $2 \times \lceil n/2 \rceil$  when  $n$  is an odd number. The embedding of  $HH_n$  into  $Q_n$  can be simulated in constant time  $O(n)$  and the embedding of  $Q_n$  into  $HH_n$  in constant time  $O(1)$ .

**Keywords** Half hypercube · One-to-many broadcasting · Embedding

## 1 Introduction

There is increasing interest in parallel processing as a technique for achieving high-performance owing to the need for many computations with real-time data processing in modern applications [1, 2]. A parallel processing system can connect hundreds of thousands of processors with their own memory via an interconnection

---

M.-H. Kim

Department of Computer Science Education, Catholic University of Daegu,  
Daegu, South Korea  
e-mail: mihyekim@cu.ac.kr

J.-S. Kim

Department of Computer Science, University of Rochester, Rochester 14627, USA  
e-mail: Rockhee7@gmail.com

H.-O. Lee (✉)

Department of Computer Education, Suncheon National University, Suncheon, South Korea  
e-mail: oklee@sunchon.ac.kr

network. The overall performance of the system is dependant on the performance of each processor and the architecture of the interconnection network used [1–3]. Many interconnection network topologies have been described in the literature, such as star, mesh, bubble-sort, and pancake graphs.

The hypercube,  $Q_n$ , is a typical topology and is an  $n$ -regular and node- and edge-symmetric graph with  $2^n$  nodes and diameter  $n$  ( $n \geq 2$ ). The hypercube  $Q_n$  has simple routing algorithms and recursive structures with maximum fault-tolerance. In addition, it has the advantage that its network structure can easily be embedded in various types of commonly used interconnection networks. With such advantages, it is widely used in various application areas [1, 4, 5]. However, its network cost is increased considerably in relation to the increased degree when the number of nodes increases. To resolve this drawback, several hypercube variations have been introduced. We have proposed the half-hypercube interconnection network, reducing its degree by approximately half, even though it has the same number of nodes as a hypercube  $Q_n$ . In this paper, we propose an algorithm for one-to-many broadcasting in an  $n$ -dimensional half hypercube,  $HH_n$ , and analyze the embedding method of a half hypercube graph into a hypercube graph and vice versa.

The most common properties for measuring the performance of interconnection networks include degree, diameter, connectivity, fault tolerance, broadcasting, and embedding [6, 7]. In [1], we analyzed the degree, diameter, connectivity, and fault-tolerance parameters of the half hypercube. Here, we examine the broadcasting and embedding properties of a  $HH_n$  to strengthen its effectiveness. Broadcasting is one of the major primitives for communication of parallel processing involving message disseminating from an origin node to all the other nodes (one-to-many broadcast) or among the nodes (many-to-many broadcast) in an interconnection network. Embedding is to evaluate the relative performance of two arbitrary interconnection networks. This is of interest, because the properties and algorithms developed in a certain topology can easily be adapted to another network at less cost [8].

The organization of this paper is as follows. Section 2 presents the definition of the half hypercube  $HH_n$ . Section 3 proposes and analyzes a broadcasting algorithm for an  $n$ -dimensional half hypercube,  $HH_n$ . Section 4 examines the embedding algorithms between hypercube and half hypercube graphs. Section 5 concludes the paper.

## 2 Definition of the Half Hypercube

The half hypercube  $HH_n$  ( $n \geq 3$ ) is defined as an  $n$ -dimensional binary cube where the nodes of  $HH_n$  are all binary  $n$ -tuples in the same way as the hypercube  $Q_n$  ( $n \geq 2$ ). That is, an  $n$ -dimensional half hypercube is denoted as  $HH_n$  and each node is represented with  $n$  binary bits. The degree and node connectivity of  $HH_n$  are  $\lceil n/2 \rceil + 1$  and  $\lceil n/2 \rceil + 1$  ( $n \geq 3$ ), respectively. An  $HH_n$  graph is expanded with recursive structures and has  $2^{\lfloor n/2 \rfloor} \lfloor n/2 \rfloor$ -dimensional hypercube structures with maximum fault-tolerance [1].

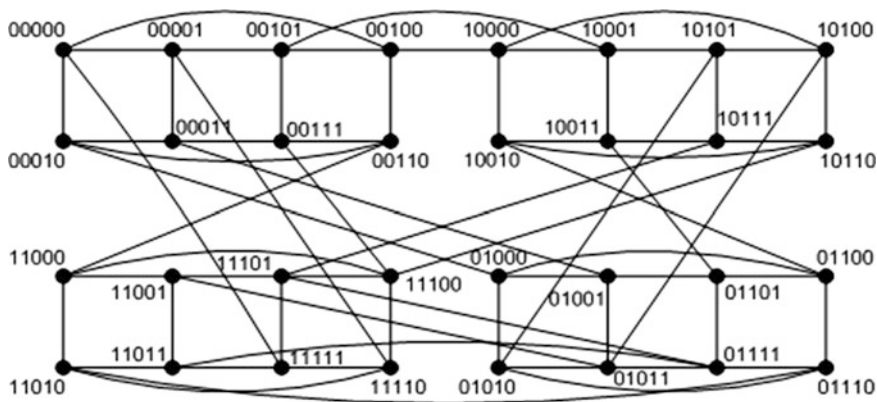


Fig. 1 Example of a 5-dimensional half hypercube ( $HH_5$ ) [1]

In  $Q_n$ , an edge exists between two arbitrary nodes  $S$  and  $S'$  if and only if their corresponding  $n$ -tuples differ in exactly one  $k$  position of the bit strings of  $S$  and  $S'$  ( $1 \leq k \leq n$ ) [9]. On the other hand, in  $HH_n$ , two types of edge exist: the  $h$ -edge, which connects node  $S$  to a node that has the complement in one bit at exactly one  $h$  position ( $1 \leq h \leq \lfloor n/2 \rfloor$ ), and the *swap*-edge (shortly, *sw*-edge), which connects node  $S$  to a node where the  $\lfloor n/2 \rfloor$  leftmost bits of the bit string of  $S$  and the  $\lfloor n/2 \rfloor$  bits on the right-side of  $S$  starting from the  $\lfloor n/2 \rfloor$  position are exchanged ( $h = \lfloor n/2 \rfloor$ ). However, if two parts of  $\lfloor n/2 \rfloor$  bits to swap in the bit string of node  $S$  are the same, the node  $S$  connects to node  $\bar{S}$ , which is the one's complement of the binary number of node  $S$  [1]. Note that  $\bar{S}$  indicates the one's complement of the binary number of node  $S$  in this paper.

At node  $S(=s_n s_{n-1} \dots s_{\lfloor n/2 \rfloor + 1} s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} \dots s_2 s_1)$  of  $HH_n$ , the address of node  $S'$  adjacent to node  $S$  via an *sw*-edge is considered using two cases depending on whether  $n$  is even or odd. For instance, when  $n$  is even, node  $S(=s_n s_{n-1} \dots s_{\lfloor n/2 \rfloor + 1} s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} \dots s_2 s_1)$  is adjacent to node  $S'(=s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} \dots s_2 s_1 s_n s_{n-1} \dots s_{\lfloor n/2 \rfloor + 1})$  where  $(s_n s_{n-1} \dots s_{\lfloor n/2 \rfloor + 1})$  and  $(s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} \dots s_2 s_1)$  of  $S$  are swapped. When  $n$  is odd, node  $S$  is adjacent to node  $S'(=s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} \dots s_2 s_n s_{n-1} \dots s_{\lfloor n/2 \rfloor + 1} s_1)$  where  $(s_n s_{n-1} \dots s_{\lfloor n/2 \rfloor + 1})$  and  $(s_{\lfloor n/2 \rfloor} s_{\lfloor n/2 \rfloor - 1} \dots s_2)$  of  $S$  are swapped. Figure 1 presents a 5-dimensional half hypercube ( $HH_5$ ) [1].

### 3 Broadcasting Algorithm for $HH_n$

Broadcasting is a basic data communication technique for interconnection networks involving message transmission between nodes and is used by parallel algorithms [7, 10]. There are two types of broadcasting communication: one-to-many transmission, which transmits messages from a node to all the other nodes,



and many-to-many transmission, which transmits messages among nodes. Here, we will demonstrate that the one-to-many broadcasting time of  $HH_n$  is  $n + 1$  when  $n$  is even and the broadcasting time is  $2 \times \lceil n/2 \rceil$  when  $n$  is odd.

**Theorem 1** *When  $n$  is even, the one-to-many broadcasting time of  $HH_n$  is  $n + 1$  and when  $n$  is odd, the one-to-many broadcasting time of  $HH_n$  is  $2 \times \lceil n/2 \rceil$ .*

*Proof* Each cluster of  $HH_n$  represents a hypercube and the one-to-many broadcasting time of a hypercube  $Q_m$  is  $m$ . A cluster is connected to all the other clusters in  $HH_n$  by an external *sw*-edge. The broadcasting process is divided into three phases as follows:

- (1) Phase 1: Node  $S$  transmits messages to all the other nodes within its cluster
- (2) Phase 2: All nodes within the cluster to which node  $S$  belongs, including node  $S$  transmit messages to an arbitrary node in all the other clusters of  $HH_n$  using an external *sw*-edge.
- (3) Phase 3: Repeat the process of Phase 1 in each cluster of  $HH_n$ .

When  $n$  is even, the one-to-many broadcasting time is as follows. As the broadcasting time of an internal cluster of  $HH_n$  is the same as the one-to-many broadcasting time of a hypercube, the broadcasting time of Phase 1 is  $n/2$ . As broadcasting is performed only once in Phase 2, its broadcasting time is 1. As Phase 3 repeats the process of Phase 1, the broadcasting time is  $n/2$ . Therefore, the one-to-many broadcasting time of  $HH_n$  is  $n/2 + 1 + n/2 = n + 1$  when  $n$  is even.

When  $n$  is odd, the one-to-many broadcasting time is as follows. As the broadcasting time of an internal cluster of  $HH_n$  is the same as the one-to-many broadcasting time of a hypercube, the broadcasting time of Phase 1 is  $n/2$ . As broadcasting is performed only once in Phase 2, its broadcasting time is 1. If  $n$  is odd, the number of the *sw*-edges connecting clusters is 2 or 4. Thus, the number of nodes that initiate a message transmission is 2 or 4 in Phase 3. If the number of start nodes is 2, the one-to-many broadcasting time of a hypercube is reduced by 1. Therefore, the broadcasting time of Phase 3 is  $\lceil n/2 \rceil - 1$ . Consequently, the one-to-many broadcasting time of  $HH_n$  is  $\lceil n/2 \rceil + 1 + \lceil n/2 \rceil - 1 = 2 \times \lceil n/2 \rceil$  when  $n$  is odd.

## 4 Embedding Between Hypercube and Half Hypercube Graphs

Numerous parallel processing algorithms are being designed to solve many problems in a variety of interconnection network structures. Whether such algorithms designed for a specific interconnection network structure can be run on different interconnection network structures is an important issue in parallel processing. One of the most widely used measuring methods for this issue is embedding [10, 11], which involves mapping the processors and communication links of an interconnection network into those of another interconnection network.

We can represent an interconnection network as a graph  $G(V, E)$ , where  $V(G)$  and  $E(G)$  are the set of nodes and edges of graph  $G$ , respectively, and the set of paths of graph  $G$  is  $P(G)$ . The embedding of an interconnection network  $G(V, E)$  into another interconnection network  $G'(V', E')$  is defined as a function  $(\Phi, \rho)$ , where  $\Phi$  maps the set of vertices  $V(G)$  one-to-one into the set of vertices  $V'(G')$  and  $\rho$  maps the set of edges  $E(G)$  into the set of paths  $P'(G')$ ; that is,  $\Phi: V \rightarrow V'$  and  $\rho: E \rightarrow P'(G')$ . The representative measurement parameters for embedding costs are dilation and congestion. Dilation is the length of the shortest path from node  $S'$  to node  $T'$  in  $G'$  when the nodes  $S$  and  $T$  of an edge  $(S, T)$  in  $G$  are mapped to nodes  $S'$  and  $T'$  of  $G'$ ; i.e., the number of edges comprising the shortest path from node  $S'$  to node  $T'$  in  $G'$ . Congestion is the number of edges in  $G$  that pass an edge  $e$  in  $G'$  when  $G$  is mapped to  $G'$  [3, 4]. In this section, we analyze the embedding between a hypercube  $Q_n$  and a half hypercube  $HH_n$  using dilation.

**Theorem 2** *An  $n$ -dimensional hypercube  $Q_n$  can be embedded into an  $n$ -dimensional half hypercube  $HH_n$  with dilation 3.*

*Proof* We can analyze the dilation of this mapping through the number of edges of  $HH_n$  required to map the  $k$ -dimensional edge ( $1 \leq k \leq n$ ), which represents the adjacent relationships of the nodes in  $Q_n$ , into edges in  $HH_n$ . Theorem 4 is proven by dividing the  $k$ -dimensional edge of  $Q_n$  into two cases depending on the dimension of  $k$ .

Case 1  $k$ -dimensional edge,  $1 \leq k \leq \lceil n/2 \rceil$

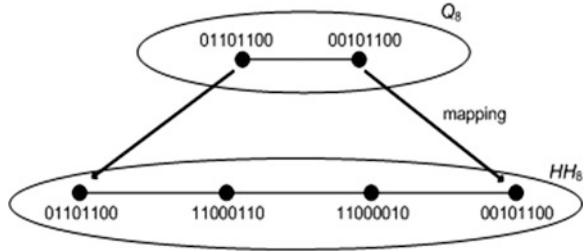
It can be easily observed that the  $k$ -dimensional edge of hypercube  $Q_n$  ( $1 \leq k \leq \lceil n/2 \rceil$ ) are the same as the  $h$ -dimensional edge of half hypercube  $HH_n$  ( $1 \leq h \leq \lceil n/2 \rceil$ ). Therefore, it is clear that the embedding of an  $n$ -dimensional hypercube  $Q_n$  into an  $n$ -dimensional half hypercube  $HH_n$  is possible with dilation 1 when the two adjacent nodes via a  $k$ -dimensional edge in  $Q_n$  are mapped to two adjacent nodes through an  $h$ -dimensional edge in  $HH_n$ .

Case 2  $k$ -dimensional edge,  $\lceil n/2 \rceil + 1 \leq k \leq n$

The address of node  $S'$  adjacent to an arbitrary node  $S(=s_n s_{n-1} s_{n-2} \dots s_k \dots s_{\lceil n/2 \rceil} \dots s_3 s_2 s_1)$  of  $Q_n$  via a  $k$ -dimensional edge has the complement at exactly one  $k$  position of the bit string of node  $S$  (i.e., bit  $s_k$ ). An edge of  $HH_n$  that has the same role as the  $k$ -dimensional edge of hypercube  $Q_n$  can be presented by sequentially applying the following edge sequence:  $\langle sw\text{-edge}, (k - \lceil n/2 \rceil)\text{-edge}, sw\text{-edge} \rangle$  ( $\lceil n/2 \rceil + 1 \leq k \leq n$ ). That is, it reaches a node with the address  $s_n s_{n-1} s_{n-2} \dots \bar{s}_k \dots s_{\lceil n/2 \rceil} \dots s_3 s_2 s_1$  when the edge sequence  $\langle sw\text{-edge}, (k - \lceil n/2 \rceil)\text{-edge}, sw\text{-edge} \rangle$  is applied sequentially to node  $S(=s_n s_{n-1} s_{n-2} \dots s_k \dots s_{\lceil n/2 \rceil} \dots s_3 s_2 s_1)$  of  $HH_n$ . Let a node  $S'$  adjacent to node  $S(=s_n s_{n-1} s_{n-2} \dots s_k \dots s_{\lceil n/2 \rceil} \dots s_3 s_2 s_1)$  of hypercube  $Q_n$  via a  $k$ -dimensional edge ( $\lceil n/2 \rceil + 1 \leq k \leq n$ ) be  $S'(=s_n s_{n-1} s_{n-2} \dots \bar{s}_k \dots s_{\lceil n/2 \rceil} \dots s_3 s_2 s_1)$ .

The address of node  $sw(S)$  adjacent to node  $S(=s_n s_{n-1} s_{n-2} \dots s_k \dots s_{\lceil n/2 \rceil} s_{\lceil n/2 \rceil + 1} s_{\lceil n/2 \rceil} \dots s_3 s_2 s_1)$  of  $HH_n$  via an  $sw$ -edge is  $s_{\lceil n/2 \rceil} \dots s_3 s_2 s_1 s_n s_{n-1} s_{n-2} \dots s_k \dots s_{\lceil n/2 \rceil + 1}$ . To invert

**Fig. 2** Embedding example between  $Q_8$  and  $HH_8$  with dilation 3



the bit  $s_k$  in the bit string of node  $sw(S)$  to the complement, we take a node  $S''(=s_{\lceil n/2 \rceil} \dots s_3 s_2 s_1 s_n s_{n-1} s_{n-2} \dots \bar{s}_k \dots s_{\lceil n/2 \rceil + 1})$  adjacent to node  $sw(S)$  via a  $(k - \lceil n/2 \rceil)$ -dimensional edge. The address adjacent to node  $S''$  through an  $sw$ -edge is  $s_n s_{n-1} s_{n-2} \dots \bar{s}_k \dots s_{\lceil n/2 \rceil + 1} s_{\lceil n/2 \rceil} \dots s_3 s_2 s_1$ . Therefore, the embedding of a  $k$ -dimensional edge of  $Q_n$  into a half hypercube  $HH_n$  is possible with dilation 3. Figure 2 presents an example of embedding between  $Q_8$  and  $HH_8$  with dilation 3.

**Theorem 3** A half hypercube  $HH_n$  can be embedded into a hypercube  $Q_n$  with dilation  $n$ .

*Proof* There exist two types of edges in half hypercube  $HH_n$ . Thus, we proved Theorem 5 by dividing it into two cases: h-edge and sw-edge.

Case 1 h-edge,  $1 \leq h \leq \lceil n/2 \rceil$

The address of node  $S'$  adjacent to node  $S(=s_n s_{n-1} s_{n-2} \dots s_{n/2} \dots s_3 s_2 s_1)$  of half hypercube  $HH_n$  via an h-edge is  $s_n s_{n-1} s_{n-2} \dots s_{n/2} \dots \bar{s}_h \dots s_3 s_2 s_1$  ( $1 \leq h \leq \lceil n/2 \rceil$ ). The address of node  $S'$  adjacent to node  $S(=s_n s_{n-1} s_{n-2} \dots s_k \dots s_{n/2} \dots s_3 s_2 s_1)$  of hypercube  $Q_n$  through a  $k$ -dimensional edge ( $1 \leq k \leq n$ ) is  $s_n s_{n-1} s_{n-2} \dots \bar{s}_k \dots s_{n/2} \dots s_3 s_2 s_1$ . Accordingly, the dilation of this embedding is 1 because the h-edge in half hypercube  $HH_n$  and the  $k$ -dimensional edge in  $Q_n$  are equivalent ( $1 \leq h, k \leq \lceil n/2 \rceil$ ).

Case 2 sw-edge

The sw-edge of half hypercube  $HH_n$  can be divided into two cases depending on the address of node  $S(=s_n s_{n-1} s_{n-2} \dots s_{n/2} \dots s_3 s_2 s_1)$ . Here, we prove the case with dilation  $n$ . If the  $n$ -address bits of node  $S$  are all binary 0, the  $n$ -address bits of node  $S'$  adjacent to node  $S$  through an sw-edge are all binary 1. The shortest path from node  $S(=s_n s_{n-1} s_{n-2} \dots s_{n/2} \dots s_3 s_2 s_1)$  to node  $S'(=s_n s_{n-1} s_{n-2} \dots s_{n/2} \dots s_3 s_2 s_1)$  in hypercube  $Q_n$  is the same as the path to which the  $k$ -dimensional edges are all applied. Therefore, a half hypercube  $HH_n$  can be embedded into a hypercube  $Q_n$  with dilation  $n$  ( $1 \leq k \leq n$ ).

## 5 Conclusion

This paper proposes a one-to-many broadcasting algorithm for the half hypercube interconnection network,  $HH_n$  that we proposed in [1], and proved that the one-to-many broadcasting time is  $n + 1$  when  $n$  is even and  $2 \times \lceil n/2 \rceil$  when  $n$  is odd in  $HH_n$ . We also showed that it is possible to embed an  $n$ -dimensional hypercube  $Q_n$  into an  $n$ -dimensional half hypercube  $HH_n$  with dilation 3, and that it is possible to embed  $HH_n$  into  $Q_n$  with dilation  $n$ . These results suggest that our half hypercube interconnection network  $HH_n$  has potential for implementation in large-scale systems for parallel processing.

## References

1. Kim JS, Kim M, Lee HO (2013) Analysis and design of a half hypercube interconnection network. ATACS 2013, LNEE. Springer, Heidelberg (will be appeared)
2. Kim M, Kim DW, Lee HO (2010) Embedding algorithms for star, bubble-sort, rotator-fabermore, and pancake graphs. HPCTA 2010, LNCS, vol 6082. Springer, Heidelberg, pp 348–357
3. Lee HO, Sim H, Seo JH, Kim M (2010) Embedding algorithms for bubble-sort, macro-star, and transposition graphs. NPC 2010, LNCS, vol 6289. Springer, Heidelberg, pp 134–143
4. Saad Y, Schultz MH (1988) Topological properties of hypercubes. IEEE Trans Comput 37(7):867–872
5. Seitz CL (1985) The cosmic cube. Commun ACM 26:22–33
6. Leighton FT (1992) Introduction to parallel algorithms and architectures: arrays, hypercubes. Morgan Kaufmann Publishers, San Francisco
7. Mendia VE, Sarkar D (1992) Optimal broadcasting on the star graph. IEEE Trans Parallel Distrib Syst 3(4):389–396
8. Feng T (1981) A survey of interconnection networks. IEEE computer 14:12–27
9. Bettayel S, Cong B, Girou M, Sudborough IH (1996) Embedding star networks into hypercubes. IEEE Trans Comput 45(2):186–194
10. Hedetniemi SM, Hedetniemi T, Liestman AL (1988) A survey of gossiping and broadcasting in communication networks. Networks 18:319–349
11. Hamdi M, Song SW (1997) Embedding hierarchical hypercube networks into the hypercube. IEEE Trans Parallel Distrib Syst 8(9):897–902

# Obstacle Searching Method Using a Simultaneous Ultrasound Emission for Autonomous Wheelchairs

Byung-Seop Song and Chang-Geol Kim

**Abstract** A method to locate an obstacle and calculate the distance to it is proposed. The proposed method utilizes multiple ultrasound emitters that generate signals of identical frequencies and intensities. Corresponding sensors detect the reflected ultrasound signals, and the position of the obstacle is calculated based on the time of flight (TOF) of the ultrasound wave. This method is suitable for autonomous wheelchairs as it facilitates detection of the nearest obstacle, and yields more accurate estimation of the position.

## 1 Introduction

Assistive technology (AT), which aids the disabled and elderly people, is garnering worldwide attention, and latest IT and robotics technologies are being integrated with AT, resulting in novel devices [1–3]. One such life-changing device for the severely disabled people who cannot transport themselves is the autonomous wheelchair, the commercialization of which is still impending.

The overall purpose of an autonomous wheelchair is to transport the user to a destination safely and precisely. To operate the device without external aids, autonomous wheelchairs employ voice recognition, automatic control, radar, navigation, and robotics technologies. These technologies were used in several blind guide systems [3–9]. An essential capability that every autonomous wheelchair must have for user safety is obstacle detection and automatic avoidance. The obstacle detection and automatic avoidance technology, which is also used in autonomous mobile robots, emits ultrasound waves, locate obstacles based on the detected reflection, and adjust the path to avoid them.

---

B.-S. Song (✉) · C.-G. Kim  
Department of Rehabilitation Science and Technology,  
Daegu University, 201 Daegudae-ro Jillyang, Gyeongsan, Gyeongbuk 712-714, Korea  
e-mail: bssong@daegu.ac.kr

An autonomous wheelchair uses multiple ultrasound sensors. Each sensor sequentially emits and detects ultrasound waves. In the traditional method, an obstacle in the direction of each sensor is detected based on the time of flight (TOF) of the wave, and a map of surrounding obstacles is generated [10, 11]. However, this method has a drawback in calculating accurate distances when the signal is emitted from a moving wheelchair because the error in obstacle detection grows as the source moves, and each emitter sends signals at different positions and times. One of many efforts to resolve this problem is by the error eliminating rapid ultrasonic firing (EERUF) technique designed at the University of Michigan [12]. EERUF controls the sequence of ultrasound emission to effectively reduce the error. This method shows significant reduction in the error; however, EERUF is not the ultimate solution as the sources of error, i.e., the delay time, still exists.

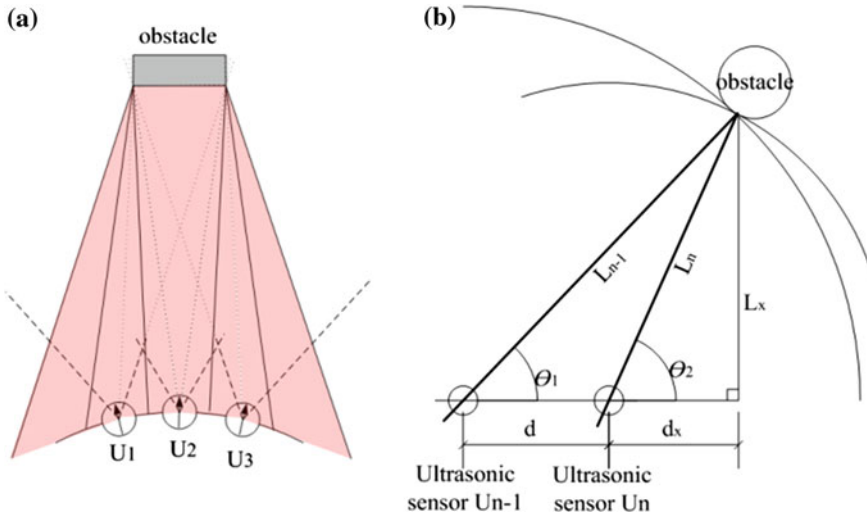
This paper proposes a simultaneous ultrasound emission technique that utilizes “crosstalk” as additional information to counter this problem. Crosstalk is an interference signal at a sensor in a multi-sensor configuration, which occurs due to simultaneous signal waves being emitted by many emitters. The proposed method emits ultrasound signals of identical frequency and intensity from all emitters at the same time, to reduce the error due to time delay that a sequential emission technique inherently generates. Each sensor detects the very first reflected signal from an obstacle, and in this way, the distance to the closest obstacle can be calculated with significantly improved accuracy levels.

## 2 Method

The proposed method emits simultaneous ultrasound signals of the same frequency and intensity from multiple emitters, and in the process detects the nearest obstacle. Each ultrasound sensor has an effective emission angle of  $60^\circ$ , and the searching system collects information from multiple sensors for the front of the wheelchair. For instance, the 3-sensor configuration shown in Fig. 1a, which considers the wheelchair motion and the overlapping area, can cover an angle of about  $120^\circ$ . The distance from the sensor to the detected object is calculated using the TOF method. The duration between the emission of the signal and the detection of the reflection is used in the distance calculation. Assuming that all the sensors emit signals at the same time, the reflected signal will be detected first at the closest sensor as shown in Fig. 1b. For an object at a longer distance, each sensor may capture reflected waves from other sensors that have longer flying time.

The traditional method regards any reflected signal from other emitting sources detected at a sensor as crosstalk noise. However, the proposed method considers crosstalk as additional information.

In the proposed method, the distance between each sensor and the obstacle is calculated as follows:



**Fig. 1** Example of obstacle detection with three ultrasound sensors: **a** Simultaneous emission and reflection pattern, **b** Distance calculation model

$$L_n = c \frac{T_{n,n}}{2} \tag{1}$$

$$L_{n-1} = c \frac{T_{n,n-1} + (T_{n,n-1} - T_{n,n})}{2} \tag{2}$$

where,  $c$  is the speed of ultrasound wave,  $L_n$  is the distance between sensor ‘ $n$ ’ and the obstacle,  $L_{n-1}$  is the distance between sensor ‘ $n-1$ ’ and the obstacle,  $T_{n,n}$  is the TOF for a wave emitted from sensor ‘ $n$ ’ to return to sensor ‘ $n$ ’ and  $T_{n,n-1}$  is the TOF for a wave emitted from the sensor  $n$  to return to the sensor ‘ $n-1$ ’.

### 3 Experiments

The proposed method was validated and compared with the traditional method using sensors installed on a wheelchair. The scan rate of the sensors used in the experiment was 30 ms, and the center frequency of the emitted ultrasound signal was 40 kHz. Five sensors were placed in the front portion of the wheelchair, 15 cm apart from each other, and a  $56 \times 62$  cm wooden panel was used as the obstacle.

The first measurements were taken using both the methods for a stationary wheelchair with the obstacle placed 2 and 3 m ahead. Next, the obstacle was placed 4 m ahead, and the distance was measured from the moving wheelchair as it approached 2 and 3 m distances from the obstacle at 90 and 120 cm/s. The test speed of the wheelchair was determined based on the speed during actual

operation, which is around 1 m/s. Each test segment was repeated three times, and the average distance was considered. For the scan method, ultrasound signals were sequentially emitted from the U1 sensor on the left to U2–U5 at an interval of 30 ms, and the distance was calculated using the measured TOF at each corresponding sensor. For the simultaneous emission method, distances from each sensor to the obstacle were calculated using the measured TOF of simultaneously emitted signals from each sensor. An 8-bit, ATmega 128 microprocessor was used to control the sensors, and to calculate the distances using the measured times. The microprocessor calculated the distances in real time and saved the data on a computer. Figure 2 shows the percentage error of the calculated distances to the measured values.

## 4 Discussion and Conclusion

As shown in Fig. 2, both methods yielded errors within 1–2 % with no significant difference for the stationary case, except for the 5th sensor that resulted in a larger error for the scan method. The moving wheelchair cases yielded larger errors with no noticeable correlation with the speed.

However, the errors still stayed within 2 % for the simultaneous emission method, while the scan method yielded generally larger errors, with some sensors showing much larger error jumps. The error jumps were significant on U1 and U5, and Fig. 2b shows the jump on U4 as well. It is inferred that the cause for the larger errors is the inherent time delay in the scan method. In general, the proposed method shows better results in terms of the error in distance calculation than the traditional scan method.

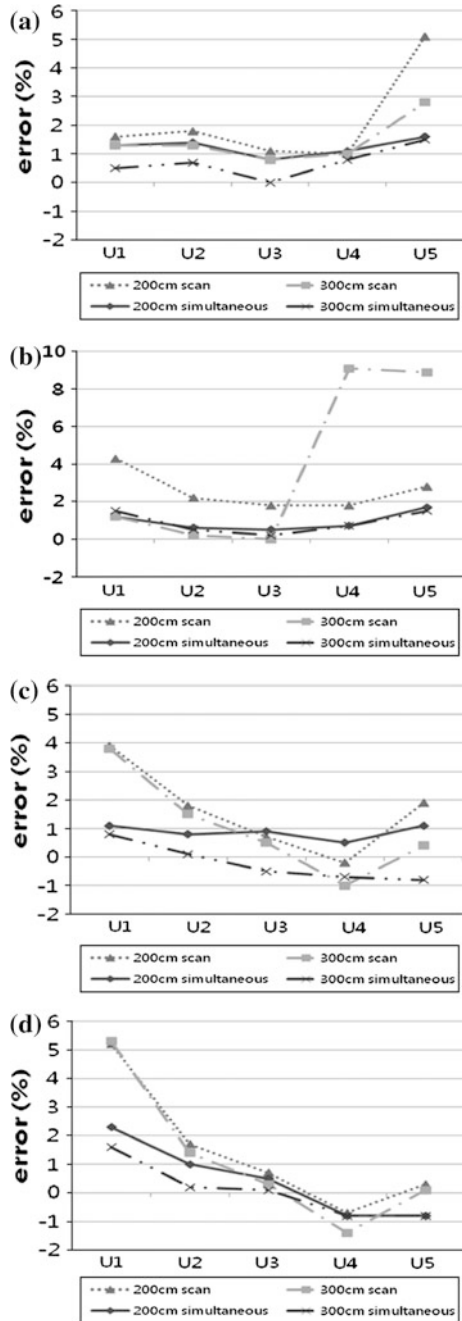
If the distances from three or more sensors to the obstacle are known, the obstacle can even be located in 3-D space. Assuming that each sensor is at the center of a sphere of which the radius is the distance between the obstacle and sensor, the intersection of the spheres that the sensors generate denotes the location of the obstacle. Therefore, with three sensors solving three sphere equations yields the location of the obstacle, which is an efficient method for an autonomous wheelchair. Since accurate estimation of distances is crucial in 3-D calculations, the proposed method is suitable for application to 3-D mapping of obstacles.

However, the proposed method is not superior to the scan method in every aspect. The proposed method detects only the closest object, while the scan method can detect multiple objects at the same time. Searching for multiple obstacles is beneficial because a wheelchair encounters many in a practical environment. On the other hand, from a user's perspective, information about the closest object is most important. Therefore, it is critical for the safe and convenient operation of autonomous wheelchairs to detect precisely and avoid the closest object.



**Fig. 2** Percentage error of calculated distances to measured values for various wheelchair speeds:

**a** Stationary, **b** 60 cm/s,  
**c** 90 cm/s, **d** 120 cm/s



In summary, the proposed obstacle detection method based on a simultaneous emission strategy exhibits excellent performances on autonomous wheelchairs.

**Acknowledgments** This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2010-013-D00091).

## References

1. Brabyn J, Gerrey W, Fowle T, Aiden A, Williams J (1989) Some practical vocational aids for the blind. In: Proceedings of 11th annual international conference of IEEE engineering in medicine & biology society, pp 1502–1503
2. de Acevedo RLM (1999) Electronic device for the blind. In: IEEE AES systems magazine, pp 4–7
3. Tachi S, Komorya K, Tanie K, Ohno T, Abe M (1981) Guide dog robot-Feasibility experiments with Meldog Mark III. In: Proceedings of 11th international symposium on industrial robots, pp 95–102
4. Tachi S, Tanie K, Komoriya K, Abe M (1985) Electrocutaneous communication in a guide dog robot (MELDOG). IEEE Trans Biomed Eng 32(7):461–469
5. Kim CG, Lee HG, Kang JH, Song BS (2007) Research of wearable walking assistive device for the blind. Korean J Vis Impair 23(1):147–164
6. Borenstein J, Ulrich I (1997) The GuideCane-a computerized travel aid for the active guidance of blind pedestrians. In: Proceedings of the IEEE conference on robotics & automation, New Mexico, pp 20–25
7. Shoval S, Ulrich I, Borenstein J (1998) The Navbelt-A computerized travel aid for the blind based on mobile robotics technology. IEEE Trans Biomed Eng 45(11):1376–1386
8. Kang JH, Kim CG, Lee SH, Song BS (2007) Development of walking assistance robot for the blind. J Korean Sens Soc 16(4):286–293
9. Rentschler AJ, Cooper RA, Blasch B, Boninger ML (2003) Intelligent walkers for the elderly: Performance and safety testing of VA-PAMAID robotic walker. J Rehabil Res Dev 40(5):423–432
10. Shoval S, Ulrich I, Borenstein J (2003) Robotics-based obstacle-avoidance systems for the blind and visually impaired, NavBelt and the GuideCane. IEEE Robotics Autom Mag 10:9–20
11. Moon CS, Do YT (2005) Design of range measurement systems using a sonar and a camera. J Korean Sens Soc 14(2):116–124
12. Borenstein J, Koren Y (1995) Error eliminating rapid ultrasonic firing for mobile robot obstacle avoidance. IEEE Trans Robotics Autom 11(1):132–138

**Part X**  
**Future Technology and its Application**

# A Study on Smart Traffic Analysis and Smart Device Speed Measurement Platform

Haejong Joo, Bonghwa Hong and Sangsoo Kim

**Abstract** In recent years, with a fast spread of smart phones, the number of users is rapidly increasing, causing the saturation of various kinds of traffic in mobile networks and access networks. This implies that although not only conventional Internet traffic but also many smart phone applications create much traffic, the development of network analysis technologies has yet to be furthered. As a method to analyze and monitor the traffic types of smart devices, the method of using signature and classifying traffic is used, and the speed measurement agent is used to measure the Internet speeds in smart devices. This paper aims to classify the smart device traffic and Internet traffic, analyze the traffic use amount, measure Internet speeds in smart phones amid traffic being created, and thus propose a platform designed to measure user service quality.

**Keywords** Smart device traffic analysis · Smart device speed measurement · Traffic and speed monitoring

---

H. Joo

Department of LINC, Dongguk University, #710, 82-1 Pil-dong 2-ga,  
Jung-gu, Seoul 100-272, Korea  
e-mail: hjjoo@dongguk.edu

B. Hong (✉)

Department of Information and Telecommunication, Kyung Hee Cyber University,  
1 Hoegi-Dong, dongdaemun-Gu, Seoul 130-701, Korea  
e-mail: bhhong@khcu.ac.kr

S. Kim

Contents Vision Corp, #613, 82-1 Pil-dong 2-ga, Jung-gu, Seoul 100-272, Korea  
e-mail: cqsky@paran.com

## 1 Introduction

In recent years, with the wireless Internet rapidly growing, there is a growing access to broadband services through new devices such as smartphones, netbooks, and mobile Internet devices (MIDs). Online video service traffic is increasingly spreading to mobile services through these devices, significantly burdening the network [1].

Studies on mobile traffic reveal that mobile traffic is characterized by a high ratio of HTTP traffic and many applications with the client–server structure [2]. To classify these mobile applications, information on each mobile application, like with conventional traffic classification, needs to be gathered and analyzed. However, mobile devices, unlike general PCs, are very slow in handling speeds, and have limited memory. Thus, this paper proposes a configuration designed to classify and analyze mobile traffic in the general PC environment, as well as a platform designed to measure the traffic handling speed in smart devices through speed measurement agents.

This study proposes a platform by which a monitoring device is installed in the access network to monitor the traffic of both smart devices and Internet so as to classify traffic and to measure the speed of the WiFi Internet services in smart devices.

## 2 The Proposed Structure of Traffic Analysis and Classification

This Chapter explains the proposed traffic analysis structure. [Section 2.1](#) describes the network traffic gathering structure, and [Sect. 2.2](#) explains the traffic analysis structure which uses the open source snort.

### 2.1 Gathering of Traffic Data

The method of extracting flow data according to the environment of measurement points is outlined as follows [3, 4].

- Agent method: the method designed to set a flow monitoring function in the data gathering device to gather flow data.
- Probe method: the method designed to use the tapping or monitoring method and extract the flow information instead of extracting flow information directly from the gathering device.

In order to gather raw data flowing in the network and analyze them, this study uses the probe method and gathers traffic data (Fig. 1).

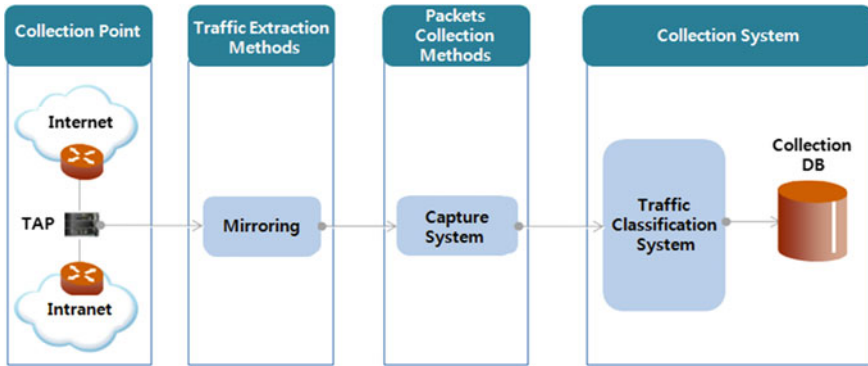


Fig. 1 Traffic data collection

## 2.2 Classification and Analysis of Smart Traffic

This study uses packet sniffer and packet logger among snort functions and classifies raw data gathered from the gathering system [5, 6].

The following explains the general functions of snort.

- Packet sniffer: the function designed to read and show packets of the network
- Packet logger: the function designed to store monitored packets and leave them in the logger
- Network IDS: the function to analyze network traffic and explore attacks

This study uses the packet sniffer function, transforms network traffic (TCP, UDP, ICMP IP, etc..) into easily analyzable patterns, and uses them in analyzing network properties (Fig. 2).

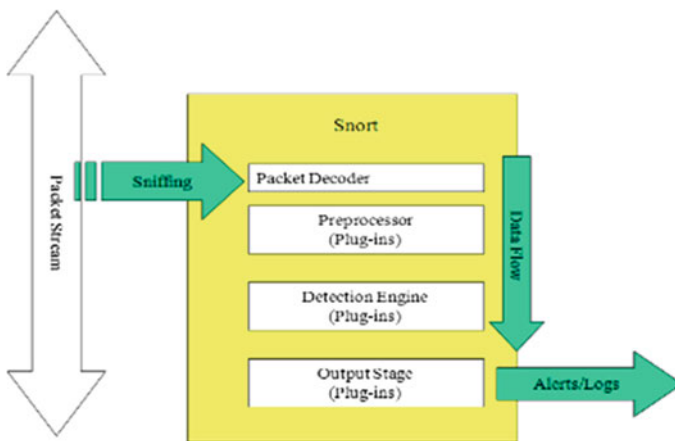


Fig. 2 Structure of SNOT

**Table 1** Classification of traffic using the snort signature

Rule header							Rule option
Action	Protocol	IP address	Port	->	IP address	Port	(Option)
1	2	3	4	5	6	7	8

### 2.3 Classification of Traffic Using the Snort Signature

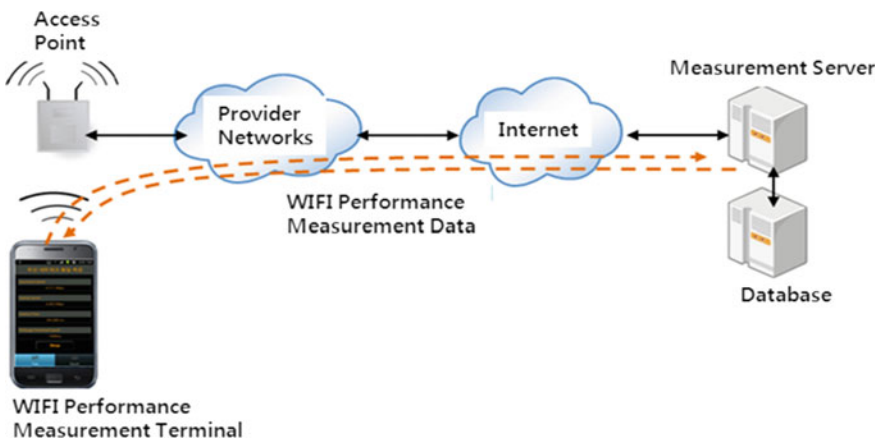
Data, processed through snort packet decoder, are classified through the preprocessor, and this study uses the snort signature’s headers and rule options and classifies data [3]. The gathered data can be classified through two methods, namely, the method of classifying packets on the payload, and the statistical method. In order to analyze the content of data, this study uses the method of classifying packets on the payload (Table 1).

## 3 Structure of Measuring Speeds in Smart Devices

This study explains the speed measurement structure in smart devices. However, since 3G and 4G networks are separated from wire data networks, only Wi-Fi service speeds are measured (Fig. 3).

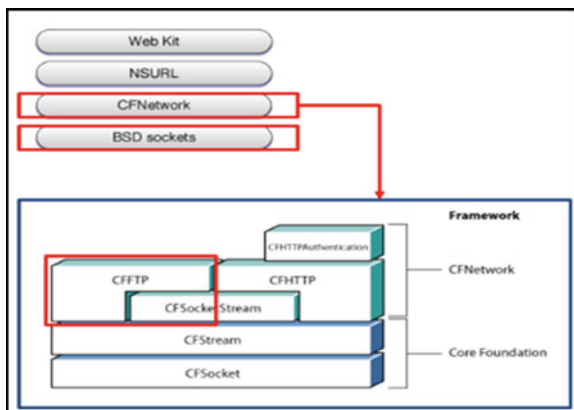
The Wi-Fi quality indices are measured using the indices provided by the National Information Society Agency.

In order to measure the speeds in I-Phone, this study uses the CFNetwork and DSC socket standard SDK (Fig. 4).



**Fig. 3** The Wi-Fi speed measurement environment in smart devices

Fig. 4 API structure



### 4 Structure of Traffic Analysis and Speed Measurement

Figure 5 shows the structure of the entire system for the traffic analysis and speed measurements in smart devices.

- Quality measurement server: Use the server, and measure downloading, uploading and delays in order to measure the quality linked to smart devices.
- TAP: Device designed to gather data flowing between the school network and the Internet.
- Monitoring system: The system designed to analyze data flowing in through TAP and to monitor the performance of network.

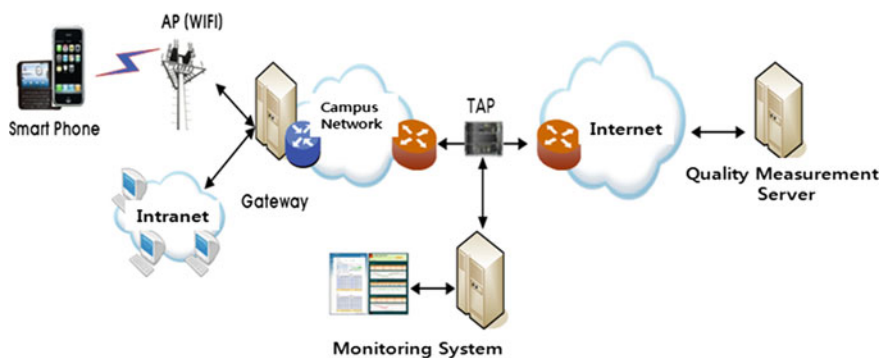


Fig. 5 The total system structure



## 5 Conclusion

With many smart device applications currently being served, diverse traffic patterns are created. Thus, it is increasingly important to analyze mobile traffic in the Internet environment.

To monitor the mobile and wire traffic, this study proposes a method designed to use the snort signature, classify and monitor data, as well as an environment in which to measure the performance of smart devices in the monitoring environment. When huge data flow into the network from smart devices, the method can analyze how the performance in smartphones will change. The proposed method is expected to apply to mobile networks (3G, 4G).

**Acknowledgments** This research was supported by the Kyung Hee Cyber University a sabbatical Year in 2013.

## References

1. National IT Industry Promotion Agency (2011) Weekly Technical Trends, July 15
2. Maier G, Schneider F, Feldmann A (2010) A first look at mobile hand-held device traffic. Passive and active measurement. Zurich, Switzerland, pp 161–170, Apr 7–9
3. WAN and Application Optimization Guide, chapter 5 ‘Traffic classification’
4. National IT Industry Promotion Agency (2012) Weekly Technical Trends, April 11
5. ITU-T Recommendation I.350 (1993) General aspects of quality of service and network performance in digital networks, including ISDNs, Mar 1993
6. ITU-T Recommendation Y.1541 (2003) Network performance objectives for IP-based services, Feb 2003
7. ITU-T Recommendation Y.1543 (2007) Measurements in IP networks for inter-domain performance assessment, Nov 2007

# Analysis and Study on RFID Tag Failure Phenomenon

Seongsoo Cho, Son Kwang Chul, Jong-Hyun Park  
and Bonghwa Hong

**Abstract** RFID tag failure analysis test involves analysis of general devices and interpretation of analysis results, which suggests failure mechanism. Hence, it can only be performed by failure analysis experts with rich expertise and experience. Interpretation of causal relationship between a failure phenomenon and explainable causes of the failure has to rely on failure analyzers' knowledge and experience. Analyzers have the capability to figure out the precise failure mechanism because they know which type of a failure is associated with which failure mechanism. Failure mechanism by major failure type frequently observed in RFID tag and analysis method should be fully recognized by analyzers.

**KeyWords** RFID · Failure site · ESC · Solvent crack

## 1 Introduction

RFID (Radio Frequency Identification) is a non-contact identification system in which a small chip is attached on an object to transmit and process the object's

---

S. Cho (✉) · S. K. Chul · J.-H. Park  
Department of Electronic Engineering, Kwangwoon University,  
20 Kwangwoon-ro, Nowon-gu, Seoul 139-701, Korea  
e-mail: css@kw.ac.kr

S. K. Chul  
e-mail: kcson@kw.ac.kr

J.-H. Park  
e-mail: world78u@hanmail.net

B. Hong  
Department of Information Communication, Kyunghee Cyber University,  
Dongdaemun-gu, Seoul 130-701, Korea  
e-mail: bhhong@khcu.ac.kr

information wirelessly. Recognized as the most dramatic data identification technology, RFID is rapidly advancing and taking a firm position in the market driven by advances in semiconductor and wireless communication as well as global standardization. The minute semiconductor chip embedded in RFID helps transmit information about an object and the surrounding environment via radio-frequency. It consists of a tag, reader and antenna, etc. RFID tag can be automatically identified and tracked anywhere, anytime and the embedded memory enables information to be updated and revised. Furthermore, unlike a bar code limited by environmental factors like rain, snow, fog and pollution and a smartcard only identified from a short distance, RFID tag can be identified even in the most constrained environment. It can even identify moving objects [1–4].

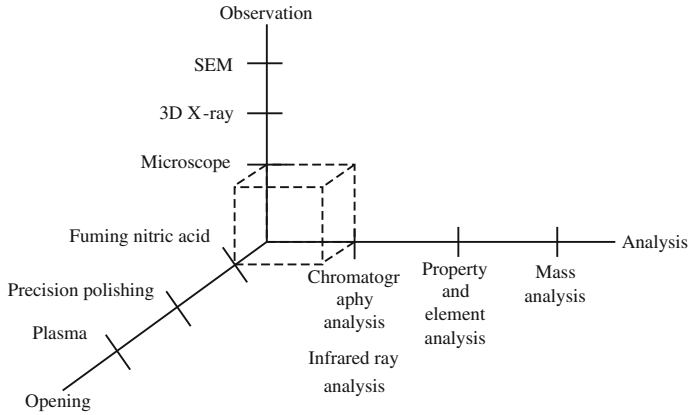
RFID tag's failure analysis, which is a specialized area with long history and evolved with advances in science and technology, has apparently made a big contribution to delivering mutually beneficial output in both physics-of-failure and reliability tests. As one of the key factors for evaluating a product's improvement process and reliability, it focuses on finding the root cause of issues regarding quality or reliability of products or parts. The ability to secure stable performance of new parts and materials is critical since it is directly related to reliability of parts and materials. But, loss cost can be kept to minimum just by forecasting the most vulnerable areas and type of failures for further improvement [5–9].

Failure analysis based on physics-of-failure adopted a comprehensive approach to figure out root causes of a failure. Failure mode and failure site can be identified through data analysis gained from performance evaluation, nondestructive analysis and destructive physical analysis technology, based on which failure mechanism, which is the mechanical, electrical and physical process causing a failure, can be revealed. The guidelines in this study proposed failure analysis methods for electric and electronic parts based on physics-of-failure.

## **2 Failure Analysis**

### ***2.1 Basic Technology of Failure Analysis***

Failure analysis technology is still far from complete. Rather, development is still ongoing in all areas and electronic parts, in particular. In advanced technology, ultramicro analyzer, atom-level analyzer and device for property of a matter and failure analysis (computational analysis) is necessary. Worse, there are many areas that do not render analysis and it even tends to fall behind the speed of advance in electronic parts. Basic elements of failure analysis are observation, opening and analysis. Observation may be taken for granted but it is actually the start of a failure analysis since it takes a special perspective. Opening is to observe the interior structure of electric and electronic parts. To date, there is no technical publication on opening as opposed to heaps of literature and studies that can be



**Fig. 1** Three elements of failure analysis

used as a reference for product manufacturing. Being such, the whole process of disassembling a part one by one without destructing for observation is still quite a challenge. Furthermore, even the most advanced technology to date cannot open all parts. Analysis is a technology required to check chemical change but there is no way to analyze a minute area and its chemical composition with a simple device. For now, an expensive device analyzer is the only available means for analysis. The basic elements are shown in Fig. 1 and the area surrounded by the dotted line is the only scope in which failure analysis can be carried out with a simple tool.

Failure analysis starts off with exterior observation of a specimen, property measurement, and observation of the interior structure on a sequential basis and finally discovers a failed site or the cause of a failure. Failure analysis technology refers to observation; opening (disassembly) to observe the structure interior; and finally analysis to investigate the cause of a failure.

## 2.2 Cause of a Failure

In general, failures don't occur by accident. Rather, major factors of a failure include stress, material, structure and geometry. In other words, a failure can occur by many causes, which can be tracked. A failure can be classified into initial failure, accidental failure and abrasion failure depending on the period failure occurs during product operation. Initial failure is mainly caused by a manufacturer, which can be prevented with quality control and screening. Accidental failure is mainly caused by user's misuse or poor design. Abrasion failure is mainly caused by poor reliability or durability that was not properly considered by the designer.

### 2.3 Data Analysis

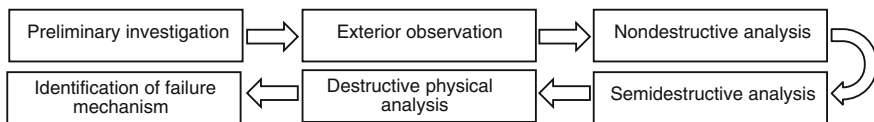
Data analysis is a process to analyze data collected using an integrated technology and put them into a system. Cause and Effect Diagram, Ishikawa Diagram or Fishbone Analysis specifically provides conditions to find the root cause of a failure or all the possible causes related to the failure. Failure mode and effect analysis is a qualitative reliability forecast and failure analysis method surveying impact of failure mode of parts implementing a system on other parts, system and users with a bottom-up approach. It is a tool used in the design and process stage. Failure tree analysis uses logic symbols (AND or OR) on causal relationship between a system's failure phenomenon and its cause to draw a failure tree shaped in tree branches. It is a quantitative failure analysis and reliability evaluation method adopted to improve system's reliability by calculating system's failure probability. Pareto diagram indicates issues that potentially require the most improvement by showing relative frequency or critical [10, 11].

### 2.4 Procedure of Failure Analysis

Order of failure analysis is extremely critical. Improper order of failure analysis could ruin analysis just as a product manufacturing requires a certain manufacturing order. Unlike products, which can be remanufactured, failure analysis cannot be conducted again, in most cases, if the analysis is not done in the right order because the work designed to identify failure mechanism includes the nonreciprocal destructive element. Order of failure analysis and its details differ from one electric and electronic part to the other but the general procedure for failure analysis is shown in Fig. 2.

## 3 Analysis of Environmental Stress Cracking

Environmental stress cracking (ESC) refers to when a stress (environmental stress) works on a solvent crack phenomenon at the same time to cause uniform destruction in a short period of time. Stress is divided into mechanical stress, which is external, and residual stress that exists inside the high molecules during



**Fig. 2** Procedure of failure analysis

forming processing, which is internal. Residual stress that occurs in the interior during formation exists in high-molecular materials and these high molecules in which residual stress exists can face destruction from even the slightest external stress when they contact chemical substances like solvent, release agent, surfactant and machine oil.

The general difference of crack shape of solvent crack and ESC lies in the shape of failure surface and high-molecular materials even though distinguishing one from the other is not straightforward. Shape of a failure surface is a solvent crack (a smooth failure surface on a mirror) and ESC (brittle fracture or mixture of brittle fracture and failure surface on a mirror). High-molecular material is solvent crack (on non-crystalline high molecules) and ESC (on non-crystalline high molecules with relatively higher crystallizability than solvent crack).

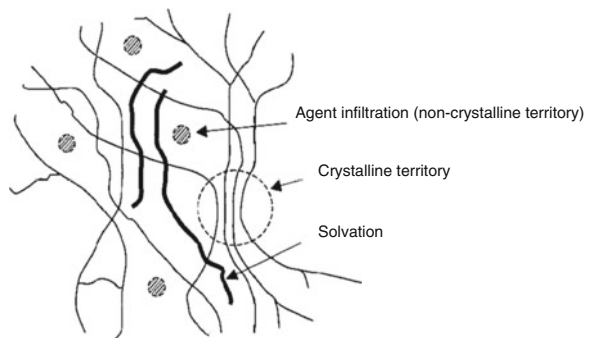
### 3.1 Failure Phenomenon

ESC incurs a crack failure mixed with solvent crack as environmental stress simultaneously acts on the solvent crack.

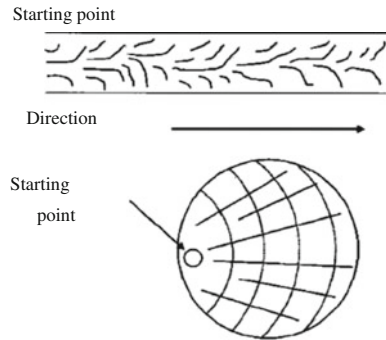
Solvent crack refers to a crack caused by contact with plastics and chemicals (oil, detergents and blooming of other high-molecular additives) under stress situation (residual stress during forming processing, etc.) and it is largely observed in non-crystalline plastic (PC, PMMA, ABS, etc.).

Craze phenomenon occurs in high molecules when a certain level of critical stress acts and molecules locally take an orientation (molecules are locally elongated and arranged towards the same direction). A micro destruction occurs in the crazed area, which later cracks. High-molecular products usually have residual stress resulting from forming processing even when there is no external stress. The mechanism working on solvent crack as shown in Fig. 3 occurs when chemical agents infiltrate into high molecules caused by weakening of van der Waals force between molecules of high-molecular substances due to environmental stress. The solvent crack shows a mirror-face shape (smooth gloss).

**Fig. 3** Type illustration of molecular structure after solvent crack



**Fig. 4** Plane shape of brittleness fracture



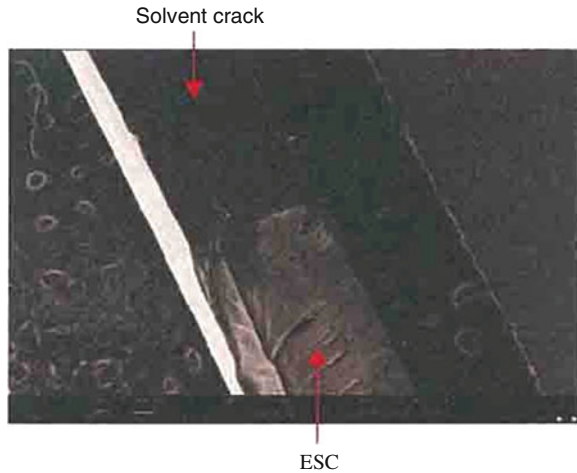
When solvent contacts with high molecule surroundings, it creates a strong bond with high-molecular substances, which separates mutual cohesion of high-molecular substances and causes solvent to infiltrate in between high-molecular substances and swell. In extreme cases, it could dissolve the substances. Naturally, the degree of solvent crack differs depending on whether chemical substances just infiltrated into high-molecular substances or actually caused dissolution. ESC behavior of high molecules varies by exposure time, exposure temperature, and density of stress destructive materials and stress load degree of high molecules.

### ***3.2 Analysis Method***

Failure analysis caused by ESC took a relatively simple process that involved analysis of the quality of material, observation of fracture plane, analysis of fracture plane's surface component and reproducibility test. Fractured plane is observed with SEM or stereoscopic microscope and interpretation of fracture plane shape can help analyze the direct cause of fracture. Brittle fracture is caused by shock fracture as shown in Fig. 4. In Chevron Pattern shaped like a chevron, the starting point is in chevron peak and crack occurs in the opposite direction to the peak as indicated in  $< \rightarrow < \rightarrow < \rightarrow$ .

Striation shaped like a shell has a unique stripe otherwise called beach mark. It shows a failure surface. Striation, which is produced by repeated stress, occurs when crack develops and recedes on a regular basis. Small solvent crack produced from a contact surface or a surface infiltrated with a solvent usually occurs vertically to the direction where a mechanical load works. The crack surface is mostly smooth like a mirror shown in Fig. 5 and is thus dubbed as mirror face.

**Fig. 5** Solvent crack and ESC section



**Fig. 6** Exterior observation of clothing tag **a** front side of clothing tag **b** back side of clothing tag



**(a)** Front side of clothing tag **(b)** Back side of clothing tag

## 4 Case of RFID Tag Failure

### 4.1 Exterior Observation

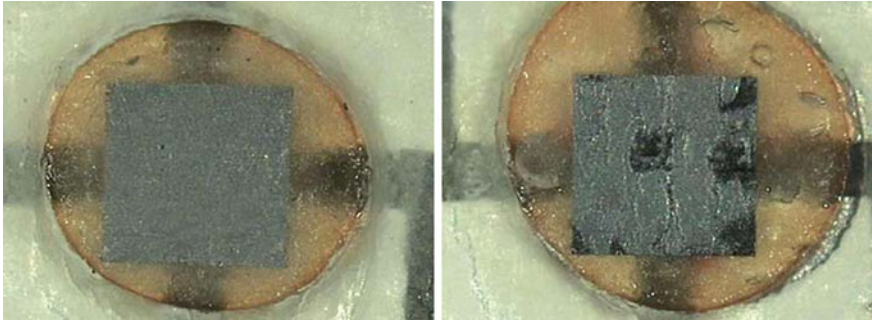
When a product code is printed on RFID tag chip, fault occurs due to chip's volume and foreign substances appear on the back of where chip is attached (Fig. 6).

### 4.2 Microscopic Observation

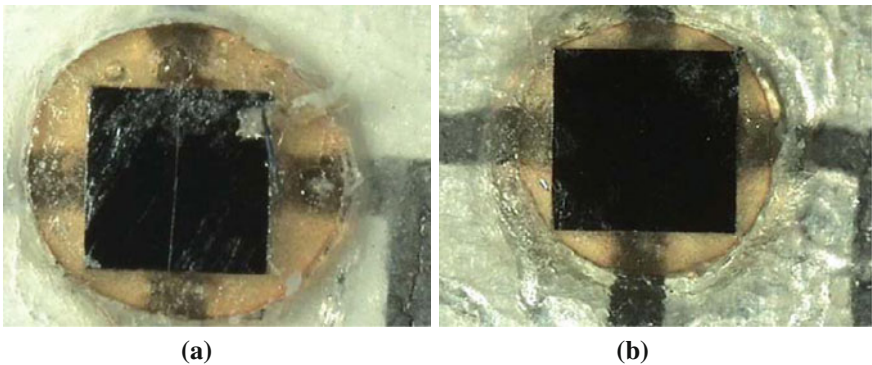
Figures 7 and 8, a detailed observation of chip condition via optic microscope.

As presented in Fig. 7, it was hard to observe chip due to adhesive used to apply ACP and attach a tag. Crack in (a) poor specimen in Fig. 8 could be observed with optic microscope but crack could not be observed with optic microscope after removing ACP and adhesive in (b) normal specimen in Fig. 8.





**Fig. 7** RFID chip observation via optic microscope



**Fig. 8** Chip surface ACP and removal of adhesive **a** poor specimen **b** normal specimen

## 5 Conclusion

The cause of RFID failure and development of the failure that prompted the cause of the failure to turn into failure phenomenon can be explained as failure mechanism. Analysis tests of failure analysis can be performed by anyone who can analyze general equipment and has good understanding of analytical chemistry. However, interpretation of analysis results to suggest mechanism of a failure is a job that can only be carried out by failure analysis experts built with expertise and experience. Interpretation on the causal relationship between a failure phenomenon and explainable failure causes has to rely on knowledge and experience of failure analyzers and thus a rich understanding on numerous failure mechanisms by each type of failure is required to come up with the precise failure mechanism. In this sense, failure mechanism by major type of failure frequently observed in RFID tag and ways of analysis should be properly understood and ways to combine and summarize results gained from procedures of failure analysis should hopefully put to good use in industries.

Series of test results associated with failure analysis are limited in suggesting improvement measures capable of suppressing or preventing failures. Furthermore, there is potential of proposing measures that have serious error since the proposed improvement measures are based on inferences not demonstrated with tests. Even so, failure analysis is meaningful since it deals with improvement measures aimed at suppressing or preventing a failure to identify causes of a failure and their development into a failure phenomenon. Hence, it plays an effective role in developing alternatives to suppressing or preventing failures.

## References

1. Subramanian V, Chang PC, Huang D, Lee JB, Molesa SE, Redinger DR, Volkman SK (2006) All-printed RFID tags: materials, devices, and circuit implications. *VLSI Design*
2. Leachman RC, Hodges DA (1996) Berkeley semiconductor manufacturing. *IEEE Trans Semicond Manuf*
3. Bloomsburg J (2002) RFID tag manufacturing. MIT UROP
4. RFID Guardian (2005) A battery-powered mobile device for RFID privacy management. In: Proceedings of the Australasian conference information security and privacy. [http://www.cs.vu.nl/~melanie/rfid\\_guardian/papers/acisp.05.pdf](http://www.cs.vu.nl/~melanie/rfid_guardian/papers/acisp.05.pdf)
5. Wagner U, Franz J, Schweiker M (2001) Mechanical reliability of MEMS-structures under shock load. *Microelectron Reliab* 41(9–10):1657–1662
6. McCluskey P (2002) Design for reliability of microelectro-mechanical systems. In: Proceedings of the electronic components and technology conference, pp 760–762
7. Müller FR, Wagner U, Bernhard W (2002) Reliability of MEMS- a methodical approach. *Microelectron Reliab* 42(9–11):1771–1776
8. Anderson TL (1995) Fracture mechanics: fundamentals and applications 2nd edn. CRC Press, Boca Raton
9. Finn J (1995) Electronic component reliability. Wiley, USA, pp 21–153
10. Yiping W (2004) RFID tag manufacturing. *Glob SMT Packag* 4(6):4–7
11. An B, Cai XH, Chu HB (2007) Flex reliability of RFID inlays assembled by anisotropic conductive adhesive. In: Proceedings of the 9th International IEEE CPMT Symposium on high density design, packaging and microsystem integration, pp 60–63

# Administration Management System Design for Smart Phone Applications in use of QR Code

So-Min Won, Mi-Hye Kim and Jin-Mook Kim

**Abstract** With the development of information and communication technology, the social infrastructure is secured in order to pass and use anytime, anywhere the information we want. Smart phone application occupies an important place in Ubiquitous Environments. More efficient and sensible way are needed across the medical field taking and getting the prescribed medications according to the disease. This paper present the smart phone application based medication management system using the QR code. We can prevent the duplication prescription and reduce the side effects such as exasperating the disease from cases of overdose or under-taking due to the failure of recognizing the times or time of taking the medication or augmenting immunity to the drug schedule through this model. Also, More efficient and sensible research are needed such as connection problems between the most basic home diagnostic equipment and smart phone application, the algorithm about avoiding drug duplication, time management algorithm of taking the medication and medication delivery algorithm base on this model.

**Keywords** Medication management system · Mobile · Smartphone application · QR code

---

S.-M. Won (✉) · M.-H. Kim

Department of Computer Science, Chungbuk National University, Cheongju-si, South Korea  
e-mail: wsm012@nate.com

M.-H. Kim

e-mail: mhkim@cbnu.ac.kr

J.-M. Kim

Division of Information Technology Education, Sunmoon University,  
Cheonan, South Korea  
e-mail: calf0425@sunmoon.ac.kr

## 1 Introduction

Together with the development of information and communications, the social infrastructure ensuring provision, receive and confirmation of information without time and locational limitations has been secured. Under such circumstances, smart phone applications have taken significantly important positions in the modern society. Among the medical business, a more efficient and rational measure is required in the overall process of drug prescription and administration in line with diseases.

This study proposes a model for drug administration management system for smart phone applications. This model is expected to prevent duplications in the drug prescription process and to reduce cases of worsening diseases due to overdose or under-dose from recognition failure for drug dose number of times or time or enhancing immunity to certain drugs.

In addition, a study is required based on this model regarding the relation issue between most basic domestic diagnosis apparatus and smart phone applications, algorithm for drug redundancy prevention, time management algorithm for drug dose and concrete and efficient algorithm for delivery system after taking prescription.

Together with the development of information and communications, the social infrastructure ensuring provision, receive and confirmation of information without time and locational limitations has been secured. Under such circumstances, smart phone applications have taken significantly important positions in the modern society. Among the medical areas, a more efficient and rational measure is required in the overall process of drug prescription and administration for diseases.

In a case where a patient gets medical services from different private clinics for several diseases, he or she might be prescribed for an identical drug by each clinic and then, redundant dose might be caused for the patient due to drug prescription redundancy. In the process of prescription drug dose, it is practically difficult to make drug dose in a punctual manner and there are many cases it is difficult to make correct judgment on whether dose times are well abided by. Such cases are unanimous regardless age and gender. Therefore various side-effects might be caused in disease treatment attributable to drug overdose from redundant dose caused by incorrect recognition drug dose time and number of times; drug dose deficiency from drug dose omission caused by recognition failure of drug dose and excessive drug dose caused by repetitive dose.

Therefore this study proposes a model for smart phone applications based on the drug administration management system using QR code which is a new method for prevention of disease worsening while providing proactive contributions to treatment of patients through effective administration management. With application of ubiquitous concept, this model enables patients to get administration information and treatment whenever and wherever; to receive individual prescription through their smart phone by using QR code; and execute administration management based on transferred data.

## 2 Relevant Researches

Largely attributable to recent development of information and communication technologies, Korea has witnessed active researches for technology convergence in each professional area with digital devices based on quality high-speed communication infrastructure better than those in other countries. The medical industry is yearning for Ubiquitous Healthcare (hereinafter referred to as U-Health Care) for monitoring individual's health status 'wherever and whenever' and providing customized health management service by utilizing development of digital device and wireless communication technologies and various kinds of bio signal measuring sensors which are small and portable. In addition, various solutions have been provided for automatic medical services. The representative example is a method using RFID with a purpose to prevent relapse of disease and following re-hospitalization with management of defined drug therapy. Researches are conducted to prevent side effects from redundant drug doses by patients attributable for redundant prescription in the process of administration prescription.

U-Health Care enables individuals not to recognize medical services since U-Health Care Service by itself makes real-time monitoring on individual's health status and automatically takes actions at a time requiring treatment or management. Therefore individuals are able to sustain best health conditions and get convenient and precise medical service without time and location restriction when it is required. Moreover for well-being which is considered most interested item of people; and for preparation against upcoming 'aged society', conventional level of medical service is not appropriate and a more sophisticated medical service focusing on management and prevention is in a desperate need.

U-Health Care refers to health management and medical service provided by collection, process, delivery, and management of health related information without time and location limits which is ensured by application of applies information and communication technologies (ubiquitous computing) to the public health and medical industry. Ubiquitous computing refers to a condition realizing computing whenever and wherever with various computers coexisting with human beings, physical objects and environment and linked each other.

In other words, it might be considered as a comprehensive medical service ranging from remote management of patient's disease to daily health management as it is ensured by application of information and communication technologies such as wired and wireless networking to the medical industry for convenience and efficiency of medical service usage [1].

U-Health Service is a business providing services such as self-diagnosis, remote monitoring and management of medical professionals based on real time transfer of health index of patients such as blood pressure, blood sugar, pulse, and body fat to medical institutions in use of wired or wireless health measurement devices linked by network of cell phones, PDA, and the Internet [2].

Studies conducted to prevent potential redundant description which might be occurred in the process of drug prescription in hospitals are as follows;

In a case where a patient gets drug prescription in more than two hospitals, redundant prescription gets highly likely with increased dose opportunities of various drugs so that the measures to secure medical safety against side effects of drugs or overdose. Medical institutions have recently tried measures to enhance quality of drug treatment with systematic supports to prevent and pre-check potential errors which might be caused in the process of drug prescription of doctors such as drug overdose, drug allergy, drug prohibited during pregnancy and breast feeding, drug side effects, drug interactions, prohibited drugs by age group and dose control for kidney patients in a way of application of Clinical Decision Support System (CDSS) to Computerized Physician Order Entry (CPOE) [3].

The representative example of provision of automatic medical services is an administration management in use of RFID. The introduction of automatic solution using RFID embedded cell phones, tag and web based servers. The target of this solution is to reduce the number of heart patients who are re-hospitalized from failure to follow defined drug therapy. To discharged patients, eMedonline cell phone attached with RFID reader is provided. The cell phone has eMedonline software application and is linked with web based server. In addition, the patient gets RFID label with names of prescribed drugs. At a time of drug dose, the patient gets information on which drug should be taken vis eMedonline cell phone. It is followed by questions on dose time, change prescription and his or her health status and those questions can be responded through the touch screen.

When a patient approaches prescribed drug bottle with a tag to RFID reader of a cell-phone in couple of inches, the reader reads ID No. of the tag and transfer to the web-based sever. EMedonline software checks whether this ID No. matches with the drug in which the patient should take and displays the photo of relevant pills on the phone to help the patient identify the correct drug while it also saves information on the level of drug therapy obeisance of the patient, time, health status and behaviors. When a patient fails to take the prescribed drug, relevant message is transferred to a hospital and the hospital is able to notify such failure to a guardian of the patient.

### 3 Proposed Model

This study proposes an administration management system which is able to provide more effective medical services through convergence of various communication devices and technologies for medical services based on high-speed communication network infrastructure.

The system in Fig. 1 calls data of patients saved in DB pool of a hospital server and transfers it to a pharmacist to help drug prescription and supports patients to get messages on alarms and information in line with the cycle of prescribed drugs via smart phone. In addition, patients are able to take drugs of their prescription at home rather than taking it in pharmacy. By doing so, inconvenience and time for sick patients to get to pharmacy to take drugs can be minimized. A patient who

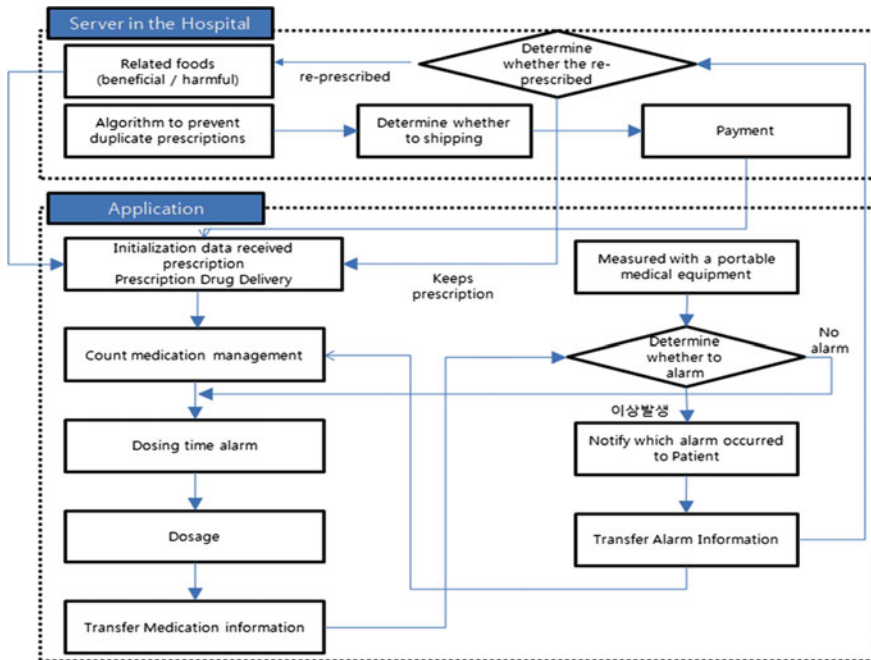


Fig. 1 Flow of administration management system

received prescription, food-related information, and cautions related to prescribe drugs also are entitled to measuring device information, prescription information, administration alarm and relevant food function services provided by smart phone applications.

Each function provided by smart phone applications is structured as data collection stage and analysis stage as in Fig. 2.

In the data collection stage, check can be made for prescription information of patients transferred from hospital server, relevant food information which should be referred to depending on disease severity of each patient, drug administration timing and contents in accordance with prescription, basic measuring information on portable medical devices like blood sugar device and link status of medical devices.

The data analysis stage can identify dose time, remaining times, omission of administration and abnormality occurrence with administration record as an information generated and provided based on information collected in the data collection stage.

The system is categorized into hospital server and smart phone applications and functions of each process in the hospital server are as follows;

Data Collection and Calibration carries out a function collecting administration timing and relevant information of patients from smart phone applications of patients. As it is the case for Data Collection, Alarm Collection and Notification

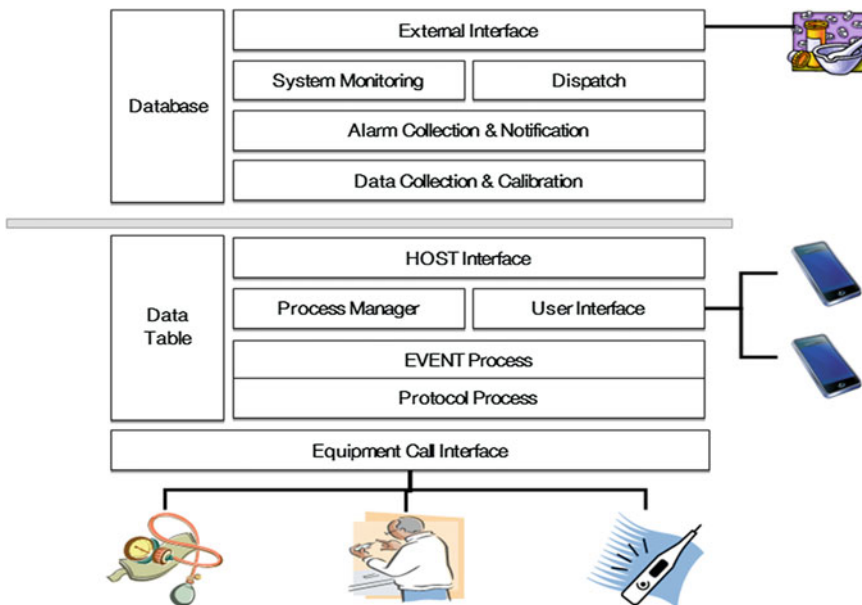


Fig. 2 System architecture

collect data from smart phone applications but for abnormality occurrence related to administration and alarm information. System Monitoring executes functions of activating whole process, operation supervising and monitoring. Dispatch check redundant prescription for prescriptions issued by doctors, and generates data on which delivery information can be transferred to pharmacists collected via smart phone applications of patients. External Interface executes a function of transferring data generated by Dispatch to pharmacists.

Next, the process diagram for smart phone is as follows;

Equipment Call Interface carries out a role of data transfer and receive between portable medical devices like blood sugar device used by patients and smart phones. Protocol Process as a sub-element of Equipment Call Interface plays a role of Protocol converter which is able to deal with individual transfer methods of each portable medical device. Event Process check abnormality based on patients data collected by Equipment Call Interface and transfer the information to hospital server though HOST Interface which is for data transfer and receive with hospital server. Process Manager activates whole processes of smart phone and executes monitoring on operation supervise. User Interface functions data illustration through which patients are able to search and check relevant data and alarm on screen.

Figure 3 is about functions in smart phone applications. It is able to save data in smart phone applications which is record by linking measuring devices for blood pressure, body temperature and blood sugar with smart phone; to analyze such data, notify hospital server when abnormality is occurred and to request re-



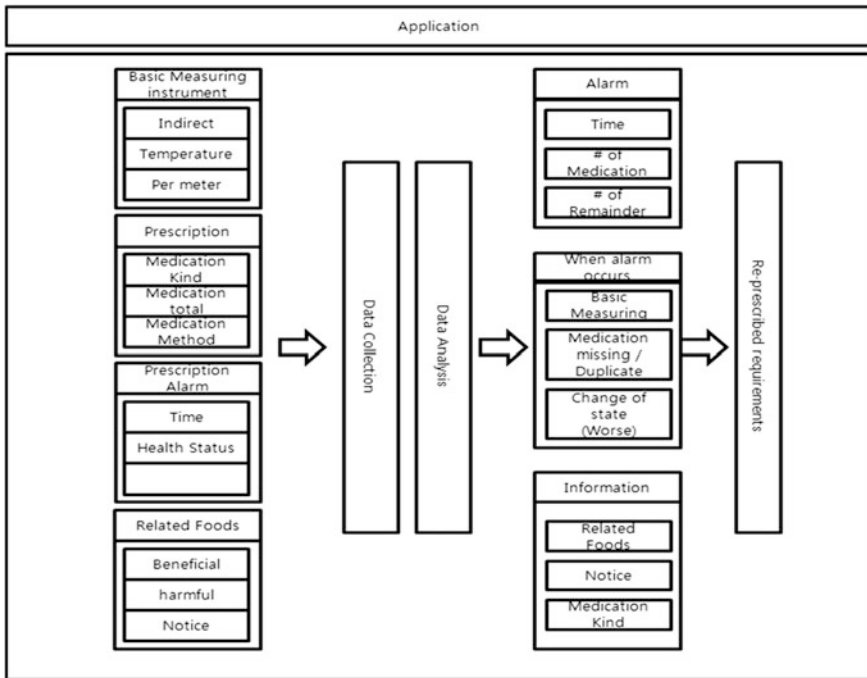


Fig. 3 Functions in smart phone application

prescriptions. Regarding to prescription, the information of administration types can be transformed, times and method of administration might be delivered and it is possible for patients to check information in a convenient manner via smart phones whenever he or she desires to do. The administration alarm function notifies patients on administration schedule and times and gives warning depending on critical mass for certain timing and times upon failure to abide by administration schedule and times while the critical mass is broken, it notifies the abnormality occurrence to hospital server and gets information related to re-prescription. In addition, It gets information on relevant foods, cautions during drug dose and foods avoided and based on such information, gives more effective supports for the recovery of the patient. Moreover, it is possible for a patient to request re-prescription. In such case, the patient feels worsening status rather than improving.

The comprehensive functions of smart phone applications are to collect and analyze information on basic measuring device, prescription, administration alarm function and relevant foods; to request re-prescription to hospital server when abnormality occurs; and to gets the prescription.

## 4 Conclusion

As smart phone health care systems have recently used in hospitals and public health clinics, various smart phone health care systems are under development; but in reality, health care system supporting patients in need of regular management is in short.

Therefore, this study proposes a model for drug administration management system for smart phone applications to prevent duplications in the drug prescription process and to reduce cases of worsening diseases due to overdose or under-dose from recognition failure for drug dose number of times or time or enhancing immunity to certain drugs. In addition, the delivery system is proposed to get rid of inconvenience of patients who should get prescribed drugs by themselves, which enables collection, management and application of individualized information by user and further optimized administration management by person through learning.

As for future research project, a study will be conducted for linking issue between most basic domestic diagnosis devices and smart phone applications, algorithm for drug redundancy prevention and concrete and efficient algorithm on delivery system after getting prescription.

## References

1. Features and applications, and the need of [U-healthcare] Ubiquitous Healthcare (u-health care), Features and meaning of U-healthcare, Development practices and the types of services. [http://www.sysbase.co.kr/s\\_faq/rs232.htm](http://www.sysbase.co.kr/s_faq/rs232.htm)
2. Direction and development of an effective response of U-health
3. Development and evaluation of drug duplication alerting (2007) 연구보고서 Ewha Woman University
4. u-Health New Business Model (2010)

# Use of Genetic Algorithm for Robot-Posture

Dong W. Kim, Sung-Wook Park and Jong-Wook Park

**Abstract** Robot-posture with genetic algorithm is presented in this paper. As a robot platform walking biped robot is used. To cope with the difficulties and explain unknown empirical laws in the robot, practical robot walking on a descending sloped floor is modeled by genetic architecture. These results from the modeling strategy is analyzed and compared.

**Keywords** Genetic algorithm · Robot posture · Comparison analysis

## 1 Introduction

To achieve walk realization, the foot should be controlled well but generally it cannot be controlled directly but in an indirect way, by ensuring the appropriate dynamics of the mechanism above the foot. Thus the overall indicator of the mechanism behavior is the point where the influence of all forces acting on the mechanism can be replaced by one single force. This point was termed the zero moment point (ZMP). Recognition of the significance and role of the ZMP in the biped artificial walk was a turning point in gait planning and control.

In this paper, practical robot walking on a descending sloped floor is employed as robot platform, neuro-fuzzy system, and evolutionary architecture are also used as intelligent modeling strategies. In addition, these results from two methods are shown, analyzed and finally compared.

---

D. W. Kim (✉)

Department of Digital Electronics, Inha Technical College,  
Incheon, South Korea

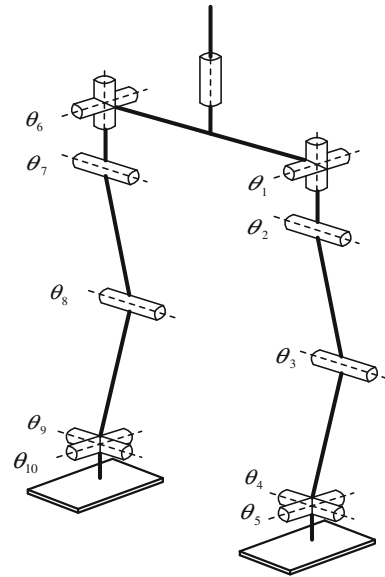
e-mail: dwnkim@inhac.ac.kr

S.-W. Park · J.-W. Park

Department of Electronics, University of Incheon, Incheon, South Korea

e-mail: jngw@incheon.ac.kr

**Fig. 1** Joint angle representation of the robot



## 2 Robot Platform

The identical robot platform employed in [7, 8] is also used. The robot has 19 joints (three DOFs are assigned to each arm, three and two DOFs are assigned to the hip and the ankles, respectively, and one to each of the two knees). But only 10 dominant joints are used for input candidates. The locations of the joints are shown in Fig. 1. The height and the total weight are about 445 mm and 3 kg, respectively. Each joint is driven by the RC servomotor that consists of a DC motor, gear, and simple controller.

## 3 Usage of Genetic Algorithm

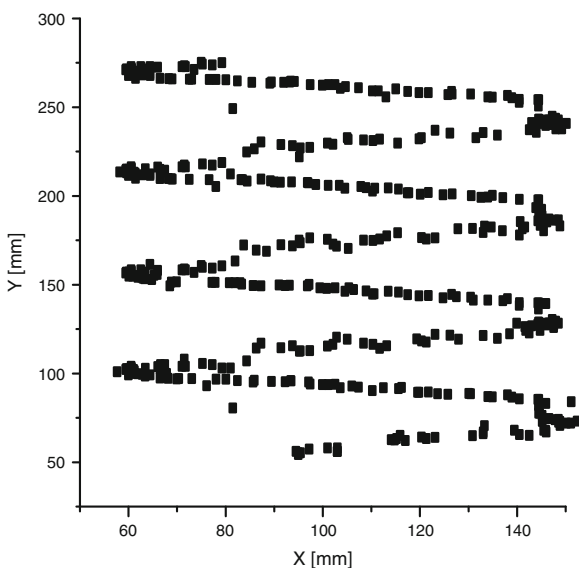
The genetic algorithms (GA) is an efficient tool for searching solutions in a vast search space. By the GA, proper type of MF, number of MF, type of consequent polynomial, set of input variables, and dominant inputs among input candidates are likely to be found in the case where the 10 input candidates have complex correlation.

These parameters stated above in designing a fuzzy system are determined in advance by the trial and error method. But in this paper, the key factors for optimal fuzzy system are specified by GA automatically. When designing a fuzzy system using the GA, the first important consideration is the representation strategy, which is how to encode the fuzzy system into the chromosome. We employ binary coding for the available design specifications. The chromosomes encoded information for

fuzzy system are made of five sub-chromosomes. The first one has one bit and presents type of membership function. Two types of MF, Triangular and Gaussian MF, are used as the MF candidates. Each is represented by a bit 0 and 1. If the gene in the first sub-chromosome contains 0, the corresponding type of MF is Triangular type. If it contains 1, the MF is Gaussian type. The second sub-chromosome has two bits for number of MF. If many number of MF is selected for certain input variables then fuzzy rules and computational complex can be increased. So we constrain the number of MF to vary only between 2 and 4 for each input variable. The 3rd sub- chromosome has two bits and represents types of polynomial. A total of four types of polynomial are used as candidates and each candidate is represented by two bits. Selection of dominant input variables which is greatly contributed to the output and number of these variables is very important. Research of appropriated method for the input selection is still under investigation. In this paper, we handle these problems using the fourth and fifth sub-chromosomes.

For the best x-coordinate, produced string information is [0, 4, 3, 4: 2, 4, 8, 9]. This string means triangular MF type, 4 MFs, Type 3, and 4 inputs, second, fourth, eighth, ninth, are selected to get good performance. 3.9877 of MSE value is obtained from this string information. For the best y-coordinate, [1, 4, 2, 4: 2, 7, 8, 9] of string information is obtained. So Gaussian MF type, 4 MFs, Type 2, and 4 inputs, second, seventh, eighth, ninth, are selected by evolutionary algorithm and 5.5385 of MSE value is obtained. The corresponding walking trajectory for a descent floor based on the model output is shown in Fig. 2.

Fig. 2 Walking trajectory on a descent slope



## 4 Conclusion

Robot posture strategy with genetic algorithm is presented and its results are analyzed in this paper. As a robot platform walking biped robot which has 19 joints is used and trajectory of the zero moment point (ZMP) is employed for the important criterion of the balance. To cope with the difficulties and explain unknown empirical laws in the robot, practical robot walking on a descending sloped floor is modeled by genetic architecture. In this paper, Fig. 2 is finally from the genetic algorithm based model.

## References

1. Vukobratovic M, Brovac B (2004) Zero-moment point—thirty five years of its life. *Int J Humanoid Rob* 1:157–173
2. Vukobratovic M, Andric D, Borovac B (2005) Humanoid robot motion in unstructured environment—generation of various gait patterns from a single nominal. In: Kordic V, Lazinica A, Merdan M (eds) *Cutting edge robotics*. In Tech
3. Hirai K, Hirose M, Haikawa Y, Takenaka T (1998) The development of Honda humanoid robot. In: *Proceedings of the IEEE international conference robotics and automation*, pp 1321–1326
4. Vukobratovic M, Brovac B, Surla D, Stokic S (1990) *Biped locomotion*. Springer, New York
5. Kim D, Seo SJ, Park GT (2005) Zero-moment point trajectory modeling of a biped walking robot using an adaptive neuro-fuzzy systems. *IET Control Theory Appl* 152:411–426
6. Kim D, Park GT (2007) Advanced humanoid robot based on the evolutionary inductive self-organizing network. *Humanoid robots—new developments*, pp 449–466
7. Kim D, Park GT (2010) Intelligent walking modeling of humanoid robot using learning based neuro-fuzzy system. *J Inst Control Rob Syst* 16(10):963–968
8. Kim DW, Silva CW, Park GT (2010) Evolutionary design of Sugeno-type fuzzy systems for modeling humanoid robots. *Int J Syst Sci* 41(7):875–888
9. Chun BT, Cho MY, Jeong YS (2011) A study on environment construction for performance evaluation of face recognition for intelligent robot. *J Korean Inst Inf Tech* 9(11):81–87
10. Shin JH, Park JG (2012) Implementation of an articulated robot control system using an On/Off-line robot simulator with TCP/IP multiple networks. *J Korean Inst Inf Tech* 10(01):37–45

# Use of Flexible Network Framework for Various Service Components of Network Based Robot

Dong W. Kim, Ho-Dong Lee, Sung-Wook Park and Jong-Wook Park

**Abstract** These days wide variety of platforms and frameworks were researched for network based robot (NBR) but it is difficult to apply those platforms and frameworks to the NBR, because the NBR and its service components keep developed. In other works, the interfaces and execution environments of the NBR and the service components are changed and upgraded very often, and at times, the service components are totally reconstructed. Consequently, a new flexible and reliable network framework that adapts to various service components is necessary. To handle these problems, a solution is suggested in this paper.

**Keywords** Network based robot (NBR) · Platform and frameworks for service component

## 1 Introduction

TCP/IP (Transmission Control Protocol/Internet Protocol) is a well known and the most popular communications protocol for Internet connection. TCP/IP is a traditional method of integrating components via a network. It is reliable and stable, and it is very easy to develop its components. Various architectures and integration schemes have been researched. In this paper, communications interfaces based on TCP/IP are designed for an NBR.

---

D. W. Kim (✉)

Department of Digital Electronics, Inha Technical College, Incheon, South Korea  
e-mail: dwnkim@inhac.ac.kr

H.-D. Lee

Korea Institute of Science and Technology, Seoul, South Korea

S.-W. Park · J.-W. Park

Department of Electronics, University of Incheon, Incheon, South Korea  
e-mail: jngw@incheon.ac.kr

**Fig. 1** Internet protocol stack

Layer	Protocol
Application Layer	TFTP, TLS/SSL, FTP, HTTP, IMAP, IRC, POP3, SMTP, SNMP, TELNET, PNRP, ...
Transport Layer	TCP, UDP, DCCP, SCTP, IL, RUDP, ...
Network Layer	IP (IPv4, IPv6)
Physical Layer	Ethernet, Wi-Fi, Token-ring, PPP, SLIP, FDDI, ATM, Frame Relay, SMDS, ...

## 2 Communications Interface Design

TCP/IP consists of an IP (Internet protocol), which is an Internet protocol that uses the packet communication method, and a TCP (transmission control protocol).

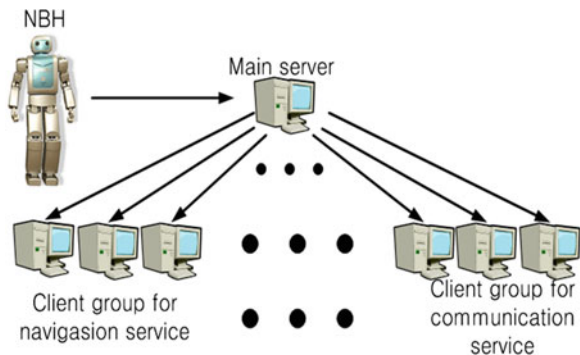
Figure 1 shows the TCP/IP protocol’s stack and related protocols. The hierarchical structure is also shown.

Environment data such as video streams and audio streams from an NBR are sent to each resource via the main server. Thus, if the number of service components increases, then the data throughput of the main server is increased. This situation is shown in Fig. 2.

To develop the network framework described previously, a network core that can be configured into the network with a tree structure was designed. Also, the network core serves as a server and a client simultaneously. In Fig. 3, the network core that was designed and implemented is shown.

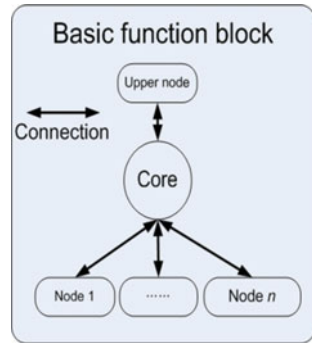
Figure 4 shows a block diagram of the network core. As shown in Fig. 4, the sender transmits data packets to other network cores. The receiver receives

**Fig. 2** Case of clients receiving information from the main server

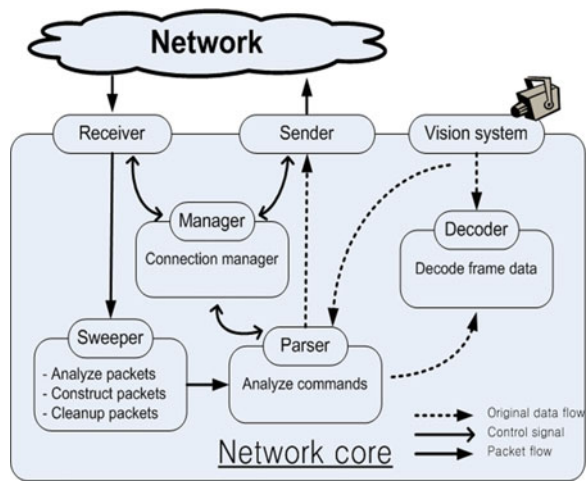




**Fig. 3** Basic concept of a network core structure



**Fig. 4** Block diagram of network core



the data packet from other network cores. The sweeper assembles the packets from the receiver to the data. The parser is responsible for processing the data. A network core has these four basic functions and management roles.

### 3 Performance Analysis and Discussion

The transfer rate is measured between cores through a wired LAN, a wireless LAN, and a local machine, respectively. It shows also quick network core data transfer. Note that during the data transfer, the network core partitions and assembles the data.

## References

1. Schuehler DV, Moscola J, Lockwood J (2003) Architecture for a hardware based, TCP/IP content scanning system. In: Proceedings of the high performance interconnects 2003, pp 89–94
2. Tang W, Cherkasova L, Russell L, Mutka MW (2001) Customized library of modules for STREAMS-based TCP/IP implementation to support content-aware request processing for Web applications. In: Advanced issues of E-Commerce and web-based information systems, WECWIS 2001, pp 202–211
3. Hansen JS, Riech T, Andersen B, Jul E (1998) Dynamic adaptation of network connections in mobile environments. *IEEE Internet Comput* 2(1):39–48
4. Lee KB, Schneeman RD (1999) Internet-based distributed measurement and control applications. *IEEE Instrum Measur Mag* 2(2):23–27
5. Liao R-K, Ji Y-F, Li H (2006) Optimized design and implementation of TCP/IP software architecture based on embedded system. In: Proceedings of the machine learning and cybernetics 2006, pp 590–594

# China's Shift in Culture Policy and Cultural Awareness

KyooSeob Lim

**Abstract** Since the modern age, China's major historical events have always been related to culture and it has been centered to discussions on the national development. Since adopting the reform and openness policy, China has been emerging as a new power and enlarging influence in international politics, economics, and military. However, China recognizes that its cultural influence is still light. Thus, the Chinese government exercises cultural policy which enables China to expand Chinese cultural clout and soft power externally and, at the same time, integrate society by establishing a clear national identity internally. China's cultural heritage, which once was target to be overthrown, is now inherited and developed. China is changing its cultural awareness.

**Keywords** Culture policy · Cultural system reform · Cultural heritage · Soft power · China

## 1 Introduction

We can say that Chinese history is in line with Chinese cultural history. Since the modern times, China's major historical events have always been related to its culture [1], and culture has been centered to the discussions. Broadly, culture exercises influence in 'people' who are main agents of politics and economics. For such reasons, culture consists of one system providing phenomenon related to

---

K. Lim (✉)

Department of Foreign Language and Culture, Institute of International Education,  
Kyung Hee University, 26 Kyunghee-daero, Dongdaemun gu,  
Seoul 130-701, South Korea  
e-mail: lks@khu.ac.kr

politics and economics [2]. Even now, in contemporary China, culture interacts with politics and economics and develops its rules and functions.

Since the new cultural movement in China, China has defined culture as comprehensive concept including not only literature and arts but also knowledge, ideology, ethics, and values. China also stressed culture saying ‘cultural transformation’ should be proceeded in order to bring a political, economic, and social change. The Chinese government estimated that to transform politics, economics, and society, overall changes in people’s value should be preceded. For that reasons, in the contemporary Chinese history, in every major transitions for political, economic, and, social change, there have been discussions on the culture [3]. There is a saying that ‘cultural change should be put first’ and the tradition is still influential. As a result, after the reform and openness policy, the Chinese government emphasizes ‘cultural reformation’ to overcome internal and external conflicts and crises and to transform its politics and society.

In the past, the Chinese government and Chinese leaders regard culture as means for propaganda and thought indoctrination. However, escaping from this recognition, the Chinese government started to rethink the value and meaning of culture in a different way. Specially, reform and openness makes progress in conjunction with cultural ‘marketization’ and ‘popularization’ and consequently the Chinese government’s cultural awareness has fundamentally changed. The central question that I intended to examine in this article is ‘why the Chinese government is responsive to culture issues?’ In the following, I will try to examine backgrounds to Chinese cultural policy and its contents and analyze China’s awareness shift on culture.

## **2 Backgrounds to China’s Culture Policy**

### ***2.1 Building Legitimacy of Communist Party and Social Integration***

China’s cultural heritage which is combined to Chinese nationalism has been major means of stabilizing the Chinese Communist regime. China’s cultural strategy has been used to remove threats on the Chinese Communist regime rather than means to overcoming national crisis. In other words, as Chinese economy rapidly expands through reform and openness, Chinese people recover national confidence while China’s national ideology was weakened in the globalization and the post-Cold War era. So, the Chinese government tried to find the Chinese national identity through succession of national culture and unite and integrate the nation through cultural heritage [4]. The Chinese government still maintains Marxism-Leninism and Maoism socialist system perfunctorily, however, it stresses Confucius idea to expand and develop the Chinese culture. The Chinese government which uses Confucian ideas tries to maintain its regime by constantly

announcing that it is a legitimate successor of the Chinese nation [5]. As ideological function of socialism which once played as a key role in social integration weakened, China intended to form a new national ideology from Confucianism in order to maintain economic development and steadily integrate society.

## ***2.2 Attempting Shift in Economic System***

Since adopting Western capitalism, the essential crisis of the cultural awareness has been not only about its values but also about economy or capital. In other words, Western culture industry with the gigantic capital strength and global competitiveness seemed to encroach on China's culture market [6]. Owing to the growing economic importance of cultural industry, the Chinese government gave shape to culture as one sector of major policies and strategies [2]. Beijing put stress on economic value of culture for economic development and brought fundamental change of 'means to the economic development in the future.'

## ***2.3 Resisting Western Culture Invasion***

Against expanding Western culture invasions such as culture standardization, China is experiencing 'cultural crisis', so it carries out aggressive culture system reformation mainly towards to Western advanced capitalism countries [6]. In the current world system under the Western political ideology and values and capitalistic market economy, China should observe western values-dominant international order. China thinks the Western powers attempt to gradually change the Chinese social system by using these values and orders. Thus, China appreciates unless it deters the Western infiltration, Chinese identity and loyalty to Chinese culture would be deteriorated and the Chinese culture would come to a crisis by being eroded assimilated with the Western culture. Beijing makes ceaseless effort to get over Western centrism. China aims to maintain identity and independence of Chinese civilization against the Western culture and its first priority aim is executing culture policy.

## ***2.4 Forming Peaceful National Image***

The western countries and neighboring nations express concerns on stronger China. These concerns greatly impede Chinese national development and its leverage [7]. For that reasons, China started to claim new ideology calling 'harmonious society' and 'harmonious world' and make better national image internationally using soft power. To eliminate negative awareness such as 'Chinese throat' and make better national image, the Chinese government highly pushes ahead with the cultural policy.

## ***2.5 Developing Toward a Culture Power***

In line with economic, military, and international political clouts from success of reform and openness policy and cultural outcomes from rich history, China hopes to develop to a culture power nation. The Chinese government wants to resolve unbalanced development between economy and culture, and also hopes to enlarge soft power in the world. In other words, China started to regard both hard power development plans and soft power national development plans as important. The Chinese government emphasizes cultural revival to become a culture power and a powerful nation at the same time.

## **3 China's Culture Policy**

### ***3.1 Culture Policy***

In the era of globalization, the Chinese government promotes 'globalizing Chinese culture' and seeks to new national identity through traditional cultural elements. Especially, since 2000, the Chinese government highlights the importance of culture in various fields and regards culture as crucial factor for the national development. As a matter of fact, the Chinese government accentuates culture-related policies such as culture security, soft power, culture power, national image, cultural diplomacy, Confucius revival, and Confucius Institute.

The Chinese government advocates it should strengthen soft power claimed by Joseph Nye in diplomatic policy. Currently, Beijing proceeds with its soft power policy by 'actively engaging with national competency.' The Beijing Olympics in 2008 and Expo 2010 Shanghai China are good examples. Particularly, the Chinese government established the first Confucius Institute on November 21, 2004, so as to publicize Chinese culture to the world and expand international leverage of the Chinese culture. Since then, until July 2012, China establishes and operates some 387 Confucius Institutes in 108 countries.<sup>1</sup>

### ***3.2 Government Official Papers***

Since the reform and openness, we could the witness awareness change in culture at the Chinese government released official papers. Since <the 12 Chinese Communist party congress>, China has emphasized the importance of 'the culture'. In <the 12th report> of the 1982 Chinese Communist Party congress, China

---

<sup>1</sup> <http://www.hanban.edu.cn/>

separated cultural construction and ideology construction from the socialist civilization construction and put an emphasis on culture saying it is a part of the socialist civilization construction. In <the 14th report> of the 1992 Chinese Communist Party congress, China released the cultural construction line. In essence, China intended to 'actively promote cultural system reform and perfectly implement economic policy relating to culture for the sake of flourishing socialist culture. At the time, the Chinese government firstly insisted that it should promote the cultural system reform on the basis of economic and political system reformation.

In <the 15th report> of the 1997 Chinese Communist Party congress, the Chinese government suggested that it should build culture construction with Chinese socialism trait, and since then, it has been the basis of the Chinese Communist party's cultural line [8]. In <the 16th report> of the 2007 Chinese Communist Party congress in October, the Chinese government explained cultural construction and cultural system reform claiming that 'according to the socialist civilization construction, it should implement cultural system reformation along with the socialistic market economy development.' China stressed that it should perfectly prepare cultural market system and its management mechanism [9]. In <the 17th Chinese Communist Party congress> in 2007, Hu Jin-Tao raised that 'we should lead greater development and prosperity of the socialistic culture.'<sup>2</sup> At the event, Hu Jin-Tao firstly included 'Chinese culture' to the Chinese Communist Party's official paper. In <the 6th general meeting of the 17th Chinese Communist Party congress> in October, 2011, Chinese government selected the cultural reformation for an agenda, and passed 'the resolution on crucial issues to intensifying cultural system reformation of central Chinese Communist Party and promote greater development and prosperity of the socialistic culture.'<sup>3</sup> At the meeting, the Chinese government decided to improve Chinese identity and confidence on the Chinese culture by keeping cultural security and strengthening its soft power in succession to the China's economic development. The report said that China is faced to the problem of protecting cultural security and improving soft power and international clouts of the native culture.

## 4 Changing Awareness in Culture

### 4.1 Cultural Status in the Traditional China

In the traditional China, culture was the basis for the Chinese order. In the warring states period (Xian Qin period), Chinese 'Mun (文, knowledge)' tradition which opposed to Mu (武, martial arts) laid the basis of the Chinese Hwaism (華夷論).

<sup>2</sup> <http://news.163.com/07/1020/07/3R7TGC240001124J.html>

<sup>3</sup> [http://www.gov.cn/jrzq/2011-10/25/content\\_1978202.htm](http://www.gov.cn/jrzq/2011-10/25/content_1978202.htm)

Chinese people thought that they are the center of the world and they enjoy the most advanced culture in the world. Chinese people developed Hwaism, another kind of Sinocentrism, which ignores and neglects other cultures [10]. In the traditional China, the concept of culture included ‘discriminatory structure’, and Chinese culture is top of other barbarous cultures. They believed Chinese culture is related to their nature spirit, and there is no comparison.

## ***4.2 Changing Status of Culture Since the Modern Age***

Since the modern times, by shocking from the Western civilization, Chinese myth that Chinese culture is unique has been collapsed. Also, all the national crises of China brought awareness change in ‘culture’ [1]. Especially, there had constantly been arousing discussions on ‘Chinese cultural heritages’ since the May Fourth movement (the new cultural movement) and it was the key issues on the national development. In fact, in the May Fourth movement and the Cultural Movement period, the Chinese cultural heritage was denial and traditional ideas and values was target to be overthrown. In the May Fourth movement period, Confucius was accepted as an ultimate cause of national crisis, and in the Cultural Movement period, the traditional heritage was considered as an obstacle for establishing socialist regime by accepting proletarian culture [3].

At the discussion of ‘the culture fever’ in 1980s, discussion on the traditional China and the contemporary China, Chinese fundamental cognition on the Chinese culture was unchanged. Chinese regarded feudal culture, before the May Fourth movement culture, as old tradition and ultraleftism culture in Mao Zedong times, as neo tradition, and considered them to be overthrown [11]. By the time, China thought tradition as an opposite to the moderns, and there was prevailing opinion that tradition should be defeated to build a modern nation.

## ***4.3 Culture Awareness During Reform and Openness Times***

After the reform and open policy, China began to overcome economy fell, but the policy caused numerous side effects i.e. mass inflow of the Western culture, deepening economic bipolarization, lacking political democratization, deficit of the new sense of values, and cultural identification [12]. Therefore, China sought for an alternative ways that could lead social stabilization and integration and constant economic development simultaneously, and started to regard highly traditional culture such as Confucius. In 1990s, from the negative standpoint on the traditional cultures, China started to recognize cultural heritage as a target to be succeed and developed. In particular, the identity crisis, evoked from the 1989 Tiananmen Massacre, and anti-West and nationalism emotion arose, and that prepared the ground for ‘the sinology fever’ [13].



The Chinese government concentrated on the Confucianism, representative Chinese traditional culture, and deemed it positive to the Chinese modernization and economic development. Especially, China considered that culture would resolve a great many negative factors caused by the introduction of western capitalistic market economy system [12]. After 2000, China heightens confidence on economic power and pays much attention to culture in both governmental and private sector.

#### ***4.4 Relationship Between Culture and Economy***

The report 'several view on enforcing the cultural market' from the Department of Culture on February 4, 1994, said 'development of the cultural market promotes the interaction between culture and economy and it establishes developing cultural industry.'<sup>4</sup> In this report, the Chinese government emphasized the relations between culture and economy on an equal footing. The culture policy, especially the culture industry, includes not only cultural but also economic traits. That is to say, it consists of the two fields, 'upper structure' and 'economic foundation.' The Chinese government made strategic choice to integrate economic and cultural value at the same time through culture industry [6]. In the past, in the emphasis on politics with class strife, culture was a tool for spreading national policy and socialist ideology under the slogan, 'serving workers · farmers · soldiers and national politics. On the other hands, in the present market economic system, culture is a tool for creating 'profit' and consuming 'culture' itself at the same time. By forming a new version of cultural market, the Chinese government's cultural awareness, unlike to the past, has fundamentally changed.

#### ***4.5 Relationship Between Culture and Politics***

Since the New China, Chinese leaders have taken culture as a secondary way for political strife. The Chinese government cognizes culture not comprehensive but limited concepts that is helpful to attain 'specific purpose'. During the Cultural Revolution period, culture worked as an ideological tool for class strife and political strife. The Chinese government considered and developed the relation between politics and culture as a master-servant relationship. However after the reform and openness in 1979, China changed its national development strategy focusing on economic development and Chinese awareness in culture gradually changed. Namely, though a master-servant relationship between politics and

---

<sup>4</sup> Department of Culture report: [http://www.34law.com/lawfg/law/6/1187/law\\_251625388934.shtml](http://www.34law.com/lawfg/law/6/1187/law_251625388934.shtml).

culture still exists, the relationship slowly escaped from ‘the one-sided relationship [14].’ The Chinese government used culture’s public ripple effect to maintain and enforce national ideology. It produced ‘culture’ and expressed cultural political acts and then increased clouts in the culture market. The Chinese government changed cultural identity to political or national identity so as to integrate China using the national emotion. The political intervention changed contexts and rules of culture, and the transformed culture once again provided the basis of legitimacy to the political power.

## 5 Conclusion

Although China attained success in reform and openness policy, there have been lots of side effects all over the politics, economy, and society. To settle contradictions in the Chinese society and consolidate the Chinese Communist party regime, the Chinese government actively promoted national development strategy in the culture area. The Chinese government carried out many different culture policies such as culture security, soft power, national image, cultural diplomacy, Confucius revival, and Confucius Institute. As a result, Chinese awareness in culture gradually changed. In other words, China started to recognize cultural heritage, which was once seen in a negative light, as a target to be succeeded and developed. Also, China’s viewpoint between culture and economy and culture and politics has been changed. Currently, it is likely that the Chinese government and the Chinese society simultaneously attempt to make cultural transformation. The ‘top-down’ cultural transformation led by China is still occurring.

## References

1. Seo K (2008) Reviewing Chinese civilization and its language in the early 20th century. *Chin Lang Overv* 27:51–74
2. Kim K (2012) China’s cultural policy. *East Asia Briefing* 7.2(24)
3. Lee Y (2006) Discussions on the Chinese culture and cultural development strategy in the age of globalization. *Contemp Chin Lit* 37
4. Lee K (2009) Chinese cultural nationalism and the practical strategy. *Korean Northeast J* 52
5. Zhao C-S, Lin Y-J (2012) Chinese Marxism and Chinese Communist Party’s interpretation on culture. *Prospect Investig* (Taipei) 5.5
6. Kwon K (2012) China’s nation vision in the 21st century and its strategy for developing cultural industry. *Res Contem China* 14.1
7. Pan B, Nam C, Chang Y (2011) The process of the Chinese culture diplomacy and problems. *Unification Res* 15.2
8. Zhang S-L (2012) Historical contexts on cultural reformation and development of the Chinese Communist Party. *Chinese Communist Party YunNan* (Yunnan), vol 13.1
9. Long X-M (2012) China’s culture system reformation and development. *BaiHuaChao* (Beijing), vol 8

10. Choi S (2011) Chinese culture, how can we understand and which approach could we take? *Sinol J* 4
11. Tang Y-J (1998) Three problems of traditional culture research. *XinHuaWenZhai* (Beijing) 1988, vol 3
12. Yeon J (2012) Implications and prospects of the restoration movement in the Tang Dynasty. *Korean Philos J* 30
13. Zhang X-C (2002) Discussion on practicing multiculturalism. *21st Century* (Hongkong), June 2002 vol 71
14. Lee J (2009) Research on the Chinese socialist market economy and cultural policy change. *Chin Lit* 60

# China's Cultural Industry Policy

WonBong Lee and KyooSeob Lim

**Abstract** In the globalized world, the cultural industry has emerged as a new promising field. Countries are accelerating their competition to become a culture power nation. In the past, China used culture in the purpose of electing ideology. Since China's entry into the World Trade Organization (WTO), China started to foster the cultural industry. Since 2001, China started to carry out the cultural industry policy. In 2009, after 'Cultural Industry Promoting Plan', China propelled cultural industry to promote policy in earnest. In 2011, the Chinese government claimed it would nurture the cultural industry as a national strategic industry. In China, cultural industry has been stressed as a part of its soft power strategy. The culture has been emerged as a mean for spreading ideology and a new growth engines for industry.

**Keywords** Chinese culture · Cultural Industry · Cultural Strategy · Cultural system reform

## 1 Introduction

In the globalized world, 'culture' is combined to 'industry.' At the same time, countries have been forming 'the cultural industry' by systemizing and industrializing culture for supply. Since 1999, the Chinese government has given an

---

W. Lee (✉)

Department of Chinese Studies, Kyung Hee Cyber University, 1 Hoegi-Dong, Dongdaemun Gu, Seoul 130-701, South Korea  
e-mail: wblee@khcu.ac.kr

K. Lim

Department of Foreign Language and Culture, Institute of International Education, Kyung Hee University, 26 Kyunghee-daero, Dongdaemun gu, Seoul 130-701, South Korea  
e-mail: lks@khu.ac.kr

important on the status of culture, and in the 2000s, the government accelerated importance of the culture. Since China's reform and openness policy, China's economy has been growing and it brought changes in consuming culture. Therefore, demands on the culture products are skyrocketed and Beijing realized lack of cultural products which should be consumed in various mass media [1]. Also, as China emerged as an economic power, the Chinese government accentuated fostering cultural industry in order to enhance China's status to the higher level in the world [2].

Although China's cultural industry largely developed in line with the economic growth, cultural industry is given little weight than other industries. In 1996, the added value of the cultural industry was 21.184 billion yuan (RMB), accounting 0.3 % of GDP. Moreover, in 2009, the added value of the cultural industry was 103.77 billion yuan (RMB), accounting 0.3 % of GDP. Since 2000, China's cultural industry has comparably maintained stable growth [3]. China sets up the cultural industry as an emerging industry and considers it as a driving force for the national development in the tide of globalization. China aims to enhance social and cultural level through the cultural industry policy and wants to emerge as a culture power. Also, the Chinese government tries to establish Chinese own cultural industry system. The focal point of our discussion will examine the forming factors of the Chinese cultural industry and its development strategies with current state of the cultural industry on the regional basis. This paper will also analyze characteristics of China's cultural industry.

## **2 Cultural Industry Policy**

### ***2.1 Culture and Cultural Industry***

Culture refers to the people's way of living or the way of thinking at a certain place. Culture is used extensively including food, clothing, shelter, language, religion, knowledge, arts, and institutions [4]. Culture enriches people's lives by changing natural state of human being artificially with skills and labors. In other words, culture endows new value to the nature. Culture is a creation of both spiritual value and material value [5]. Recently, countries put an importance on culture as a key factor when it comes to evaluating comprehensive nation's power. Every nation tries to heighten its cultural soft power.

Generally, the United Nations Educational, Scientific, and Cultural Organization (UNESCO) defines the cultural industry as 'one business's production, reproduction, storing, and distribution of cultural products or services and its steering those cultural products or services in commercial way. In other words, the cultural industry means using culture for economic interests rather than cultural development itself [4].' The cultural industry connotes one nation's value comprehensively. Therefore, culture plays important role in building one country's

national identity. The spiritual culture provides foundation for the cultural industry and the cultural industry exercises a great effect on people's mindset.

The cultural industry is a new phenomenon which is evoked in the process of integrating culture, economy, and technology. The cultural industry creates synergy effect through integrating environment-friendly industries, higher value-added businesses, and other industries. Also, the cultural industry develops new markets and jobs with small money. As the cultural industry does not harm the environment, the cultural industry draw attention to many nations as a new emerging industry. At the same time, the cultural industry is a promising industry with higher practical ability with relations to other industries and higher synergy effect. The cultural industry became a key industry for social development and economic growth. The competition among nations to become a culture power has become cutthroat. The cultural industry market in the world over 1 trillion dollar, and it will be one of industry upgrade in the future.

## ***2.2 Decision Factor of the Chinese Cultural Industry Policy***

Before the reform and openness, China recognizes culture as a mean for spreading ideology rather than industry. Also Beijing did not put in any efforts to consider culture as an industry. Culture was managed by the government budget and used only for the ideological purpose. Namely, culture was accentuated only for the key propaganda tool for the Chinese politics. As a result, China's culture could not sharpen its competitiveness.

According to the China's rapid economic growth, national power also enhanced. China's international standing also largely surged. However, the competitiveness of the Chinese culture was not commensurate with China's international standing. Therefore, we could say that the Chinese leadership awareness change on culture started from outside factors. China's entry into the World Trade Organization (WTO) is a good example. Since China's entry into the World Trade Organization (WTO), China was necessary to open its cultural market. According to the change, China needed to protect its cultural sovereignty. Unless China grew its competitiveness of the cultural industry, it destined to face consumer market from outside influences. Since the entry into the World Trade Organization (WTO), the Chinese government has consistently issued policies to nurture cultural industry.

Therefore, along with the economic growth, China experienced changes of consumption patterns. With the changes, the Chinese government keeps enlarging demands on pop culture. Based on the economic ability, the Chinese government started to promote cultural industry aggressively. Beijing has strong will to combining its 5000 years rich history, which was an incomparable culture power in Asia, with culture and nurturing cultural industry.

### **3 Development Process and Development Strategy of China's Cultural Industry Policy**

#### ***3.1 Categorizing Cultural Industry Policy***

Cultural industry includes some 10 types of industry i.e. publishing, radio and television, newspapers and magazines, commercial display, entertainment, exhibition, and network. Those industries share commons as well as different characters. The Bureau of Statistics of China categorizes cultural industry according to their industrial fields for statistics process. The key field of cultural industry is four parts: news paper, publishing and copyright, movie and drama, and cultural arts. Departments in charge are the Bureau of Culture, the Bureau of Newspapers and publishing, and the Bureau of Photoelectric.

In 2012, the Bureau of Statistics of China reissued 'Category for Culture and relating industries' and separates the cultural industry with two parts depending on the industrial relevance.<sup>1</sup> The first sector is 'production of cultural goods' including newspapers, publishing, broadcasting, TV, movie, cultural arts, culture and information release, cultural creation and establishment, leisure and entertainment service, and crafts arts. The second sector is 'production of culture-related goods' including production of auxiliary materials for cultural goods (publication right, print, and copy etc.) and production of cultural goods (office stationary, music instrument, and plaything etc.).

#### ***3.2 Development Process of China's Cultural Industry Policy***

Before 1999, China used the term 'cultural business' rather than 'cultural industry.' The term 'cultural industry' firstly emerged in the Chinese society after 2000. China's cultural industry policy corresponds with development stages of its cultural industry. China's cultural industry had its earliest beginning at 1978. We can call the period the simple supporting stage (1978–1992). In the 1990s, as the Chinese economic development accelerated, development of the cultural industry also became faster. This period is the promotion stage. Since 2000, the Chinese government has implemented policies so as to supporting cultural industry strategically.

The simple supporting stage was the first step to the cultural industry. In the period, culture manufacturing businesses and cultural service industries emerged with some advertisement companies. In the promotion stage, the Chinese government intentionally encouraged the development of the cultural industry. In this

---

<sup>1</sup> [http://www.stats.gov.cn/tjbz/t20120731\\_402823100.htm](http://www.stats.gov.cn/tjbz/t20120731_402823100.htm)

period, policies focusing on reforming cultural system established. Also, all sorts of regulations were enacted. At the 5th Plenary Session of the 15th Central Committee of the Communist Party of China, the Chinese official papers firstly used the definition of 'the cultural industry' and 'the cultural industry policy.'

Since 2001, we call this period the strategically supporting period. After the twenty first century, cultural competition among nations has been deepening. Since China's entry into the World Trade Organization (WTO) in 2001, China started to firmly consider the cultural industry's statistic standing. Since then, the Chinese government released announcements to lead cultural system reformation and the development of the cultural industry.

In 'the 10th Five-Year Plan (2001), the Chinese government put an emphasis on electing the cultural industry policy and promoting development of the cultural industry. In 'the 11th Five-Year Plan (2006), the Chinese government established ordinances and regulations to develop cultural industry. In 'the 12th Five-Year Plan (2011), the Chinese government decided on to nourish cultural industry. In 'the 11th Five-Year Plan and cultural development planning', the Chinese government mentioned it would promote movie, publishing, printing, advertisement, entertainment, exhibition, digital contents and character, and animation industries. The statement intended to specify cultural industry relating policies. 'The cultural industry promoting plan' was issued in September 2009, and it provided a basis for the cultural industry promoting policy. The 12th Five-Year Plan contained plan for nourishing cultural industry as a major industry. To achieve the goal, the Chinese government implemented several supportive policies such as inviting major enterprises and investors [6] (Table 1).

### 3.3 Regional Distribution of the Cultural Industry

China's cultural industry is largely different from one region to another. On added value basis in 2009, Shanghai, Zhejiang province, Guangdong province, Jiangsu province, Sichuan province, Shandong province are of great importance when it comes to culture. We could say added value of the cultural industry in East coast region is relatively greater than middle or western region [3].

**Table 1** Development process of China's cultural industry policy

	Period	Major issues
Initial stage	10th Five-Year Plan (2001–2005)	Establishing cultural industry policy Categorizing cultural-related industry
Embodiment stage	11th Five-Year Plan (2006–2010)	Development of cultural industry Strengthening governmental control
Deepening stage	12th Five-Year Plan (2011–2015)	Fostering culture as a major industry Electing policies and strategies to be a 'culture power'



In order to develop regionally and ethnically characteristic cultural industry complex, China try to its competitiveness of the culture industry by enlarging its industrial volume. In the ‘Table 2’ we could find out 8 regions with high added value of the cultural industry during the period of ‘the 11th Five-Year Plan’ (2006–2010). Through the Chinese government’s management systems and policies which are suited for each region, most regions develop the cultural industry in different fields.

## 4 Characteristics of the Chinese Cultural Industry

### 4.1 *Future Industry Combining Value of Economy and Culture*

At the 6th Plenary Session of the 17th Central Committee of the Communist Party of China, the Chinese government set up a goal to nourishing cultural industry as a key industry in economy by 2020 in order to realizing the nation’s vision. In China, cultural industry enjoys its standing not only as a key industry but also so called ‘future industry’ or ‘next-generation industry’. The strategic value of the cultural industry lies not just in creating economic value, but in reforming

**Table 2** During the 11th Five-Year Plan, Characteristics of cultural industry major cities

Region	Main characteristics	Major industry
Beijing	Securing strength on resources with rich history and culture Providing abundant cultural assets on the development of cultural culture creating industry	Publishing, news paper, movie, TV, entertainment industry
Shanghai	Developing culture focused on advertisement and service	Record and video, TV, play, mobile game
Guangdong	Taking the No. 1 ranking in China in printed material, broadcasting, digital publishing, printed publishing	Media cultural industry, amusement, TV
Hubei	Development of animation, online game, online optional service, new media	Book, publishing, record and video, exhibition
Zhejiang	Having the largest private cultural enterprises, with the largest movie studio established by the private capital China’s largest animation industrial complex	Education, animation, movie
Shan dong	Geographically taking the advantageous position to spread culture in another region	Broadcasting, publishing, record and video
Sichuan	Western cultural industrial region with the center market	Internet game, media
Shanxi	Promoting cultural policy aiming to cultural industry and tourism	Play, Internet, amusement

industrial structure. That is the fundamental transformation of 'the way of the economic development [7]. Moreover, China intends to promote the cultural industry combining 'the economic development' and 'the proud 5000 years cultural power [8].' China considers cultural industry is intensely related to the China's national strategy in the future as well as China enhance strategic value of the cultural industry.

#### ***4.2 Changing Attitude from Control and Management to Revival***

In 1998, the Chinese government reorganized the Bureau of Culture and established 'the Department of cultural Industry.' It meant that the Chinese government's awareness on 'culture' had been changed from 'the market' to 'the industry.' Using the word 'industry' in 'culture' means the Chinese government's attention to the culture pays more attention on the production of cultural goods and service and its circulation and distributions to the market. In the end, the process is centered on the sales to the customers and the economic market. The related policies also stressed on the 'industrial aspects' of culture rather than 'its cultural aspects.' The changing viewpoint represents the cultural policy is more 'promoting prosperous' rather 'imposing control.'

#### ***4.3 The Chinese Communist Party and the Chinese Government's Leadership***

In China, the government encourages the cultural market [6]. Especially, the Chinese cultural industry is led by 'national enterprises', and Beijing is enlarging the number of private and public joint cultural corporations. Depending on the government's progressive reformation of the cultural system, most national enterprises claim free competition [8]. In the future, the major actor of the Chinese cultural industry will gradually move on from the government to the private sector such as individuals and private businesses.

#### ***4.4 Enlarging Publicness and Public Interest***

China aims to reform overall cultural system in the end. Beijing wants to transform national cultural company into business management system. However, as China is a communist nation and is high lightening universal welfare system, it wants to enlarge cultural industry for publicness and public interest.

#### ***4.5 Supporting Policy***

China's cultural industry policy is more of supportive rather than control. The Chinese government considers the cultural industry as a promising industry and basic industry in the future. The government sees cultural industry as a national strategic industry and implements various supportive policies to promote it.

#### ***4.6 Gradual Opening Strategy***

The spiritual culture provides foundation for the cultural industry and the cultural industry exercises a great effect on people's mindset. Therefore, the Chinese government is very cautious when it comes to the external opening of the cultural industry. The Chinese government attentively accepted other cultures and cultural industries and made deliberate choice on induction of foreign capital and inviting foreign capital. In order to protect its own cultural industry, China gradually carries into openness policy [8].

#### ***4.7 Promoting Different Industrial Policy in Accordance with Industries***

In China, there is none 'unified cultural industry policy' which can lead the development of the whole cultural industry [9]. Most industrial policies aim to specific fields, and it has certain regulation so as to developing the fields. It is closely related to the separated development of the Chinese cultural industry for a long time.

#### ***4.8 Implementing the Cultural Industry Policy with Regional Traits***

Through management systems and policies that are suited for each region, most regions is developing cultural industry differently according to their traits [6]. Through developing cultural industry with regional and ethnic characteristics, China aims to enlarge industrial scale and enhance competitiveness. China's industrial policy is different from regions to regions.

## 5 Conclusion

In the current international society, each nation tries to enhance national image through the cultural industry. At the same time, countries make an effort to earn economic benefits from cultural industries. In the globalized world, competitions among nations to become a culture power are deepening. Cultural industry serves as the important foundation for forming national identity. The spiritual culture provides foundation for the cultural industry and the cultural industry exercises a great effect on people's mindset.

Since China's entry into the World Trade Organization (WTO), China consistently put an emphasis on the development of cultural industry. The Chinese government accentuated culture as a pivotal factor for the national competitiveness as cultural ability represents for the ethnic vital force and cohesiveness. The Chinese government considers cultural industry as a significant foundation in establishing ethnic identity. Also, China deems cultural industry as a major factor in forming people's mindset. In 2009, after 'Cultural Industry Promoting plan', Beijing's cultural industry promoting policy was implemented in earnest.

For China, cultural industry is not only creating new values, but also resolving weaknesses of existing industries. Cultural industry in China is emerging as a part of soft power strategy. In other words, culture is not only the means for spreading ideology, but also driving forces for the national development. The Chinese leadership not only elevates strategic value of culture industry, but also considers it is closely related to the China's future national strategy.

## References

1. Park J (2006) China's cultural industry and Korean wave. *Asia Pac Trend* 30
2. Kim B, Lee J (2012) China's cultural industrial policy and human resources science the reform and openness. *Int Labor Brief* 15
3. Kim S (2011) Comparing national productivity of China's cultural industry. *China Stud* 57
4. NAVER <http://terms.naver.com>, 2013.1.2-2013.1.30
5. Cho C *World cultural history*. Parkyoungsa 2079
6. Oh H (2012) Research on the policy development and major characteristics of China's cultural industry science the reform and openness. *China Center in Busan University, China Research*, 13
7. Kwon K (2012) China's nation vision in the 21 century and its strategy for developing cultural industry. *Res Contemp China* 14(1)
8. Seon J (2011) Characteristics and strategical goal of China's cultural industry. *Korean Research Center of Korean University, Korean Studies* 37
9. Yang J (2007) On defects in cultural industry policy. *Bus Adm (Beijing)* 3(3)
10. Seon J (2012) Current state and prospect of China's cultural. *Sinol J* 37
11. Yang G (2011) Research on the Chinese cultural industry and knowledge production. *Chin Lit* 57
12. Jeong W (2009) Research on market participant in China's cultural: focusing on 3C (company, customer, and competitor). *Curr China Res* 11:1

# Development of Mobile Games for Rehabilitation Training for the Hearing Impaired

Seongsoo Cho, Son Kwang Chul, Chung Hyeok Kim  
and Yunho Lee

**Abstract** This research is to suggest a mobile game program for rehabilitation of the hearing impaired. The suggested program is based on the characteristics of hearing-impaired children and the classification of hearing loss. The voice recognition technology is the most intimate way of delivery of information. The suggested program does not require any additional learning or training course, but enables the hearing impaired to do vocal exercises through games and helps them to enjoy rehabilitation training.

**Keywords** Auditory training · Voice recognition · Rehabilitation training · Hearing-impaired children · Mobile game

## 1 Introduction

All the games have background stories, helping developing objectives (goals), rules, adaptability, problem-solving skills and interaction [1–3].

---

S. Cho (✉) · S. K. Chul · C. H. Kim  
Department of Electronic Engineering, Kwangwoon University,  
20 Kwangwoon-ro, Nowon-gu, Seoul 139-701, Korea  
e-mail: css@kw.ac.kr

S. K. Chul  
e-mail: kcson@kw.ac.kr

C. H. Kim  
e-mail: hyeokkim@kw.ac.kr

Y. Lee  
Department of Social Welfare, Kyung Hee Cyber University, 1 Hoegi-Dong,  
Dongdaemun-Gu, Seoul 130-701, Korea  
e-mail: anne6@khcu.ac.kr

Among the hearing-impaired children, those who are completely deaf and must use sign language for communication are only 10 %. Children with low hearing impairment can have a command of a language if they receive auditory training and language reinforcing training during the critical period for language development (36–40 months old). Hearing-impaired children receive various kinds of treatment, including play psychotherapy, art therapy, music therapy and cognitive therapy, as well as auditory training and language reinforcing training, at rehabilitation centers for the disabled, clinics and hospitals, and rehabilitation clinics, but it is found that their satisfaction levels with those therapies have not been high [4].

This research describes characteristics of hearing-impaired children, classifies them based on level of hearing loss, provides plan and design for mobile games through analysis of needs, and finally suggests the direction of future research.

## **2 Related Studies and Technologies**

### ***2.1 Characteristics of Hearing-Impaired Children***

Children who lost their hearing severely before the language acquisition period have very limited accommodation of language through speech sound, so internalization of a language can be delayed in comparison with normal children. Schum (1991) found that 18-month-old hearing-impaired children use 0–9 words only in average, while 22-month-old hearing-impaired children receiving an oral training program can command approximately ten spoken words. Griswold and Cummings (1974) found that hearing-impaired children at the age of 4 command approximately 158 spoken words, while normal children at the same age use approximately 2,000 words. In a research on the gap of language functioning between hearing-impaired children and normal children, it was found that only 75 % of 5-year-old hearing-impaired children know the words which are known by all 4-year-old normal children. Hearing-impaired children at the age of 3 and 4 use gestures and hand signs. In a pointing test, it was revealed that only a part of 5-year-old hearing-impaired children give answers for questions which are answered by all 4-year-old normal children, and it was not possible for hearing-impaired children at their age of 3 and 4 to give answers for the same questions. The results show that hearing-impaired children have a limited vocabulary, and 5-year-old hearing-impaired children have no sufficient understanding of abstract words [5, 6].

### ***2.2 Grades of Hearing Impairment***

Hearing impairment is divided into grades based on the degree of loss of hearing. Loss of hearing is measured with internationally-approved audiometers. Table 1

**Table 1** Classification of hearing-impaired children based on the degree of loss of hearing

Grade	Average loss of hearing (dB)	Characteristics
Low hearing impairment	27–40	Difficult to speak clearly and understand words spoken in a low voice
Medium hearing impairment	41–55	Understand a face-to-face conversation with a person at a distance of 1 m, but do not understand about 5 % of a group discussion
Medium–high hearing impairment	56–70	Understand a conversation in a loud voice only, but difficult to participate in a group discussion
High hearing impairment	71–90	Understand words spoken loudly in ears Distinguish some vowels but no consonants at all
Deafness (top hearing impairment)	91 or higher	Sense a loud voice through vibration rather than tone, and have a language defect Depend on vision rather than sound for communication

shows the classification of hearing-impaired children based on the standard established by International Organization for Standardization (ISO).

### 3 Voice Recognition Technology and its Development

The voice recognition technology is a kind of pattern-recognition process, through which input voice is analyzed by a computer, and is converted into a similar command through the voice model database. Because each person has his/her own tone, pronunciation and intonation, the standard pattern is created based on common characteristics extracted from voice data from people as many as possible.

Research on voice recognition technology was first started when AT&T Bell Laboratories developed the single-voice number recognition system called ‘Audrey’. In 1963, IBM released the world’s first voice recognition device called Shoebox, which recognized 16 English words and supported simple calculation. Since then, government laboratories of the United States, the United Kingdom, Japan and the Soviet Union have developed exclusive hardware devices that can recognize human verbal output, extending the voice recognition technology to four vowels and nine consonants. From 1971 to 1976, Defense Advanced Research Projects Agency (DARPA) under US Department of Defense implemented the largest voice-recognition research project ever in history, called Speech Understanding Research. As voice portal services became popular in 2000s, Voice VML (VXML) 1.0 was established as the standard language for voice-based Internet use [7]. The voice recognition technology has been widely adopted in various fields, including home appliances, computer and information device.

### 4 Implementation of Mobile Game Design

The suggested game is developed based on Android as OS, JDK 1.7, Eclipse 3.72 and Android SDK as development tools, and Android OS SmartPhone as the execution environment.

The suggested functional game is designed for speaking training and phonetic correction. It is designed for the hearing-impaired to enjoy the rehabilitation training through smart phone games. This Android-based game gradually increases its level of difficulty, enabling users to exercise, while competing with each other.

Figure 1 shows the principles of the voice recognition technology and Fig. 2 shows the sources for the voice recognition. The rehabilitation training game convinces users of its treatment effect, and helps users to overcome a sense of uneasiness and fear.

Players start the game at the basic level, and as the level grows, acquire cognitive skills while competing with each other.

At the starting level of the game, the number of bugs which obstruct the play increases as the play time gets longer. At the second and the third level, longer and harder-to-pronounce words are given as the play time gets longer. The game layout consists of the main screen, game-selection screen, and game-play screen. On the

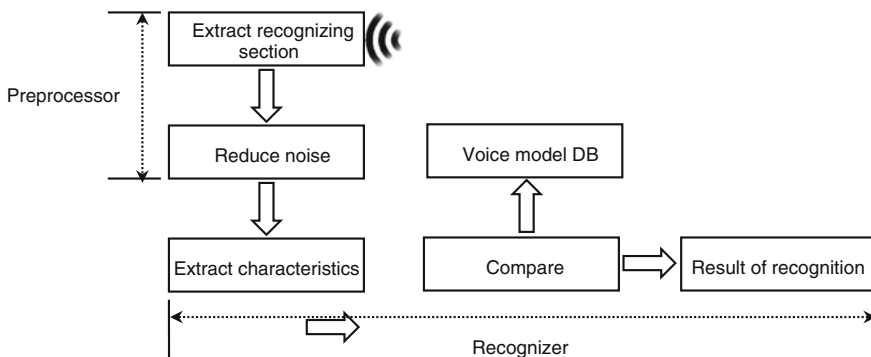


Fig. 1 Principle of voice recognition technology

```

BEGUIN
Insert I = new Intent(RecognizerIntent); //Create voice recognition intent
I.putExtra(getPackageName()); //Set called package data
I.putExtra(LANGUGE, 'ko-KR'); //Set voice recognition language
mRecognizer =SpeechRecognizer.create;//Voice recognition object
mRecognizer.start; //Start voice recognition
END
    
```

Fig. 2 Sources for voice recognition



main screen, users select the game title and the athletic location as the background. Touch the screen to display the game-selection screen. The game-selection screen enable users to select a game with a flip gesture, and lists the best score of the game. The first game screen displays the graph on the right to measure the decibel and the score at the top of the screen. The second and the third game screen show Life and Score at the top left of the screen.

## 5 Conclusion

The mobile functional game program proposed in this research is designed for rehabilitation training for the hearing-impaired. The functional game is designed for rehabilitation treatment through phonetic correction and speaking training, and is implemented to run on smart phones so that users can play anytime and anywhere. The functional games are expected to be very effective for rehabilitation. To further enhance the effect of rehabilitation treatment, it is required to develop a hearing training program, and to have the hearing impaired to finish the hearing training course before phonetic correction and speaking training course. It is required to develop phonetic correction games for language training, in addition to the functional games, through research in the hearing training field. This research is expected to promote utilization of the positive aspects of mobile games.

## References

1. Federation of American Scientists (2006) R&D challenges in games for learning. Report of the learning federation. <http://www.fas.org>
2. Gee JP (2003) What video games have to teach us about learning and literacy, computers in entertainment (CIE)—theoretical and practical computer applications in entertainment archive, vol 1(1). Palgrave/Macmillan, New York
3. McFarlane A, Sparrowhawk A, Heald Y (2002) Report on the educational use of games. TEEM (Teachers Evaluating Educational Multimedia). <http://www.teem.org.uk>
4. Nickes L, Howard D (2004) Dissociating effects of number of phonemes, number of syllables, and syllabic complexity on word production in aphasia: it's the number of phonemes that counts. *Cogn Neuropsychol* 21(1):57–78
5. Metz DE, Samar VJ, Schiavetti N, Sitler RW, Whitehead RL (1985) Acoustic dimensions of hearing-impaired speakers intelligibility. *J Speech Hear Res* 28:345–355
6. Beate P, Carol S-G (2005) Timing errors in two children with suspected childhood apraxia of speech (sCAS) during speech and music-related tasks. *Clin Linguist Phonetics* 19(2):67–87
7. Juang B-H, Furui S (2000) Special issue on spoken language processing. *Proc IEEE* 88(8):1139–1141

# A Study to Prediction Modeling of the Number of Traffic Accidents

Young-Suk Chung, Jin-Mook Kim, Dong-Hyun Kim  
and Koo-Rock Park

**Abstract** Traffic accidents are one of the big problems of modern society. The social damage caused by the traffic accidents are increasing. So, there have been a variety of research analyses to predict the traffic accidents. But there are few studies to predict the frequency of traffic accidents. In this paper, the modeling proposes applying the Markov chain modeling to predict the traffic accidents. In this paper, it is expected that the proposed traffic accident prediction modeling to predict the number of traffic accidents.

**Keywords** Simulation · Crime statics · Predictive model · Traffic accident

## 1 Introduction

Despite the efforts to reduce traffic accidents, the Korea still has many traffic accidents occur.

Looking at the current status of the OECD in case of an accident in 2009, the number of traffic accident deaths per 100,000 is 12.0 people. Greece following up

---

Y.-S. Chung · K.-R. Park (✉)  
Division of Computer Science and Engineering, Kongju National University,  
Cheonan, Korea  
e-mail: ecgrpark@kongju.ac.kr

Y.-S. Chung  
e-mail: merope@kongju.ac.kr

J.-M. Kim  
Division of Information Technology Education, Sunmoon University, Asan, Korea  
e-mail: calf0425@sunmoon.ac.kr

D.-H. Kim  
Department of IT Management, Woosongn University, Daejeon, Korea  
e-mail: dhkim@wsu.ac.kr

of the OECD countries is high [1]. The study was conducted in order to reduce traffic accidents. After confirming the relationship of a traffic accident occurs, the type and severity of traffic accidents, the risk of type is presented, and the characteristics of the driver and the studies were to investigate the relationship between traffic accidents [2]. There is a traffic accident forecasting model based on the curve radius, gradients, such as road traffic accident that occurs at the highway turnoff for linear elements to Study [3]. Prediction and detection system design for the proposed studies have the bridge section of the road freezing [4]. However, so far, the progressed traffic accidents studies analyzed the type of traffic accidents. And the study was conducted according to the forms of the road traffic accidents relating to the prediction. There are a few studies to predict the incidence of traffic accidents.

This paper proposes modeling to predict the number of traffic accidents. The implementation is being used to studies the various predictions by applying a Markov chain modeling to predict the number of traffic accidents. This paper is organized as follows. Section 2, related to the studies of Markov chains is discussed. Section 3, there will be discussed proposed of a traffic accidents prediction modeling, in this paper. Finally, conclusions and future research are discussed.

## 2 Markov Chain

Markov processes in discrete stochastic process representing the Markov chain are called [5]. In any case, the previous state to the current state of the Markov chain will affect and from the state's past does not affect the probability of the process.

A random time  $t_1 < t_2 < \dots < t_k < t_{k+1}$ ,

About  $X(t)$  when discrete value, the Markov chain.

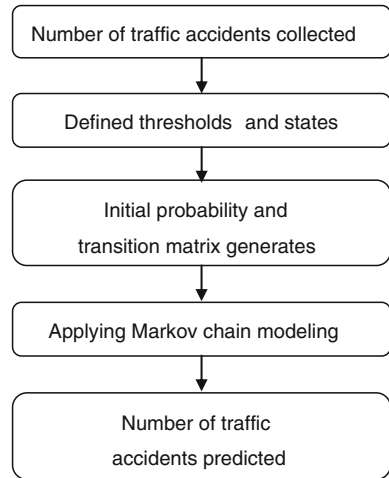
The following formula (1), expressed as

$$\begin{aligned} P[a < X(t_{k+1}) = x_{k+1} | X(t_k) = x_k, \dots, X(t_1) = x_1] \\ = P[X(t_{k+1}) = x_{k+1} | X(t_k) = x_k] \end{aligned} \quad (1)$$

Equation (1) then  $t_k$ : the current point in time,  $t_{k+1}$ : Future point in time,  $t_1, \dots, t_{k+1}$ : Past the point.

Markov chain is a set of states (group of the state), the initial probability (the initial probability vector), each transition matrix (transition between each state) the configuration [6]. Markov chain to predict the future has been used in various fields. To predict the movement of the occupants within the House of Commons in order to maintain comfortable indoor air, has been applied to studies [7]. And Rubber tired AGT system of vehicle operating condition has been applied to studies modeling to predict the reliability and availability [8].

**Fig. 1** The number of traffic accidents occur predictive modeling



### 3 The Number of Traffic Accidents Occurs Predictive Modeling

The number of traffic accidents occurs predictive modeling is shown in Fig. 1.

Each step is executed as follows.

First, collect the number of traffic accidents.

Second, Thresholds are set to the number of traffic accidents by analyzing the collected.

And define the status of each of the thresholds.

Third, the number of traffic accidents and the state defined mapping.

Fourth, Mapped, using state of the initial and transition probability matrix is generates.

Fifth, Initial probability and transition matrix is applied to the Markov chain modeling.

And, calculate the probability of traffic accidents occur in the future.

Finally, the predicted value of the probability of occurrence traffic accidents occurred recently.

Applied to the average of the number of traffic accidents occur. And predict the number of Traffic accidents.

### 4 Conclusion

In this paper using a traffic accident statistics, the number of traffic accidents prediction modeling is proposed. Prediction utilized in the foreseeable future study applied the Markov chain. If you predict the number of traffic accidents is expected

to be helpful for policy formulation to reduce the damage caused by a traffic accidents.

## References

1. Traffic accident statistics. <http://taas.koroad.or.kr/index.jsp>
2. Shim K-B (2009) The determination of risk group and severity by traffic accidents types—focusing on Seoul City. *J Korea Soc Road Eng* 11(2):195–203
3. Choi Y-H, Oh YT, Choi K, Lee CK, Yun I (2012) Traffic crash prediction models for expressway ramps. *J Korea Soc Road Eng* 14(5):133–143
4. Sin G-H, Song Y-J, You Y-G (2011) Ridge road surface frost prediction and monitoring system. *J Korea Contents Assoc* 11(11):42–48
5. Grinstead CM (1997) Introduction to probability, 2nd revised edn. American Mathematical Society, Providence, pp 405–406 (in press)
6. Kim Y-G, Baek Y, Peter In H, Baik D-K (2006) A probabilistic model of damage propagation based on the Markov process. *J KIISE* 33(8):524–535
7. Kim Y-J, Park C-S (2008) Prediction of occupant's presence in residential apartment buildings using Markov chain. Korea Institute of Architectural Sustainable Environment and building System. 2008 autumn conference, pp 116–121
8. Ha C-S, Han S-Y (2004) Reliability evaluation of AGT vehicle system using Markov chains. 2003 autumn conference and annual meeting of the Korean society for railway, pp 91–96

**Part XI**  
**Pervasive Services, Systems and**  
**Intelligence**

# A Wiki-Based Assessment System Towards Social-Empowered Collaborative Learning Environment

Bruce C. Kao and Yung Hui Chen

**Abstract** The social network has been a very popular research area in the recent years. Lot of people at least have one or more social network account and use it keep in touch with other people on the internet and build own small social network. Thus, the effect and the strength of social network is a very deep and worth to figure out the information delivery path and apply to digital learning area. In this age of web 2.0, sharing knowledge is the main stream of the internet activity, everyone on the internet share and exchanges the information and knowledge every day, and starts to collaborate with other users to build specific knowledge domain in the knowledge database website like Wikipedia. This learning behavior also called co-writing or collaborative learning. This learning strategy brings the new way of the future distance learning. But it is hard to evaluate the performance in the co-writing learning activity, researchers still continue to find out more accurate method which can measure and normalize the learner's performance, provide the result to the teacher, assess the student learning performance in social dimension. As our Lab's previous research, there are several technologies proposed in distance learning area. Based on these background generation, we build a wiki-based website, provide past exam question to examinees, help them to collect all of the target college or license exam resource, moreover, examinees can deploy the question on the own social network, discuss with friends, co-resolve the questions and this system will collect the path of these discussions and analyze the information, improve the collaborative learning assessment efficiency research in social learning field.

---

B. C. Kao (✉)

Department of Computer Science and Information Engineering, Tamkang University,  
New Taipei, Taiwan, People's Republic of China  
e-mail: acebruce@gmail.com

Y. H. Chen

Department of Computer Science and Networking Engineering, Lunghwa University  
of Science and Technology, Taoyuan, Taiwan, People's Republic of China  
e-mail: cyh@mail.lhu.edu.tw

**Keywords** Social learning · Social network · Wiki · Past exam · Co-writing · Collaborative learning · Assessment

## 1 Introduction

‘Wiki’ is the Hawaiian word for ‘quick’. Broadly, wiki is a open and convenient editing tool for participants to visit, edit, organize and update website. The most well-known public wiki is Wikipedia [1], which is an online encyclopedia. The most distinguished feather of wiki is that anyone can create knowledge on wiki at any time anyplace, and the intelligence of more creators is much greater than individual creators, it offers multidirectional Communications among creators [2].

After Web 2.0 technology has been proposed, wikis have been widely used in the realm of education, and serve as a medium for collaborative learning. In a scenario of wiki-based collaboration, students are divided into groups and assigned tasks. Liu et al. [3], and, everyone who use internet service in the world usually have one or more accounts of social network site in recent years, social network behavior in cyber world is become more and more important part in peoples life, if we can combine these two different but have same basic ideas web service, apply on blended learning to improve the learning efficiency.

In Taiwan, the higher education’s examinations past question just provide the questions but without answers, so the examinees must calculate the answer by itself. Or spend lots of money to join the tutorial, even search the answer on Wikipedia, also can not understand the problem solving process and learning with friends. Thus this MINE wiki service provide a community and customized edit tool to help examinees discuss and discover, make sure the answer and record the

The main method we apply in this research is based on the Prof. Trentin’s research [4], this paper is focus on the evaluation of collaborative learning project, he design a formula to calculate the contribution of each student in the learning group, also use the Wiki-like system to do the experiment, but most of this evaluation procedure needs the manual assessment by teacher. So there are many procedures can be improved, like system assist data mining, recording and calculating, reduce the assessment time for teacher, and help the student to learn the co-writing skill in social dimension then assess the contribution correctly.

In this study, we describe a Wiki Based Web system which provide online past exams about admission of master’s degree or PhD degree. And apply the co-writing assessment theory provide by [4] in this system to help other user understand which answer is the best one of the questions, is not learning in the group, but to use the social network, Students may search and have discussion with other register users in this system, if users doesn’t know how to resolve some questions, he or she can deliver to own social network to other friends who maybe know the answers or forward this message to their own social network until the question is resolved.



## 2 Related Work

In this section, we will discuss the past research about social learning or the system use wiki-based system to enhance distance learning.

In Web 2.0, one of the emerging visions is the “collective intelligence” where the folks are motivated to contribute their knowledge to solve common problems or to achieve common goal [5].

A Wiki is a type of social software that allows users to write, share and edit content real-time, with only rudimentary skills in web page creation. Anyone can edit and manage the content of Wiki, coordinate and create knowledge in collaboration with other members. The essence of Wiki consists of opening up, cooperation, equality, creating, and sharing. As a collaborative authoring platform or an open editing system, the most fundamental characteristic of Wiki is the teamwork and open editing [6].

In Marija’s research, He described and evaluated two consecutive trials of the use of wiki technology as a support tool for curriculum delivery and assessment, as well as for students’ learning [7]. The common characteristic of all trials was that they were based on weekly wiki (MediaWiki) updates by students that were triggered by tutor-set questions and assessed. The details of the assessment strategy for wiki contributions have been discussed in [8].

Wiki is considered the latest web innovation on content management and sharing [4]. Using any web browser, a user can visit a Wiki site—a web site running Wiki software, and by using simple Markup text, the user can create new pages, edit existing pages, or restructure page hierarchy and links. The simplicity and flexibility of Wiki make it an appealing tool for content sharing and online collaboration [9].

In Chang’s research [1], a case study is based on an optional curriculum called social technology and tools in Beijing Normal University. The participants are fresh students or sophomore students with various majors. There are totally 75 students participating in this activity. This study is based on the open software ‘Mediawiki’. They also choose a definite subject for this activity which is ‘google products’, and providing several different products for the students to choose which one they intend to edit [2].

Research on using computers to assess and support the development of social and emotional skills has focused on a range of populations including children with particular needs [10], and a major purpose of fully collaborative writing is to ease the dysfunctional anxiety of the individual solitary student when confronted with a blank piece of paper [11]. Traditional learning hopes student can learning spontaneously, but ignored the social dimension, but how to evaluate the individual contribution of each student is the first question we must surmount, but the assessment method of this area still not have enough related research can fully support the teacher to evaluate the students contribution automatically or more conveniently.

In view of these related researches, we can comprehend the collaborative learning and Wiki-based learning process or framework can help learner improve their learning efficiency through collaborative learning; social network also can integrate with wiki service to promote the scale of social learning.

### 3 System Designs

In this section, we will present the system designs of this Wiki-based past exam system.

$$P_{total} = \sum P_{norm} = P_{forum, norm} + P_{peer-review, norm} + P_{links, norm} + P_{content, norm} \quad (1)$$

The main assessment formula we applied in this system is shown above, it is from the Trentin's research [4], and the system will accord this method and the assessment data table but improve the scoring process from manual to the system automatic, try to exclude the human factor from the teacher or other users, the main scope of this study is assess the separate contribution of each question through this method and display it to other users.

The MINE Wiki site is built by Mediawiki: the open source wiki-site software, as shown in Fig. 1, is the portal page of MINE wiki, when the user access the web site, the main pager will shows the broadcast message in middle of the windows, and the navigation area on the left side, can allow users access into the main page to start the learning sequence.

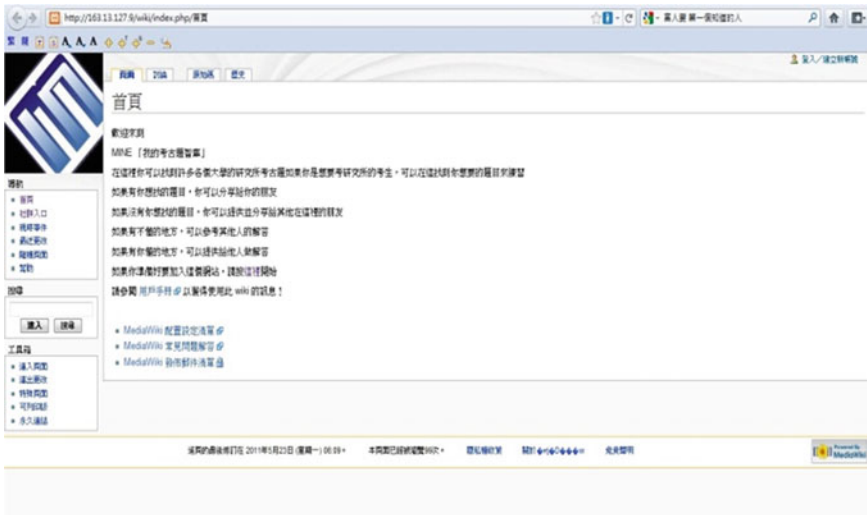


Fig. 1 The portal page of MINE Wiki



Fig. 2 Subject select page

If the user is a unregistered, the user only can allow to search and watch the past exam questions, can not edit or deliver any answer on the page and wiki, even user was invited by register user, still need create an account in this wiki.

When user login the community of MINE Wiki, the first page entered was the list of the four regions in Taiwan, in this research, MINE Wiki's default content we provide is the every universities in Taiwan, in other words, this site have the all past examination question file about Master degree and PhD. Degree's admission, users can search all examination's file in this wiki, in Taiwan, the higher education's examination's past question just provide the questions but without answers, so the examinees must calculate the answer by itself.

After user choose the region, it will enter the university select list page, when user choose an university, the system will navigate to the collage list page shown in Fig. 4, this page shows the whole colleges in this university.

When user selects the college, as shown in Fig. 2, system will list the all subject in this collage at least past five years, user may choose any subject which he needs here.

The most important part in this wiki based site, is the past examination papers, as shown in Fig. 3, the PDF files will shows in the middle of the page, users can not watch the questions in this page, and tag or mark the specific question in this file, also will save in user's learning profiles, allow users to share or take down. The bottom of this file, is the message board, any users who ever access this files and leave the question, answers or messages all will save in this place to modeling the learning process of this past examination paper.

The last part of this site, shows in Figs. 4 and 5, the ELGG social site, a open source social network website, just like the Facebook, in our Lab., we use this open source software to build a social network learning website, test and collect the social learning data to analyze the learner's learning efficiency.



Fig. 3 Past examination paper

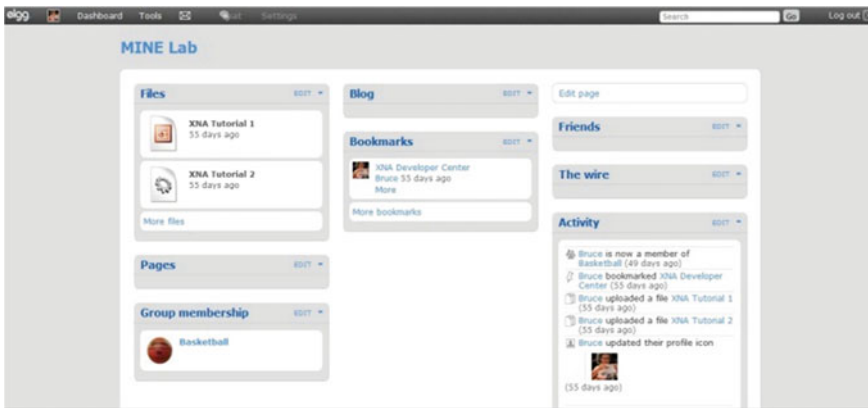


Fig. 4 ELGG social site

The MINE Wiki also merge this social network website, when users can't figure out the answers in the Wikis, he may publish the message to the wall in ELGG, the friends of the users will see the issue in their wall, if anyone know the answers or meet someone who may have a direct thinking way to the right answers, just need to link or forward to other friend's walls, the system will record this path of spread, build a social learning model in this Wiki.

Figure 5 shows the assessment result of each past exam paper, the system will record the contribution and the interact with other users, this result will save in the account profile, every user can access the contribution made by the user in each past exam subject, teacher also can according this result to evaluate the student performance in this social co-writing learning process.

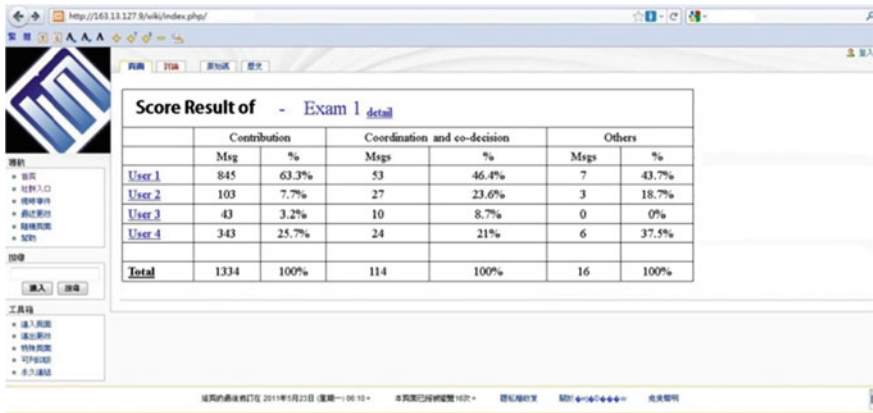


Fig. 5 Assessment result

### 4 Conclusion

In this study, we proposed a Wiki-based past exam System which can provide examiners search and discuss with other people in the website, this Wiki system provide lots of past exam questions of all Taiwan University’s Master Degree admission, if user can not find the past examination questions which he need, user can upload any question PDF files by himself, and build the wiki page about these questions, then deploy to his own social network website like Facebook, twitter, pluk, Elgg...etc. to ask friends who can resolve the questions or join the discussion, even they are all can not figure out the answer, they still can deploy to their own social network to ask another friends, use the strength of the social network, extend the co-writing learning field to the internet.

The system will record this process, and build a learning process of this past examination questions file, analyze the efficiency of learning. Construct the analyze result to the teacher or other users, allow teacher to use these resource to understand what kind of skills and gain student learned during this co-writing learning activity, it also can let other users can understand their knowledge contribution turn into normalization numbers, compare with others, comprehend their knowledge level they reach.

Our future work includes: (a) find out more appropriate user model or education theory to support this wiki based web service even blended social learning; (b) after build a robust user model of this learning process, we may design a assessment mechanism with strong education background support to evaluate users learning efficiency more accurate.

## References

1. Chang B, Zhuang X (2008) Wiki-based collaborative learning activity design: a case study. In: International conference on computer science and software engineering
2. Tseng S-S, Weng J-F (2009) Wiki-based design of scientific inquiry assessment by game-based scratch programming. In: IEEE international conference on advanced learning technologies (ICALT)
3. Liu B, Chen H, He W (2008) Wiki-based collaborative learning: incorporating self-assessment tasks. In: Proceedings of the 4th international symposium on Wikis (WikiSym), ACM
4. Trentin G (2009) Using a wiki to evaluate individual contribution to a collaborative learning project. *J Comput Assist Learn* 25(1):43–55
5. Lu Q, Chen D, Hu H (2010) Wiki-based digital libraries information services in China and abroad. In: Wireless communications networking and mobile computing (WiCOM)
6. Leuf B, Cunningham W (2001) *The wiki way: quick collaboration on the web*. Addison-Wesley, Boston
7. Cubric M (2007) Wiki-based process framework for blended learning. In: Proceedings of the 4th international symposium on Wikis (WikiSym), ACM
8. Cubric M (2007) Using wikis for summative and formative assessment. In: International online conference on Re-engineering assessment practices (REAP), May 2007
9. Xu L (2007) Project the wiki way: using wiki for computer science course project management. *J Comput Sci Coll* 22(6)
10. Jones A, Issroff K (2005) Learning technologies: affective and social issues in computer-supported collaborative learning. *J Comput Educ* 44(4):395–408
11. Sutherland JA, Topping KJ (1999) Collaborative creative writing in eight-year-olds: comparing cross-ability fixed role and same-ability reciprocal role pairing. *J Res Reading* 22(2):154–179
12. <http://www.en.wikipedia.org>

# Universal User Pattern Discovery for Social Games: An Instance on Facebook

Martin M. Weng and Bruce C. Kao

**Abstract** With the population of social platform, such as Facebook, Twitter and Plurk, there are lots of users interest in playing game on social platform. It's not only the game style they want to play, but also high interaction with other users on social platform. Hence, the development of social network makes social games as a teaching tool becomes an emerging field, but in order to use the social network games for teaching, it needs to understand the game flow as an appropriate reference to be judged. In this paper, we use simulation games and social utility games of Facebook as examples. Using the flowcharts and the triangulation methods theory to analyze the characteristics of the flows by finding verbs and goals, and obtaining the differences by comparing social behavior, accumulate of experience, items collection system and tasks.

**Keywords** Social game · Flowchart · Simulation game · Social utility game

## 1 Introduction

Web 2.0 application, such as Facebook and Twitter, have dramatically increased within the last 5 years for social purposes. The provision of common channel for increasing social interactions has revealed the major attractive issue, The survey [quote the related papers] also presents that the entertainment part (e.g. add-on games) has the potential for raising the interactions.

---

M. M. Weng · B. C. Kao (✉)

Department of Computer Science and Information Engineering, Tamkang University,  
New Taipei, People's Republic of China  
e-mail: acebruce@gmail.com

M. M. Weng

e-mail: wm25@hotmail.com

We found many attractive features in social games. But how to use the advantages of online social network for education? And how to use social games as a new game-based learning method? It is still an untapped research issues. However, there are many types of social games in social community. If we want to find out the developing prototype in various of social games which consistent with the educational purposes, we need to comprehend the process in different type of social games first.

In this paper, we use simulation games and social utility games from Facebook as examples, and generalize the game frameworks of the two types we mentioned before. Finally, we make analysis and comparison. Facebook is one of the emerging social websites in recent years. A social network is a set of clustered nodes (group of people) or single nodes interconnected to transmit the information from one cluster or node to another. If we compare Facebook with other social websites, we can find that Facebook offers many services and applications. And one of the popular application and service is social game. Because of the information transformant and sharing functions in social networks are cooperative, social games are designed to use the way of cooperative games [1]. In this contention, author consider that social games have the social property of interaction and cooperation.

## 2 Related Works

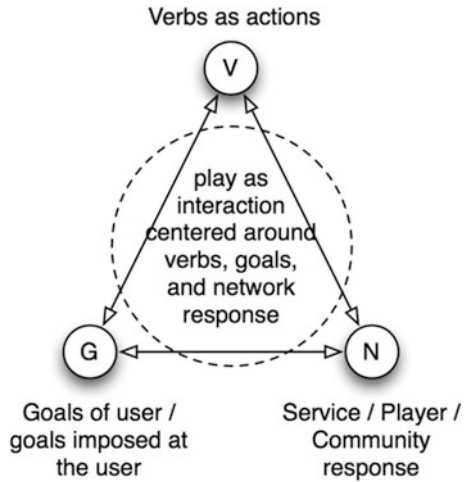
### 2.1 *Triangulation Method*

Aki Järvine proposeda Verbs–Goals–Network play model of triangulation, which helps to define ‘game mechanics’ [2]. The verbs as mechanics are linked with the goals of the game, which are the means to reach the ends. In social network games, the system of the social network is as a whole, consisting of the service, individual players, and the community. The dynamic within these elements can be conceptualized as a triangle with the three elements, around the user experience that start to emerge as play (Fig. 1).

This paper refers to the model as the method which analyzes the features of the game process. On social platform, we only focus on Facebook which is the most popular social platform in social community. Then looking for the commonality among player’s actions or behaviors in flowcharts, sort out the corresponding verbs to reach the goals. Verbs and goals are the basis in this research which use to judging whether game processes are similar or not.



**Fig. 1** Verbs-goals-network play model



### 3 Features of the Games

Since the current game types are becoming more and more composite. The basis for subjects' selection in this study is selecting the games which do not mix with the elements in other game types significantly. This paper selects three simulation games: Restaurant city, My Fishbowl, Happy Harvest, and three social utility games: Hugged, Give Hearts, The King of Kidnappers, to be the subjects of this study We use those selecting games to sort out their own game flowcharts, extracts common features and compares them.

#### 3.1 Simulation Games

The following contents are the three simulation games which we select in this paper:

- Restaurant city

Restaurant city is a restaurant simulation game, players open their own restaurant, and find some ways to collect ingredients, increase their menu and keep the operation of the restaurant. When the level up, it increase the scale of the restaurant gradually.

- My Fishbowl

My Fishbowl is a fish raising game. Players can see their own fishbowl when login to the game, and obtain treasures in the process of raising fish or access friends' fishbowls to steal treasure. Treasures in the game can be sold into money

for buying more kinds of fish or fish feeds, which continue to operate their own fishbowl.

- Happy Harvest

Happy Harvest is a farming game, players' main work is planting crops and selling them into money. They can also go to friends' house to steal crops or hinder crops grow. The money can be use to buy items needed for farm operation.

### ***3.2 Social Utility Games***

The following contents are three social utility games which select in this paper:

- Hugged

Hugged is a simple game to communicate emotions with friends. Players select their friends and send them a type of hug. And the receiver will receive the hug and the associated picture by message.

- Give Hearts

Give Hearts is a game of heart gifting, players select hearts and gift to their friends. In the game, we can see the types of hearts he collect and the ranking of hearts number friends' get.

- The King of Kidnappers

The King of Kidnappers is a game of kidnapping and rescuing friends. Players use all kinds of funny ways and tools to kidnap their friends. Players may become captives by their friends, and use the way which guess the hiding place to escape or rescue friends.

### ***3.3 Features of the Games***

- The use of currency

The reason why virtual currency is developed have been mentioned by Hui Peng and Yanli Sun. Then main reason is mentioned as below: 'The demands of virtual economy development', 'Technical progress and internet development', 'Demands from users', and 'The profit maximization of issuers' [3]. Those games are developed on Facebook platform. It will certainly be developed and use virtual currency due to the demands from users and issuers, with sufficient support of network technology.

- Social behavior

The nature of social networks is built during play [4]. Because the games are developed on the social networking platform, spontaneously be inseparable with community interaction. But they stressed that the players interact with the user must have any relationship with other users (such as friends). This means that social games have some standard to the scope of relationships between interaction players and the user. Only players who have a certain relationship with the user can participate. This can avoid unacquainted players allot the incomes which get from interactions in the games (steal resources). Nabeel Hyatt has pointed out how social network games can ‘rely heavily on social context (namely school, department, and residence loyalties) to provide a framework for alliances, game playing and motivation’ [5]. Social context is the interpersonal relationships in the player’s real life, and social networking platform extends the relationships into the games. This also becomes the main scheme of the interactions in simulation games.

According to Aki Järvinen’s argument, game design techniques are needed to integrate competition, challenge, and tension into those acts of socializing, and mostly of the integration would improve the playfulness of the games [2]. The competition in social games are established in the interactions with friends, but maintain the idea of non-zero-sum games basically, with certain risk control. Friends can compete with each other, but don’t come to win or lose directly; friends can steal each other’s resources, but there are certain quantitative restrictions, this protects the players’ incomes.

- Accumulate experience points to upgrade and get reward

Kirman, Lawson, and Linehan have mentioned that in *Fighters’ Club* and *Familiars*, player score (Street Credit in FC) is a function of the social behavior of the person within the game [4]. This paper considers that in most social games which use experience points, one of the sources of experience points is also a function of the social behavior of the person within the game. In addition, the experience points are also the incomes of a player’s labor in the game. As the same with score, the more a player works, the more experience points he will gets.

Valentina Rao lists fast rewards for player actions, abundance of positive feedback, no negative consequences for exploration [6]. Experience point is an income in fast rewards, which won’t be reduced by player actions. This encourages players to act in the game positively. The value of the experience points that player gains will enhance his status and reputation in social network games. As the triangulation method mentioned, this cyclic process of development plays the core mechanics of the games [2]. The incomes of the players’ operations will reach the goals. Social behavior such as visiting friends, is one of the source of experience points, and in this process, exchange and compare with friends will be spontaneous.

- Item collection system

Kim has identified certain core game mechanics, i.e. player actions, such as collect and exchange [7]. Those actions promote the games to run. Both of the items collected and accumulated are experience points that are incomes from players' actions in the game, the difference is that items collected in some games may burden risk because the players cannot always keep in the game, for example, goods are stolen by friends, but experience points cannot be stolen by using any channel; This is the difference between item collection and experience points accumulation.

Item collection also led the need to add features to the game. Hence, it would evolve with players' needs, achievements in the form of different badges, new types of cars, etc. [2].

- Task

Task has a feature that is various situations. No matter how many changes are in the form of the tasks, generally around the verbs, goals theories of triangulation method, simply changing the way of the tasks to add more fulfillment.

## 4 Conclusion

In recent year, social network issues have become very popular and the application of social network also play an important role on related research issues.

In this paper, we used the flowcharts and triangulation method theory to analyze and compare the characteristics of simulation games and social utility games in Facebook. As a contribution, the discovered pattern can be applied onto the design of social learning games by integrating customized factors. The result will be provided as suggestions when developing the related social games with educational purposes. In the future, we will focus on more social game on Facebook or other social platform, and analyze the flowchart of game and show the compare results of different games in similar categories.

## References

1. Reddy YB (2009) Role of game models in social networks. In: Conference on computational science and engineering (CSE), pp 1131–1136
2. Järvinen A (2009) Game design for social networks: interaction design for playful dispositions. In: Proceedings of the ACM SIGGRAPH symposium on video games, pp 95–102
3. Peng H, Sun Y (2009) Network virtual money evolution mode: moneyness, dynamics and trend. In: Conference on information and automation (ICIA), pp 550–555

4. Kirman B, Lawson S, Linehan C (2009) Gaming on and off the social graph: the social structure of facebook games. In: Conference on computational science and engineering (CSE), pp 627–632
5. Hyatt N (2008) What's wrong with facebook games? <http://nabeel.typepad.com/brinking/2008/01/whats-wrongwit.html>
6. Rao V (2008) Playful mood: the construction of facebook as a third place. In: Proceedings of the 12th international conference on entertainment and media, Mindtrek 2008, pp 8–12. <http://portal.acm.org/citation.cfm?id=1457199.1457202&coll=Portal&dl=GUIDE&CFID=24746181&CFTOKEN=85617762>
7. Kim AJ (2008) Putting the fun in functional, applying game mechanics to functional software. <http://www.slideshare.net/amyjokim/putting-the-fun-infunctiona?type=powerpoint>
8. Sharabi A (2007) Facebook applications trends report, 19 Nov 2007. <http://no-mans-blog.com/2007/11/19/facebookapplications-trends-report-1/>
9. Järvelin K, Kekäläinen J (2004) IR Evaluation Methods for Retrieving Highly Relevant Documents. In: ACM international conference on information retrieval
10. Jensen D, Neville J (2002) Data mining in social networks. In: Proceedings of national academy of sciences symposium on dynamic social network analysis
11. Thomas LS (1990) Decision making for leaders—the analytic hierarchy process for decisions in a complex world. RWS Publications, Pittsburgh
12. Andersson N, Broberg A, Bränberg A, Janlert L-E, Jonsson E, Holmlund K, Pettersson J (2002) Emergent interaction—a pre-study. UCIT, Department of Computing Studies, Umea University, Umea, Sweden
13. Caillois R (1962) Man, play and games. Thames and Hudson, London, p 12
14. Carroll JM, Aaronson AP (1988) Learning by doing with simulated intelligent help. Commun ACM 31(9):1064–1079
15. Chalmers M, Galani A (2004) Seamful interweaving: heterogeneity in the design and theory of interactive systems. In: Proceedings of the ACM designing interactive systems (DIS2004)
16. Dourish P (2004) What we talk about when we talk about context. Pers Ubiquit Comput 8(1):19–30
17. Prensky M (2001) Digital game based learning. McGraw-Hill, New York
18. Sierra JL, Fernández-Valmayor A, Fernández-Manjón B (2006) A document-oriented paradigm for the construction of content-intensive applications. Comput J 49(5):562–584

# Ubiquitous Geography Learning Smartphone System for 1st Year Junior High Students in Taiwan

Wen-Chih Chang, Hsuan-Che Yang, Ming-Ren Jheng  
and Shih-Wei Wu

**Abstract** In recent years, ubiquitous learning becomes more and more popular. Geography learning can be adapted into smart phone platform to be very useful learning system. Junior high school students can assess the smartphone to study geography in Taiwan. With simple test items, the system will generate individual learning profile and test analysis report for students.

**Keywords** Ubiquitous learning · Geography learning · Smartphone

## 1 Introduction

Over the years, the progress of the E-Learning shows the significant development in content digitalization and learning technologies obviously. During this progress, the kernel value of E-Learning is created by constructing new learning style with new technologies such as network or multimedia. Meanwhile, some traditional pedagogic theories shifted to underpin the development of learning style.

Many studies tried to apply auxiliary elements to conventional e-learning for improving the learning efficiency and enriching the learning motivation. In addition, the mobile technology is now widely used and the related mobile facilities are

---

W.-C. Chang · M.-R. Jheng · S.-W. Wu  
707, Sec.2, WuFu Rd, Hsinchu, 30012 Taiwan, People's Republic of China  
e-mail: yilan.earnest@gmail.com

M.-R. Jheng  
e-mail: jhengmingren@gmail.com

S.-W. Wu  
e-mail: sware1786@gmail.com

H.-C. Yang (✉)  
152, Sec. 3, Beishen Rd, Shenkeng dist, New Taipei City, 222 Taiwan People's Republic  
of China  
e-mail: hsuanche.yang@mail.tnu.edu.tw

also affordable to the public. As a result, more and more research put emphasis on the integration of e-learning and mobile technology, and the terms of “m-learning” (mobile learning) and “u-learning” (ubiquitous learning) then become the magic words while talking about the technology enhanced learning. The mobile learning allows the learning activities to be performed no longer limited to specific location, and accordingly provides more learning opportunities for learners with the mobile technology. Churchill and Churchill [1] pointed out the mobile devices play the roles of multimedia-access, connectivity, capture, representational, and analytical tools for mobile learning activities. Eschenbrenner and Nah [2] also revealed that the learning performance and efficiency can be much more improved with the benefits brought by mobile technologies. They also found that the mobile learning activity can be applied to encouraging the ability of problem solving, and to providing the opportunity of self-regulation learning. Furthermore, with the mobile technology supported, it's more practical to achieve the collaborative learning. The mobile learning is not only changing the way of traditional learning style and behaviors, but also impacting the way in future learning. Most of the m-learning applications and researches mainly focus on the “mobility”, the first half of “m-learning”, including the accessibility, the transferability, and the content delivering through the mobile learning devices, which makes learning activities no longer limited to specific time and space. However, there is the other half in m-learning, and that is the “learning” which is usually disregarded in the learning context. As a result, it's hard for people to find out the interrelationship between mobility and learning activity. An approach to this issue is typically called the location-aware learning that integrates both location-based and contextual learning activities.

Geography learning needs map and geographic background knowledge for beginner. Geography forms the basis for understanding our political and physical realities. A great of geography education can be effectively taught through geography interactive learning materials and maps. Combining the mobile devices, the geography can be more fun and attractive.

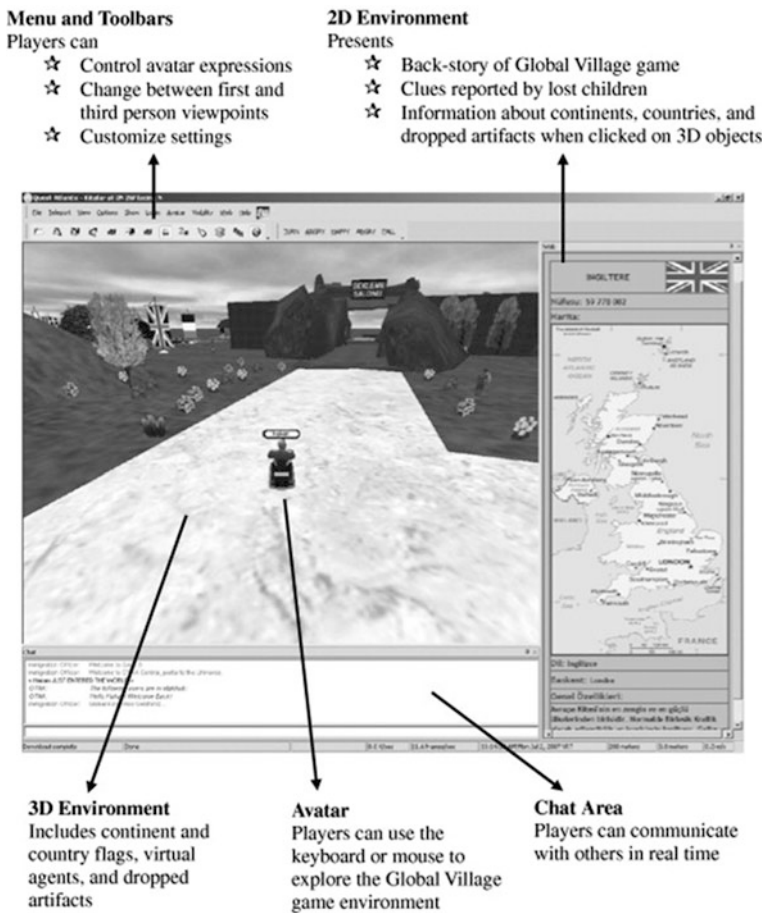
The organization of this paper is as follows. [Section 2](#) introduces related technologies and background knowledge in this work. In [Sect. 3](#), we discuss the design of the smart phone learning applications for geography learning topic, and we also take account some significant pedagogical methods to support our proposed ubiquitous learning environment. In [Sect. 4](#) comes the conclusion.

## **2 Related Works**

### ***2.1 Geography Learning***

Geography is a Obligatory course from elementary education. Geography learning makes people learn human, physical and geography environment in the real life. In traditional geography learning, teacher used the map and figures assist learners in the classroom.

Hakan et al. [3] designed and developed a three-dimensional educational computer game for learning about geography by primary school students. Twenty four students in fourth and fifth grades in Ankara, they learnt about world continents and countries through this game for 3 weeks. The effects of the game environment on students' achievement and motivation made significant learning gains by participating in the game-based learning environment. These positive effects on learning and motivation, and the positive attitudes of students and teachers suggest that computer games can be used as an ICT tool in formal learning environments to support students in effective geography learning (Fig. 1).



**Fig. 1** MAP game and geography game. *Source* <http://www.sciencedirect.com/science/article/pii/S0360131508000985>



## 2.2 Smartphone Platforms

Generally speaking, the mobile devices include notebook, Tablet PC, Ultra Mobile PC (UMPC), Smartphone, Personal Digital Assistant (PDA) and other portable devices with computing capability. With the phenomenal growth of GPS technology, the mobile devices equipped with the GPS functionality are now wildly used to take a huge leap toward location-aware computing. It also facilitates the integral services of complex computing and personal information seamlessly. In order to realize our proposed ubiquitous learning environment, we use the Smartphone device as the location-aware learning platform due to its flexibility and expansibility.

Some Smartphone provide open architecture as desktop computer with standard Application Program Interfaces (APIs) to allow the varied developments from the third parties. Therefore the Smartphone is so-called an open operating system and is also a mobile phone with the capability of running applications. Typically the Smartphone has the network capabilities and is often equipped with a build-in or slide-out QWERTY keyboard (Fig. 2).

## 2.3 Mobile Learning

In environment of traditional classroom, students are restricted in closed space. The way students to absorb knowledge is from what Teacher teaches, And context of books are boring to students, which lessens the efficiency of learning. The above-mentioned learning way is typical passive learning [3]. Learning is no more restricted to traditional classroom and the knowledge resources are no more to

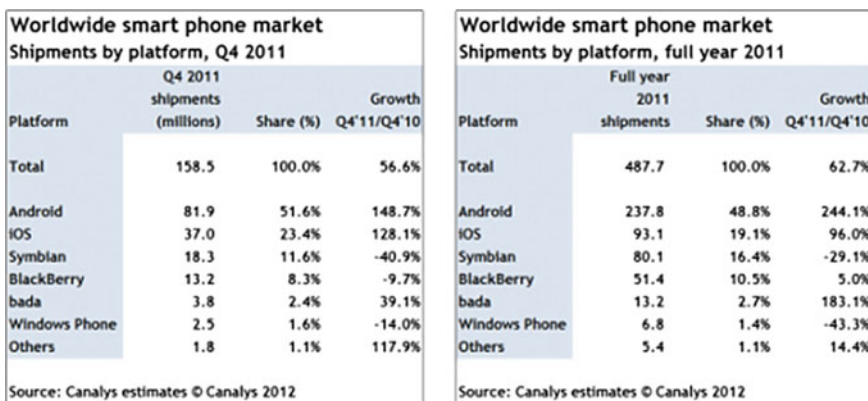


Fig. 2 Analysis on market shares of smart phone by operating system in full year 2011. Source <http://paidcontent.org/2012/02/04/419-canalis-worldwide-smartphone-shipments-overtake-pc-tablet-market/>

textbooks. Nowadays, people can learn ubiquitously via internet, which makes learning more flexible and enjoyable [4]. Technological hardware is a must for school learning environment; meanwhile, we cannot ignore the importance of learning in natural environment because technological learning cannot be a substitute for interaction with nature while learning [5].

## ***2.4 Ubiquitous Learning***

With the development of embedded system and sensor technology, people obtain useful information through sensor network around our environment. According to individual need, the systems provide the adaptive needs automatically which called ubiquitous computing [6, 7]. Taiwan government promotes “U Taiwan” which uses RFID and wireless services integrating on digital family and internet. Under the infrastructure and hardware support, ubiquitous learning becomes a new learning trend. The ubiquitous computing makes ubiquitous learning more easily. Based on ubiquitous learning system support, students can learn more around the living world. U-learning (ubiquitous learning) connects the back end database; analyze knowledge cognition distribution and supports immediate learning feedback after students respond.

## **3 The Ubiquitous Geography Learning System**

We proposed the geography ubiquitous learning system which is composed of population, industry (I), industry (II) and traffic learning content for junior high school students. The following shows the learning content.

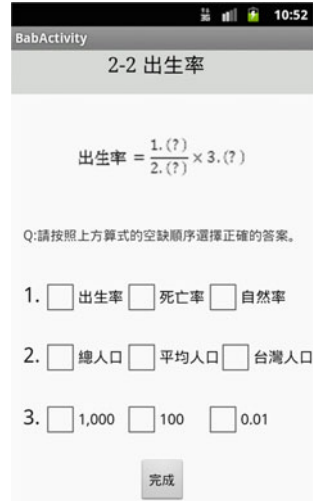
### **(1) Population**

It can not only provide adequate human resources but also be good development of a country or region by appropriate population, healthy population structure and excellent quality of the population. Due to the high population density, living environment is getting more and more pressure and potential impact. In recent years, the changes of population structure is caused by “the low birth rate”, “aging” and “international migration”. The main content is composed of Taiwan’s population size and distribution, Taiwan’s population growth, Taiwan population migration, Taiwan’s demographic composition and Taiwan’s population problems.

### **(2) Industrial (I)**

Industrial is in response to a variety of human needs, of which agriculture is the basis of human beings for living. It provides good growth environment for crop by variety of terrain, and warm and humid climate in Taiwan. With the rapid economic development, Taiwan’s agriculture has been towards the refinement and

**Fig. 3** Population choice test item



**Fig. 4** Population choice test item



development of leisure and tourism. A wide variety range of agricultural products enhance the added value of agriculture and create a new style for Taiwan’s agriculture. The main content is composed of Taiwan’s industrial structure, Taiwan’s primary industrial and Globalization of Taiwan’s agricultural.

(3) Industrial (II)

In the past, Taiwan had the name “banana kingdom” and then it is loud of this name “Leather shoes kingdom”. Nowadays we strengthen industrial restructuring actively and effort to create “Boutique of Taiwan”. It is the direction of Taiwan’s industrial development to build brand, pursuit of high value-added and further

Fig. 5 Industry (I) choice test item

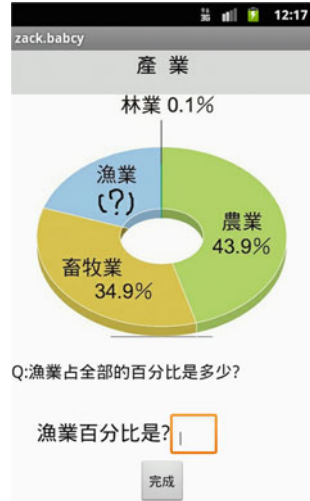


Fig. 6 Chapter score summary



more promote economic growth. The main content is composed of Industry, Characteristics of Taiwan's industrial development, Service sector and Globalization of Taiwan's service sector.

#### (4) Traffic

With technological advances, transport is an indispensable part of modern life. Modern transport brings us comfort and convenience, as well as the pace of life is more compact. Taiwan's transportation develops rapidly and its efficacy is rising in many forms. It is an important foundation for Taiwan's internal and external contact. The main content is composed of traffic types, Taiwan's transport and Taiwan's communications (Figs. 3, 4, 5, 6).

## 4 Conclusion

The Geography ubiquitous learning system provides smart phone platform for junior high school students learning. We completed four chapters for first year junior high schools in Taiwan. It is composed of population, industry (I), industry (II) and traffic. The system adapted multiple choice problem, fill-in blank problem, match problem types. Students can assess this system offline. It promotes students learning motivation and learning confidence.

**Acknowledgments** We would like to thank the NSC for funding this research under grants NSC-97-2511-S-032-006.

## References

1. Churchill D, Churchill N (2008) Educational affordances of PDAs: a study of a teacher's exploration of this technology. *Comput Educ* 50(4):1439–1450
2. Eschenbrenner B, Nah FF-H (2007) Understanding highly competent information system users. In: *SIGHCI 2007 proceedings*
3. Tüzün H, Yılmaz-Soylu M, Karakuş T, Inal Y, Kizilkaya G (2009) The effects of computer games on primary school students' achievement and motivation in geography learning. *Comput Educ* 52(1):68–77
4. Lin T, Chen T (2007) The study of instructional design and learning performance in context-aware mobile learning environments. In: *2007 Information education and technological applications conference (IETAC)*
5. Chen C-H, Chen Y-M (2006) The study of integrating global positioning system into mobile learning. *National Pingtung University E-Learning 2006, Pingtung*
6. Chen T-S, Chiu P (2007) A study of learner's behavioral intentions in a context-aware mobile learning environment. In: *2007 National computer symposium*, pp. 20–21, Dec 2007
7. McDonald DS (2004) The influence of multimedia training on users' attitudes: lessons learned. *Comput Educ* 42(2):195–214
8. Mark W (1999) Turning pervasive computing into mediated space. *IBM Syst J* 38:677–692

# Housing Learning Game Using Web-Based Map Service

Te-Hua Wang

**Abstract** Nowadays E-learning is widely used in teaching and training purposes, and it also facilitates conventional classroom-based learning activities with advanced information technologies. On the other hand, such advanced information technologies also make digital learning content easily applied to some acknowledged learning theories, such as Behaviorism, Cognitivism, Constructivism, and make e-learning being more practical and diverse in many pedagogical purposes.

**Keywords** Game-based learning · Web map service · Self-regulated learning · Learning motivation

## 1 Introduction

Many acknowledged learning theories, such as Behaviorism, Cognitivism, and Constructivism, can be realized with the improving cutting-edge e-learning technologies. The aim of E-learning has gradually turned into the process-oriented learning approach from the traditional content-based learning materials. In this paper, we proposed an online game-based learning platform utilizing the web-based map service technology. The game element lies in the completion and challenges of specific housing game missions. In addition, given the diversity of online maps mash-up services and the rise of online virtual community gaming platforms, we merge the learning content with the game-based housing missions into the proposed digital map. Learners are able to make an arrangement for acquiring the assigned game-based learning object for specific game-based learning activities according to their own preferences. With the accomplishment of

---

T.-H. Wang (✉)

Department of Information Management, Chihlee Institute of Technology,  
New Taipei City, Taiwan  
e-mail: tehua@mail.chihlee.edu.tw

the learning activities, learners then can get visualized learning feedback on the game-based e-map. By introducing essential game elements to web-based learning activities, we aim at providing an attractive learning platform to motivate learners to get more involved and engaged during the online learning process. Furthermore with the storyline of the housing missions, we expect learners to develop attitudes toward active and self-regulated learning, and to realize the importance of accumulating knowledge is similar to accumulating the housing budgets, just as the old saying goes, “many a little makes a mickle”.

## 2 Related Works

One of the most significant issues for e-learning lies in the way of enriching the learning motivation. Accordingly many outstanding e-learning researches proposed various mechanisms to motivate learners via multimodal learning activities and found that the essential elements of gaming, including challenge, curiosity, fun, instant feedback, and achievement, provide the best solution.

Gaming itself attracts people with interesting storyline and fantastic interactions, which make the players to be willing to dedicate the attention to specific gaming scenario. Game-based learning can be applied to enhance the interactivity and richness of digital learning content. By combining the game situation and the aims of education, the learning motivation can be stimulated and the attractiveness of the learning process can be enhanced as well [8]. Another interesting study also pointed out that the different game types will have different effects on the pedagogical objectives [1]. All the first and second year college students in biology department participated in the experiment; four independent groups of subjects each were applied to different types of game-based learning, including simulation games, strategy games, narrative-driven adventure games and first person shooter games. The study concluded that the adventure games and strategy games are much more acceptable for learners to get better learning performance, since these two types of games require more logical thinking, memory, imagination, and problem-solving ability. And these factors also play the essential roles of game-based learning. The research in [3] defines and conceptualizes the key factors of successful game-based learning, such as Identity, Interaction, Risk Taking, Customization, Situated Meanings, System Thinking and etc. Another work proposed a comparative analysis between entertainment and learning and found that the most challengeable and attractive game for learning is puzzle game. Puzzle game is good for understanding the appearance of the object, and the process of the puzzle develops learners' organizing ability, as well as the ability for pattern analysis [4]. Squire found that the incompleteness of game might be due to the insufficient prerequisite knowledge and is helpful for motivating learning to conquer the game barriers [10]. A significant example of successful game-based learning can be found in [7]. The authors pointed out light game for learning should contain some essential characteristics as serious game, including awarding, challenging, curiosity, fantasy,

objectives, competition, cooperation and achievement, to improve intrinsic motivations for learning. Another point to note is the possible mash-up services covered in web-based technologies. Such mash-up applications can be easily found in Web-based Map Service (WMS), such as e-commerce, traffic, broadcasting, parking, online society, and etc. Map provides information by instinct, and has the attribute of leading direction. People can learn from environment about the location-aware information, including landmark knowledge, route knowledge and survey knowledge [6, 9]. To avoid learning astray, a map could be used to represent the overall interrelation among specific subjects.

### 3 Educational Game Design

The current trend of e-learning is towards edutainment, providing an interactive and attractive learning environment. In this work, we target at the intuition of the e-map service to enrich both the learning motivation and the learning accomplishment. With respect to the game design, the aggressiveness of being rich or having lots of houses would become a symbol of great achievement. Accordingly, we put the aggressiveness of getting rich into the game-based learning activities on the e-map. Learners are able to learn various topics on the e-map, and the acquisition of knowledge and the feedback of learning competence can be considered to accumulate the property as the housing funds for different types of buildings. Eventually, we look forward to realizing the old sayings “reading brings us everything,” by using our game-based e-map learning platform. Furthermore, the learning sequence in the e-map can be self-regulated according to the learning interests. Learners are able to select the appropriate sequence of learning activities and develop the ability to seek knowledge. So that learners can intuitively understand the learning targets by using e-map and plan appropriate sequence to complete the learning tasks. And not only the objectives of learning content are obtained, but the logical thinking skills and self-learning ability are also enhanced. The task-oriented learning strategy is a constructivist teaching theory suitable for developing self-learning ability, as well as the problem-solving ability. Bae et al. revealed that learners are able to achieve predefined learning objectives via specific tasks, missions or barriers in game scenario [2]. As a result, activity-based and task-oriented learning strategy can be easily applied to edutainment. Another research defined activity theory as a philosophy process and interdisciplinary approach describing human development [5]. In other words, activity theory utilizes various exercises and practices to acquire knowledge, and affiliates learning objects with learning experience.

The abovementioned game-based learning research and practical teaching strategies highlight the process of the game and provide training activities allowing learners to find a better way to achieve the objectives according to gaming rules and storylines. A well-designed educational game provides various ways to achieve the learning objectives, and thus, different gaming process will



lead to corresponding learning competency. In this study, we set up a game-based e-map to facilitate web-based learning activities by using housing mission. Eventually, by providing such learning platform, learners are able to develop the ability with self-regulated learning strategies through the task-solving process.

## 4 System Architecture and Implementation

The proposed learning platform mainly delivers learning content, and in addition, during the task solving process, learners are able to develop problem-solving ability, and gradually to accumulate the knowledge of the learning content.

### 4.1 System Architecture

The system architecture is discussed in three components. The first component is design and management of the backend database, which contains three main data tables to maintain the learning content, learning activity, and the learning portfolio. The second one includes the functionalities and services of the game-based e-map, such as the interactive e-map module, learning resource management module and learning activity management module. The third component aims at the analysis of learning competency and performance. It provides instructors and learners a review module respectively to examine the degree of the accumulation of obtained knowledge. The system architecture and functionalities are illustrated in Fig. 1, and each component can be discussed in detail as follows.

#### Design and Management of the Backend Database

- Learning Content: To record the information of learning content, including title, description, author, routing information, additional learning resource, difficulty, hierarchy, and knowledge domain.
- Learning activity: To connect learning content and the specific house type. Each learning subject can be considered as an activity unit, and instructors can assign the relationship between learners and learning activities.
- Learning portfolio: To track the corresponding feedback in the proposed game-based e-map learning platform, including learner id, obtained house, amount of house, the longitude and latitude information, time, duration, and the unsolved activities.

#### Functionalities and Services of the Game-Based E-Map

- Interactive E-Map Module: to provide the interactive event and to manage the game barrier using the Google Maps APIs. Learners are able to receive learning

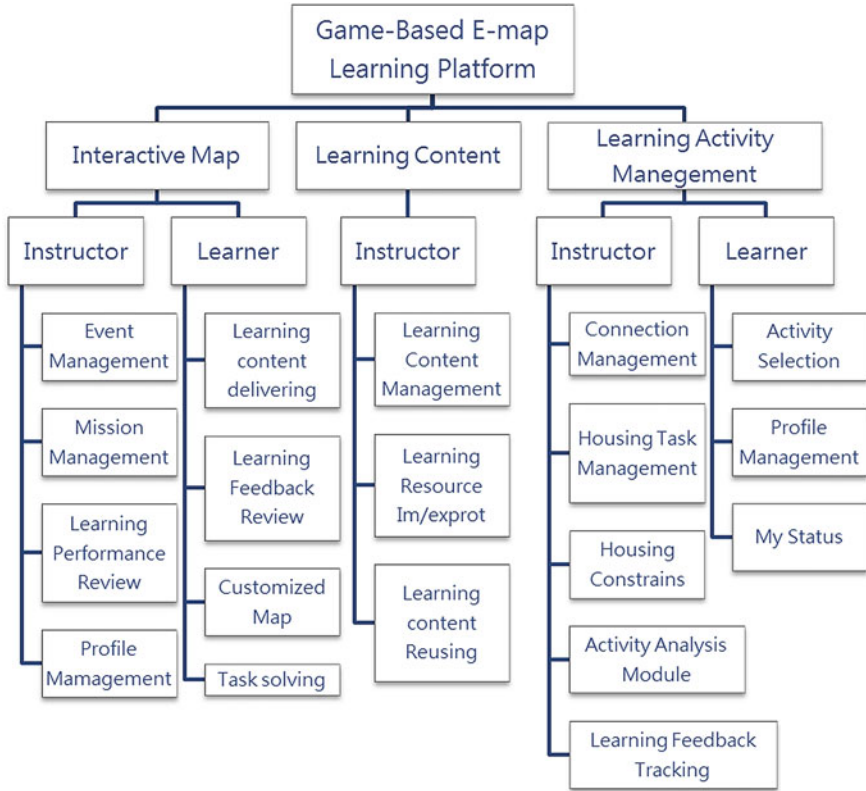


Fig. 1 Game-based e-map system architecture

content, to review the learning outcomes, and eventually to customize the look-and-feel of e-map.

- Learning Resource Management Module: to manage learning resource and to allow instructor to upload, to update, to delete the specific learning resource. Most of the learning resource is based on multimedia, such as audios, videos, and web pages. As a result, the learning resource can be sharable and reusable in this platform.
- Learning Activity Management Module: to relate the learning content and the learning activity. In addition, this module confines the housing task to self-regulated learning sequence. And learners can choose particular learning activity according to their needs.

**Analysis of Learning Competency and Performance**

- Learning Feedback Tracking Module: To examine the learning performance of each involved learner. And the learning process will be sent to the backend server and kept in the learning portfolio module. Instructors are able to analyze and summarize the learning performance.

- Learning Activity Analysis Service: to ensure the quality of service in the game-based e-map learning platform. Also this service can be applied to analyze the performance of learning content delivering.

## ***4.2 Game Design of Housing Learning Missions***

The educational game design in the proposed system is discussed in two directions. The first one is about the learning content design, and the other is about leading gaming factor into the learning activities. We chose “Data Structure” as the experimental subject in the proposed learning platform, and the learning content on the e-map is delivered to learners with the methods supported in Google Maps APIs. To realize the evaluation of learning performance, a corresponding assessment will be triggered when a learning activity ends up, and the result will then be sent to the backend server for updating the learning portfolio.

Another point to note is the way leading gaming factors into the learning activities to improve the learning motivation. To serve this goal, we allow instructors to arrange corresponding house type according to the difficulty and importance of the learning content. And the arrangement represents the relationship between learning content and the requirement of purchasing house on the e-map. Learners have to start the learning activity from the elementary housing level, and accumulate adequate learning property (i.e. houses on the e-map) in successive learning activities. For instance, learners should get a brick house before having three thatched cottages gained from corresponding learning content. Learners can make their own arrangements of the learning activities according to their needs. If learners try to enter the restricted activity without having adequate learning property, the learning management system will alert learners with a popup dialog box. Eventually learners can review the status of housing mission and the amount of different type of obtained houses. In addition, learners are able to view the learning status of other learner with permission.

## ***4.3 Interaction and Learning Feedback***

The interaction and feedback can be considered as the key factors of a successful game. Learners in a game-based learning system might have different skills and sequences to complete the assigned mission. Therefore, in the proposed game-based e-map platform, we allow learners to determine the order for acquiring the same type or the same level of houses according to their preference or arrangement (Fig. 2).

In the proposed e-map, the instant learning feedback can be displayed on the e-map in the form of various house icons (as shown in Fig. 3). These icons on the e-map represent the results after passing corresponding learning activity. As long

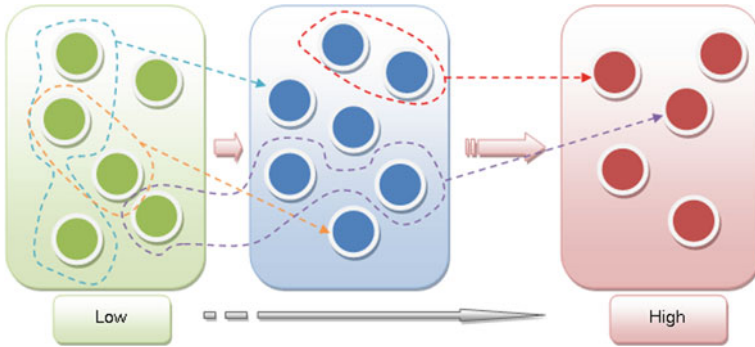


Fig. 2 Self-regulated routing strategy

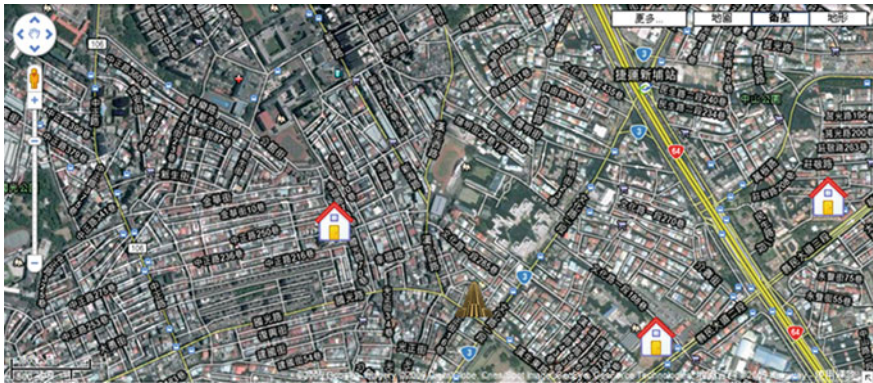


Fig. 3 Game-based e-map in satellite view

as a learner enters a new learning activity, the learning content not only improves the property in knowledge but also the property in housing.

## 5 Conclusion

A game-based e-map learning platform is proposed in this study. Learners are able to acquire learning content after having the predefined learning activities on the e-map with housing missions. The gained knowledge content can be transformed into the property for corresponding buildings or houses. The participating learners can view personal learning status and compare the learning results as a contest with others.

Game-based learning to current educational methods is proven to improve learning motivation, situated cognition and problem-solving skills by providing a seamless virtual world affiliated with learning experience. In real life, housing for

most of the young students is an unattainable activity, and they need to accumulate sufficient assets to achieve the purpose of housing. In the proposed learning platform with the WMS technology, learners can gradually accumulate their knowledge assets. And through the game scenario, learners can be easier to realize the philosophy of a penny saved is a penny earned.

## References

1. Amory A, Naicker K, Vincent J, Adams C (1999) The use of computer games as an educational tool: identification of appropriate game types and elements. *Br J Educ Technol* 30(4):311–321
2. Bae Y, Lim J, Lee T (2005) Work in progress—a study on educational computer, games for e-learning based on activity theory. In: *Frontiers in education, 2005. FIE '05. Proceedings 35th annual conference*, pp F1C-18
3. Gee JP (2003) *What video games have to teach us about learning and literacy*. Palgrave Macmillan, New York
4. Koster R (2005) *A theory of fun for game design*. Paraglyph Press, Scottsdale
5. Kuutti K (1995) Activity theory as a potential framework for human-computer interaction research. In: Nardi B (ed) *Context and consciousness: activity theory and human-computer interaction*. MIT Press, Cambridge, pp 17–44
6. Lynch K (1960) *The image of the city*. MIT Press, Cambridge
7. Malone TW, Lepper MR (1987) Making learning fun: a taxonomy of intrinsic motivations for learning. *Aptitude Learn Instr* 3:223–253
8. Sandford R, Williamson B (2005) *Games and learning: a handbook*. NESTA Futurelab, Bristol
9. Siegel AW, White SH (1975) The development of spatial representations of large-scale environments. *Adv Child Dev Behav* 10:9–55
10. Squire K (2005) Changing the game: what happens when video games enter the classroom? *J Online Educ*

# Digital Publication Converter: From SCORM to EPUB

Hsuan-pu Chang

**Abstract** The resources and applications about ebook have been changing people's way of reading due to the popularization of ebook readers. It also has significantly changed the traditional rules and concepts of publication. Therefore, an important issue we have to face consequentially is how to create qualified digital publications for readers. In fact, there are a lot of excellent e-learning contents have been produced and stored in repositories or management systems. But unlike the convenience of enjoying ebooks with ebook readers, these excellent learning contents have much more complicated design issues and have to be put on specific learning management systems (LMS) due to the conformance of learning strategies and management requirements. Moreover, many excellent learning contents and courses are produced by the teachers who spent a lot of time and energy. They also have the expectation of converting these learning contents to their own private publications. As the result, we propose a file converter which is able to convert the SCORM compliant courses into EPUB publications. The system consist of four modules; Presentation Transforming Module, Metadata Transforming Module, Sequencing and Navigation Transforming Module and Packaging Transforming Module. We look forward to seeing these excellent SCORM learning contents can be wildly distributed and enjoyed with the EPUB format and publications.

**Keywords** Ebook • Digital publication • SCORM • EPUB • Converter

---

H. Chang (✉)

Department of Information and Library Science, Tamkang University, Taipei, Taiwan  
e-mail: musicbubu@gmail.com

## 1 Introduction

Every university department has its particularly professional knowledge and skills applied for employment. Accordingly, department professional competences are set as learning targets that students are expected to possess after graduating from school. Many scholars have attempted to define competence and numerous studies have developed, examined or applied assessment and validation techniques for evaluating performance by analyzing competences [1].

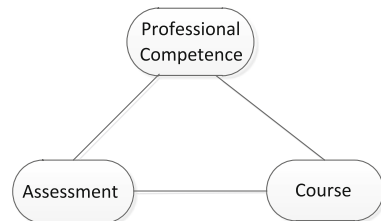
Some scholars [2, 3] described the core intent of competency as “an underlying personal characteristic which results in effective and/or superior performance in a job” or “an underlying characteristic of an individual that is causally related to criterion-referenced effective and/or superior performance in a job or situation.” Competency is defined as the “underlying characteristics” that can be used to predict the job performance of specific professionals. Competences are long-lasting characteristics that comprise motives, traits, self-concept, knowledge and skills, and can be demonstrated in thinking, behaviors or onsite responses. Furthermore, researchers defining competences agreed that they can be trained and improved [3–6]. The trainability of competencies has major implications for training departments and educational institutions.

## 2 The Components of the Evaluation System

The major purpose of this system is assessing students’ department professional competences after a serial of courses designed and arranged by department course developing committee. Generally, in order to represent a student’s learning performance a score is used to express whether the student is qualified for this course’s learning objectives. But in fact a course may include plural professional competences that a student need to learn and pursue. In order to inspect whether a student is mastering professional competences in a course more or less, the relation among course, assessment and professional competence as the Fig. 1 shows.

The following sections detail how we connect the three parts.

**Fig. 1** Three parts for constructing the evaluation system



## 2.1 Build the Relation Between Course and Professional Competence

In this section, a Course-Competence table will be introduced to describe the relation between course and professional competence that a department course committee needs to conduct a discussion for accomplishing it.

### 2.1.1 Construct Course-Competence Table

The first step of carrying out the competence evaluation system is constructing the Course-Competence table that describes the relation between courses and competences. Because a course may contain different professional knowledge and skill for learning, which means more than one professional competences may exist in a course only the matter of ratio. The course committee needs an entire picture of department development for constructing the Course-Competence table. For instance, Table 1 is the department of Digital Information and Library Science (DILS) in Tamkang University (TKU) in Taiwan, the first column lists the a few example course titles of the department and the rest of columns are professional competence index and their ratio contained in these courses. The details of the eight competences A–H are described in the Table 2.

### 2.1.2 Add Competence Information in Course Syllabus

As Fig. 2 shows, while teacher is designing a course syllabus on line, system automatically retrieves the competence information from database and adds it to teacher’s syllabus. It is a significant step that reminds teachers or instructors what competences should be learned after taking this course. The competence information gives teachers a direction to prepare teaching content and design learning activities. Meanwhile it’s also an opportunity to reconsider whether the relation set between the competences and the course is properly matched.

**Table 1** Course-competence

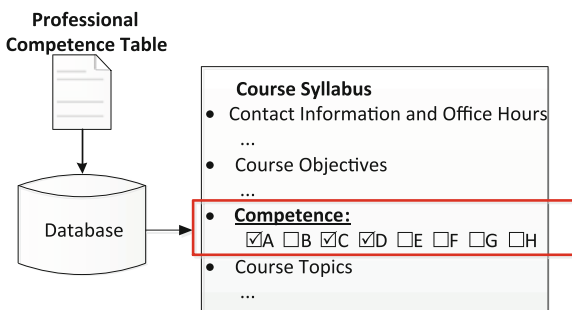
Course	Competence ratio (%)							
	A	B	C	D	E	F	G	H
Introduction to librarianship & information science	20	20	15	20	20	0	5	0
Statistics for library science	0	30	35	0	35	0	0	0
Introduction to innovative publishing industry	0	0	0	20	0	0	40	40
Archive management and development	0	25	15	0	20	40	0	0



**Table 2** Professional competences of DILS, TKU

Competence	Description
A	Have the competences to know the library and information science principles and trends
B	Have the competences to develop, organize, archive and integrate various information resources.
C	Have the competences to realize information theories and apply information systems
D	Have the competences to communicate and coordinate information services
E	Have the competences to manage information services in various libraries and institutions
F	Have the competences to manage digital documents and file archives.
G	Have the competences to integrate traditional publication affairs and library works
H	Have the competences to integrate library works and digital content industry

**Fig. 2** System retrieves competence information from database then adds it to a course syllabus



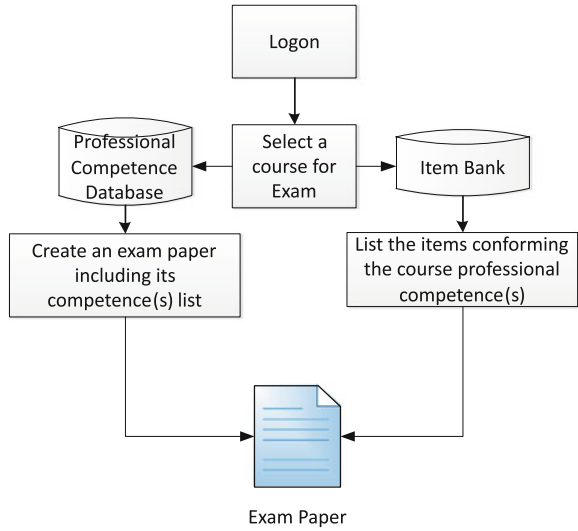
## 2.2 Build the Relation Between Competence and Assessment

In the previous section, the Course-Competence table has been built to describe the relation between course and competence. The competences are not only the specific learning objectives but also taken into account when we are evaluating student’s learning. In this section we focus on how to infuse the competence factors into traditional exam process.

### 2.2.1 Providing Exam Authoring Tool

The connection between exam and competence includes two parts; first is connecting a competence-oriented item bank, the other is an exam authoring interface which primarily helps teachers picking up questions from item bank. The authoring tool architecture is illustrated in Fig. 3.

**Fig. 3** Exam authoring tool architecture

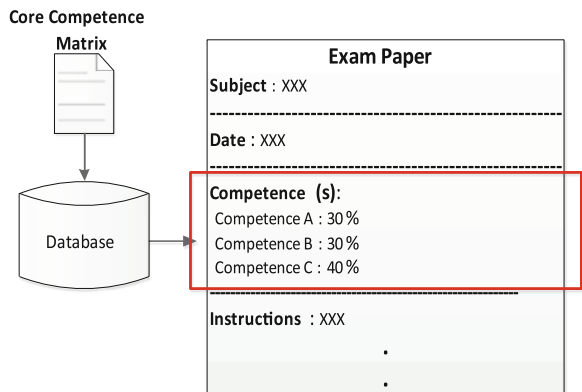


### 2.2.2 Provide Competence Ratio Information for Teacher Editing Exam Paper

As Fig. 4 shows, that’s similar to the competence information added in a syllabus. While teacher is starting from a blank exam paper, the professional competence ratio information of this course is automatically retrieved from database then added on an exam paper that reminds a teacher the competences should be evaluated in this course.

But the notable point here is that the competence ratio information is not used to instruct teachers to create an exam completely conforming the ratio but remind teacher the ratio could be a reference kept in mind while editing the exam. The teacher still has the flexibility to edit the exam paper content according to their teaching progress and professional knowledge.

**Fig. 4** A course professional competence ratio added on exam paper



				50.0	50.0	100.0
No.	Department	Student Number	Name	Mid	Final	
O11	DILS	498000016	Jeremy Lin	80	90.0	85.0
C				12.00	13.50	25.5
B				12.00	13.50	25.5
A				16.00	18.00	34.0

Competence C:30 %  
 Competence B:30 %  
 Competence A:40 %

Fig. 5 Grade sheet with competence calculation

### 2.3 Build the Relation Between Competence and Grade Sheet

A teacher generally proposes a few testing activities to evaluate student’s learning, such as paper works, presentations, exams etc. In order to express how students have learned the competences, a grade is separated to its corresponding competence grades.

Take Fig. 5 for example, a student who gets 80 in midterm which can be extended to display his/her each competence score. When a teacher fills in the score which will be automatically separated to the three competences grade according to ratio defined in the Course-Competence table. The grade of competence A in midterm is 16 because the midterm contains 50 % of the entire course grade and the ratio of the competence A is 40 %, so the result is multiplying 80 by 50 % and 40 %.

The grade sheet also allows a teacher to manually fill in or adjust the grade for each competence; the default competence score is calculated according to the ratio of competences of this course.

## 3 Conclusion and Future Work

We provide a prototype system of evaluating college student’s professional competences that they have learned in their department. Students may take a serial of training courses for acquiring the professional knowledge and skills in a particular department, but the results can’t be completely represented or measured by a grade. Through constructing Course-Competence table, the relation between a course and its corresponding competence(s) can be assured as well as the designing of the course syllabus. Through constructing the exam authoring tool, teachers can easily pick up required questions corresponding to competences set in this course and have the information of proper competence ratio distribution while creating the exam paper. Through constructing the relation between competence

and grade sheet that can not only present students learning grade but also a description of how they have learned the competences in this course. So the proposed evaluation system can used to understand a student professional competences learning situation. Moreover the evaluation feedbacks can not only help a teacher inspects the relation between their teaching and these professional competences but also provide valuable information for reviewing the entire department course structure. Our future works will include the visualizing the evaluation results and resolving the calculation problems caused by elective courses and data normalization.

**Acknowledgment** The authors would like to thank the anonymous reviewers for their insightful comments on an earlier version of this paper. The work described in the paper has also been supported by the National Science Council of Taiwan under Grant NSC 101-2221-E-032-064.

## References

1. Parry SB (1998) Just what is a competency? (And why should you care?). *Training* 35(6):58–64
2. Boyatzis RE (1982) *The competent manager: a mode for effective performance*. Wiley, New York
3. Spencer LM, Spencer SM (1993) *Competence at work: models for superior performance*. Wiley, New York
4. Clarke N (2010) The impact of a training programme designed to target the emotional intelligence abilities of project managers. *Int J Project Manag* 28(5):461–468
5. Parry SB (1996) The quest for competencies. *Training* 33(7):48–56
6. Yeomans WN (1989) Building competitiveness through HRD renewal. *Train Dev J* 43(10):77–82

# An Intelligent Recommender System for Real-Time Information Navigation

Victoria Hsu

**Abstract** People like to attend exhibition activities, but hard to enter into the information effectively. We build new system with wireless internet and mobile device to guide visitor into the core information initiatively and effectively. The mobile guide system could classify visitor base on exhibition information and personal information that provide more suitable for users. Our system combined with semantic web technology to connect items data which users' markup the type or property information in our system to created human portfolio. Our system is in compliance with human portfolio and metadata method to provide user information automatically and appropriately.

**Keywords** Mobile guide • Mobile device • Wireless internet • Semantic web • Human portfolio

## 1 Introduction

Many people visit the exhibitions or museums for their leisure time. Most of the museums and the exhibitions will provide the corresponding information to people for their visiting. Now, the technologies of mobile devices and wireless network could provide visitors their own style visiting via mobile devices which devices may be provided from the organizers or their own.

This research aims to propose a scheme of the guide system which first proposes a data storage format such that the exhibition organizers can store all the exhibition data simply. Second, the visitors can simply describe some personal information, which will be evaluated by the best appropriate recommendation

---

V. Hsu (✉)

Department of Computer Science and Information Engineering,  
Tamkang University, Taipei, Taiwan  
e-mail: saintvoice.1981@gmail.com

method (BAR) we proposed in this paper, and this guide system will provide the contents or information that is fit for the visitor by wireless network technique. And this information will be shown on the mobile device to the visitor.

Mobile devices are small computational equipment [1]. Users can get information from internet or telecommunication networks and execute some program by these devices, e.g. cell phone, PDA, notebook, iPad and so on. Metadata of the information is very important to mobile devices. Metadata is first defined in the conference Metadata workshop [2] and applied to data storage, data retrieving and so on. Dublin Core is a simple [3], efficient and popular metadata standard. It can fast organize the network resources, improve the precise of data search and retrieving, provide a metadata format to describe the network resources by many experts from different areas, and the network resources will be divided to 15 categories.

Categories for the description of works of art, CDWA, are a popular metadata definition to art exhibitions and museums categories [4]. It is proposed by Art Information Task Force, AITF, of J. Paul Getty Trust. CDWA provides a scheme to describe the content of works of art such that we can establish a database of the works of art by these describes. There are 27 main categories and 233 subcategories in CDWA.

After establishing the metadata, the ontology and the semantic web will be the critical techniques to develop our BAR method. Ontology was used to some specified and existed type or the well-described statements in philosophy [5, 6]. In computer science, ontology represents knowledge as a set of concepts within a domain, and the relationships between those concepts. Common components of ontologies include individuals, classes, attributes, relations, function terms, restrictions, rules, axioms and events. Semantic web is the concept proposed by Berners-Lee in W3C [7]. The main idea of semantic web is to let computers can “understand” the text files on the internet, that is, to know the semantics of the text files. By using the techniques of semantic web, the search engine can use a unique and precisely vocabulary and mark to the text files they searched without confusing.

## **2 Intelligent Recommender System**

### ***2.1 System Procedure***

For an exhibition or a museum, we first establish the database of the works of art by Dublin Core and CDWA. A visitor has to describe some of his/her personal data to the recommendation system before using this recommendation system. The recommendation system evaluates these personal data by the BAR method and finds some works of art will be recommended to the visitor. Then, the visitor will get the information about the recommended works of art via the mobile device he/she takes.

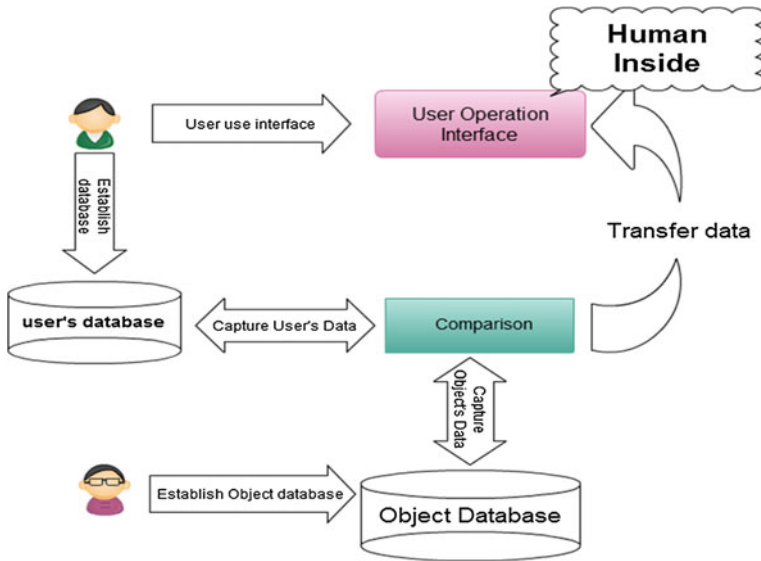


Fig. 1 Procedure of the recommendation system

With this recommendation system in the mobile device, the visitor can mark the works of art he/she likes during the visiting. The recommendation system will record the works of art to the database of the visitor’s profile. The more marks the visitor made, the more precisely our recommendation system will be, by the BAR evaluation. Figure 1 is the procedure of the recommendation system.

### 2.2 Database Establishment

The database of the recommendation system includes the following tables: art\_detail, art\_relation, type, human\_relation, human.

By CDWA, the table art\_detail stores the information about the works of art including: title, author, date, format, material, and description. The table human stores the visitors’ personal information including: name, sex, birthday, telephone number, e-mail address, address, and education degree. The table type stores the information of all kinds of types in this system. The table art\_relation stores the information about the types of the works of art in the exhibition. The table human\_relation stores the relationship of the visitor and the work of art. The visitor likes a work of art and mark it in the system that will be store in this table. Figure 2 represents the database structure of the recommendation system.

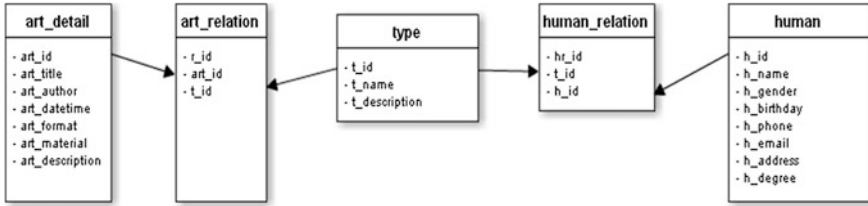


Fig. 2 Database structure of the recommendation system

### 2.3 Best Appropriate Recommendation Method

The BAR method includes two phases: (1) when the visitor establishes his/her personal data initially, BAR evaluates to determine the initial weights for recommendation; (2) when the visitor mark a work of art, BAR evaluates the attributes of the work of art and the personal data to update the weights to make more precisely commendation.

In phase 1, we first evaluate the initial weights by the following variables of personal data:

- UA user's age
- UD user's education degree
- UI user's interest
- WUA weight of user's age
- WUD weight of user's education degree
- WUI weight of user's interest
- $P = \{H_i | 1 \leq i \leq n\}$  a set of human\_portfilio database where  $H_i$  means the  $i$ th human data from human\_portfilio database.

The user's age is divided into 3 intervals: less than 19, from 20 to 40, greater than 40, and we assign weight 1, 2, 3, to each interval respectively. The user's education degree is classified by primary school, junior high school, senior high school, collage, educated school, and we assign weight from 1 to 5 for each degree respectively. The user's interest options include art, music, sport etc. and are divided into art related and not art related, and assign weight 2 and 1 respectively.

Input user's personal data values in database

Output user's weight values WP

Set new user,  $WP = 0$ ,  $WUA = 0$ ,  $WUD = 0$ ,

$WUI = 0$

//evaluate weight of user's age

If  $0 < UA \leq 19$

$WUA = WUA + 1$

else If  $20 \leq UA \leq 39$

$WUA = WUA + 2$



```

else
  WUA = WUA + 3
//evaluate weight of user's education degree
If UD = ``primary school``
  WUD = WUD + 1
else If UD = ``junior high school``
  WUD = WUD + 2
else If UD = ``senior high school``
  WUD = WUD + 3
else If UD = ``collage``
  WUD = WUD + 4
else
  WUD = WUD + 5
//evaluate weight of user's interest
If UI = art related options
WUI = WUI + 2
else
WUI = WUI + 1
//add all weights
WP = WUA + WUD +WUI
Store WP to user's profile

```

The initial weight is evaluated in phase 1 and the recommendation system recommends the works of art to the visitor depending on it before the visitor marks some works of art he/she likes. Then, when the visitor starts to visit the exhibition and marks some works of art he/she likes, the recommendation system receives the marks and BAR begins evaluating to update the weight of the visitor. That is the phase 2 of BAR method. Some variable definitions used in phase 2 BAR is shown as follows.

$A = \{A_i | 1 \leq i \leq n\}$  set of art database where  $A_i$  means the  $i$ th author data from author database.

$S = \{S_i | 1 \leq i \leq n\}$  a set of art database where  $S_i$  means the  $i$ th style data from style database.

$MAI = \{VA_i | VA_i \in A, 1 \leq i \leq k\}$  author's identification of the work of art which user marked

$MSI = \{VS_i | VS_i \in S, 1 \leq i \leq k\}$  style identification of the work of art which user marked

$RAI = \{RA_i | 1 \leq i \leq n\}$  Record the frequency of author's identification of the work of art which user marked

$RSI = \{RS_i | 1 \leq i \leq n\}$  Record the frequency of style identification of the work of art which user marked

In phase 2 of BAR method, when the visitor marks a work of art, the recommendation system retrieves the attributes author and style from database. After the

visitor finishing this visiting, the recommendation system evaluates that which author and which style the visitor marked most, then store this author and this style information in the visitor database. Therefore, when the visitor visits another exhibition next, the recommendation system can make good recommendation by these data.

## 2.4 Linking Semantic Web

The mobile device which is taken by the visitor during the visiting can receive the marks he/she made and provides the information and recommends of the works of art to the visitor. Moreover, by using the techniques of semantic web and wireless network, the visitor can get more information by the mobile device by connecting to other websites.

```

Input user marked author MAI,
user marked style MSI
Output the highest author and style value
For i = 1 to i = k do
Set RAI = 0, RSI = 0
//count the frequency of authors and styles that the visitor
marked
For i = 1 to i = k do
For j = 1 to j = n do
IF VAI = Aj do RAI += 1
IF VSI = Sj do RSI += 1
//find which author and which style that the visitor likes
most

```

$$RAI_{\max} = \arg \max \{RA_i | 1 \leq i \leq n\}$$

$$RSI_{\max} = \arg \max \{RS_i | 1 \leq i \leq n\}$$

Store RAI<sub>max</sub>, RSI<sub>max</sub>

## 3 Concluding Remarks

Using navigation system by mobile devices is very popular for exhibitions and museums recently. The wireless network technique, the semantic network technique and personal mobile devices are also well-developed. Visitors can get more information by different mobile devices than before. The appropriate recommendation system we designed can provide suitable information to visitors fast and convenient by their own mobile devices.

The appropriate recommendation system can be improved by considering the visitors' own experiences in the BAR method evaluation such that the system can recommend works of art to visitors more precisely. The appropriate recommendation system can be extended to be a community system. The visitors can share and exchange their experiences and make more commands to the exhibitions or museums on the system. These are all the future researches.

## References

1. Roschell J (2003) Unlocking the learning value of wireless mobile devices. *J Comput Assist Learn* 19:260–272
2. <http://zh.wikipedia.org/wiki/Metadata>
3. Hillman D (2005) Using dublin core. <http://dublincore.org/documents/usageguide/#whatis>
4. Agbabian MS, Masri SF, Nigbor RL, Ginell WS (1988) Seismic damage mitigation concepts for art objects in museums. In: *Proceeding of ninth world conference on earthquake engineering*
5. Gruber TR (1993) A translation approach to portable ontology specifications. *Knowledge Systems Laboratory, Palo Alto*, pp 199–220
6. Arvidsson F, Flycht-Eriksson A (2008) Ontologies I. <http://www.ida.liu.se/~janma/SemWeb/Slides/ontologies1.pdf>. Accessed 26 Nov 2008
7. Berners-lee T, Connolly D, Kagal L, Scharf Y, Hendler J (2008) *N3logic: a logical framework for the world wide web. theory and practice of logic programming*, vol 8. Cambridge University Press, New York, pp 249–269

**Part XII**  
**Advanced Mechanical and Industrial**  
**Engineering, and Control I**

# Modal Characteristics Analysis on Rotating Flexible Beam Considering the Effect from Rotation

Haibin Yin, Wei Xu, Jinli Xu and Fengyun Huang

**Abstract** This paper deals with modal frequencies of rotating flexible beam. To investigate effects on the modal frequencies from rotation, the mathematical models are derived by using three descriptions on deformation: the conventional approach, the quadratic approach, and a synthetical approach. The theoretical solutions of modal frequencies based on the three methods are used to compare and draw some summaries.

**Keywords** Flexible beam · Dynamic modeling · Modal characteristics · Rotation

## 1 Introduction

Because of light weight, small inertia, high operating speed, and low energy consumption, flexible beam has many promising applications such as helicopter propellers, flexible robot, etc. However, flexible beam also has its shortages one of which is vibration. There are a lot of studies on modeling and vibration control of flexible manipulators [1].

So far, there are three classifications on description of elastic deformation for flexible beam during modeling. The common and most widely used method is the conventional linear deformation method [2]. In the past decades, some papers discussed the quadratic deformation approach, such as Abe investigation on trajectory planning based on dynamic model, which adopted the quadratic method to describe the elastic deformation of flexible beam [3]. In 2005, a synthetical method had been proposed to derive the dynamic model of flexible robots by Lee [4]. In 2011, the synthetical method was extended to two-link flexible manipulator by Yin

---

H. Yin (✉) · W. Xu · J. Xu · F. Huang  
School of Mechanical and Electronic Engineering, Wuhan University of Technology,  
122 Luoshi Road, Wuhan, Hubei, People's Republic of China  
e-mail: chinaliuyin@whut.edu.cn

et al. where the synthetical method was deemed to be a better approach for flexible manipulator at high speed than conventional method [5].

Above three approaches are used to derived dynamic model of a rotating flexible beam and solve the modal frequencies in consideration of the effect from rotation in this report. In recent years, some researchers have investigated the effects on modal characteristic from rotation of flexible beam. Mei proposed differential transform method (DTM) to analyze the modal shape functions of a centrifugally stiffened Timoshenko beam, where author concluded as the modal shape functions were affected by rotation [6]. Gunda addressed a rational interpolation functions to analyze rotating beam and concluded that the shape functions were not only functions of positions but also functions of rotational speed [7]. Additionally, Kaya et al. studied the modal frequencies of rotating cantilever Bernoulli–Euler beam by using DTM, where the modal frequencies increase with angular velocity [8]. However, these published modal analyses on rotating flexible beam would focus on digital method rather than base on theoretical solution. This paper proposed the theoretical solutions of modal frequencies based on three models in consideration of the rotating effects.

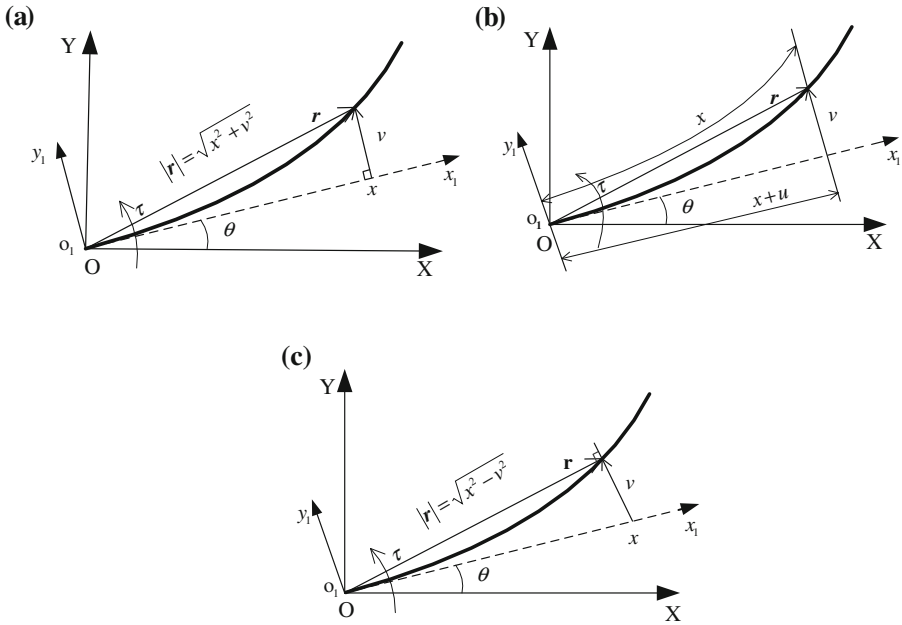
## 2 Mathematical Modeling

Figure 1 shows the schematic diagram of deformation for the flexible beam. In the conventional deformation, the displacement  $v$  of an arbitrary point at the flexible beam is vertical to initially undeformed beam shown as Fig. 1a. In the quadratic deformation, the length from the arbitrary point at the flexible beam to original point remains unchanged before and after deformation, shown as Fig. 1b. The synthetical deformation describes that the displacement  $v$  of an arbitrary point at flexible beam is vertical to the vector  $\mathbf{r}$  direction, which firstly proposed by Lee shown as Fig. 1c.

The coordinate O-XY represents the inertial reference frame with the original point O at the center of the hub, while the  $o_1-x_1y_1$  is the local coordinate system fixed to the root of the flexible beam rotating around the hub. The  $x$ -axis is oriented along the beam in undeformed configuration. The radius of the hub is assumed to be negligible and the hub rotates around the global Z-axis, with angle  $\theta$ . The three models have uniform beam with cross-section  $A$  and density  $\rho$ , the length  $l$  and flexural rigidity  $EI$ . The vector  $\mathbf{r}$  denotes the position vector of arbitrary point at flexible beam after deformation. The flexible beams are modeled as Bernoulli–Euler beam.

**Conventional Deformation.** In the Fig. 1a, the vector  $\mathbf{r}$  in the global frame is expressed as:

$$\mathbf{r} = \begin{bmatrix} x \cos \theta - v \sin \theta \\ x \sin \theta + v \cos \theta \end{bmatrix}. \quad (1)$$



**Fig. 1** Schematic diagram on description of deformation. **a** Conventional deformation. **b** Quadratic deformation. **c** Synthetical deformation

The terms related to the kinetic energy can be expressed as:

$$\dot{\mathbf{r}}^T \mathbf{r} = (x^2 + v^2)\dot{\theta}^2 + \dot{v}^2 + 2x\dot{v}\dot{\theta}, \tag{2}$$

where  $\cdot$  denotes derivative with respect to time  $t$ .

The dynamic equation associated with flexible deformation is derived through the Hamilton’s principle as follows:

$$\int_0^l [\rho A \ddot{v}(x, t) + EI v''''(x, t) - \rho A v(x, t) \dot{\theta}^2 + \rho A x \ddot{\theta}] dx = 0, \tag{3}$$

where  $'$  denotes derivative with respect to position  $x$ .

**Quadratic deformation.** In the Fig. 1b, the parameter  $u$  denotes the axial shortening. The vector  $\mathbf{r}$  in the global frame is expressed as:

$$\mathbf{r} = \begin{bmatrix} (x + u) \cos \theta - v \sin \theta \\ (x + u) \sin \theta + v \cos \theta \end{bmatrix}. \tag{4}$$

The terms related to the kinetic energy can be expressed as:

$$\dot{\mathbf{r}}^T \mathbf{r} = [(x + u)^2 + v^2]\dot{\theta}^2 + \dot{v}^2 + 2(x + u)\dot{v}\dot{\theta} + \dot{u}^2 - 2v\dot{u}\dot{\theta}. \tag{5}$$

The flexible dynamic equation is derived through the Hamilton’s principle as follows:

$$\int_0^l \{ \rho A \ddot{v}(x, t) + EI[1 + (v')^2]v'''(x, t) - \rho Av(x, t)\dot{\theta}^2 + \rho A(x + u)\ddot{\theta} + \rho A\dot{u}\dot{\theta} + EI[4v'v''v''' + (v'')^3] \} dx = 0. \tag{6}$$

**Synthetical deformation.** In the Fig. 1c, the position  $\mathbf{r}$  is expressed as:

$$\mathbf{r} = \begin{bmatrix} (x - \frac{v^2}{x}) \cos \theta - \frac{\sqrt{x^2 - v^2}}{x} \sin \theta \\ (x - \frac{v^2}{x}) \sin \theta + \frac{\sqrt{x^2 - v^2}}{x} \cos \theta \end{bmatrix}. \tag{7}$$

The terms related to the kinetic energy can be expressed as:

$$\dot{\mathbf{r}}^T \mathbf{r} = (x^2 - v^2)\dot{\theta}^2 + \frac{x^2}{x^2 - v^2}\dot{v}^2 + 2\sqrt{x^2 - v^2}\dot{v}\dot{\theta}. \tag{8}$$

Considering the displacement  $v$  is very small. So Eq. (8) can be simplified as:

$$\dot{\mathbf{r}}^T \mathbf{r} = (x^2 - v^2)\dot{\theta}^2 + \dot{v}^2 + 2xv\dot{\theta}. \tag{9}$$

The flexible dynamics is derived through the Hamilton’s principle described as:

$$\int_0^l [\rho A \ddot{v}(x, t) + EIv''''(x, t) + \rho Av(x, t)\dot{\theta}^2 + \rho Ax\ddot{\theta}] dx = 0. \tag{10}$$

### 3 Theoretical Solution on Modal Frequencies

According to Eqs. (3) and (10), the governing differential equations of vibration including rotational effect and non-conservative force are respectively described as follows:

$$\rho A \ddot{v}(x, t) + EIv''''(x, t) - \rho Av(x, t)\dot{\theta}^2 = f_1, \tag{11a}$$

$$\rho A \ddot{v}(x, t) + EIv''''(x, t) + \rho Av(x, t)\dot{\theta}^2 = f_1, \tag{11b}$$

where non-conservative force is represented by:

$$f_1 = -\rho Ax\ddot{\theta}. \tag{12}$$

Neglecting the high order infinitesimal such as  $(v')^2$ , the governing differential equation of vibration including rotational effect and non-conservative force. Based on Eq. (6) is expressed as:

$$\rho A \ddot{v}(x, t) + EIv''''(x, t) - \rho Av(x, t)\dot{\theta}^2 = f_2, \tag{13}$$



where non-conservative force is defined as:

$$f_2 = -\rho A[(x + u)\ddot{\theta} + \dot{u}\dot{\theta}]. \tag{14}$$

To obtain the homogeneous solutions of above three differential equations of vibration, the  $f_1$  and  $f_2$  are set equal to zero in Eqs. (11a) and (13) as following unified equation:

$$\rho A \ddot{v}(x, t) + EI v''''(x, t) \pm \rho A v(x, t) \dot{\theta}^2 = 0, \tag{15}$$

where the operator “+” of the third term in left side is based on the synthetical deformation, the operator “-” of the third term in left side is based on the conventional and quadratic deformations.

Assuming the harmonic vibration with centrifugally affected angular frequency  $\omega$  in flexible dynamics, the unified differential equations are written as:

$$EI v'''' - \rho A (\omega^2 \mp \dot{\theta}^2) v = 0. \tag{16}$$

Defining an equivalent angular frequency  $W$ , which is angular frequency of flexible beam in static structural dynamics and the equivalent differential equation of vibration is described as:

$$EI v'''' - \rho A W^2 v = 0, \tag{17}$$

where  $W$  is dependent on boundary conditions and mechanical parameters of flexible beams, the centrifugally affected angular frequency is represented by:

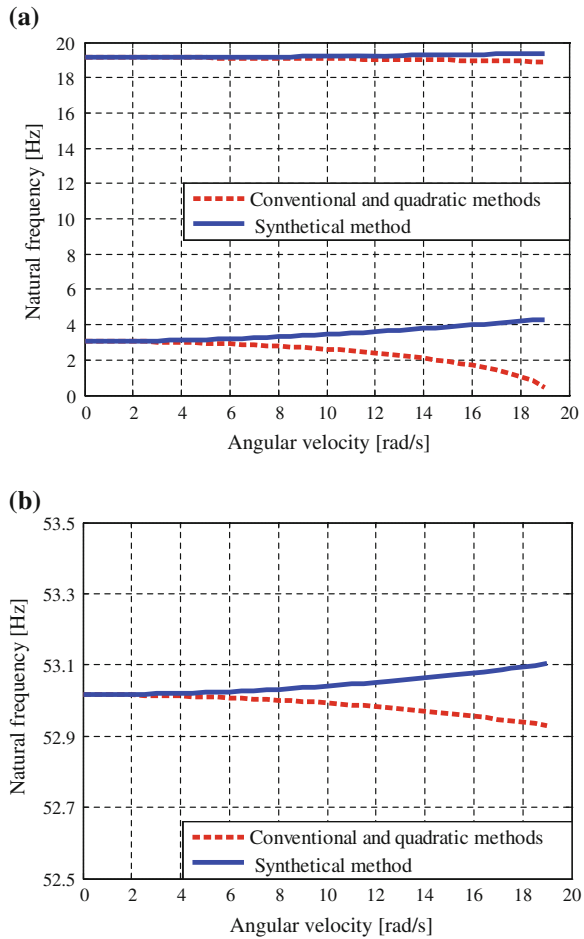
$$\omega_i = \sqrt{W_i^2 \pm \dot{\theta}^2}, \tag{18}$$

where  $i$  is the modal order number; the operator “+” is based on the synthetical method, the operator “-” is based on the conventional and quadratic methods.

### 4 Calculated Sample

This section gives calculated sample based on following parameters of a flexible arm: the mass density of the beam  $\rho = 7.70 \times 10^3 \text{ kg/m}^3$  and the sectional area of the beam  $A = 1.85 \times 10^{-5} \text{ m}^2$ . The elasticity modulus is  $E = 200 \text{ Gpa}$ , and the moment of inertia  $I = 5.40 \times 10^{-13} \text{ m}^4$ . The beam length is  $l = 0.40 \text{ m}$ . In static structural dynamics, the first three frequencies ( $i = 3$ ) of cantilever beam are  $W_1 = 19.22 \text{ rad/s}$ ,  $W_2 = 120.28 \text{ rad/s}$  and  $W_3 = 333.11 \text{ rad/s}$ , respectively. Considering the rotational effect from angular velocity, the centrifugally affected natural frequencies based on three models are shown in Fig. 2. Figure 2a, b denote the first two and third modal frequencies, respectively. The dash lines show the negative variation with angular velocity based on the conventional and quadratic

**Fig. 2** Centrifugally affected natural frequencies from the angular velocity. **a** The first and second mode. **b** The third mode



methods, the solid lines depict the positive relationship with angular velocity based on the synthetical method. The first frequencies are most significantly affected by rotational velocity.

## 5 Summary

This research has derived the differential equations of flexible vibration based three models, and the theoretical solution of centrifugally affected frequencies was proposed. The modal frequencies based on conventional and quadratic methods decrease with rotational velocity, but the modal frequencies based on synthetical method increase with rotational velocity. The theoretical solution was based on conditional simplification; consequently, further study is required in future work.

**Acknowledgments** The authors acknowledge the support of the National Natural Science Foundation of China (Grant No. 11202153).

## References

1. Dwivedy SK, Eberhard P (2006) Dynamic analysis of flexible manipulators, a literature review. *Mech Mach Theory* 41:749–777
2. Edelstein E, Roson A (1998) Nonlinear dynamics of a flexible multirod system. *J Dyn Syst Meas Control* 120:224–231
3. Abe A (2009) Trajectory planning for residual vibration suppression of a two-link rigid-flexible manipulator considering larger deformation. *Mech Mach Theory* 44:1627–1639
4. Lee HH (2005) New dynamic modeling of flexible-link robots. *J Dyn Syst Meas Control* 127:307–309
5. Yin H, Kobayashi Y, Hoshino Y (2011) Modeling and vibration analysis of flexible robotic arm under fast motion in consideration of nonlinearity. *J Syst Des Dyn* 5:219–230
6. Mei C (2006) Differential transformation approach for free vibration analysis of a centrifugally stiffened Timoshenko beam. *J Vib Acoust* 128:170–175
7. Gunda JB, Ganguli R (2008) New rational interpolation functions for finite element analysis of rotating beams. *Int J Mech Sci* 50:578–588
8. ÖZdemir Ö, Kaya MO (2006) Flapwise bending vibration analysis of a rotating tapered cantilever Bernoulli-Euler beam by differential transform method. *J Sound Vibr* 289: 413–420

# The Simulation Study on Harvested Power in Synchronized Switch Harvesting on Inductor

Jang Woo Park, Honggeun Kim, Chang-Sun Shin,  
Kyungryong Cho, Yong-Yun Cho and Kisuk Kim

**Abstract** Different piezoelectric harvester interface circuits are demonstrated and compared through SPICE simulation. The simulations of the effect of switch triggering offset and switch on time duration on SSHI's power are performed. The inductor's quality factors in synchronized switch harvesting on inductor interface have important effect on the harvested power. Parallel SSHI shows the optimal output voltage to harvest the maximum power varies according to the  $Q$  severely. It is concluded that switch triggering offset has more impact on the s-SSHI than p-SSHI and the switch on-time duration is more important in case of the p-SSHI. p-SSHI shows when the on-time duration becomes more than 1.3 times or less than 0.7 times of exact duration time, the harvested power gets negligible. s-SSHI reveals the characteristics that when less than 1.5 times exact on-time duration, the harvested power varies significantly with the on-time duration.

**Keywords** Piezoelectric · Energy harvesting · SSHI

---

J. W. Park (✉) · H. Kim · C.-S. Shin · K. Cho · Y.-Y. Cho  
Department Information and Communication Engineering,  
Suncheon National University, Suncheon 540-950, Republic of Korea  
e-mail: jwpark@sunchon.ac.kr

H. Kim  
e-mail: khg\_david@sunchon.ac.kr

C.-S. Shin  
e-mail: csshin@sunchon.ac.kr

K. Cho  
e-mail: jkl@sunchon.ac.kr

Y.-Y. Cho  
e-mail: yycho@sunchon.ac.kr

K. Kim  
Power Engineering Co., Ltd, Gwangyang 540-010, Republic of Korea  
e-mail: marohyun@hanmail.net

## 1 Introduction

The recent development of ultra-low power applications in ubiquitous sensing and computing demands low cost, long lifetime, small volume and light weight and especially eliminating the battery. Some ubiquitous applications can reduce the average power consumption to the level of tens to hundreds of microwatts, which results in energy harvested from environments to be used as an alternative power [1, 2]. Sustainable power generation can result from converting ambient energy into electrical energy. Mechanical energy conversion is one of the common sources for energy harvesting applications and exists almost everywhere. It is estimated that mechanical vibrations inherent in the environment can provide a power density of tens to hundreds of microwatt per  $\text{cm}^3$ , which is sufficient to sustain operations of a sensor node [3]. In Mechanical energy conversion, while electromagnetic and electrostatic generators have been developed [4, 5], piezoelectric generators [6, 7] are of major interest due to solid-state integration abilities.

While conventional power supplies and batteries typically have very low internal impedance, internal impedance of the piezoelectric generators is relatively high, which restricts the amount of output current driven by the piezoelectric source to the micro-amp range. The relatively low output voltage of the piezoelectric device is another challenge of this power source. This low output voltage poses a difficult on developing efficient rectifier circuits. The piezoelectric element subjected to a vibration generates the alternating voltage. However most of the electronic sensor nodes and circuits need the DC voltage. So called the standard interface has been widely used, where the interface consists of full-bridge rectifier and storage element. Some techniques to increase significantly the amount of energy by piezoelectric harvesters have been proposed, which are derived from called “synchronized switching damping (SSD) [8]”. The SSD technique is based on a non-linear processing on the voltage delivered by the piezoelectric element. This process increases the electrically converted energy resulting from the piezoelectric mechanical loading cycle. From SSD, parallel [8, 9] and serial [8] synchronized switching harvesting on inductor (SSHI) have been proposed. The techniques have increased the harvested power several times more than the standard technique. The nonlinear processing of SSHI consists in inductor and a switch in series and then needs the strict switching action of the switch.

In this paper, different harvesting interface circuits including standard interface, standard interface with a switch, parallel SSHI, and serial SSHI are simulated and compared with LT-SPICE® [10]. The voltage and current waveforms helps the comprehension of the interfaces. Then, effect of the switching time of SSHI on the harvested power is examined through simulation. Especially, it is also studied how the switch on time offset and on-time duration deviation have an effect on the power harvested from SSHIs.

## 2 Standard Interface Circuits for Piezoelectric Harvesters

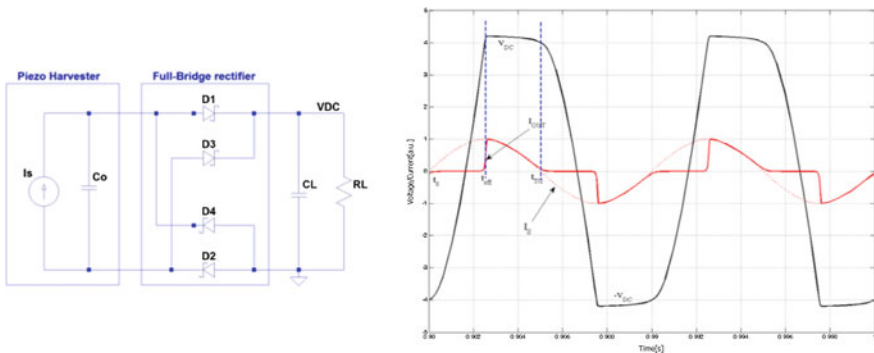
An input vibration applied on to a piezoelectric material causes mechanical strain to develop in the device which is converted to electrical charge. The piezoelectric laminate is mechanically forced to vibrate and thus works as a generator to transform the mechanical energy into electrical energy for micro-power generation. At or close to resonance, the piezoelectric element can be modeled in electrical domain. When excited by sinusoidal vibrations, the piezoelectric element can be modeled as a sinusoidal current source in parallel with a blocking capacitance  $C_0$  which represents the plate capacitance of the piezoelectric material. The amplitude  $I_0$  of current source depends on a displacement and frequency of the vibration.

$$i_s = I_0 \sin \omega_0 t \tag{1}$$

where  $\omega_0 = 2\pi f_0$  and  $f_0$  is the frequency with which the piezoelectric harvester is excited.

Because the power output by the piezoelectric harvester is not in a form which is directly usable by load circuits, the voltage and current output by the harvester needs to be conditioned and converted to a form usable by the load circuits. The power conditioning and converting circuits should also be able to extract the maximum power available out of the piezoelectric energy harvester.

Figure 1 shows the standard interface using full-bridge rectifier and the simulated voltage and current of the piezoelectric harvester. For the sake of this analysis, assume that the value of  $C_L$  is so large compared to  $C_0$  that the voltage at the output of the rectifier ( $V_{DC}$ ) is essentially constant. During the interval from  $t_0$  to  $t_{off}$ , the piezoelectric current source is charging its capacitor  $C_0$  to the  $V_{DC}$  and all diodes in the bridge rectifier are reverse-biased. And then, in the interval between  $t_{off}$  and  $t_{T/2}$ , the bridge rectifier will be on, the piezoelectric source provides the current to the load. We can know the reduction in the duration for charging the  $C_0$  can allows the power delivered to the load to be maximized.



**Fig. 1** Standard interface to extract power: circuit schematic and waveforms

To increase the power from the piezoelectric harvester, several interfaces have been proposed. Synchronized switching harvesting on inductor (SSHI) consists of a non-linear processing circuit. There are two types of SSHI, one is parallel-SSHI (p-SSHI) where the non-linear processing circuit is connected across the piezoelectric harvester and a full-bridge rectifier, and the other is series-SSHI (s-SSHI) where the non-linear processing circuit is connected between the piezoelectric harvester and a full-bridge rectifier in series. The non-linear processing circuit is composed of an inductor and a switch in series. This interface utilizes the synchronous charge extraction principle which consists in removing periodically the electric charge accumulated on the blocking capacitor  $C_0$  of the piezoelectric element, and to transfer the corresponding amount of electrical energy to the load or to the energy storage element. Figures 2 and 3 shows the two interface circuits and the voltage and current waveforms in them.

The electronic switch is briefly turned on when the current source of the piezoelectric elements crosses zero. This moment is when the mechanical displacement reaches maxima. At these triggering times, an oscillating electrical circuit  $L - C_0$  is established, where the electrical oscillation period is chosen much smaller than the mechanical vibration period  $T$ . The switch is turned off after a half

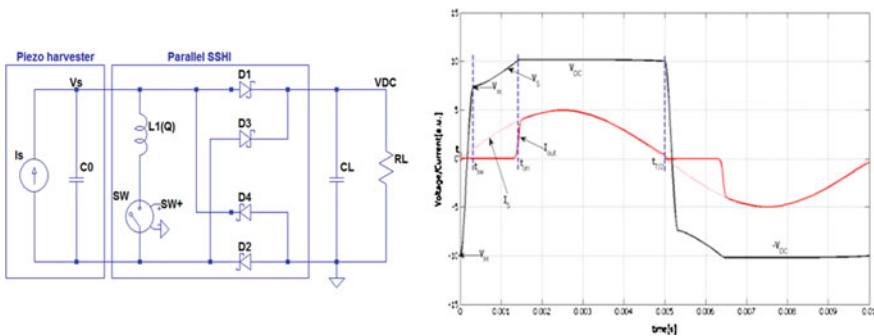


Fig. 2 Parallel synchronized switch harvesting on inductor interface

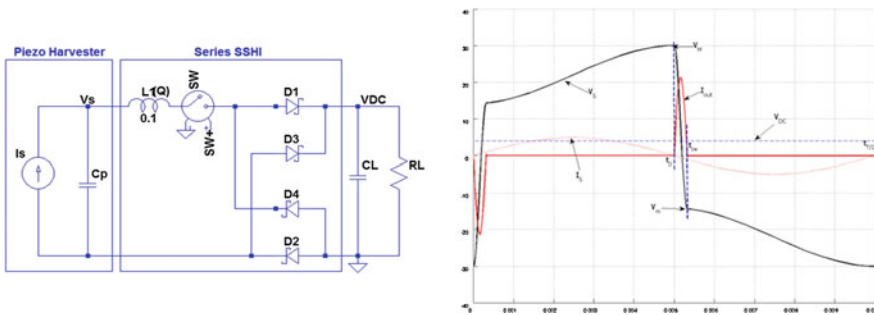


Fig. 3 Parallel synchronized switch harvesting on inductor interface

electrical oscillating period, resulting in a quasi-instantaneous inversion of the voltage  $V$ . The time interval,  $t_{sw}$  during the switch is on is expressed as:

$$t_{sw} = \pi\sqrt{LC_0} \tag{2}$$

The voltage relation between before the switch is on and after the switch is off depends on the quality factor  $Q$  of inductor.

$$V_m = -V_M e^{-\pi/2Q} \text{ for p-SSHI} \tag{3}$$

$$(V_m + V_{DC}) = -(V_M - V_{DC})e^{-\pi/2Q} \text{ for s-SSHI} \tag{4}$$

where  $V_m$  is the voltage after the switch is off,  $V_M$  is the voltage right before the switch is on and  $Q$  is the quality factor of inductor.

### 3 The Effect of Switching Time on Harvested Power of SSHI

As expected, in SSHI interfaces, the operation of the switch is very important. The switch has to turn on exactly when the displacement reaches maxima and then has to stay on only very short duration, a half period of  $L - C_0$  oscillation period. The switch triggering offset which is the switch on time deviation from the ideal on time and the on-time duration deviation have an important effect on the harvested power.

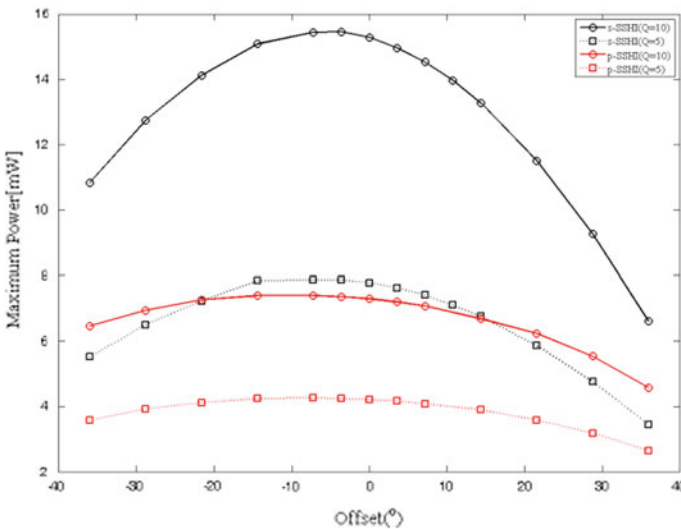


Fig. 4 The harvested power as a function of the switch triggering offset



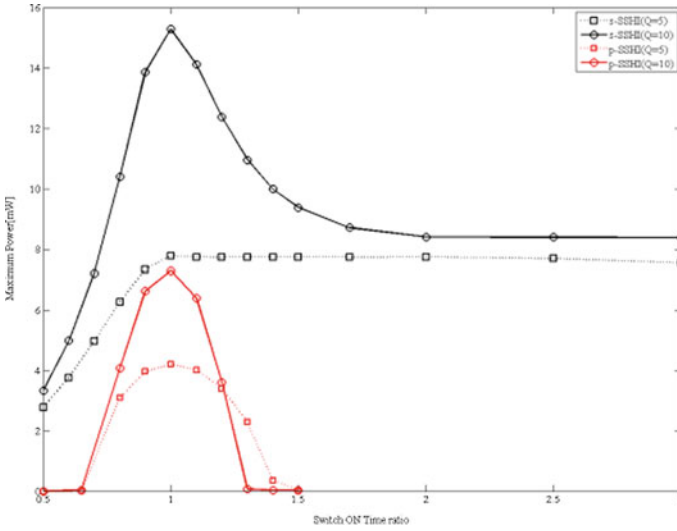


Fig. 5 Harvested power as a function of the switch on time duration deviation

In Fig 4, the effect of the switch triggering offset on harvested power is shown. This figure shows s-SSHI is more dependent on the triggering time than p-SSHI. Figure 5 shows the harvested power depending on the switching on-time duration. In this figure, switching on time duration ratio is the ratio of the real on-time duration and exact on-time duration as Eq. 2. The power harvested in p-SSHI interface is very dependent on the on-time duration, where when the on-time duration becomes over 1.3 times of exact on-time duration expressed by Eq. 2, the harvested power goes zero. However, s-SSHI shows when the deviation gets larger, the harvested power does not depend on the on-time duration. In s-SSHI, on-time duration gets larger, piezoelectric voltage gets smaller, however when on-time duration becomes over two times of exact duration, the voltage waveform does not change.

### 4 Conclusion

In this paper, we demonstrate and compare different piezoelectric harvester interface circuits using SPICE simulation. We consider the standard interface, the standard interface with a switch, parallel and serial synchronized switching harvesting on inductor. Switch triggering offset and switch on-time duration are very interest in calculating the power in SSHI. It is conformed that switch triggering offset has more impact on the s-SSHI than p-SSHI. It is recommended that the switch triggering offset is kept smaller than 10 % of piezoelectric element vibration period. The switch on-time duration is more important in case of the

p-SSHI. p-SSHI shows when the on-time duration becomes more than 1.3 times or less than 0.7 times of exact duration time, the harvested power gets zero. Because on-time duration is very small, careful consideration on the duration will be needed. s-SSHI reveals the characteristics that when less than 1.5 times exact on-time duration, the harvested power varies significantly with the on-time duration, however larger than 1.5 times exact on-time duration has scarcely influenced on the harvested power. Inductor's  $Q$  also has contributed on these characteristics.

**Acknowledgments** This work was supported by the Industrial Strategic technology development program, 10041766, Development of energy management technologies with small capacity based on marine resources funded by the Ministry of Knowledge Economy (MKE, Korea) and this work (Grants No. R00045044) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration.

## References

1. Mateu L, Moll F (2005) Review of energy harvesting techniques and applications for microelectronics. In: The proceedings of the SPIE microtechnologies for the new millenium, pp 359–373
2. Roundy S, Wright PK, Rabaey J (2003) A study of low level vibrations as a power source for wireless sensor nodes. *Comput Commun* 26(11): 1131–1144
3. Arms SW, Townsend CP, Churchill DL, Galbreath JH, Mundell SW (2005) Power management for energy harvesting wireless sensors. In: SPIE international symposium on smart structures and smart materials. San Diego, CA
4. Glynne-Jones P, Tudor MJ, Beeby SP, White NM (2004) An electromagnetic vibration-powered generator for intelligent sensor systems. *Sens Actuators A* 110:344–349
5. Mitcheson PD, Green TC, Yeatman EM, Holmes AS (2004) Architectures for vibration driven micropower generators. *IEEE J Microelectromech Syst* 13:429–440
6. Roundy S, Wright PK (2004) A piezoelectric vibration based generator for wireless electronics. *Proc Smart Mater Struct* 13:1131–1142
7. Ottman G, Hofmann H, Bhatt A, Lesieutre G (2002) Adaptive piezoelectric energy harvesting circuit for wireless remote power supply. *IEEE Trans Power Electron* 17(5):669–676
8. Lefeuve E, Badel A, Richard C, Petit L, Guyomar D (2006) A comparison between several vibration-powered piezoelectric generators for standalone systems. *Sens Actuators A* 126:405–416
9. Guyomar D, Badel A, Lefeuve E, Richard C (2005) Toward energy harvesting using active materials and conversion improvement by nonlinear processing. *IEEE Trans Ultrason Ferroelectr Freq Control* 52:584–595
10. LTspice IV, <http://www.linear.com/designtools/software/>

# An Approach for a Self-Growing Agricultural Knowledge Cloud in Smart Agriculture

TaeHyung Kim, Nam-Jin Bae, Chang-Sun Shin, Jang Woo Park,  
DongGook Park and Yong-Yun Cho

**Abstract** Typically, most of the agricultural works have to consider not only fixed data related with a cultivated crop, but also various environmental factors which are dynamically changed. Therefore, a farmer has to consider readjust the fixed data according to the environmental conditions in order to cultivate a crop in optimized growth environments. However, because the readjustment is delicate and complicated, it is difficult for user to by hand on a case by case. To solve the limitations, this paper introduces an approach for self-growing agricultural knowledge cloud in smart agriculture. The self-growing agricultural knowledge cloud can offer a user or a smart agricultural service system the optimized growth information customized for a specific crop with not only the knowledge and the experience of skillful agricultural experts, but also useful analysis data, and accumulated statistics. Therefore, by using the self-growing agricultural knowledge cloud, a user can easily cultivate any crop without a lot of the crop growth information and expert knowledge.

---

T. Kim (✉) · N.-J. Bae · C.-S. Shin · J. W. Park · D. Park · Y.-Y. Cho  
Department of Information and Communication Engineering, Suncheon National University,  
413 Jungangno, Suncheon, Jeonnam 540-472, Korea  
e-mail: taehyung@sunchon.ac.kr

N.-J. Bae  
e-mail: bakkepo@sunchon.ac.kr

C.-S. Shin  
e-mail: csshin@sunchon.ac.kr

J. W. Park  
e-mail: jwpark@sunchon.ac.kr

D. Park  
e-mail: dgpark6@sunchon.ac.kr

Y.-Y. Cho  
e-mail: yycho@sunchon.ac.kr

**Keywords** Ubiquitous agriculture · Agricultural cloud · Smart service · Knowledge-based

## 1 Introduction

Recently, to improve labor-intensive working environments, to secure economic feasibility, and to enhance productivity and quality in the fields of agriculture, many researchers are concentrating on the convergence of information technologies into the agricultural environments. Now, the agricultural environment is preparing new advancement. In agricultural environment, many studies about the IT-agriculture convergence have included optimum growth monitoring and growth environmental controlling system. The works are generally based on situation conditions from various sensors on ubiquitous sensor networks, which are deployed around the cultivation facilities or grounds. Existing studies for the optical crop growth information are based on a few fixed environmental data, which are temperature, humidity, illuminations, etc. However, crops are living organisms. Because environmental conditions are affected with each other in a very detailed and complex relationship, the crop growth status may be different in the same environmental conditions. Therefore, these studies about the optical crop growth information are underway constantly. That is, for smart agricultural environments, we need a method which can control dynamically and efficiently the environmental information about the crops.

A smart service in agricultural environments has to be able to use various environmental conditions automatically and organically as decision conditions on the service execution without human's interference. However, because it is so difficult to orchestrate the useful information from a lot of data related with specific crops and so hard to apply to other crops, getting stable and meaningful information to cultivate a specific crop is not simple. Recently, there have been a few of interesting researches to apply situation information into agricultural environments. However, most of the current agricultural systems using the situation information cannot make the best use of the great store of agricultural knowledge.

In this paper, we introduce an approach for self-growing agricultural knowledge cloud in smart agriculture. Knowledge DBs in the self-growing agricultural knowledge cloud have a lot of defined situation conditions, which are called contexts [1]. In smart agricultural environments, a context is one of very important elements to make an agricultural service autonomous and smart. The proposed self-growth agricultural knowledge supports a knowledge cloud architecture based on the various kinds of the knowledge DBs, which may be very far apart from each other and can offer a user or a smart agricultural service system the optimized growth information customized for a specific crop. The information contains not only the knowledge and the experience from users, researchers, and experienced farmers, but also useful analysis data and statistics accumulated into the knowledge DBs. Especially, the knowledge cloud can be plentiful more and more after

the lapse of time, and can be growing by itself. Therefore, users can access to the knowledge cloud system through the Internet, and obtain various and useful information for agricultural works.

## 2 Related Work

**Smart service in agriculture environments.** Generally, an agricultural environment that provides smart services is called ubiquitous agricultural environment. And there have been many studies about the convergence of state of the art information technology to build a new agricultural environment in many countries well aware about importance of agriculture. At its most basic, the monitoring services based on the wireless sensor network are designed. One of them [2] uses only the data from sensors in agricultural environment and another service [3] uses weather monitoring network and on-farm frost monitoring network and another one [4] uses Geographic Information Systems (GIS). And there are services for precision agriculture.

Smart services are well suited to apply the greenhouse. And these smart services in agricultural environment, designed by context-aware service models for the systematical definition and extensibility [1].

Many studies are underway to provide smart services in agricultural environment, but all have limitations in the domain that can be used. Farmers want to receive services that given the best choice regardless of the type of crops and the environmental characteristics. For this purpose, we need a service that can offer the knowledge and the experience of skillful agricultural experts rather than relying on system algorithms.

**Knowledge-based services.** A knowledge-based is a special kind of database for knowledge management. A knowledge base provides a means for information to be collected, organized, shared, searched and utilized. Therefore, the knowledge-based services help machines to have a decision-making like a human's decision-making. Recently, many knowledge-based services are provided in various fields. And it makes it possible to provide an appropriate service without human's interference. To do that, the knowledge-based services are designed by the context models based-on the ontology language [5]. For the smart home service, commonsense knowledge base is defined and designed by context modeling [6]. The field of e-learning investigates the integration of e-Learning systems and knowledge management technology to improve the capture, organization and delivery of both traditional training courses and large amounts of corporate knowledge [7].

To provide services without human intervention, it should be based on the knowledge-base. Therefore, the smart services in agricultural environment also offer the knowledge-based services for the ubiquitous agricultural environment.

**A cloud service in smart environments.** Cloud computing provides a new way to build applications on on-demand infrastructures instead of building applications

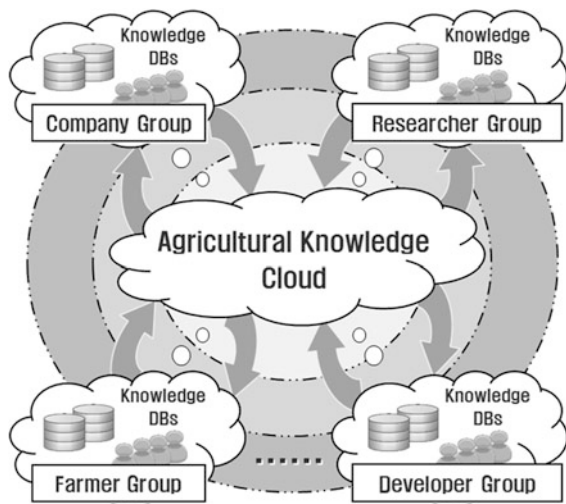
on fixed and rigid infrastructures. Cloud computing solutions only requires access to the Internet and a Web browser, and the heavy lifting of the software and hardware of the individual computer workstation is removed [8]. Therefore, cloud service is provided to users through the Internet anytime and anywhere. Cloud service is the ideal way to build ubiquitous environment. So Many cloud services are provided in various ubiquitous environments. Recently, A few researches have begun to cloud services in agricultural environment. A service model based-on an agricultural expert cloud is introduced [9]. This service is based an expert system, in which the knowledge and the experience of the various fields related in agriculture is accumulated. Because cloud service in agricultural environment is in the beginning step, more research is needed.

A self-growing agricultural knowledge cloud is designed as a unified architecture between knowledge-based service and cloud service. Through this, user can successfully and stably cultivate any crop by adopting the automatically suggested service to their agricultural systems anytime and anywhere.

### 3 A Self-Growing Agricultural Knowledge Cloud

Figure 1 shows a conceptual view for a self-growing agricultural knowledge cloud in smart agriculture. As shown in Fig. 1, the various groups can access the self-growing agricultural knowledge cloud just through the Internet anytime and anywhere. Therefore, a user in these groups can take the valuable agricultural information from users in other groups. Then, this information is applied to the user’s agricultural environment. In addition, users can upload their own information of experience and knowledge to Knowledge DBs in the self-growing

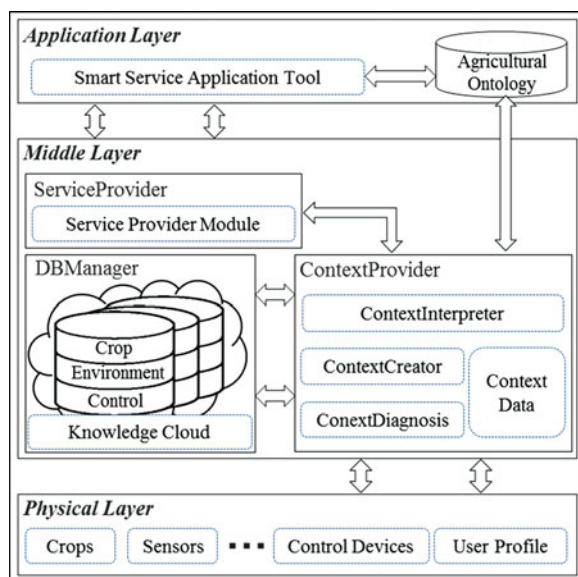
**Fig. 1** A conceptual view for self-growing agricultural knowledge cloud



agricultural knowledge cloud. As time passes, the agricultural knowledge cloud will grow more valuable itself because of the accumulating the various information of experience and knowledge from the various groups continually. So, we named it the self-growing agricultural knowledge cloud.

Figure 2 shows a brief conceptual architecture of the proposed approach for a self-growing agricultural knowledge cloud in Smart Agriculture. The architecture consists three layers, which are the Application Layer, the Middle Layer, and Physical Layer. The Physical Layer generates low-level data, which are crop conditions, real sensed data, status of control devices and user profile, etc. The Middle Layer consists of the ServiceProvider, the ContextProvider and the DBManager. The ContextProvider defines low-level data from the Physical Layer into Context data and uploads it to a self-growing agricultural knowledge cloud DBs. And it can also take agricultural information which is experience and knowledge from various other groups in knowledge DBs. A self-growing agricultural knowledge cloud can be accessed through the Internet anytime and anywhere. The ServiceProvider provides the best service through the ContextInterpreter based-on context data from user’s agricultural environment and agricultural knowledge cloud. The Application Layer supports that a developer composes a smart agricultural service application easily and quickly. To do this, the layer offers a GUI-based development toolkit and an agricultural ontology, which is used to make a smart service scenario by a developer and a context by a ContextProvider appeared in Fig. 2. The service scenario may focus on the growth environment of crops, a growth rate of crops and a consumption of energy, etc.

**Fig. 2** A brief conceptual architecture of the proposed approach



## 4 A Sample Smart Service Scenario

In this section, we show a sample smart service scenario for the self-growing agricultural knowledge cloud in smart agriculture. The following example scenario in the Table 1 is Mark's situation in the common agricultural environment.

To meet the hope of Mark, a smart service system using the self-growing agricultural knowledge cloud may need various contexts as situation conditions. Of course, the contexts have to include not only simple sensed contexts, but also high-level contexts composed from knowledge and experience which he has. Now, let's suppose that Mark's greenhouse needs a temperature control system in order to maintain an optimal growth environment of lettuce. Because Mark has no experience of lettuce cultivation, he can take Contexts of lettuce cultivation from the agricultural knowledge cloud. This is shown in Fig. 3a. Then, the temperature control service will be executed when the current temperature is sensed by a temperature sensor and recognized as a context by the smart service system. Again, let's suppose when the temperature is low enough to start the heating service but is a little bit high to do that in the seasonal and geographical aspect. Again, let's suppose when the temperature is low enough to start the heating service but not affects lettuce in the seasonal and geographical aspect. In this case, for the cultivation goal, Mark accesses to the agricultural knowledge cloud using the 'economical way' as a keyword. In this case, for the cultivation goal, Mark accesses to the agricultural knowledge cloud using the 'economical way' as a keyword. So, he can take a lot of information of environment, growth and control, etc. from experienced farmers who cultivate lettuce in the similar situation. Finally, he cannot have to operate the heating service using the best economical control context. This is shown in Fig. 3b.

**Table 1** The example scenario

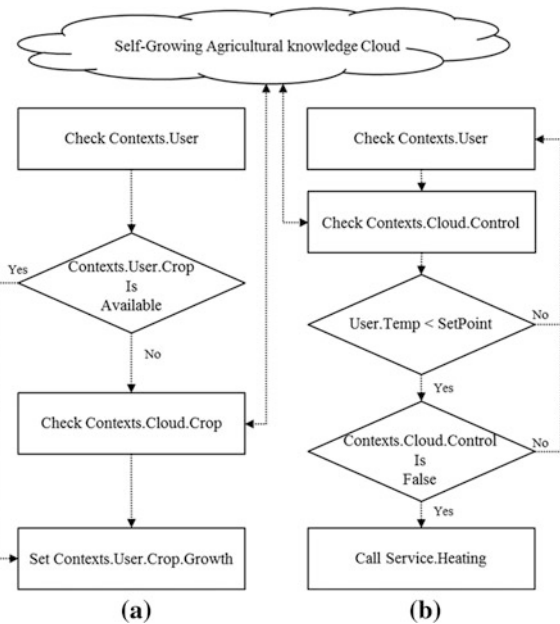
---

Mark is an experienced farmer who cultivates various crops, except the lettuce in his greenhouse.
Now he wants to cultivate lettuce
He should refer to the method of cultivation or the advice of an experienced farmer
He frequently comes into his greenhouse and checks the environmental information and the growth condition of lettuce
Then, he has to take proper action according to the conditions
If the temperature is below the normal values, he has to turn on the temperature control systems.
In this case, the operation on heating system affects the humidity in greenhouse. So, the windows open and shut system and the irrigation system should also be considered
He should invest a lot of money in order to maintain an optimal growth environment of lettuce.
Therefore, He wants a smart growth service system to do the tiresome works automatically without his intervention and conserve cost using the economical way

---



**Fig. 3** A process path for the crop searching and heating service



### 5 Conclusion and Future Works

Typically, most of the works in the agricultural environments are affected by various conditions, which tend to be unsystematically and dynamically changed. For the smart agricultural service in ubiquitous agricultural environment, the knowledge and experience about crops growth and agricultural skills have to be offered to users immediately through the Internet and networks. In this paper, we propose an approach for a self-growing agricultural knowledge cloud in smart agriculture. To support various and valuable agricultural information anytime and anywhere to service users, the introduced approach uses an agricultural knowledge cloud, which is various knowledge DBs stored by other expert group users through the Internet or networks. Therefore, the proposed approach can be used for users to provide appropriate various smart services for higher productivity and better quality in ubiquitous agricultural environments without service user’s interference.

As a future work in this paper, we will focus on the studies about real implementation of an efficient smart service framework or system based on the suggested architecture in the field of ubiquitous agricultural environments. And, in order to sufficiently testify to efficiency of the implemented smart service system using the suggested architecture, we will try to compose useful smart service applications and to adopt them into real agricultural environments with various sensors and computing devices.

**Acknowledgments** This work was supported by the Industrial Strategic technology development program, 10040125, Development of the Integrated Environment Control S/W Platform for

Constructing an Urbanized Vertical Farm funded by the Ministry of Knowledge Economy (MKE, Korea). And this research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) founded by the Ministry of Education, Science and Technology (2011-0014742).

## References

1. Cho Y, Moon J, Kim I, Choi J, Yoe H (2011) Towards a smart service based on a context-aware workflow model in u-agriculture. *IJWGS* 7:117–133
2. Zhou Y, Yang X, Guo X, Zhou M, Wang L (2007) A design of greenhouse monitoring and control system based on ZigBee wireless sensor network. In: international conference on wireless communications, networking and mobile computing, WiCom 2007, pp 2563–2567, Sept 2007
3. Pierce FJ, Elliott TV (2008) Regional and on-farm wireless sensor networks for agricultural systems in Eastern Washington. *Comput Electron Agric* 61:32–43
4. Ayday C, Safak S (2009) Application of wireless sensor networks with GIS on the soil moisture distribution mapping. In: Symposium GIS Ostrava 2009—seamless geoinformation technologies, Ostrava, Czech Republic
5. Kumar H, Park P (2010) Know-ont: a knowledge ontology for an enterprise in an industrial domain. *IJDTA* 3(1):23–32
6. Kawasar F, Shaikh M, Park J, Mitsuru I, Nakajima T (2008) Augmenting user interaction in a smart home applying commonsense knowledge. *IJSH* 2(4):17–31
7. Qwaider W (2011) Integrated of knowledge management and e-learning system. *IJHIT* 4(4):59–70
8. Caytiles R, Lee S, Park B (2012) Cloud computing: the next computing paradigm. *IJMUE* 7(2):297–302
9. Cho Y, Cho K, Shin C, Park J, Lee E (2012) An agricultural expert cloud for a smart farm. *FutureTech* 164:657–662

# Determination of Water-Miscible Fluids Properties

Zajac Jozef, Cuma Matus and Hatala Michal

**Abstract** The paper presents the main specifications for a monitoring of metalworking fluids and cleaners for mass industries. The implementation of fluid outsourcing in these industries. There is presented method of laboratory determination and diagnostic in industry of water-miscible fluids.

**Keywords** Fluid management · Waste management · Fluid contamination · Metalworking cutting fluids

## 1 Introduction

Concentration possesses a quality of leading parameter amongst other parameters of water-miscible fluids in their applying in manufacturing processes. For operational detection of product concentration two time-modest methods are used:

- *Refractometric method*
- *Titrimetric method*

Refractometric method is a simple way of operational concentration detection with manual or desk-top refractometer. The method is based on light beam refraction in optical refracting prism. Refractometer can be easily calibrated with a

---

Z. Jozef (✉) · C. Matus · H. Michal

Department of Manufacturing Technologies, Faculty of Manufacturing Technologies,  
Technical University in Košice, Bayerova 1, 080 01 Presov, Slovakia

e-mail: jozef.zajac@tuke.sk

C. Matus

e-mail: matus.cuma@tuke.sk

H. Michal

e-mail: michal.hatala@tuke.sk

given water sample and adjuster screw. For water-miscible fluids it is sufficient to use refractometers scaled from 0 to 15 %. Concentrations of these fluids exceeding 15 % are very rare. If the fluid is contaminated by higher percentage of tramp oil, it is very difficult to obtain relevant information about real value of product concentration in used mix. In such a case the titrimetric method is applied.

The titrimetric method is based on oil-contaminated water titration. For measuring of total alkalinity titrimetric set is often used, e.g. TA-kit. The set contains two solutions labeled L (HCL) and K and two syringes for measuring out of the product and for measuring out of alkaline concentrate. Resultant value must be modified.

Monitoring and maintaining fluid quality are crucial elements of a successful fluid management program. A fluid must be monitored to anticipate problems. Important aspects of fluid monitoring include system inspections and periodic measurements of fluid parameters such as concentration, biological growth, and pH. Changes from optimal fluid quality must be corrected with appropriate adjustments (such as fluid concentration adjustments, biocide addition, tramp oil and metal cuttings removal, and pH adjustment). It is important to know what changes may take place in your system and why they occur. This allows fluid management personnel to take the appropriate steps needed to bring fluid quality back on-line and prevent fluid quality problems from recurring.

The pH value determines acidity (0–7) or alkalinity (7–14) number. For metalworking water-miscible fluids pH values range from 8.8 to 9.5 in order to ensure corrosion and bacteriological protection. The pH value is a fast indicator of applied mix condition.

If pH value is under 8.8 it can be assumed that the fluid contains bacteria and mix becomes unstable, corrosion protection reduces, and an odor may occur... To increase pH value it is vital to apply additive (e.g. Additive 63). If value is above 9.5 the mix is contaminated by alkaline compounds (e.g. washing and cleaning media).

In Germany it is required before implementing water-miscible fluids to carry out test TRGS611 every week because nitrites react with secondary amines into nitroamines which are included in the list of carcinogenic agents in #2 category. Progressive manufacturers do not use nitrites (e.g. soda). The only contaminations are nitrites already present in applied water. Maximum value of nitrites quantity in processing fluids is 20 ppm. In the case the value is higher, it is necessary to change water source or modify the water.

The new metalworking cutting fluids generation brings also higher requirements for the system of fluid performance monitoring and control.

The simplest way for bacteria quantity identification is applying of the set for complex determination of mould, bacteria and fungi content with straps that are immersed for 5 s in measured medium and successively for 72 h are these so called "dipslides" placed in incubation apparatus with constant temperature. When fungi, mould or bacteria are present in greater amount, cultivation on dipslides would grow, consequently the cultures can be identified and quantified. Based on the result a proper additive can be applied (e.g. Kathon, Additive 63, etc.).

Oil used for machine parts lubrication and hydraulic oils are common contaminants in water-miscible fluids and they can drastically change performance of metalworking fluids especially their washing up and cleaning abilities.

If the operating system contains more tramp oils than 1.8 %, following problems from operating contaminated medium can be expected:

- *Cooling capability decrease*
- *Degradation of filtrating*
- *Reduced stability of the mix*
- *Smoke creation in cutting zone*
- *Growth of bacteria, mould and fungi*
- *Problems with determination of refractometric concentration*
- *Deterioration of workplace environment and increase of fluid skin-aggression*
- *Tramp oils that leak into the systems with water-miscible fluids can be divided to:*
  - *Free oils*
  - *Emulsifiable oils*

The main task of water-miscible processing fluids is to dissipate heat from the cutting zone. Effective removal of the heat increases tools life-cycle and product dimensions stability. Water has a better capability to take away heat than oil, however water in cutting zone causes corrosion of machined parts as well as the machine components which are in contact with the water during manufacturing process. Corrosion may occur whole year round but the higher probability is when there is a great temperature difference between day and night, or when both temperature and humidity are high. When temperature is rising, so is chemical activity including oxidation processes. According to long-time experience with applying water-miscible fluids the critical season is late April—early May and September (“Indian summer”). Prevention is accomplished by concentration increase of approximately one third. If concentration increase is not possible (e.g. foam creation, worsened wash up or workmen skin problems), it is necessary then to use additives for better anti-corrosion protection.

Corrosion protection tests are lengthy and could be realised only in laboratory conditions. Corrosion problems can be avoided by observing pH values, quality of input water, bacteria content and mix concentration.

Metalworking fluids having pH values of 9.0 and higher should be satisfactory for ensuring short-term anti-corrosion protection (three weeks) for iron-based alloys as well as for non-ferrous metals alloys (aluminium, copper, tin, etc.).

Water containing more than 80 ppm of chlorides and more than 250 ppm of sulphides is considered to be aggressive. These compounds rapidly decrease anti-corrosion abilities of fluids. It is necessary to check out the compounds content every week and implement proper measures when their concentration increases (adding de-mineralised or distilled water).

For observing manufacturing process mix quality it is necessary to measure hardness of the mix in the system. After certain operating time the fluid hardness

increase (e.g. when central system volume is 30 m<sup>3</sup>—for bearing rings grinding—the hardness increase of cooling semi-synthetic fluid with 31 % mineral oil is from 15 to 20°GH in a month). Water hardness above 20°GH causes decrease of anti-corrosion protection ability of water-miscible fluids.

If the mix contains higher amount of bacteria that “feed” on its components and constantly create organic acids, pH value decreases and anti-corrosion properties of fluids reduce. When pH values in individual tanks drop below value of 8.8, it is possible to apply alkaline system cleaner with concentration 0.1 %, when pH in central system drops, it is inevitable to apply biocides and makrobiocides for bacteria growth regulation.

## 2 Proposal for Method of Processing Fluids Monitoring

Based on theoretical analyses, customer requirements and operational state of processing medium survey, it is vital to observe on regular basis:

- *Appearance of processing medium*
- *pH value*
- *Total alkalinity of mix in %*
- *Anionic mix concentration*
- *Refractometric mix concentration*
- *Overall oil content in the mix*
- *Free oil content in the mix*
- *Bacteria, fungi, mould content*
- *Bacteriocide content*
- *Corrosion test applied at final user*
- *Corrosion test according DIN 51 360/2*
- *Nitrites content*
- *Nitrates content*
- *Overall hardness*
- *Borates*
- *Chlorides*
- *Contaminants content*

For monitoring of fluid condition it is required to determine limit content values of particular characteristics and contaminants in processing medium (see Table 1). When limits are overrun, it is essential to carry out measures concerning fluid condition, filtration effectiveness and according to overall oil content in the fluid it is possible to determine leak of tramp oil from machinery or other sources. For the observation of fluids in central systems it is appropriate to use tabular graphic fluid conditions outputs.



### 3 Summary

Verification of fluid management in bearings factories shows that proposed method of processing media care decreases overall costs by one third. In next year the expected gain comparing to last for these factories from fluid management project is 15 % in servicing two central systems with water-miscible products and one central system with non-water-miscible product.

### References

1. Gots I, Zajac J, Vojtko I (1995) Equipment for measuring the degree of wear to cutting tools. *Tech Mess* 1:8–11
2. Zajac J (2003) Accession at answer of synergy grinding and fluids in grinding. *Manuf Eng* 2–3:14–16
3. Zajac J (2007) Develop trends in research of processing fluids. *Manuf Eng* 2:36–38
4. Chao Wu et al (2009) Study on green design and biodegradability of B-containing water-based cutting fluid. *Key Eng Mater* 407–408:309–312
5. Harnicarova M, Zajac J, Stoic A (2010) Comparison of different material cutting technologies in terms of their impact on the cutting quality of structural steel. *Tehnicky Vjesnik* 3:371–376
6. Dima IC, Gabrara J, Modrák V, Piotr P, Popescu C (2010) Using the expert systems in the operational management of production. In: *MCBE '10*, p 307–312
7. Mohamed WANW et al (2011) Thermal and coolant flow computational analysis of cooling channels for an air-cooled PEM fuel cell. *Appl Mech Mater* 110–116:2746
8. Dima IC, Modrák V, Duică A, Goldbah IR (2011) The method of optimisation of the service of several tools, using the “mechanisation coefficient”. In: *IC-SSSE-DC '11*, pp 152–160
9. Wang Fei et al (2012) Management of drilling waste in an environment and economic acceptable manner. *Adv Mater Res* 518–523:3396–3402
10. Čuma M, Zajac J (2012) The impact analysis of cutting fluids aerosols on working environment and contamination of reservoirs. *Tehnicky Vjesnik* 2:443–446
11. Novak-Marcincin J, Novakova-Marcincinova L, Janak M (2012) Simulation of flexible manufacturing systems for logistics optimization. In: *LINDI-2012*, vol 631950, pp 37–40



# Influence of Technological Factors of Die Casting on Mechanical Properties of Castings from Silumin

Stefan Gaspar and Jan Pasko

**Abstract** Die casting represents the highest technological level of metal mould casting. This technology enables production of almost all final products without necessity of further processing. The important aspect of efficiency and production is a proper casting parameters setting. In the submitted paper following die casting parameters are analyzed: casting machine plunger speed and increase pressure. The studied parameters most significantly affect a qualitative castings dimension and they influence the most a gained porosity level  $f$  as well as basic mechanical properties represented by tensile strength  $R_m$  and ductility  $A_5$ .

**Keywords** Die casting · Technological factors · Casting · Mechanical properties

## 1 Introduction

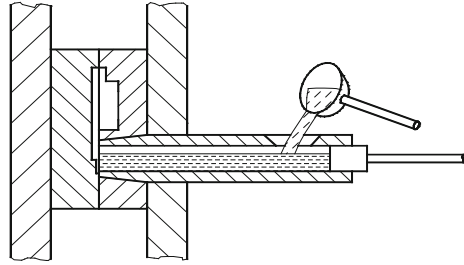
The die casting is a method of precise casting and it meets requirements for transformation from basic material into a ready product. Liquid metal is pressed at high speed ( $10\text{--}100\text{ m}\cdot\text{s}^{-1}$ ) into a cavity of a divided recursive mould (Fig. 1). The liquid metal high speed can be obtained by making intake groove more narrow at high pressure [13]. When liquid metal thoroughly fills the mould cavity, during a short time before it completely hardens, a increased pressure—increase pressure affects on it. It has to substitute gravitational setting of melting into empty mould cavities and thus suppress agglutinating and expansion of gas bubbles during cast

---

S. Gaspar (✉) · J. Pasko  
Faculty of Manufacturing Technologies, Technical University Kosice,  
Slovakia, Bayerova 1, 080 01 Presov, Slovakia  
e-mail: stefan.gaspar@tuke.sk

J. Pasko  
e-mail: jan.pasko@tuke.sk

**Fig. 1** Die casting machine with cold horizontal chamber



crystallization process (the necessary condition is a sufficient hydraulic connection of the mould with an intake system). The die casting products are of very precise dimensions, smooth surface, thin walls and very good mechanical properties [1, 2].

The die casting products quality is influenced by basic technological factors, i.e. pressing speed during casting cycle, melting specific pressure—increase pressure, filling chamber temperature and mould temperature [1–5].

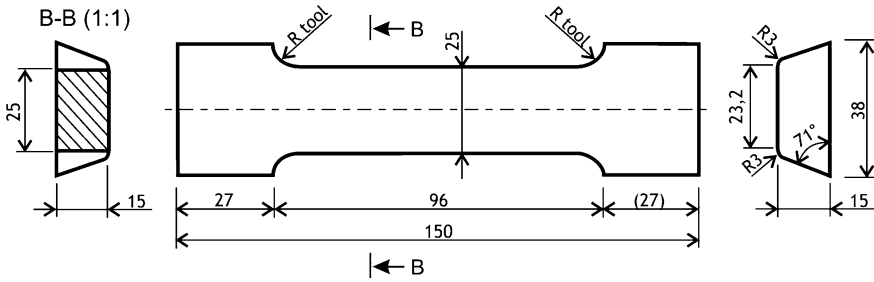
Increased attention is paid also to casting internal quality which is characterized by a kind and extend of foundry faults. The most important casting faults are exogenous cavities being generated by gases and air capturing which pass through the melt in turbulent stream. Pressing plunger speed in filling chamber of die casting machine is the most important factor of die casting. Metal stream speed in intake groove, which determines mould filling mode and thus affects both internal and external casts quality, depends on the plunger's velocity. If the main technological parameters are considered, then the die casting products quality affects also pressure acting on the cast being hardened in the casting cycle last phase—increase pressure [1, 6–8].

The aim of the paper is to analyze basic technological factors (plunger pressing speed and increase pressure of die casting and to set relations of their influence on selected mechanical properties: porosity  $f$ , tensile strength  $R_m$  and ductility  $A_5$ . Knowing these relations can be used in a new product design phase as well as in production process which can help increase manufacturing productivity and quality.

## 2 Material for Experiments, Methodology of Experiments and Used Devices

Influence of filling mode based on plunger pressing speed and increase pressure changes on selected mechanical properties study was performed on experimental samples (Fig. 2) appointed for static tensile test. For this experiment, a melting process had been carried. Its chemical composition is given in Table 1 and is in accordance with Standard EN 1706.

During the test, a wide extend of plunger speed in filling chamber within five levels was set: ( $v_1 = 1.9 \text{ m.s}^{-1}$ ;  $v_2 = 2.3 \text{ m.s}^{-1}$ ;  $v_3 = 2.6 \text{ m.s}^{-1}$ ;  $v_4 = 2.9 \text{ m.s}^{-1}$ ,



**Fig. 2** Scheme of experimental sample

**Table 1** Chemical composition of the experimental cast of the applied alloy

Chemical composition of the experimental cast of the applied alloy % of elements content											
Al	Si	Fe	Cu	Mn	Mg	Cr	Ni	Zn	Pb	Sn	Ti
85.27	12.02	0.71	1.19	0.21	0.13	0.02	0.02	0.35	0.02	0.03	0.03
<i>According to EN 1706</i>											
The	10.5–13.5	Max.	0.7–	Max.	Max.	Max.	Max.	Max.	Max.	Max.	Max.
rest		1.5	1.2	0.55	0.35	0.1	0.3	0.55	0.2	0.1	0.2

$v_5 = 3.2 \text{ m.s}^{-1}$ ). The increase pressure values extend was set also within five levels: ( $p_1 = 13 \text{ MPa}$ ;  $p_2 = 15 \text{ MPa}$ ;  $p_3 = 18 \text{ MPa}$ ;  $p_4 = 22 \text{ MPa}$ ;  $p_5 = 25 \text{ MPa}$ ). Beside factors which effect was monitored experimentally, there were factors whose effect had been reduced by keeping on the constant level. They were: melting temperature  $708 \text{ }^\circ\text{C}$  and mould temperature  $199 \text{ }^\circ\text{C}$ .

The static tensile test was performed on device ZDM 30/10 with jaw shift speed  $10 \text{ mm.min}^{-1}$ . The ductility test was then performed on these rods. Near fracture areas generated after static tensile test (Fig. 3), samples for a microscopic analysis were taken. The analysis was performed with a microscope Olympus GX51, magnifying 100 x. The samples were processed by PC with ImageJ program, which evaluated a proportional part of porosity from the studied cut (Fig. 4).

### 3 Reached Results and Their Analysis

Table 2 shows recorded reached values of porosity  $f$ , tensile strength  $R_m$  and ductility  $A_5$  in relation of studied casting parameters, i.e. plunger pressing speed and increase pressure. Figures 5, 6, 7, 8, 9 and 10 represent individual shapes of relations between selected mechanical properties and studied casting parameters.

Figures 11, 12, 13, 14, 15 and 16 represent metallographical cuts of selected macroscopic samples where dark spots show pores in the cast.

Measured porosity values  $f$ , breaking tensile strength  $R_m$  and ductility  $A_5$  presented in Table 2 show considerable variance in relation to plunger pressing speed and to increase pressure.

The result figures confirmed the fact that the lowest rate of porosity of A-category samples was found in samples with the lowest plunger speed in the filling chamber of pressing cast machine. Simultaneously, these samples represent the highest values of breaking tensile strength and ductility. Increasing of plunger speed means also increasing of porosity and decreasing of breaking tensile strength as well as ductility. Thus it can be said that the melt speed in the mould depends on plunger speed in the filling chamber. The wide range of filling plunger speed was related to test different filling modes. On the base of achieved results, studied macrostructures of the samples from point of view of porosity size and placement, the mould filling mode can be defined. At plunger speed in the filling chamber  $1.9 \text{ m.s}^{-1}$  the filling mode was laminar (Fig. 11), i.e. the front of flowing melt was continuous, homogenous and without whirlpools. When speed increases to  $2.3\text{--}2.6 \text{ m.s}^{-1}$  the filling mode changes from laminar to turbulent (Fig. 12). Turbulence of melt pulling air from the filling chamber out and its locking in cast walls. Further plunger speed increasing ( $2.9\text{--}3.2 \text{ m.s}^{-1}$ ) creates dispersive mixture of air and liquid metal, so called dispersive filling mode (Fig. 13).

When studying affect of increase pressure to porosity, breaking tensile strength and ductility, its positive influence on analyzed parameters can be unambiguously confirmed. At the lowest increase pressure values, the highest rate of porosity (Fig. 16) and the lowest values of breaking tensile strength and ductility were reported. So, increasing of increase pressure means reducing porosity (Figs. 14 and 15) and increasing breaking tensile strength and ductility. Increase pressure is recently one of the most discussed factors of press casting. On the one side, its high value reduces lifetime of the moulds and increases downtime periods, but on the other side it improves casting, reduces air volume locked in solid casts (porosity) and thus increases their quality (strength, tightness etc.).

## 4 Conclusion

The measured values of porosity  $f$ , breaking tensile strength  $R_m$  a ductility  $A_5$  of analyzed pressure cast samples produced on the pressure cast machine with cold horizontal chamber confirmed affect of plunger pressing speed and increase pressure on cast quality. It can be said that cavity filling speed defining mould filling mode is depended on plunger speed in the filling chamber. The casts made under higher pressure during hardening present better mechanical properties with the lowest porosity proportional rate. Performed studies of internal porosity dimensional evaluation and distribution enables to compare measured values with these parameters. Firstly, the studies showed that breaking tensile strength and ductility correlate with pores size which make the cast cross-section more weak.

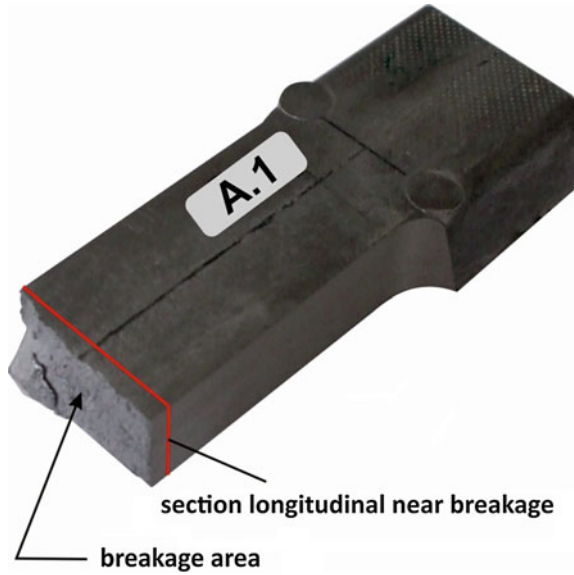


Fig. 3 Experimental sample

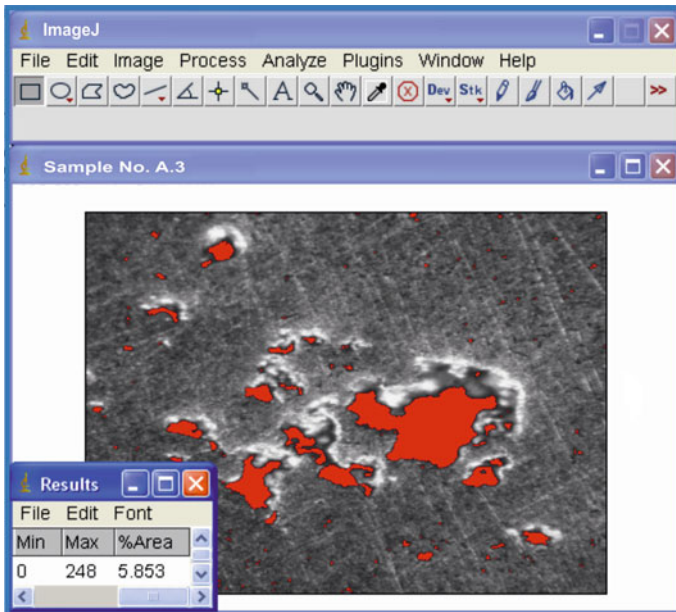
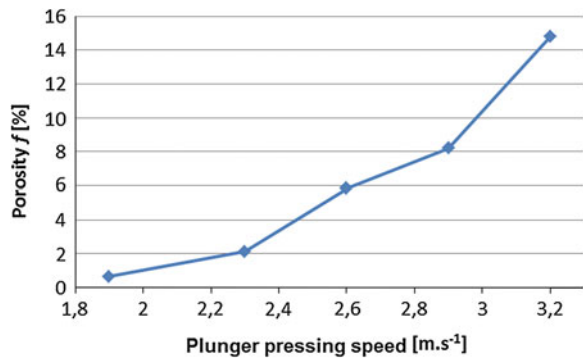


Fig. 4 Evaluation of the porosity of the sample no. A.3 s sample porosity evaluation by the Image J software porosity 5.85 %

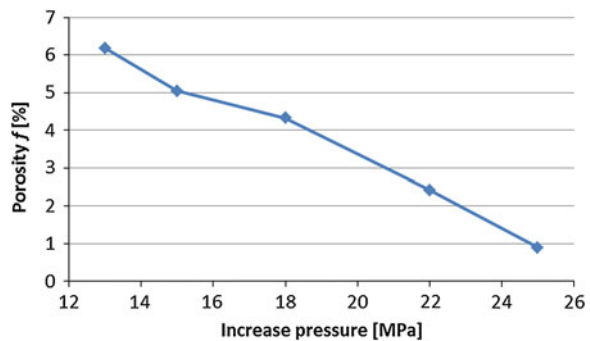
**Table 2** Reached results of studied mechanical properties

Studied technological casting parameters		Studied mechanical properties measured values			
		Porosity “f” [%]	Breaking tensile strength “ $R_m$ ” [MPa]	Ductility “ $A_5$ ” [%]	
Plunger pressing speed [m.s <sup>-1</sup> ]	No. A.1	1.9	0.65	169	2.8
	No. A.2	2.3	2.13	143	2.7
	No. A.3	2.6	5.85	124	2.5
	No. A.4	2.9	8.2	115	2.3
	No. A.5	3.2	14.8	103	2.2
Increase pressure [MPa]	No. B.1	13	6.18	121	2.3
	No. B.2	15	5.04	127	2.4
	No. B.3	18	4.32	133	2.5
	No. B.4	22	2.42	140	2.6
	No. B.5	25	0.89	153	2.7

**Fig. 5** Relation between plunger pressing speed and porosity “f”



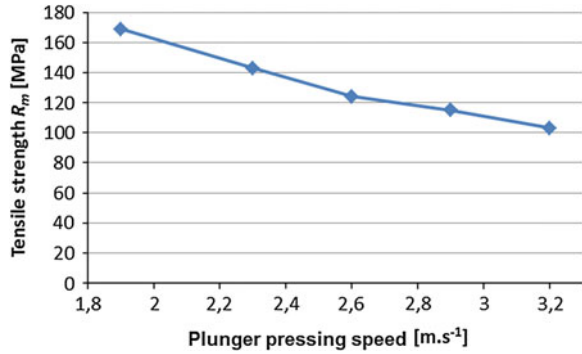
**Fig. 6** Relation between increase pressure and porosity “f”



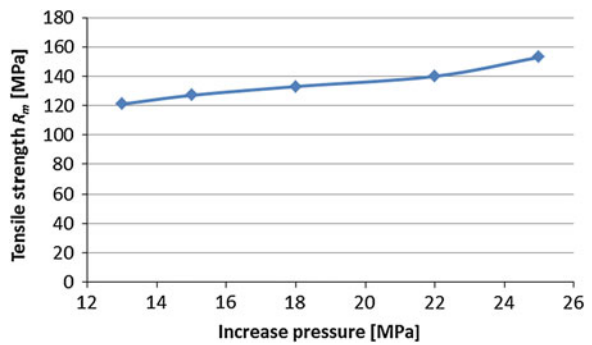
Fulfilling of pressing plunger defined speed and of increase pressure values is decisive for quality casts production.

This article has been prepared within the project VEGA No. 1/0593/12: Research of Technological Parameters Influence of Die Castings and Design

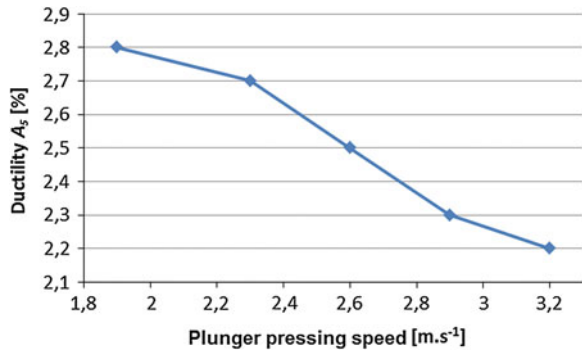
**Fig. 7** Relation between plunger pressing speed and breaking tensile strength “ $R_m$ ”



**Fig. 8** Relation between increase pressure and breaking tensile strength “ $R_m$ ”

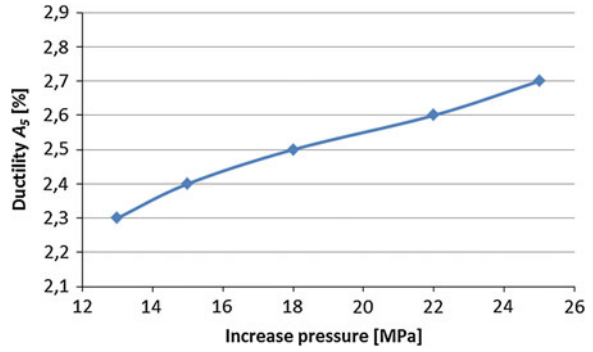


**Fig. 9** Relation between plunger pressing speed and ductility “ $A_5$ ”

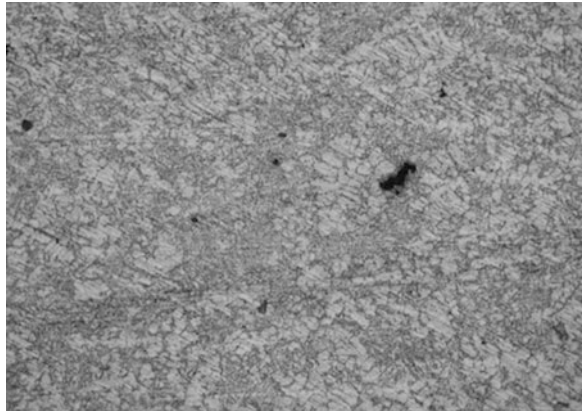


Modification of the Die System of the Casting Machine on Mechanical Properties of Die Castings of Lower Mass Category Made of Silumin.

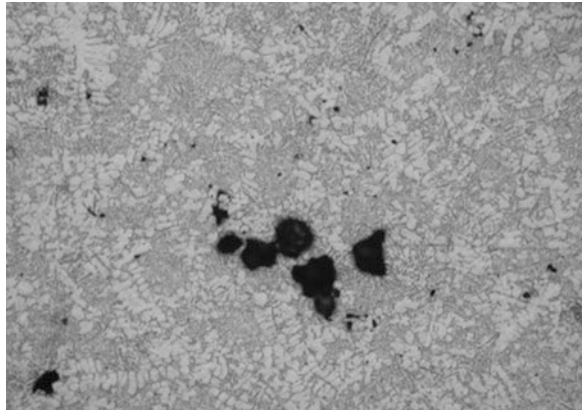
**Fig. 10** Relation between increase pressure and ductility “A<sub>5</sub>”



**Fig. 11** Sample no. A1 porosity 0.65 %

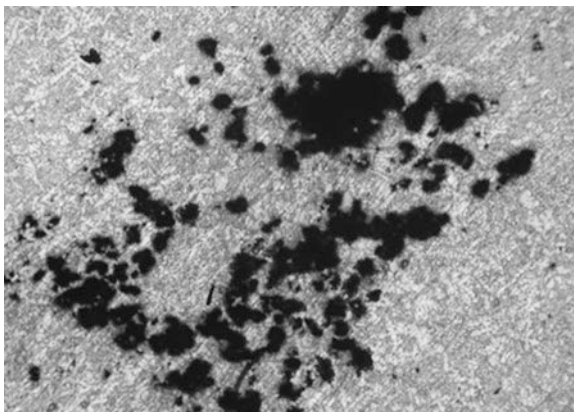


**Fig. 12** Sample no. A2 porosity 2.13 %





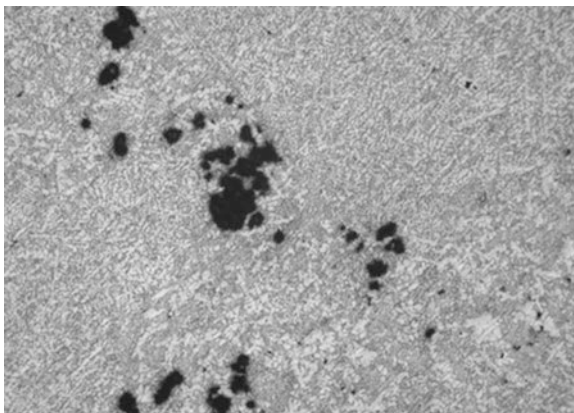
**Fig. 13** Sample no. A.5  
porosity 14.08 %



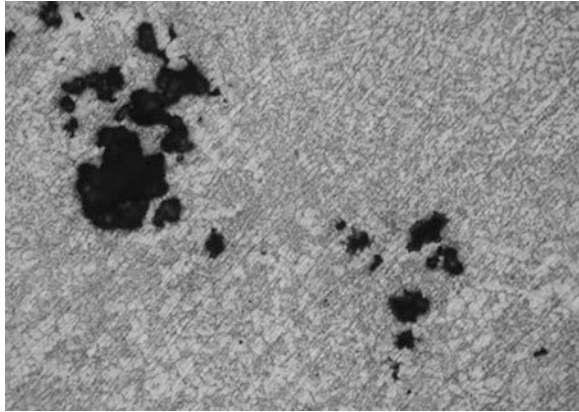
**Fig. 14** Sample no. B5  
porosity 0.89 %



**Fig. 15** Sample no. B4  
porosity 2.42 %



**Fig. 16** Sample no. B1  
porosity 6.18 %



## References

1. Gaspar S, Mascenik J, Pasko J (2012) The effect of degassing pressure casting molds on quality of pressure casting. *Adv Mater Res* 428:43–46
2. Semanco M, Fedak M, Rimar M, Ragan E (2012) Equation model to evaluate fluidity of aluminium alloys under pressure die-casting condition. *Adv Mater Res* 505:190–194
3. Stailcek L, Bytysev AI, Caplovic L, Batysov KA (2007) Metallographic verification of the model of the flow enforced during solidification under high external pressure. *Die Cast Eng* 51:56–60
4. Puskar M, Bigos P (2012) Method for accurate measurements of detonation in motorbike speed racing engine. *J Int Meas Confed* 45:529–534
5. Ragan E (2007) *Liatie kovov pod tlakom*, Presov
6. Vaskova I, Malik J, Futas P (2009) Tests of moulding mixture by using various clay binder granularity. *Arch Foundry Eng* 9:29–32
7. Kocisko M, Novak-Marcincin J, Baron P, Dobransky J (2012) Utilization of progressive simulation software for optimization of production systems in the area of small and medium companies. *Tech Vjesn* 19:983–986
8. Choi JC, Kwon TH, Park JH, Kim JH, Kim CH (2012) A study on development of a die design for die casting. *Int J Adv Manuf Technol* 20:1–8

# Active Ranging Sensors Based on Structured Light Image for Mobile Robot

Jin Shin and Soo-Yeong Yi

**Abstract** In this paper, we propose a ring array of active structured light image-based ranging sensors for a mobile robot. Since the ring array of ranging sensors can obtain omnidirectional distances to surrounding objects, it is useful for building a local distance map. By matching the local omnidirectional distance map with a given global object map, it is also possible to obtain the position and heading angle of a mobile robot in global coordinates. Experiments for omnidirectional distance measurement, matching, and localization were performed to verify the usefulness of the proposed ring array of active ranging sensors.

## 1 Introduction

Localization is the estimation of current position and heading angle, i.e., the posture of the mobile robot. Ranging sensors for the measurement of distances to surrounding objects are required for localization. There exist many kinds of ranging sensors, such as ultrasonic sensors, infrared laser sensors, laser scanners, stereo cameras, and active structured light image-based sensors [1]. Among these sensors, the structured light image-based sensor can effectively acquire distance

---

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0009113).

---

J. Shin · S.-Y. Yi (✉)

Seoul National University of Science and Technology, Seoul, Republic of Korea  
e-mail: suylee@seoultech.ac.kr

J. Shin

e-mail: gomlands@naver.com

information [2]. Bulky laser equipment and long image processing time have discouraged the use of the structured light image-based method in the past, but recent advancements in semiconductor laser equipment and faster processors have made this system more viable and economical.

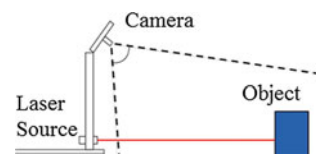
In this paper, a new ring array of ranging sensors is presented for the localization of a mobile robot. The ring array sensor has four active structured light image-based ranging sensors attached to the mobile robot. It is clear that omnidirectional distance acquisition is much more useful for a mobile robot than unidirectional distance acquisition. A ring array structure of ranging sensors that covers all directions had been used with ultrasonic ranging sensors [1]. In case of the ultrasonic ring array of sensors, however, it is impossible to activate multiple ultrasonic sensors at the same time because of signal crosstalk, which slows the distance measurement rate. In contrast, the structured light image-based sensors in a ring array can measure omnidirectional distances in one shot, without any mutual interference. To alleviate the computational burden in the main controller of a mobile robot, we developed structured light image-based ranging sensor modules by embedded image processor and arranged them in a circular pattern on the mobile robot. Each ranging sensor module transmits distance data to the main controller of the mobile robot after structured light image processing, and the main controller estimates the posture of the robot by matching the omnidirectional distance data with a given global object map.

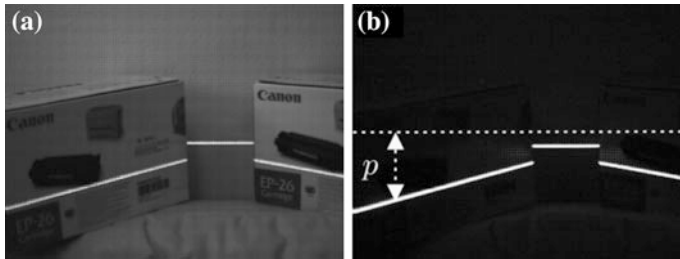
## 2 Structured Light Image-Based Distance Measurement with a Ring Array

As shown in Fig. 1, a structured light image-based ranging sensor consists of a camera and a structured light source. In order to obtain horizontal object distances under the assumption of robot motion on two-dimensional ground, a horizontal sheet of structured laser light is used in this study. By using the time difference of two images with modulated structured laser light, it is possible to extract the structured light pixel image, as shown in Fig. 2 [3].

Figure 2 shows the image illuminated by structured light and the extracted structured light pixel image obtained through image processing. From the center line of the image in Fig. 2b, structured light pixel distance  $p$  is detected in the vertical direction. From the pixel distance  $p$ , measurement angle  $\rho$  is given as follows:

**Fig. 1** Distance measurement based on structured light image





**Fig. 2** Extraction of structured light pixel image: **a** structured light image. **b** Extracted structured light pixel image (solid line)

$$\rho = \tan^{-1}\left(\frac{p}{\lambda}\right) \tag{1}$$

where  $\lambda$  represents the focal length of the camera. From the distance measurement model in Fig. 3a, the distance  $l$  to an object can be obtained as follows:

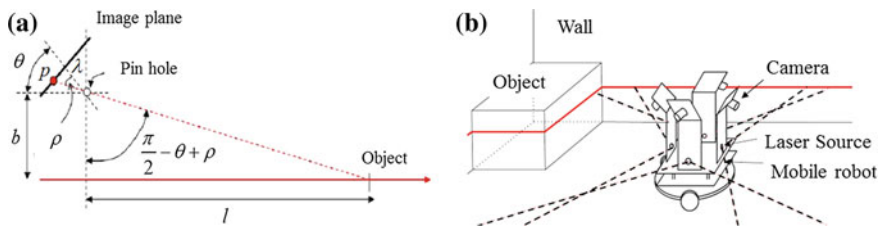
$$l = b \cdot \cot\left\{\theta - \tan^{-1}\left(\frac{p}{\lambda}\right)\right\} \tag{2}$$

In Fig. 3a,  $\lambda$  is the camera focal length,  $\theta$  represents the camera view angle, and  $b$  denotes the baseline.

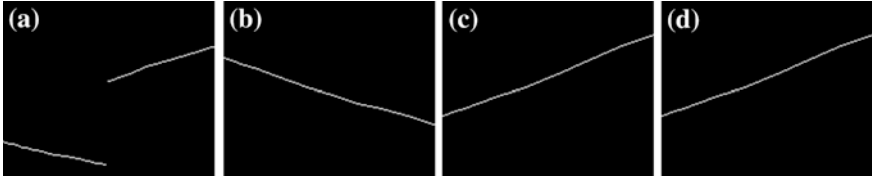
The well-known CMUcam3 [4] and a 660 nm wavelength infrared semiconductor laser are adopted to develop the structured light image-based ranging sensor module. The embedded processor in CMUcam3 performs all of the image processing and only transmits the distance data to the main controller of the robot. Figure 3b shows the ring array of the ranging sensor modules attached to the mobile robot, which can measure omnidirectional object distances.

### 3 Posture Estimation from Omnidirectional Distance

From the structured light pixel images of the ranging sensor array shown in Fig. 4 and Eq. (2), it is possible to acquire a local omnidirectional distance map, as

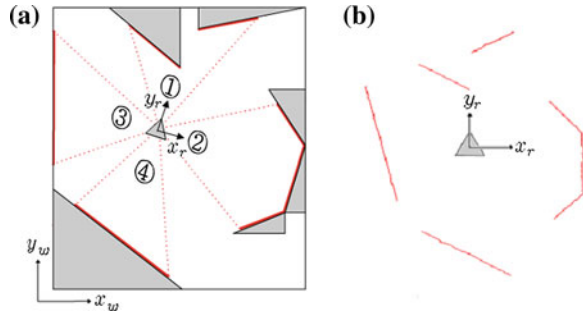


**Fig. 3** Distance measurement model and omnidirectional ranging through ring array. **a** Distance measurement model. **b** Ring array of ranging sensor module



**Fig. 4** Structured light pixel images from ranging sensor array. **a** From camera 1. **b** From camera 2. **c** From camera 3. **d** From camera 4

**Fig. 5** Omnidirectional distance data. **a** Mobile robot environment. **b** Measured local distance map



shown in Fig. 6. In Fig. 5a, the circled numbers denote each camera in the ring array corresponding to each image in Fig. 4. Figure 5b shows the local distance map in the moving coordinates of the mobile robot. The local distance map consists of a set of measured points  $(x_m, y_m)$  in the moving coordinates. When the estimated posture of the robot is  $(\hat{x}_r, \hat{y}_r, \hat{\theta}_r)$  in world coordinates, the measured local distance data can be transformed into world coordinates as follows:

$$\begin{bmatrix} x_w \\ y_w \end{bmatrix} = R(\hat{\theta}_r) \begin{bmatrix} x_m \\ y_m \end{bmatrix} + T(\hat{x}_r, \hat{y}_r) \tag{3}$$

where  $R(\theta)$  and  $T(x, y)$  represent the rotation and translation, respectively, as follows:

$$R(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}, T(x, y) = \begin{bmatrix} x \\ y \end{bmatrix} \tag{4}$$

Posture estimation should be updated by matching the real-time omnidirectional distance map with a given global object map. There have been many studies of the matching problem. In [5] and [6], a least-squared-error-based matching algorithm was suggested to associate real-time distance data with a given global map. Since the matching algorithm considers every measured points individually, it requires a significant number of computations. In order to improve computational efficiency, a matching algorithm is developed in this paper by modifying the algorithms in [5] and [6]: line segments are obtained from the measured local

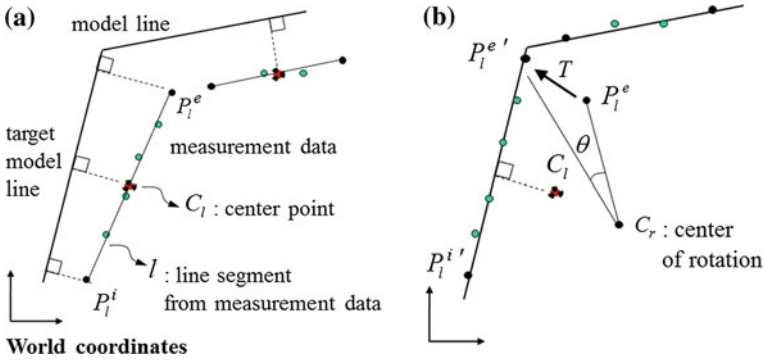


Fig. 6 Matching algorithm. a Before matching. b After matching

distance map first and only two end points of a line segment are matched with the global map, rather than all of the measured points.

The matching algorithm is described in Fig. 6 where  $P_l^i$  and  $P_l^e$  denote the two end points of a line segment  $l$  from the measured local distance map and  $P_l^c = (P_l^i + P_l^e)/2$  is the center point of the segment. Here, we assumed that those points are described in world coordinates by the transformation (3). Among the all model line segments from the given global map, the target model line segments nearest to the center point of a segment can be found. The target line segment satisfies the following (5).

$$P \cdot \mathbf{u}_l = r_l. \tag{5}$$

where  $P$  is a point on the target line segment,  $\mathbf{u}_l$  is the unit normal vector, and  $r_l$  is a real number.

Rotation by  $\Delta\theta$  about the present estimated position,  $C_r = [\hat{x}_r \ \hat{y}_r]^t$ , of the robot and translation by  $(\Delta x, \Delta y)$  makes two end points of segment  $l$  as follows:

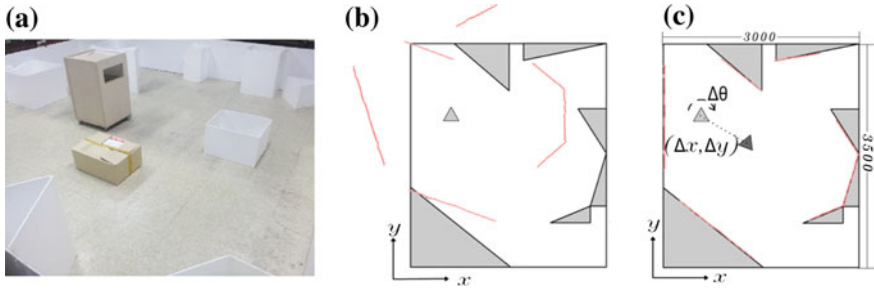
$$P_l' = R(\Delta\theta)(P_l - C_r) + C_r + T(\Delta x, \Delta y). \tag{6}$$

where  $P_l$  and  $P_l'$  represent two end points before and after the transformation in world coordinates. Then, the total matching error is defined by the sum of the squared distance between the transformed points  $P_l'$  of the all line segments  $l$  and the target line (5) as follows:

$$S = \sum_l (P_l' \cdot \mathbf{u}_l - r_l)^2 + (P_l^e \cdot \mathbf{u}_l - r_l)^2 \tag{7}$$

By the well-known gradient method to minimize the total matching error (7), it is possible to get the transformation parameters,  $(\Delta x, \Delta y, \Delta\theta)$ , which is used to update the estimation of robot's posture as follows:

$$(\hat{x}_r, \hat{y}_r, \hat{\theta}_r) \leftarrow (\hat{x}_r + \Delta x, \hat{y}_r + \Delta y, \hat{\theta}_r + \Delta\theta) \tag{8}$$



**Fig. 7** Data matching experiment. **a** Experimental environment. **b** Before matching. **c** After matching

## 4 Experimental Results

We performed experiments to verify the effectiveness of the proposed ring array ranging sensor and the matching algorithm. As shown in Fig. 7a, some polygonal objects are placed in the mobile robot's environment. The omnidirectional distance data measured at an unknown robot posture are depicted in Fig. 7b, c and shows the resultant robot posture after matching and updating by the algorithm described in (5) through (8). In Fig. 7b, the transformation parameters to update the robot's posture are  $(\Delta x, \Delta y, \Delta \theta) = (630, 320, 16.54^\circ)$ , as obtained from the matching algorithm.

## 5 Conclusion

The ring array of the structured light image-based ranging sensors proposed in this paper is able to obtain omnidirectional distances in one shot for fast localization of a mobile robot. Compact cameras with embedded processors used for the ring array of the ranging sensor in this paper send only final distance data to the main controller of the robot, thereby lowering its computational burden. Matching between the omnidirectional distance data from the proposed ranging sensors and the given global object map is required for localization of the mobile robot. A least-squared error-based algorithm was developed in this paper to associate line segments extracted from the omnidirectional distance data with the polygonal model of the global object map. Since the matching algorithm in this paper uses only two end points of a measured line segment to associate with the reference line segment of the polygonal world model, efficiency of computation is greatly improved. The proposed ring array of active structured light image-based ranging sensors and the matching algorithm in this paper were verified through experiments on local omnidirectional distance data acquisition, localization.



## References

1. Cameron S, Probert P (1994) *Advanced guided vehicles-aspects of the Oxford AGV Project*, World Scientific, London
2. Noh D, Kim G, Lee B (2005) A study on the relative localization algorithm for mobile robots using a structured light technique. *J Inst Control, Robot Syst* 11(8):678–687
3. Jain R, Kasturi R, Schunck BG (1995) *Machine vision*. McGraw-Hill, New York
4. <http://www.cmucam.org>
5. Cox I (1991) Blanche-an experiment in guidance and navigation of an autonomous robot vehicle. *IEEE Trans Robot Autom* 7(2):193–204
6. Cox I, Kruskal J (1988) On the congruence of noisy images to line segment models. In: *Proceedings of international conference on computer vision*, pp 252–258

# Improved Composite Order Bilinear Pairing on Graphics Hardware

Hao Xiong, Xiaoqi Yu, Yi-Jun He and Siu Ming Yiu

**Abstract** Composite-order bilinear pairing has been applied in many cryptographic constructions, such as identity based encryption, attribute based encryption, and leakage resilient cryptography. However, the computation of such pairing is relatively slow since the composite order should be at least 1024 bits. Thus the elliptic curve group order  $n$  and the base field are large. The efficiency of these pairings becomes the bottleneck of the schemes. Existing solutions, such as converting composite-order pairings to prime-order ones or computing many pairings in parallel cannot solve the problem. The former is only valid for certain constructions and the latter is only helpful when many pairings are needed. In this paper, we make use of the huge number of threads available on Graphics Processing Units (GPUs) to speed up composite-order computation, both between pairings and within a single pairing. The experimental result shows that our method can speed up pairing computation. Our method can also be ported to other platforms such as cloud systems.

**Keywords** Composite-order bilinear pairing · GPU · Cryptography construction

## 1 Introduction

A bilinear pairing is in the form of  $e : G \times G \rightarrow G_T$ , a bilinear pairing is said to be over a composite-order group if the order of  $G$  (and  $G_T$ ) is composite. Composite-order bilinear pairing has been used in many cryptographic constructions, such as identity based encryption, attribute based encryption, functional encryption and leakage resilient cryptography. However, computing a pairing over a composite-

---

H. Xiong · X. Yu · S. M. Yiu

Department of Computer Science, The University of Hong Kong, Hong Kong, Hong Kong

Y.-J. He (✉)

R&D Centre for Logistics and Supply Chain Management, Hong Kong, Hong Kong

order group is much more expensive compared to its prime-order counterpart and other building blocks. For example, according to [4], to achieve the same 80 bits AES security level, the composite order should be at least 1024 bits, while only 160 bits are needed in its prime-order counterpart. Since more bits are needed in composite order pairing, the underlying finite field, elliptic curve operations and the pairing evaluating algorithm itself become much slower. Freeman [4] shows that the composite-order pairing is 50X slower than prime-order pairing. Due to the above reasons, composite-order pairing computation easily becomes the bottleneck of a cryptographic scheme, especially where large amounts of such pairings need to be computed (e.g., [6]). Several solutions have been proposed to solve this problem recently. Freeman [4] proposed a method to convert a scheme based on composite-order pairing to a prime-order pairing based scheme. The shortcoming of this method is that it is only valid for certain cryptographic constructions. In fact, [8] points out that some schemes inherently require composite-order groups. Another solution is to compute composite-order pairing in parallel, either in the way where many pairings are computed in parallel or compute one single in parallel. In the early era, the hardware needed in parallel computing is expensive to most organizations and make this solution unreasonable. However, due to the rapid development of technology recently, we can apply GPU (Graphics Processing Unit) which is a relatively cheaper hardware in parallel computing. The huge number of threads available on GPU can be leveraged to speed up the composite-order bilinear pairing computation. Zhang et al. [11] proposed the first composite-order bilinear pairing computation algorithm based on GPU. It shows that it can achieve a 20-fold speedup compared to the state-of-the art CPU implementation. However, the method in [11] only computes many pairings in parallel and little work is done on the parallel computation of one single pairing. In this paper, we consider the parallelism both within and between pairings. We compute each pairing on a block of threads and compute each part of the pairing in parallel on these threads within this block, while we concurrently run many blocks to compute many pairings in parallel. We implemented 32 bit modular addition, subtraction and multiplication on each thread. The corresponding operations (addition, subtraction and multiplication) on  $F_q$  are conducted on a block of threads via Residue Number System (RNS). The corresponding multiplication and square operations on extension field  $F_{q^2}$ , addition and double operations on an elliptic curve are implemented upon  $F_q$  operations, which are based on a block of threads. Besides, we compute the part of  $g_{u,v}(\varphi(Q))$ , which is the most time consuming part in the main loop of one single pairing in parallel. Combining all these techniques, we are able to speed up the computation. Our method is a general method to compute the composite-order pairing in parallel and can serve for all cryptographic schemes constructed in composite-order pairing. As far as we know, our work is the first one which can truly achieve the parallel computation both within and between composite-order pairings on GPU. It is non-trivial to convert

existing CPU-version into a GPU-version due to the different levels of parallelism provided by GPU and CPU. Besides, it is difficult to divide one single pairing operation into several parts which can be computed in parallel.

## 2 Mathematics of Composite Order Bilinear Pairing

This section focuses on the basics of bilinear pairing and the group on which a bilinear pairing is defined. We describe the relationship between group size and security level and explain the reason why composite order in a bilinear pairing should be much larger than a prime order.

Let  $G_1$  and  $G_2$  be two cyclic additive groups and  $G_T$  a cyclic multiplicative group. A bilinear map (of order  $l \in N$ ) is defined as follows with three properties:

$$e_l : G_1 \times G_2 \rightarrow G_T \tag{1}$$

(a) bilinearity: for all  $P \in G_1$ , and  $Q \in G_2$ ,  $e_l(aP, bQ) = e_l(P, Q)^{ab}$ ; (b) non-degeneration:  $e_l(P, Q) \neq 1$  for some  $P$  and  $Q$ , where 1 is the identity element of  $G_T$ ; and (c) computable: there is an efficient algorithm to compute  $e_l(P, Q)$  for any  $P \in G_1$  and  $Q \in G_2$ . If there exists a distortion map:  $G_1 \rightarrow G_2$ , we can define a symmetric bilinear pairing  $e_l : G_1 \times G_2 \rightarrow G_T$  so that  $e_l(P_1, P_2) = e_l(P_1, \varphi(P_2))$  for any  $P_1, P_2 \in G_1$ .

Let  $E$  be an elliptic curve which is defined over a finite  $F_q$  where  $q = p^m$ ,  $p, m \in N$  and  $p$  is the characteristic of  $F_q$ . Let  $O$  be the point at infinity for  $E$ . For a nonzero integer  $l$ , the set of points  $P$  in  $E(F_q)$  such that  $lP = O$  is denoted as  $E(F_q)[l]$ . The group  $E(F_q)[l]$  is said to have security multiplier or embedding degree  $k$  for some  $k > 0$  if  $l|q^k - 1$  and  $l \nmid q^s - 1$  for any  $0 < s < k$ . The Tate pairing of order  $l$  is a map

$$e_l : E(F_q)[l] \times E(F_{q^k})[l] \rightarrow F_{q^k} \tag{2}$$

The pairing-friendly elliptic curve used in a composite order bilinear pairing is a super singular elliptic curve in the following form defined over a prime field.

$$E : y^2 = x^3 + (1 - b)x + b, \quad b \in \{0, 1\} \tag{3}$$

The group order  $l$  is composite and the embedding degree  $k$  is 2. There exists a distortion  $\varphi : E(F^q) \rightarrow E(F_{q^k})$  which allows us to define a symmetric bilinear map as

$$e_l : E(F_k)[l] \times E(F_q)[l] \rightarrow F_{q^k} \tag{4}$$

So that  $e_l(P, Q) = e_l(P, \varphi(Q))$  for any  $P, Q \in E(F_q)[l]$ . The order of  $E(F_q)$  is  $\# E(F_q) = q + 1$ . This curve is named as A1 curve in the PBC software library [7].

**Finite Field Size versus Security Level.** The security of pairing-based cryptosystems generally rely on two hard problems, elliptic curve discrete logarithm problem (ECDLP) in  $G$  and logarithm problem in the extension field  $F_{q^k}$ , that is,  $G_T$ . When a pairing-based cryptosystem requires 1024 bits security, the size of the extension field  $F_{q^k}$  should at least be 1024 bits long and the group order of  $G$  should at least be 160 bits long [9]. Besides, the security of most composite-order pairing-based cryptographic constructions also relies on the intractability of a problem called Subgroup Decisional Problem (SDP) [2]: for a bilinear map  $e : G \times G \rightarrow G_T$  of composite order  $l$ , without knowing the factorization of the group order  $l$ , the SDP is to decide if an element  $x$  is in a subgroup of  $G$  or in  $G$ . For the intractability of SDP, the group order  $l$  of  $G$  should be at least 1024 bits long. As  $l|q + 1$ ,  $q$  should also be at least 1024 bits long. As the embedding degree  $k$  is 2, the size of the extension field  $F_{q^2}$  is at least 2048 bits long [11].

### 3 Bilinear Pairing Algorithm

The arithmetic operations involved in our algorithms consists of the operations in the extension field  $F_{q^2}$  and the elliptic curve  $E(F_q)$  which are based on the base field operations in  $F_q$ . We mainly construct the operations in base field in RNS using the RNS Montgomery multiplication algorithm. The concrete algorithms are described in [11], reader can refer it for the details.

In our paper, we use Barreto et al.'s algorithm [1] based on the composite-order bilinear pairing in  $F_q$  to realize the bilinear pairing, because the computation flow of this algorithm relies on the system parameters instead of the input values, which is suitable for SIMD on GPU. In this algorithm, all operations are feasible on GPU, such as double operations in  $E(F_q)$  and multiplication operations in  $F_{q^2}$  discussed in [11].

We propose an improved algorithm, which transforms the calculation of  $g_{U, V\phi}$  [5] into separate steps. We take two lists of the elements in elliptic curve field *in1* and *in2* as the input values and construct *cacheVP* list to store the results in the algorithm. *CacheVP* is a defined structure which contains the followings:  $V$  in elliptic curve field,  $VV$  in elliptic curve field, and a boolean *hasVP* indicating which branch the flow in line 9 of Barreto et al.'s algorithm [1] will go into. When  $n_1 = 0$ , *hasVP* will be set to true, otherwise false. In line 11–15 of *computeVP*, we record the information of the parameters instead of computing the value of  $g_{U, V\phi}$  directly.

In  $g_{U, V\phi}$  algorithm [5], the input values become the list *cacheVP*, which is conducted in the *computeVP* algorithm. From the tuple in *cacheVP*, we can get the input value of  $U$  and  $V$ , and whether we should compute the result once or twice. Traversing all tuples in *cacheVP*, the computing of  $g_{U, V\phi}$  can be done, and get all the results stored in valuable *cacheG* that are needed in the subsequent steps. We call the phase above preprocessing.

**Algorithm 1:** computeVP

**Require:**  $heVP * cacheVP$ , elliptic curve field  $in1[], in2[]$

**Ensure:**  $E : y^2 = x^3 + x, q > 3$  and  $q \equiv 3 \pmod{4}$

```

1   $(x, y) \in E(F_q)[n], i \in F_{q^2} (i^2 = -1), \phi(x, iy) = (-x, y) \in E(F_{q^2})[n]$ 
2:   $n = (n_t, \dots, n_0), n_i \in \{0, 1\}, n_t = 1$ 
3:   $V \leftarrow P$ 
4:  for  $i \in [t-1, 0]$  do
5:     $idx = kernel\_index * len_n + i$ 
6:     $V \leftarrow 2V$ 
7:    if  $i = 0$  then
8:      break;
9:    end if
10:   if  $n_i = 1$  then
11:      $cacheVP[idx], VV \leftarrow V;$ 
12:      $cacheVP[idx], hasVP \leftarrow true;$ 
13:      $V \leftarrow V + P$ 
14:   else
15:      $cacheVP[idx], hasVP = false;$ 
16:   end if
17: end for
18: return
```

## 4 Implementation

We implement our solution using CUDA. We use a block of 67 (or 131) threads to represent an element in  $F_q$ . The thread number on one block is decided on the security level and the word length of GPU. For example, for the 1024/2048 bit composite order and word length of 32 bits,  $1024/32 = 32(64)$  bases are the least needed to represent a number (actually 33(65)). Considering another set of bases for the extension operation,  $33 + 33 + 1(65 + 65 + 1)$  are used to represent one element in  $F_q$ . Elements of the extension field and elliptic curve are represented based on base field  $F_q$ . For example, the element in extension field can be presented by a 2D tuple as  $(x, y)$ , with  $x$  and  $y$  be the elements in base field. Zhang et al. [11] discusses computation between many pairings at one time which is called basic scheme in our paper.

Global memory is time-consuming to load and communicate. Since the reduction operations will be computed over and over again, we utilize the texture memory to store the intermediate values. Texture memory is better for some specific data structures and we set it as 2D in our scheme. We bind some frequently-used data precomputed for the reduction operation to the 2D texture memory called *tex\_ref*. Besides, shared memory is suitable for sharing

information in the same block, so we allocate some share memory to store the intermediate information among threads in one block.

The basic scheme in [11] only considers the parallelism among many pairings and neglects the parallelism within the operations of extension field and elliptic curve and within the bilinear algorithm. The computation on one single pairing is still sequential. Through further analysis of the flow, we propose a method to improve the efficiency of the parallelism within one single pairing. According to the algorithm,  $n = E(F_q) = q + 1$ . We assume that  $n = (n_t \dots n_0)$ ,  $n_i \in \{0, 1\}$  and  $n_t = 1$ , then the main loop will go over  $t$  times. In each round of the computation, about 20 times of multiplication should be done plus to the reduction operations. It is computational expensive. We describe the solution to this problem in the following.

As presented in the computation details of  $g_{U, V}(\varphi(Q))$ , the computation of  $a$ ,  $b$  and  $c$  only relies on the input value of  $U$  and  $V$  and independent on the results of the previous step. It is suitable for parallelism model. We apply each computation to a different thread and then read the data and compute the return value of  $(c - ax_3) + by_3i$ , which will be stored in a global array. With the help of this preprocessing, we only need to read the pre-computing result according to the index  $i$  when calculating the value of  $g_{U, V}(\varphi(q))$ . In terms of the communication of different blocks, we use the global and shared memory to store the shared results. It is feasible to obtain the index  $i$  according to the data structures of CUDA (which are *threadID*, *blockId*, *blockDim* and *gridDim*). By applying this improvement, the computation efficiency of composite-order pairing will be greatly improved.

## 5 Experimental Result and Analysis

We compare the efficiency of the CPU scheme, the basic scheme [11] and our improved GPU scheme. The security level is 1024 bits. The CPU scheme was run on Intel Core 2 E8300 CPU at 2.83 GHz and 3 GB memory. We chose NVIDIA GTX 480 as the GPU running environment. When implementing the experiment on CPU, we adopt the Pairing Based Cryptography (PBC [7]) library.

**Analysis and Evaluation** As shown in Table 1, the basic-scheme is slower (49.9 ms) than the CPU version (38.4 ms) when the number of pairings is small. Parallel computation does not make significant contribution to the efficiency. It is time-consuming to do the pre-computing. However, as the number of pairings increases, the result reveals the advantage of using GPU. In the case of 100 pairings, the average time used by the improved scheme is only 16.23 % of the CPU version. When the number of pairings increases (from 20 pairings), the improved scheme starts to win over the basic scheme. We also analyze the loading time of memory for schemes based on GPU. Though the first access of the memory in GPU is expensive, the average time is quite short as the number of pairings

**Table 1** Results on CPU, basic and improved scheme

Time(ms)	Total-time			Average-time		
	CPU	Basic	Improved	CPU	Basic	Improved
1	38.4	49.9	52.3	38.4	49.9	52.3
20	783.1	201.8	128.8	39.2	10.1	6.4
100	3884.9	859.9	631.2	38.8	8.6	6.3
200	7698.7	1716.0	1262.3	38.5	8.5	6.3

increases. When the number of pairings grows large enough, the average time will become comparatively small. It can easily be deduced from the experimental results. In the basic scheme, when the pairing number is small, the time used is larger than that of the CPU version because the average loading time is large while it is nearly zero in the CPU scheme. Since we add some preprocessing in our improved scheme, the loading time will be longer than the basic scheme. However, according to the experimental result, the advantage becomes obvious when the number of pairing becomes large.

**Acknowledgment** This project is partially supported by the Small Project Funding of HKU (201109176091).

## References

1. Barreto PSLM, Kim HY, Lynn B, Scott M (2002) Efficient algorithms for pairing-based cryptosystems. In: CRYPTO, pp 354–368
2. Boneh D, Goh E-J, Nissim K (2005) Evaluating 2-dnf formulas on ciphertexts. In: TCC, Lecture notes in computer science, vol 3378. Springer, Heidelberg, pp 325–341
3. Fleissner S (2007) Gpu-accelerated montgomery exponentiation. In: International conference on computational science, pp 213–220
4. Freeman DM (2010) Converting pairing-based cryptosystems from composite-order groups to prime-order groups. In: EUROCRYPT, pp 44–61
5. Guillermin N (2010) A high speed coprocessor for elliptic curve scalar multiplications over  $F_p$ . In: CHES, Lecture Notes in Computer Science, vol 6225. Springer, Berlin, pp 48–64
6. Lewko AB, Rouselakis Y, Waters B (2011) Achieving leakage resilience through dual system encryption. In: TCC, pp 70–88
7. Lynn B Pbc: the pairing-based cryptography library. <http://crypto.stanford.edu/pbc/>
8. Meiklejohn S, Shacham H, Freeman DM (2010) Limitations on transformations from composite-order to prime-order groups: the case of round-optimal blind signatures. In: ASIACRYPT, pp 519–538
9. NIST (2007) Recommendation for key management
10. NVIDIA Corporation (2010) Nvidia CUDA C programming guide
11. Zhang Y, Xue CJ, Wong DS, Mamoulis N, Yiu SM (2012) Accelerating bilinear pairing on graphics hardware. In: ICICS, pp 341–348



# Deployment and Management of Multimedia Contents Distribution Networks Using an Autonomous Agent Service

Kilhung Lee

**Abstract** This paper introduces an agent application service and shows its application through the implementation of a multimedia data distribution service. This service is broadly comprised of agents, agent systems, an agent master and agent manager components. A software component of relaying multimedia data traffic is developed as an agent application, and is created dynamically with the distribution of the content client in the network. By using this service, it is possible to deploy a new multimedia data distribution with greater speed, efficiency and convenience.

**Keywords** Management · Multimedia content distribution · Autonomous agent service

## 1 Introduction

With the development and adoption of the Internet, many new and hitherto unimagined application services have been introduced through this revolutionary medium. Multimedia data service will be the major application of the future Internet and Intranet environments. The peer-to-peer (P2P) method has strong merits given its scalability, and has the potential to serve as an applicable means of tackling errors more easily. Likewise, mobile agent technologies that can conduct critical operations at precise locations are of significant importance in the overall software technology and development progress [1]. An agent is an autonomous programming object that can perform by itself those functions that have been entrusted to it by the software user [2]. This operational method is an innovative

---

K. Lee (✉)

Department of Computer Science and Engineering, Seoul National University of Science and Technology, 172 Gongnung2-dong, Nowon-gu, Seoul 139-743, Korea  
e-mail: khlee@seoultech.ac.kr

way of implementing the traditional mobile and distributed computing system and constitutes a new computing paradigm [3]. Through proper utilization of the functions of an agent, which can be used to operate in specific places at precise times, we can achieve significant reductions in network traffic and contribute to increased operational efficiencies [4, 5].

## 2 Component of Autonomous Agent Service

Our Autonomous Agent Service is an agent system framework developed Java language and consists of several components. Each component is a package of classes that has a role in agent application environment. The agent framework component is composed of the agent, agent system, agent master system, and agent manager [6].

**Agent:** An agent is a software component and is a small program that is designed to fulfill certain operations. An agent is also capable of moving around the different systems and once it arrives at its desired location it begins specific operations. An agent is one of the classes devised for certain interface in Java.

**Agent System:** An agent system manages the agent operation system and provides the place (i.e. the operation part of the agent). After receiving the service code, the agent system creates the agent object and provides the operation environment using services such as begin, stop, restart, and termination.

**Agent Master:** An agent master controls the local agent system, intercedes with clients who are service users for the provided agent service, and provide management interface to the agent manager. The agent system signals the start of its service operations by beginning it at the same time it registers in the agent master. When an agent from the client connects with the service to be used, the agent master returns a system list containing available agents.

**Agent Manager:** An agent manager manages the environments of the overall agent application framework. An agent manager can help achieve management efficiency, flexibility and scalability with fewer complications by managing the agent master. The agent master embodies agent management information that is necessary to manage how agents behave. The agent manager can carry out monitoring activities by reading a Management Information Base (MIB) and service management functions by creating needed values.

**Client:** A client request and control the service and can offer service codes through the agent master. A client can either provide the requested service operation code directly or the agent system can remove the address that the client provides to the service storage. The client can designate the agent system directly or help the agent master choose the appropriate agent to carry out the service. In the case of P2P applications, the agent system and the client can be combined and operated together [7].

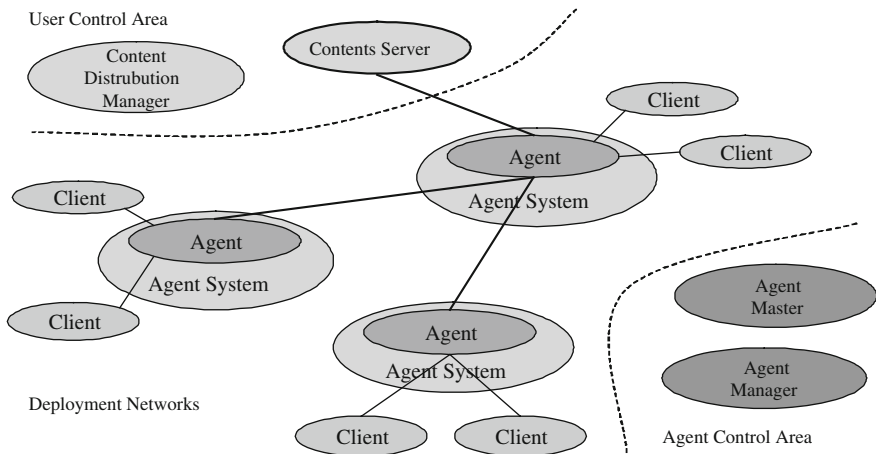


Fig. 1 Content distribution application environment

### 3 Management of Multimedia Contents Distribution Network

The content distribution network in Fig. 1 is controlled by the distribution manager. The distribution manager is one type of agent client in our agent application framework. Through the agent master system, the distribution manager initiates the service agent in the necessary agent system after searching the profile of the agent system. When agents commence new operations, they transmit the message to the distribution manager and register themselves. The distribution manager, according to its location, locates the new agents into their proper position within the distribution tree. When the content clients near the new agents request a content service, the distribution manager makes it possible to provide the service by connecting new agents to the proper agents. The functions of the components in the content service model are as follows.

**Content Server:** The content server is the data source that provides the information to the content client. The services provided by the content server include broadcasting, chatting, Video on Demand (VOD), video conferencing, stock information services, real-time news services, and more. In addition, depending on the types of services provided, there are various kinds of servers to be used, such as a push server where the client gathers information, and pull servers where the server provides information [8].

**Content Distribution Agent:** A distribution agent takes the contents from the content server and distributes them again to the client or other distribution agents. Distribution agents, depending on the content types served, take different forms. Therefore, the agents locate themselves where the service providers want them to place and are served in specific manner depending on the particular services and type of server. Like servers, agents also provide services with the client list and

depending on the quality and forms of service, one or more threads are involved in the operations. Agents monitor the quality of service while at the same time providing services.

**Content Distribution Manager:** A content distribution manager is the component that controls the data service between the server and the client. A content distribution manager possesses all of the information regarding the server, distribution agent, and the client served, and manages the types of services provided by coordinating the distribution tree. In addition, the content distribution manager is involved in the preparation of the list and location of new agent through the service quality that is monitored and reported, and readjusts the distribution tree. This system is the component that performs the function of a client in the agent application framework environments.

**Content Client:** A content client receives the content from the agent distributor and sends it to users in the proper forms. In general, the communication handling and information handling portions constitute the client system.

Figure 2 is the snapshot of the content distribution manager application. After multimedia data source server activates, content distribution manager initiates relaying agents in some important points of networks. Thereafter, new clients are then attached to the appropriate relaying agent that meets the traffic requirements, and receives data services from them. When client are increased and the quality of service is decreased in specific sectors, content distribution manager detects and

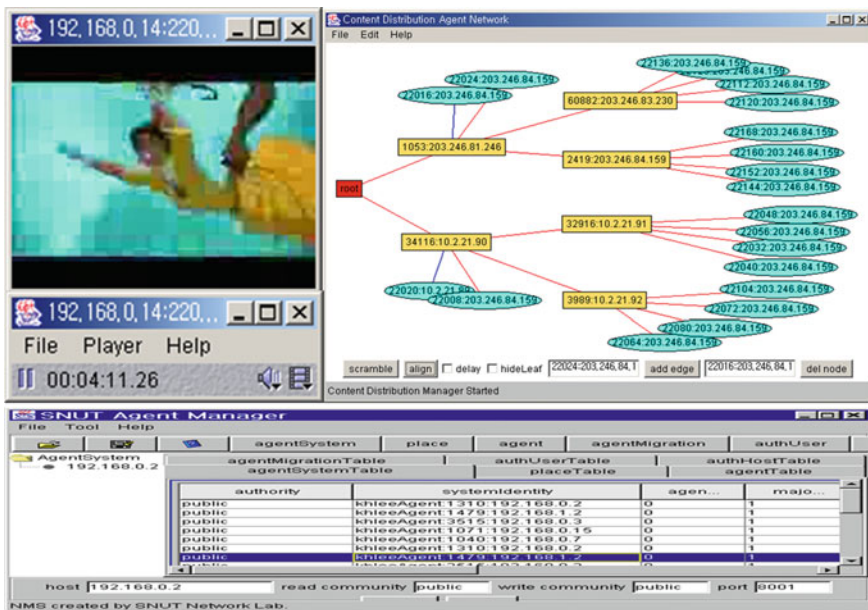


Fig. 2 Content distribution manager application

**Table 1** Transmission characteristic of the content distribution service

Hop count	2	3	4	5
Average delay (ms)	32	48	62	81
Jitter (ms)	+58/−14	+64/−21	+68/−25	+71/−28

deploy a new relaying agent to an agent management system resides in that sectors [9].

Table 1 shows the property of the data traffic delivered in a multicast tree design using the agent application framework. The transmission property in the Intranet is very satisfactory and is subject to significant property changes depending on the status of number of client and host load. As the transmission hop of the tree is increased, delays are also proportionally increased. Despite this delay increasing property, the average error in delays tends not to be dramatically changed. Jitters also show slight increasing properties.

Between autonomous systems, the sending and receiving transmission property had 2973 ms in average delay and an average jitter of +349/−572 ms. The delays in the Intranet and the value of the jitter are not subject to significant changes.

## 4 Summary

This paper has introduced the management of an agent application service and the example development of content distribution network as an application. The multimedia delivery performance of the agent application framework is acceptable in this test. A multimedia data service network can be created and controlled by this framework with simplicity, efficiency, accessibility, increased controllability and manageability properties.

## References

1. Yang X, Zhang Y, Niu Q, Tao X, Wu L (2007) A mobile-agent-based application model design of pervasive mobile devices. In: 2nd International conference on pervasive computing and applications, pp 1–6
2. Karmouch A, Pham VA (1998) Mobile software agents: an overview. *IEEE Commun Mag* 36(7) 26–37
3. Wang YH, Keh HC, Hu TC, Liao CH (2005) A hierarchical dynamic monitoring mechanism for mobile agent location. In: 19th International conference on advanced information networking and applications, vol 1, pp 351–356
4. Glietho RH, Olougouna E, Pierre S (2002) Mobile agents and their use for information retrieval: a brief overview and an elaborate case study. *IEEE Network* 14:34–41
5. Baek JW, Yeom HY (2003) d-Agent: an approach to mobile agent planning for distributed information retrieval. *IEEE Trans Consum Electron* 49(1):115–122
6. DC00087C (2002) Mobile agent facility specification support for mobility specification

7. Stolarz D (2001) Peer-to-peer streaming media delivery. In: Proceedings of the first international conference on peer-to-peer computing, pp 48–52
8. Yang S, Yang H, Yang Y (2003) Architecture of high capacity VOD server and the implementation of its prototype. *IEEE Trans Consum Electron* 49(4):1169–1177
9. Chen JS, Shi HD, Chen CM, Hong ZW, Zhong PL (2008) An efficient forward and backward fault-tolerant mobile agent system. In: Eighth international conference on intelligent systems design and applications, vol 2, pp 61–66

**Part XIII**  
**Advanced Mechanical and Industrial**  
**Engineering, and Control II**

# Design Optimization of the Assembly Process Structure Based on Complexity Criterion

Vladimir Modrak, Slavomir Bednar and David Marton

**Abstract** This paper focuses on configuration design optimization of the assembly supply chain network. It is intended to use this approach to select an optimal assembly process structure in early stages of manufacturing/assembly process design. For the purpose of optimization, structural complexity measures as optimality criteria are considered. In order to compare alternatives in terms of their complexity, a method for creating comparable process structures is outlined. Subsequently, relevant comparable process structures are assessed to determine their structural complexity.

**Keywords** Complexity indicator · Assembly model · Vertex · Arc · Tier

## 1 Introduction

The design optimization of assembly networks is one of the challenging issues for the practitioners and researchers in order to get high performances with low prices. To react to the trend of agile manufacturing, companies are endeavoring to provide a wide variety of modular products. A major advantage of this strategy is larger quantity of standard modules, which contribute to cost reduction and reduce total cycle time. However, the assembly process can become quite complex as the

---

V. Modrak (✉) · S. Bednar · D. Marton  
Faculty of Manufacturing Technologies with Seat in Presov, Technical University of Kosice,  
Bayerova 1, Presov, Slovakia  
e-mail: vladimir.modrak@tuke.sk

S. Bednar  
e-mail: slavomir.bednar@tuke.sk

D. Marton  
e-mail: david.marton@tuke.sk



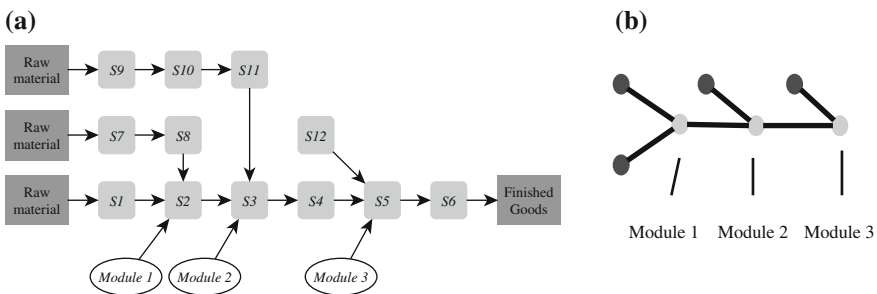
product variety increases. Therefore, by optimizing the assembly process structure it is aimed to reduce complexity and, thereby, reduce assembly times. This work applies network complexity indicators in order to compute the structural complexity measures of a simple case of the assembly process. To frame this problem we apply here a method for exact generating of all possible assembly process structures based on a number of initial nodes.

**Related works**

Probably the most important optimization criterion of assembly lines as parts of manufacturing processes has long been focused on reducing costs in each phase from product development to market achievement [1]. Many such efforts have been focused on use of modern managerial tools with aim to ensure high product quality standards, volume and mix flexibility, and delivery speed and reliability [2–5]. In configuration design, there is a number opportunities to adopt methods and tools for the evaluation of assembly process structures. Undoubtedly, novel metrics for assessing the structural complexity of manufacturing system configurations are a demanding challenge [6]. In general, the complexity of assembly supply chain networks can be characterized in terms of several interconnected aspects. Some of these aspects were described by, e.g. [7–10]. Three basic dimensions of structural complexity that links the uncertainty with performance were identified in the work presented by Milgate [11]. An innovative complexity measure for assembly supply chains has been proposed by Hu et al. [12]. This complexity measure is based on Shannon’s information entropy and is closely related to Index of vertex degree [13] that was used in presented research.

**2 Problem and Method Description**

In this work, we are interested in optimization of the assembly process structure using minimal complexity level as criterion. For this purpose we have selected a simple real assembly process, model of which is shown in Fig. 1a.



**Fig. 1** a Original assembly process structure [14], b Simplified process structure

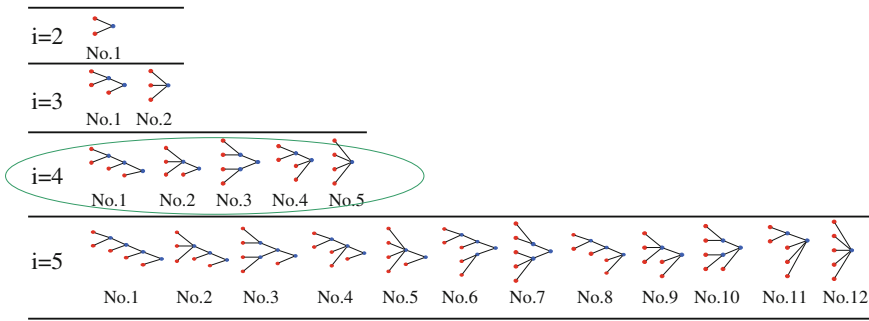


Fig. 2 Fragment of assembly process structure classes [15]

In order to compare this process structure with alternative ones we firstly transformed this model into simplest network (see Fig. 1b). Then we can formally describe this assembly network as a structure with the following elementary parameters: Number of initial nodes-4; number of all nodes-6, number of modules -3, number of tiers-4. A fragment of this classification is shown in Fig. 2. In this stage we can compare structural complexity level of all relevant process networks. For this purpose we used set of different complexity metrics to be able to objectively determine complexity differences among alternative structures of given class. These complexity indicators are described in the next paragraph.

### 3 Description of Complexity Indicators

#### 3.1 Index of Vertex Degree $I_{vd}$

Bonchev and Buck [13] by adopting Shannon’s information theory propose the following indicator to characterize complexity of a network:

$$I_{vd} = \sum_{i=1}^V \text{deg}(v)_i \log_2 \text{deg}(v)_i \tag{1}$$

where  $\text{deg}(v)_i$  represents number of the nearest-neighbors of a vertex  $i$ .

#### 3.2 Modified Flow Complexity $MFC$

Modified flow complexity indicator [16] combines FC together with Multi-Tier ratio (MTR) and index (MTI), and Multi-Link ratio (MLR). Using MTI, MTR and MLR we can determine  $\alpha$ ,  $\beta$  and  $\gamma$  coefficients. MFC basically counts all Tiers (including Tier 0), Nodes and Links and adds all these counts, weighted with

determined  $\alpha$ ,  $\beta$  and  $\gamma$  coefficients. In MFC indicator, Nodes and Links are counted only once, even if they are repeated in graph. Presence of Nodes and Links repetition is included in coefficients. In mathematical term, the MFC indicator can be expressed as follows:

$$MFC = \alpha \cdot T + \beta \cdot N + \gamma \cdot L, \quad (2)$$

$$\alpha = MTI = \frac{TN - N}{(T - 1) \cdot N}, \quad (3)$$

$$\beta = MTR = \frac{TN}{N}, \quad (4)$$

$$\gamma = MLR = \frac{LK}{L}. \quad (5)$$

where: N—Number of Nodes, TN—Number of Nodes per i-th Tier Level, L—Number of Links, LK—Number of Links per i-th Tier Level, T—Number of Tiers.

### 3.3 Supply Chain Length LSC

Németh and Foldesi [17] described Supply Chain Length (LSC) indicator and its extended definition. The LSC indicator takes besides number of nodes also number of links weighted by the complexity of links into consideration. It is mainly focused on material flows. The equation formula of LSC is expressed by the equation:

$$LSC = c_1 \cdot \sum_{i \in P} w_s \cdot V_i + c_2 \cdot \sum_{(i,j) \in P} f(D_{ij}) \cdot A_{ij} \quad (6)$$

where:  $c_1$ —constant represent the technical and managerial level of vertices,  $c_2$ —constant represent the technical and managerial level of edges,  $w_s$ —weight corresponding the nature of node, P—path from the origin to the destination,  $V_i$ —the vertices (nodes) in the path,  $A_{ij}$ —the arcs (edges) in the path,  $D_{ij}$ —distance in logistic terms (in this study it equals 1),  $f(D_{ij})$ —the weight determined by the distance in logistic terms.

### 3.4 Links Tiers Index LTI

When comparing two or more structures with the same number of tiers “t” and nodes “n” but with different number of links “l” the following argument can be constructed. The structure with the smallest number of links is topologically less complex than other one(s). Then, it is proposed to measure structural complexity by formula Links/Tiers Index [15]:

$$LTI = \sum_{j=1}^p \sum_{l=1}^m l_j \cdot t_l \quad 0,1 \tag{7}$$

### 3.5 Flow Complexity FC

The FC is proposed by Crippa [18]. It can be expressed by Eq. 8 and it counts all Tiers (including Tier 0), Nodes and Links and adds all these counts, weighted with arbitrary chosen  $\alpha$ ,  $\beta$  and  $\gamma$  coefficients. Nodes are counted only once, even if they are repeated in Tiers. Presence of repetition is included in Links count.

$$FC = \alpha \cdot \sum_{i=1}^n T_i + \beta \cdot \sum_{s=1}^m N_s + \gamma \cdot \sum_{i=1}^n \sum_{j=1}^k LK_{ij} \tag{8}$$

where:  $T_i$ —ith Tier,  $N_s$ —sth Node,  $LK$ —ith and jth Link.

### 3.6 Complexity Degree $\kappa$

Maksimovic and Petrovic [19] described a Complexity degree ( $\kappa$ ) indicator and its extended definition based on two fundamental constituents of each structure. The  $\kappa$  indicator takes besides number of elements and the interrelation between elements within the structure. It is mainly focused on flows in a system. Formally  $\kappa$  is expressed by the formula:

$$\kappa = \frac{\sum_{i=1}^{i=m} m_i}{m} \tag{9}$$

where:  $m_i$ —number of links,  $m$ —number of nodes.

### 3.7 Average Shortest Length ASP

The ASP is a network indicator which is applicable for determination distance of network between every pairs of nodes. Alex and Efstathiou [20] used it for interpretation of robustness complex networks as fragmentation of network. Formally can be described as follows:

$$ASP = \frac{1}{N \cdot (N - 1)} \cdot \sum \sum d_{ij}. \tag{10}$$

where:  $d_{ij}$ —is the shortest path in the network for all nodes from i till j.

**Table 1** Computational results of individual indicators

Graph No.	Indicators						
	Ivd	MFC	LSC	LTI	FC	K	ASP
No. 5	8	9	9	0,8	11	0,8	0,2
No. 2	9,51	11	11	1,5	14	0,83	0,36
No. 4	10	11	11	1,5	14	0,83	0,3
No. 3	11,51	13	13	1,8	16	0,86	0,33
No. 1	11,51	13	13	2,4	17	0,86	0,48

## 4 Testing of Alternative Process Structures

Using the above mentioned indicators for complexity levels computation of alternative assembly structures we have obtained values summarized in Table 1. As we can see individual indicators assign an approximate complexity values to individual structures. It means that even if individual indicators use different complexity calculation methods they still show comparable results. For us it is important to know that the process we want to optimize can theoretically be replaced by the other four structures except for structure No. 3, where topology is not transferable into structure No. 1 and vice versa. The simplest transformation way of structure No. 1 is provided by structures No. 2 and 4. It is because the transformation only needs a single integration of two modules into one. For the replacement of structure No. 1 by structure No. 5 it is necessary to integrate 3 modules into one. This type of reduction is no effective from our perspective. For that reason we consider structure No. 5 as irrelevant for purpose of substitution. From our perspective structures No. 2 and 4 are comparable. Taking in mind the specifics of the structure we want to optimize we only need to integrate two modules to obtain structure No. 2 and we would have to integrate two modules together with one operation to obtain structure No. 4.

## 5 Conclusions

Presented approach showed that complexity reduction of the assembly process structure can be effectively achieved through fusion of only two modules into one. All other attempts leading to integration of modules would give less effective results. As described above, principally, there exist only three possible ways of structural complexity reduction of given process structure. Such a method can be used as a supportive tool for designers in optimal process designing of any assembly networks.

## References

1. Patterson KA, Grimm CM, Corsi TM (2003) Adopting new technologies for supply chain management. *Transp Res E-Log* 39:95–121
2. Thomas DJ, Griffin PM (1996) Coordinated supply chain management. *Eur J Oper Res* 94:1–15
3. Holmes G (1995) Supply chain management: Europe's new competitive battleground. EIU Research report
4. Gots I, Zajac J, Vojtko I (1995) Equipment for measuring the degree of wear to cutting tools. *Tech Mes* 1:8–11
5. Cuma M, Zajac J (2012) The impact analysis of cutting fluids aerosols on working environment and contamination of reservoirs. *Tech Gaz* 19:443–446
6. Kuzgunkaya O, ElMaraghy HA (2006) Assessing the structural complexity of manufacturing systems configurations. *Int J Flex Manuf Sys* 18:145–171
7. Deshmukh AV, Talavage JJ, Barash MM (1998) Complexity in manufacturing systems. Part 1: analysis of static complexity. *IIE Trans* 30:35–44
8. Modrak V (2006) Evaluation of structural properties for business processes. In: 6th international conference of enterprise information systems ICEIS, Porto, pp 619–622
9. Calinescu A, Efstathiou J, Schirn J, Bermejo J (1998) Applying and assessing two methods for measuring complexity in manufacturing. *J Oper Res Soc* 49:723–733
10. Modrak V (2007) On the conceptual development of virtual corporations and logistics. In: Symposium on logistics and industrial informatics, Wildau, pp 121–125
11. Milgate M (2011) Supply chain complexity and delivery performance: an international exploratory study. *Sup Ch Manag Int J* 6:106–118
12. Hu SJ, Zhu XW, Wang H, Koren Y (2008) Product variety and manufacturing complexity in assembly systems and supply chains. *CIRP Ann Manuf Technol* 57:45–58
13. Bonchev D, Buck GA (2005) Quantitative measures of network complexity. In: Bonchev D (eds) *Complexity in chemistry, biology and ecology*, Springer, pp 191–235
14. Wang S, Bhaba RS (2005) An assembly-type supply chain system controlled by kanbans under a just-in-time delivery policy. *Eur J Oper Res* 162:153–172
15. Modrak V, Marton D, Kulpa W, Hricova R (2012) Unraveling complexity in assembly supply chain networks. In: 4th IEEE international symposium on logistic and industrial informatics LINDI, Smolenice, pp 151–155
16. Modrak V, Marton D (2012) Modelling and complexity assessment of assembly supply chain systems. *Proc Eng* 48:428–435
17. Németh P, Foldesi P (2009) Efficient control of logistic processes using multi-criteria performance measurement. *Act Tech Jaur Log* 2:353–360
18. Crippa R, Bertacci N, Larghi L (2006) Representing and measuring flow complexity in the extended enterprise: the D4G approach. In: RIRL international congress for research in logistics
19. Maximovic R, Petrovic S (2009) Complexity of production structures. *Fact Univer Mech Eng* 7:119–136
20. Alex KSNg, Efstathiou J (2006) Structural robustness of complex networks. *Phys Rev* (3):175–188

# Kinematics Modelling for Omnidirectional Rolling Robot

Soo-Yeong Yi

**Abstract** A ball-shaped mobile robot, called a ballbot, has a single point of contact with the ground. Thus, it has low energy consumption for motion because of the reduced friction. This paper presents the systematic kinematics modelling for a type of ballbot with omnidirectional motion capability. This kinematics modelling describes the velocity relationship between the driving motors and the robot body for the motion control of the robot.

## 1 Introduction

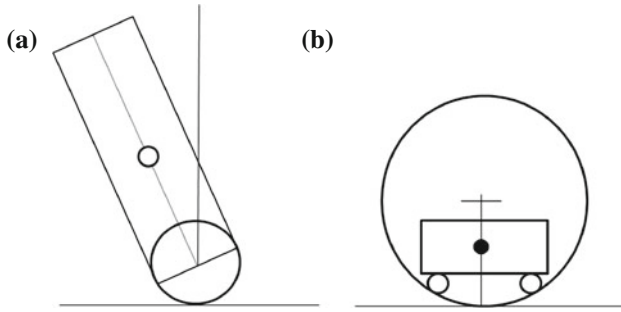
A ball-shaped robot has a single point of contact with the ground, which reduces its friction with the ground. A ball-shaped robot is generally called a ballbot. In comparison with a conventional wheeled mobile robot, a ballbot consumes less energy for motion because of the reduced friction [1]. There are two types of ballbots, as illustrated in Fig. 1. The ballbot shown in Fig. 1a has a cylindrical body on the top of a ball [2, 3]. This cylindrical body has driving motor and wheel assemblies in contact with the exterior of the ball to exert a driving force. In contrast, the ballbot shown in Fig. 1b has a pure spherical shape and contains a driving mechanism inside the ball [4]. The driving mechanisms can be classified into two types: (i) the wheeled platform type (Fig. 1b) [4] and (ii) the pendulum type (Fig. 2) [1].

The ballbot shown in Fig. 1a has many motion control difficulties because its posture is essentially unstable. In contrast, the pure ball-shaped robots shown in Fig. 1b and Fig. 2 are inherently stable, so the motion control is relatively stable.

---

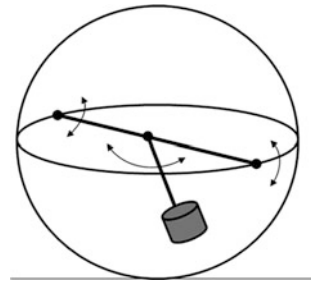
S.-Y. Yi (✉)

Department of Electrical and Information Engineering, Seoul National University  
of Science and Technology, Seoul, Republic of Korea  
e-mail: suylee@seoultech.ac.kr



**Fig. 1** Types of the ballbot. **a** Cylindrical robot body. **b** Spherical robot body

**Fig. 2** Pendulum type driving mechanism



However, these ballbots still have a motion control problem because the driving mechanism cannot provide omnidirectional motion capability.

In this paper, the systematic kinematics modelling for a ballbot with a wheel-type driving mechanism inside a ball is addressed for the motion control of the ballbot. More specifically, the driving mechanism is a platform with three Swedish wheels, so the ballbot has omnidirectional motion capability without nonholonomic constraints. Thus, the motion control of the ballbot becomes comparatively simple.

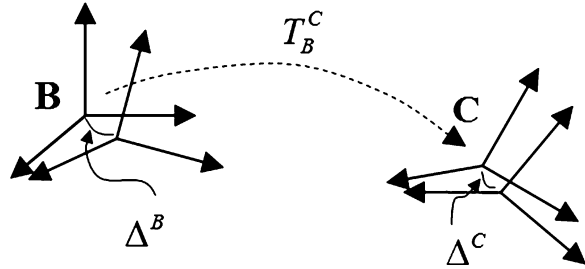
## 2 Differential Motion Between Coordinate Frames

The kinematics modelling can be described by the velocity relationship between the active driving motor and the robot body. When the transformation between two coordinate frames, **B** and **C**, is given as  $T_B^C$ , the relationship of the differential motions between the coordinate frames shown in Fig. 3 is described as

$$\Delta^C = T_B^{C-1} \cdot \Delta^B \cdot T_B^C, \tag{1}$$



**Fig. 3** Relationship of differential motions between coordinates frames



where  $\Delta^B$  and  $\Delta^C$  denote the differential motions in the corresponding coordinate frames [5]. The differential motion ( $\Delta$ ) implies a velocity transform if the motion occurs in a small time interval,  $\delta t$ , and can be written as

$$\Delta = \begin{bmatrix} 0 & -\delta_z & \delta_y & d_x \\ \delta_z & 0 & -\delta_x & d_y \\ -\delta_y & \delta_x & 0 & d_z \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2}$$

where  $\bar{\delta} = [\delta_x \ \delta_y \ \delta_z]^t$  and  $\bar{\mathbf{d}} = [d_x \ d_y \ d_z]^t$  denote the rotational and translational differential motions, respectively.

The transformation  $T_B^C$  is represented by column vectors as (3).

$$T_B^C = [\bar{\mathbf{n}} \ \bar{\mathbf{o}} \ \bar{\mathbf{a}} \ \bar{\mathbf{p}}]. \tag{3}$$

Then, from (1) and (2), each component of the differential motions in coordinate system C becomes (4-1) and (4-1).

$$\bar{\delta}^C = \begin{bmatrix} \delta_x^C & \delta_y^C & \delta_z^C \end{bmatrix}^t = \begin{bmatrix} \bar{\delta}^B \cdot \bar{\mathbf{n}} & \bar{\delta}^B \cdot \bar{\mathbf{o}} & \bar{\delta}^B \cdot \bar{\mathbf{a}} \end{bmatrix}^t, \tag{4-1}$$

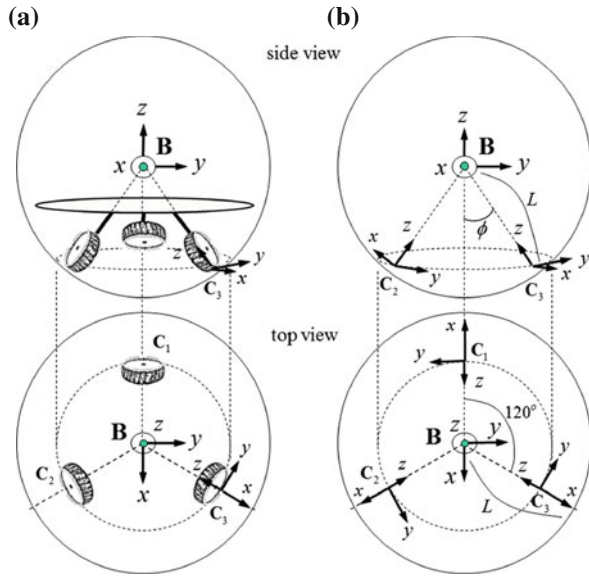
$$\begin{aligned} \bar{\mathbf{d}}^C &= \begin{bmatrix} d_x^C & d_y^C & d_z^C \end{bmatrix}^t \\ &= \begin{bmatrix} \bar{\mathbf{n}} \cdot (\bar{\delta}^B \times \bar{\mathbf{p}} + \bar{\mathbf{d}}^B) & \bar{\mathbf{o}} \cdot (\bar{\delta}^B \times \bar{\mathbf{p}} + \bar{\mathbf{d}}^B) & \bar{\mathbf{a}} \cdot (\bar{\delta}^B \times \bar{\mathbf{p}} + \bar{\mathbf{d}}^B) \end{bmatrix}^t. \end{aligned} \tag{4-2}$$

In the above equations, “.” and “ $\times$ ” denote the inner product and outer product, respectively, of the vectors.

### 3 Structure of Ballbot and Assignment of Coordinate Frames

The structure of the ballbot considered in this paper is shown in Fig. 4a. The driving system inside the ball is an omnidirectional mobile platform having three Swedish wheels with 120° spacing. Each wheel is normal to the interior tangential

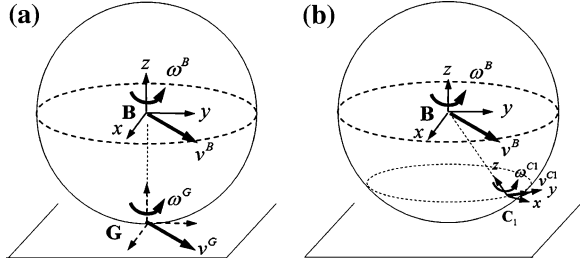
**Fig. 4** Structure and coordinate assignment of proposed ballbot. **a** Structure of ballbot. **b** Coordinate frames



plane of the ball at the point of contact. The driving force of the ball comes from the friction between the wheel and the interior surface of the ball. The coordinate frame assignment of the ballbot is depicted in Fig. 4b. In this figure, the coordinate frame at the centre of the ball is denoted as **B**, which is the inertial coordinate frame attached to the driving platform. The coordinate frames at the contact points of the wheels on the inside surface of the ball are represented as  $C_i, i = 1, 2, 3$ .

To derive the kinematics model, it is assumed that the motion of the ballbot is quasi-static. This quasi-static motion implies that the motion has a constant velocity and, as a consequence, the driving platform inside the ballbot maintains level always when in motion. This assumption simplifies the motion of the ballbot by disregarding the dynamics effects and gives the velocity relationship between each driving wheel and the robot body. The motion of the ballbot can be described by the translational and rotational velocities at ground contact **G**. Here, the translational velocity implies the differential motion on the horizontal  $x$ - $y$  plane, and the rotational velocity denotes the differential motion about the vertical  $z$  axis at the ground contact (Fig. 5). It should be noted that the translational and rotational velocities ( $v_{xy}^G, \omega_z^G$ ) of the ballbot are the same as the velocities of the inertial coordinate frame, **B**, of the platform ( $v_{xy}^B, \omega_z^B$ ). Thus, the motion kinematics of the ballbot can be represented by the velocity relationship between coordinate frame **B** and each wheel coordinate frame  $C_i, i = 1, 2, 3$ .

**Fig. 5** Motion of ballbot ( $v_{xy}, \omega_z$ ): **a** motion at ground contact **G** and at centre of ball **B**; **b** motion at centre of ball **B** and at each driving wheel **C<sub>i</sub>**



### 4 Velocity Relationship Between Wheel and Robot Body

From Fig. 4b, coordinate transformations from **B** to **C<sub>i</sub>**,  $i = 1, 2, 3$  are given as follows:

$$\begin{aligned}
 T_B^{C_1} &= Rot(z, \pi) \cdot Rot(y, -\phi) \cdot Trans(z, -L) \\
 &= \begin{bmatrix} -c\phi & 0 & -s\phi & -Ls\phi \\ 0 & -1 & 0 & 0 \\ s\phi & 0 & c\phi & -Lc\phi \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{5-1}
 \end{aligned}$$

$$\begin{aligned}
 T_B^{C_2} &= Rot\left(z, -\frac{\pi}{3}\right) \cdot Rot(y, -\phi) \cdot Trans(z, -L) \\
 &= \begin{bmatrix} -\frac{1}{2}c\phi & \frac{\sqrt{3}}{2} & -\frac{1}{2}s\phi & \frac{1}{2}Ls\phi \\ -\frac{\sqrt{3}}{2}c\phi & \frac{1}{2} & \frac{\sqrt{3}}{2}s\phi & -\frac{\sqrt{3}}{2}Ls\phi \\ s\phi & 0 & c\phi & -Lc\phi \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{5-2}
 \end{aligned}$$

$$\begin{aligned}
 T_B^{C_3} &= Rot\left(z, \frac{\pi}{3}\right) \cdot Rot(y, -\phi) \cdot Trans(z, -L) \\
 &= \begin{bmatrix} \frac{1}{2}c\phi & -\frac{\sqrt{3}}{2} & -\frac{1}{2}s\phi & \frac{1}{2}Ls\phi \\ \frac{\sqrt{3}}{2}c\phi & \frac{1}{2} & -\frac{\sqrt{3}}{2}s\phi & \frac{\sqrt{3}}{2}Ls\phi \\ s\phi & 0 & c\phi & -Lc\phi \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{5-3}
 \end{aligned}$$

In (5-1)–(5-3),  $L$  denotes the radius of the ball,  $\phi$  is the zenith angle as shown in Fig. 4b, and  $s\phi$  and  $c\phi$  represent  $\sin(\phi)$  and  $\cos(\phi)$ , respectively. The transformations in (5-1)–(5-3) can be written by column vectors as (3). In the case of  $i = 1$  for example, the relationship between each component of the differential motion of (2) can be obtained as follows. It should be noted that  $d_z^B = 0$  and  $\delta_x^B = \delta_y^B = 0$  at first. From (1), (2), (3), and (5-1) through (5-3), the differential motion components are given as follows by equating both sides of (1):

$$\bar{\delta}^{C_1} = \begin{bmatrix} \delta_x^{C_1} & \delta_y^{C_1} & \delta_z^{C_1} \end{bmatrix}^t = \begin{bmatrix} s\phi \delta_z^B & 0 & c\phi \delta_z^B \end{bmatrix}^t, \tag{6-1}$$

$$\bar{d}^{C_1} = \begin{bmatrix} d_x^{C_1} & d_y^{C_1} & d_z^{C_1} \end{bmatrix}^t = \begin{bmatrix} -c\phi d_x^B & Ls\phi \delta_z^B - d_y^B & -s\phi d_x^B \end{bmatrix}^t. \tag{6-2}$$

It should be noted that the active motion at each wheel is only  $v_y^{C_1}$  according to the coordinate assignment in Fig. 4b, which can be generated by the driving motor of the wheel. From (6-2),  $v_y^{C_1}$  is given as follows:

$$v_y^{C_1} = Ls\phi \omega_z^B - v_y^B. \tag{7}$$

Equation (7) represents the velocity relationship between a wheel and the robot body.

Similarly, from (1)–(6-2), the relationship between the velocity motion at **B** and the active motion of each wheel can be obtained as

$$v_y^{C_2} = Ls\phi \omega_z^B + \frac{\sqrt{3}}{2} v_x^B + \frac{1}{2} v_y^B, \tag{8}$$

$$v_y^{C_3} = Ls\phi \omega_z^B - \frac{\sqrt{3}}{2} v_x^B + \frac{1}{2} v_y^B. \tag{9}$$

Finally, from (7)–(9), the velocity kinematics of the ballbot in matrix form is described as

$$\begin{bmatrix} v_y^{C_1} \\ v_y^{C_2} \\ v_y^{C_3} \end{bmatrix} = \begin{bmatrix} 0 & -1 & Ls\phi \\ \frac{\sqrt{3}}{2} & \frac{1}{2} & Ls\phi \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} & Ls\phi \end{bmatrix} \begin{bmatrix} v_x^B \\ v_y^B \\ \omega_z^B \end{bmatrix} \rightarrow \tag{10}$$

$$\begin{bmatrix} v_x^B \\ v_y^B \\ \omega_z^B \end{bmatrix} = \begin{bmatrix} 0 & -1 & Ls\phi \\ \frac{\sqrt{3}}{2} & \frac{1}{2} & Ls\phi \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} & Ls\phi \end{bmatrix}^{-1} \begin{bmatrix} v_y^{C_1} \\ v_y^{C_2} \\ v_y^{C_3} \end{bmatrix}.$$

## 5 Conclusions

The ballbot considered in this paper uses an omnidirectional mobile platform with three Swedish wheels inside it as a driving mechanism and has several advantages over conventional ballbots: free motion without nonholonomic constraints, an inherently stable posture, and low energy consumption for motion. Kinematics modelling as an equation of motion is an essential prerequisite for motion control. Systematic kinematics modelling was addressed in this paper, which described the velocity relationship between the driving motors and the robot body for the motion control.

## References

1. Kim J, Kwon H, Lee J (2009) A rolling robot: design and implementation. In: Proceedings of 7th Asian control conference, pp 1474–1479
2. Lauwers T, Kantor G, Hollis R (2006) A dynamically stable single-wheeled mobile robot with inverse mouse-ball drive. In: Proceedings of IEEE international conference on robotics and automation
3. Kumagai M, Ochiai T (2008) Development of a robot balancing on a ball. In Proceedings of international conference on control, automation and systems, pp 433–438
4. Bicchi A, Balluchi A, Prattichizzo D, Gorelli A (1997) Introducing the sphericle: an experimental testbed for research and teaching in nonholonomy. In: Proceedings of IEEE international conference on robotics and automation, pp 2620–2625
5. McKerrow P (1990) Introduction to robotics. Addison-Wesley, Reading

# Design of Device Sociality Database for Zero-Configured Device Interaction

Jinyoung Moon, Dong-oh Kang and Changseok Bae

**Abstract** Nowadays people connected to the Internet are using more than six mobile connected personal devices, such as smartphones, smart pads, and laptops. To provide a simple way to share multiple devices owned by themselves or by their family and friends without configuration, this research aims at building and managing social relationships of personal devices by using human relationships obtained by social networking services. This paper proposes the design of device sociality database on the basis of ER diagrams, which is a critical step to store and manage data and information required for zero-configured device interaction. The database design is made up of the resource specification of personal devices and device sociality including device ownership, human relationships, and access permission of device resources.

**Keywords** Database design · Device sociality · Human relationship · Device collaboration

## 1 Introduction

Nowadays people own multiple connected mobile personal devices, such as laptops, smart pads, and smartphones. According to the white paper of Cisco [1], the number of mobile devices connected to the Internet was 12.5 billion in 2010 and

---

J. Moon (✉) · D. Kang · C. Bae  
Eelectronics and Telecommunications Research Institute, 218 Gajeong-ro,  
Yuseong-gu, Daejeon 305-700, South Korea  
e-mail: jymoon@etri.re.kr

D. Kang  
e-mail: dongoh@etri.re.kr

C. Bae  
e-mail: csbae@etri.re.kr

will be 25 billion by 2015. In addition, the number of connected devices per person was 1.84 and will be 3.47 by 2015. Because about 2 billion people among world population use the Internet actually, the real number of connected devices per person can be regarded as 6.25 in 2010 instead of 1.84. Therefore people need a simple and convenient way to share resources of devices owned by them or by their family and friends without explicit setting.

The human relationships are disclosing online through Social Networking Services (SNS) [2], such as Twitter, Facebook, and MySpace. Now the 65 % of adult Internet users use a social networking service, which increased dramatically compared to 5 % of adult Internet users in 2005 according to the statistics of SNS usage in [3]. The prevalence on SNS usage has speeded up online human relationships. The categorized human relationships and list of friends included in each relationship group can be retrieved by open Application Programming Interfaces (APIs) provided by commercial SNSs. If there are no explicit relationship groups provided by the SNSs, affinity-based group can be generated by analyzing SNS activities between friends, such as replying, commenting or representing emotions on a post, photo, or video, and sending and receiving private messages. Therefore, the SNS can be the feasible source for obtaining human relationships.

The purpose of our study is to build and manage social relationships between the personal devices with zero-configuration of device interaction by either extracting human relationship groups from commercial SNSs or by inferring human relationship groups from the history of device interactions [4], as shown in Fig. 1. The social relationship between devices is called device sociality and enables the devices to interact each other without manual setting of their device resources. To store and manage data required for the zero-configured device interaction, we design a device sociality database including specification of each device, human relationships between device owners, ownerships of devices, access permission of device resources.

**Fig. 1** Concept of device sociality on the basis of human sociality extracted from commercial SNS



## 2 ER Modeling

We generate an Entity Relationship (ER) model for the device sociality by using the ER modeling, which is the primary method for database design. The ER model includes entities with their attributes and relationships between the entities optionally with their attributes [5]. The model is made up of the part for device specification and the part for device sociality on the basis of human relationships. Each part is shown by an ER diagram.

Figure 2 shows the ER diagram describing the part for device specification. The ER model for the device specification describes device resources, which are a hardware device, an operating system, and other components including a CPU, storage, display, input interface, power, and sensors.

In the ER model, the device entity has two many-to-one relationships with device hardware and operating system entities because a device has one device hardware and one operating system and device hardware and an operating system can be employed to multiple devices. In addition, the device entity has a one-to-many relationship with the location log entity because a device can file zero or many location logs. The device entity has two dependent relationships with data storage and service entities because a device can share its data storage or provides its service for device interaction.

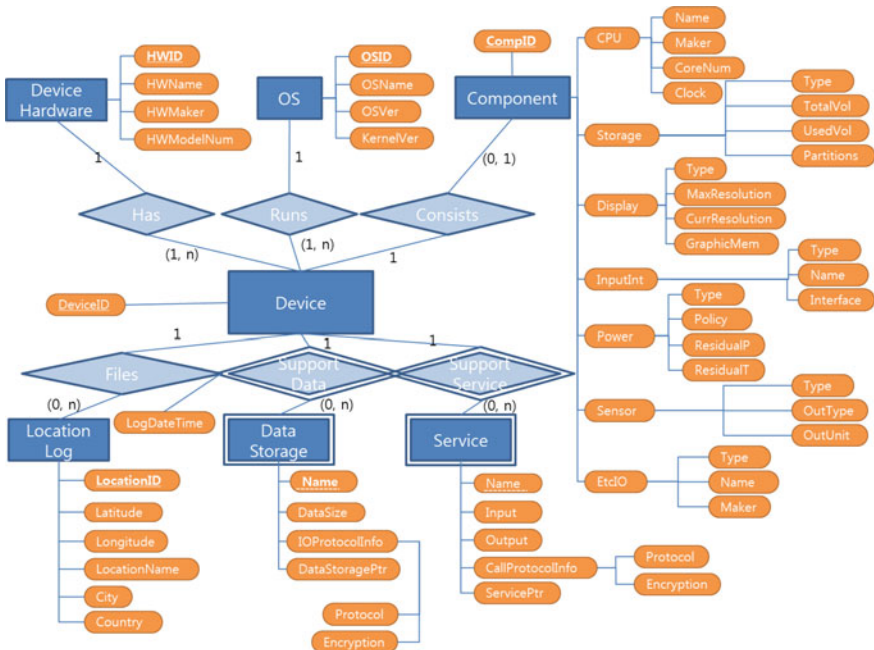
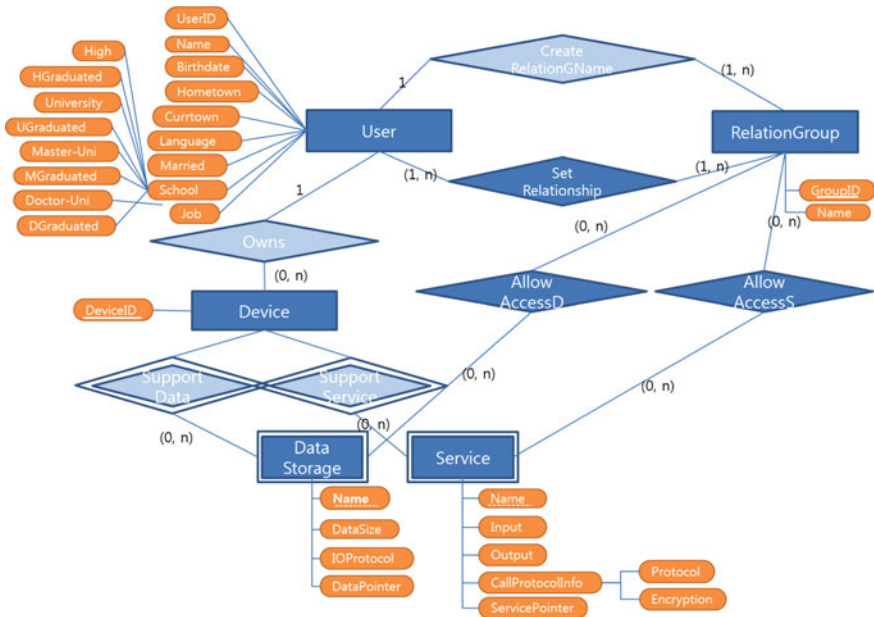


Fig. 2 ER diagram of device sociality design about device specification





**Fig. 3** ER diagram of device sociality design about device relationships

Figure 3 shows the ER diagram for the part for device sociality on the basis of human sociality. In the ER model, the user entity describes all the data related to users, such as name, birthdate, hometown, main language, schools, and current job. The user entity has a one-to-many relationship with the relation group entity. A relation group corresponds to either an extracted human relationship groups from a commercial SNS or a self-group for the user. The user entity has a many-to-many relationship with the relation group entity. The set relationship lists all the users included in each relation groups. In the ER model, the relation group can have access permission to shared resources like data storages or services. The relation group entity has many-to-many relationships with the data storage and service entities because a data storage or service can be allowed to be accessed by multiple relation groups and a relationship group can have access permission to multiple data storages and services.

### 3 Schema Design and Database Implementation

By using the proposed ER model, we obtained a schema design for device sociality database according to mapping rules of ER modeling [6]. All the entities were mapped into corresponding tables in the database schema. The attributes of an entity were mapped into columns of a corresponding table. An identifying attribute

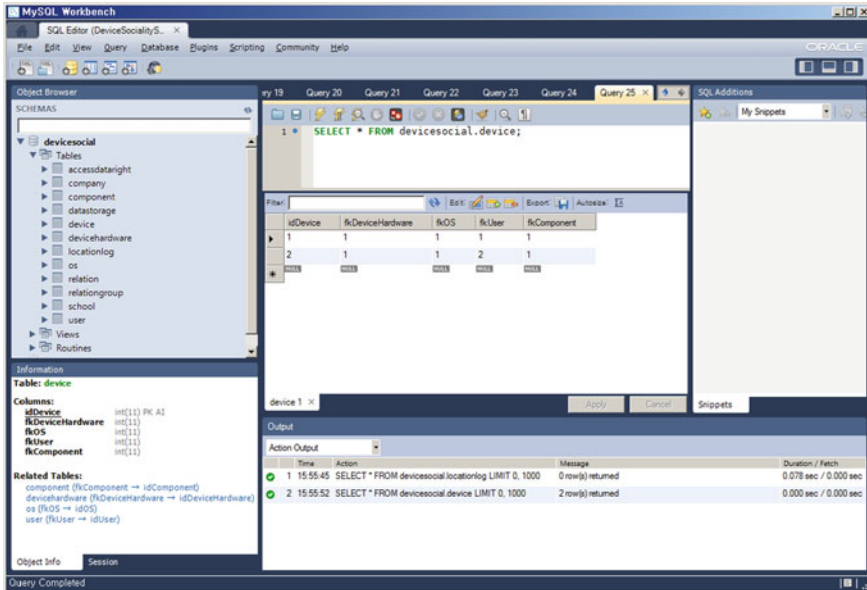


Fig. 4 Screenshot of database implementation in a commercial database system

of an entity was mapped to the primary key of the table. A one-to-many relationship between two entities was represented by a foreign key from one table, which is the primary key of the other table. A binary many-to-many relationship between two entities is mapped into a table with two foreign keys from tables corresponding to the entities.

After build the database schema from the proposed ER models, we implemented the database for device sociality by using one of commercial relational database systems, as shown in Fig. 4. By inserting data into the implemented database, we validated the proposed design for device sociality.

## 4 Conclusion

Because people connected to the Internet are nowadays using more than six mobile connected personal devices, they need a simple and easy way to share multiple devices owned by them or by their family and friends with zero-configuration. Our research aims at building and managing social relationships of personal devices by using human relationships obtained by open APIs from commercial SNSs. Therefore, this paper proposes the design of device sociality database by using ER modeling, which is a critical step to store and manage data for zero-configured device interaction.

The proposed ER model for device sociality includes the part for device ownership, human relationships, and access permission of resources as well as the part for the device specification. To validate the proposed ER model, we obtained the database schema by using the mapping rules and implemented the database in a commercial database system.

For the future work, we are going to collect and analyze real data for device sociality database from device sociality management servers and associated devices.

**Acknowledgments** This work was supported by the IT R&D program of MKE/KEIT (K10041801, Zero Configuration Type Device Interaction Technology using Device Sociality between Heterogeneous Devices).

## References

1. Dave E (2011) The internet of things: how the next evolution of the internet is changing everything. White paper, CISCO Internet Business Solutions Group (IBSG), April 2011
2. Wikipedia, Social Networking Service, [http://en.wikipedia.org/wiki/Social\\_networking\\_service](http://en.wikipedia.org/wiki/Social_networking_service)
3. Mary M, Kathryn Z (2011) 65% of online adults use social networking sites. The Pew Research Center's Internet & American Life Project surveys, 26 Aug 2011
4. Kang K, Kang D, Bae C (2013) Novel approach of device collaboration based on device social network, consumer electronics (ICCE). In: Proceedings of IEEE international conference on 11–14 Jan 2013, pp 248–249
5. Wikipedia, Entity-relationship model, [http://en.wikipedia.org/wiki/Entity-relationship\\_model](http://en.wikipedia.org/wiki/Entity-relationship_model)
6. Elmasri R, Navathe S (2010) Fundamentals of database systems, 6th edn. Addison Wesley, Boston

# Image Processing Based a Wireless Charging System with Two Mobile Robots

Jae-O Kim, Chan-Woo Moon and Hyun-Sik Ahn

**Abstract** This paper presents the image processing algorithm for wireless charging between each mobile robot. The image processing algorithm converts Red Green Blue (RGB) format of inputted image to detect edge. It calculates a specific area using Hough Transformation (HT) in detected edge and judges correct charging antenna using Speeded-Up Robust Features (SURF). Accordingly, the image processing algorithm can control position and direction of mobile robot and antenna for wireless charging. The image processing algorithm is implemented wireless charging systems, which are set up on each two mobile robot and it is verified with experiment.

**Keywords** Wireless charging · HT · SURF · Mobile robot · Color edge

## 1 Introduction

A typical mobile robot has actuators which are operated by battery. As the robot moves for a long time, the battery is exhausted, and recharging is needed. Currently, automatic recharging is general trend for an intelligent robot. And usually, contact type of recharging with a guide mechanism is widely used. But in a circumstance on which guide mechanism is unavailable, for example, recharging between different type of robots, recharging outdoor and recharging when a robot is still moving, a wireless power transmission can be considered.

---

J.-O. Kim · C.-W. Moon (✉) · H.-S. Ahn  
Department of Electronics Engineering, Kookmin University, Seoul, Korea  
e-mail: mcwnt@kookmin.ac.kr

J.-O. Kim  
e-mail: futurejo@paran.com

H.-S. Ahn  
e-mail: ahs@kookmin.ac.kr

In this paper, wireless power transmission method for two mobile robots is investigated. To obtain the maximum efficiency of power transmission, vision based position control of a transmission antenna is implemented. To recognize the antenna, Color Edge, Hough Transform and Speeded-Up Robust Features (SURF) methods are used. The information of pose of antenna is used to control the mobile robot.

## 2 Inductive Power Transmission

The basic principle of an inductively coupled power transfer system consist of a transmitter coil and a receiver coil. Both coils form a system of magnetically coupled inductors. An alternating current in the transmitter coil generates a magnetic field which induces a voltage in the receiver coil. This voltage can be used to power a mobile device or charge a battery. In this research, near range direct induction method is used.

## 3 Color Edge Detection

Common image processing is the processing of the input image is converted to grayscale and binary coded. At this time, a lot of ambient lighting, and the impact on the environment, it is difficult to extract the straight edge. As a way to solve this problem, changing the format of the input images from RGB to HSV is widely used, which changes only the size of the color difference image after edge detection.

## 4 Hough Transform (HT)

Hough Transform is widely used for searching a figure such as line and circle in an image pixel data. Equation 1 denotes a basic equation of line and Eq. 2 is another formulation of line which is represented in the parameter space. If a line becomes vertical, the slope becomes infinity, then, Eq. 3 is used alternatively [2, 5].

$$y_i = ax_i + b \quad (1)$$

$$b = -x_i a + y_i \quad (2)$$

$$x \cos \theta + y \sin \theta = \rho \quad (3)$$

### 5 Speeded-Up Robust Features (SURF)

SURF is faster than the video from one of the algorithm to find the feature points invariant to scale, lighting, point to changes in the environment such as SIFT and similarity, however, faster than SIFT when compared with key points finds. We base our detector on the Hessian matrix because of its good performance in computation time and accuracy. However, rather than using a different measure for selecting the location and the scale (as was done in the Hessian-Laplace detector), we rely on the determinant of the Hessian for both. Given a point  $x = (x, y)$  in an image  $I$ , the Hessian matrix  $H(x, \sigma)$  in  $x$  at scale.  $\sigma$  is defined as follows

$$H(x, \sigma) = \begin{bmatrix} L_{xx}(x, \sigma) & L_{xy}(x, \sigma) \\ L_{xy}(x, \sigma) & L_{yy}(x, \sigma) \end{bmatrix} \tag{4}$$

### 6 The Structure of Experimental System

Figure 1 shows block diagram of the system used in the experiment.

The image processing algorithm for wireless charging converts inputted data format of camera. Converted image is used to detect edge by difference of color signal of image. Position of Antenna is judged by HT method and image is cropped. Cropped image is confirmed by SURF and KNN for alignment with target. The wireless charging control system decides target and direction using this process. And it is implemented on mobile robot task for motion control of mobile robot.

### 7 Experimental Result

Experimental environment consists of a mobile robot for wireless charging, a mobile robot for transmission power, test antenna for wireless charging and main control program. It is as shown in Fig. 2.

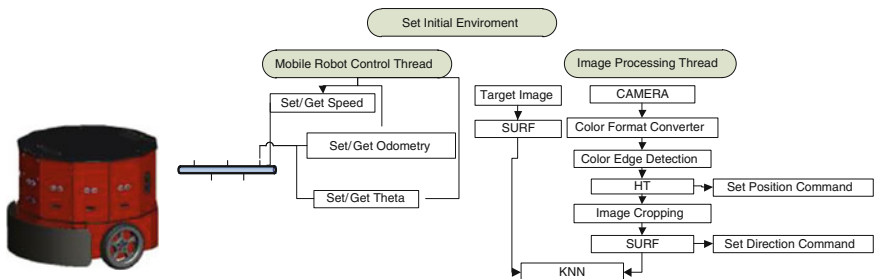


Fig. 1 Block diagram



Fig. 2 Wireless charging system and mobile robot and main control program

Figure 3 is typed as it seems the main routine for processing the data.

Experimental result of edge detection in inputted image using color format is as shown in Fig. 3. Compare image with result of edge detection of binary image is as shown in Fig. 4.

The system detects straight line in input image using HT method, extracts feature and compares with target. Position of antenna is confirmed by this process. It is shown by Fig. 5.

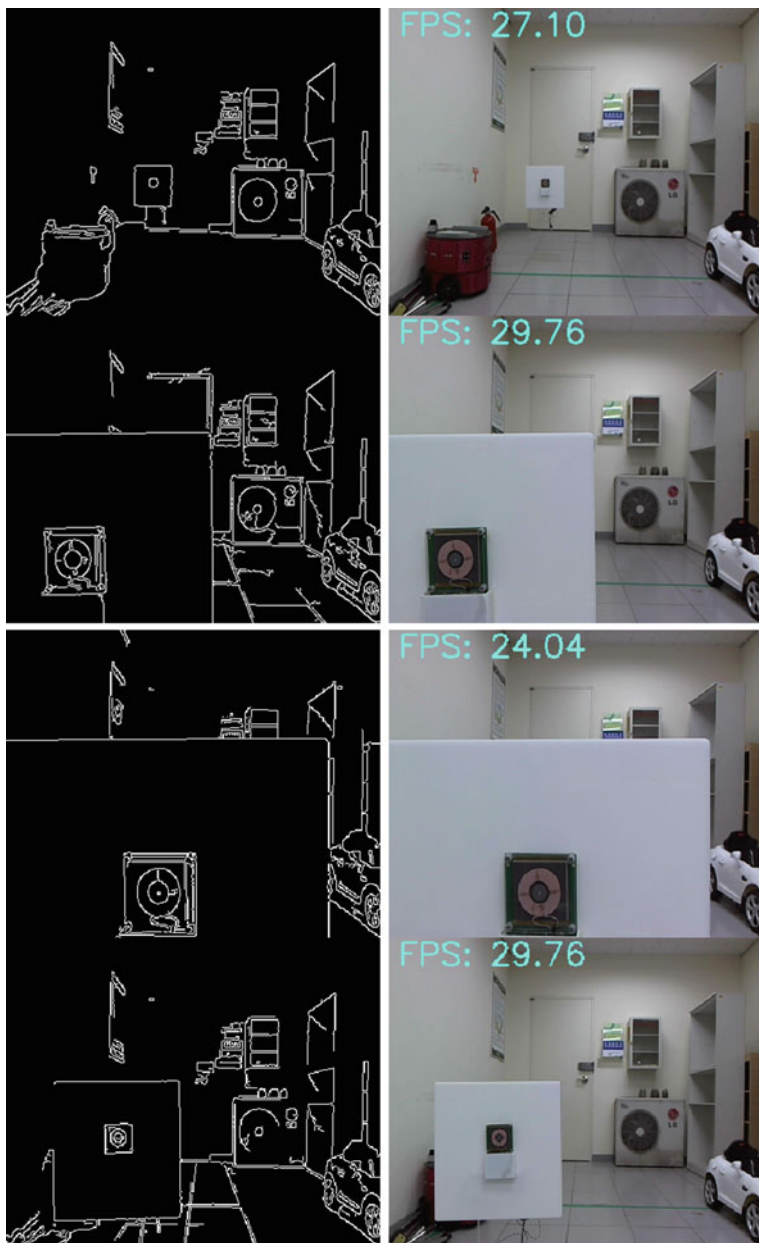


Fig. 3 Converted image and captured image





Fig. 4 The result of edge detection image used grayscale and binary data

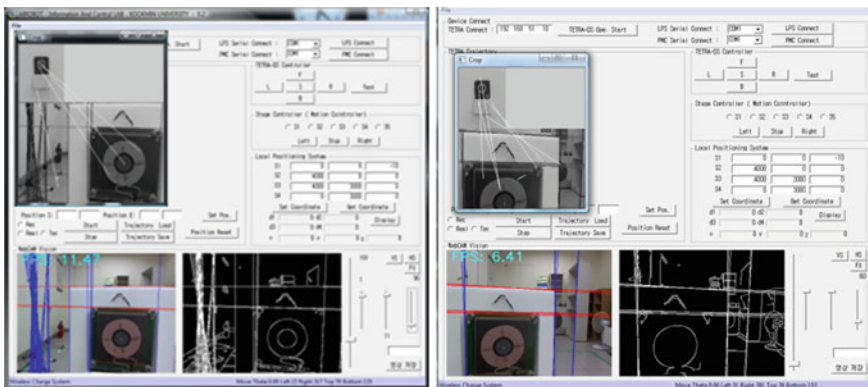


Fig. 5 The detected image

### 8 Conclusion

In this paper, we implemented a wireless charging system between two robots using image processing. Image processing algorithm is implemented with color edge method to detect position and direction of antenna, HT method for extraction of an interested area and SURF to judge correct feature of antenna. Thus wireless charging system could control position and direction of mobile robot.

### 9 Summary

A wireless charging system between two mobile robots using image processing methods in which difference of color for edge detection, calculation of antenna position and direction using HT, SURF is used is implemented.

**Acknowledgments** This research was supported by the The Ministry of Knowledge Economy (MKE), Korea, under the Information Technology Research Center (ITRC) support program (NIPA-2012-H0301-12-2007) supervised by the National IT Industry Promotion Agency (NIPA).

## References

1. Wireless Power Consortium. <http://www.wirelesspowerconsortium.com>
2. Hough PVC (1962) Methods and means for recognizing complex patterns. US Patent 3,069,654
3. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) SURF speeded up robust features. *Comput Vis Imag Underst* 110(3):346–359
4. TETRA-DS DasaRobot, DasaRobot (2009)
5. Kim J-O, Rho S, Moon C-W, Ahn H-S (2012) Imaging processing based a wireless charging system with a mobile robot. *Computer applications for database, education, and ubiquitous computing. Communications in computer and information science*, vol 352. Springer, Heidelberg, pp 298–301

# Design of a Reliable In-Vehicle Network Using ZigBee Communication

Sunny Ro, Kyung-Jung Lee and Hyun-Sik Ahn

**Abstract** This paper presents a new configuration for in-vehicle networks to increase the reliability of the communication between electronic control units (ECU) and to improve the safety level of a vehicle. Basically, the CAN (Controller Area Network) protocol is assumed for the data communication in vehicles but more reliable communication can be guaranteed by adding the ZigBee communication function to each ECU. To show the validity and the performance of the presented network configuration for some network faults, the Electronic Stability Control (ESC) operation is analyzed by using an ECU-In-the-Loop Simulation (EILS). The experimental set-up for EILS of ESC system consists of two 32-bit microcontroller boards which can be communicated with the CAN or the ZigBee protocol. A 7-DOF (Degrees Of Freedom) vehicle model and ESC algorithm is implemented on each microcontroller. It is shown by the experimental results that ESC using the high reliability CAN system can achieve the same performance as using only CAN protocol without disconnected CAN bus.

**Keywords** In-vehicle network · Controller area network · Reliability · Fault tolerance · Electronic stability control

---

An erratum to this chapter is available at [10.1007/978-94-007-6738-6\\_148](https://doi.org/10.1007/978-94-007-6738-6_148)

---

S. Ro · K.-J. Lee · H.-S. Ahn (✉)  
Department of Electronics Engineering, Kookmin University, Jeongneung-dong,  
Seongbuk-gu, Seoul, Korea  
e-mail: ahs@kookmin.ac.kr

S. Ro  
e-mail: sunyda88@nate.com

K.-J. Lee  
e-mail: streizin@nate.com

## 1 Introduction

To enhance the handling performance and the safety of vehicles, many active chassis control systems to ensure the vehicle stability have been consistently studied. ESC is a stability enhancement system designed to electronically detect and assist drivers in critical driving situation and under adverse conditions automatically [1].

In addition, ECUs are distributed throughout the vehicle to perform a variety of different vehicle functions. Accordingly, In order to transmit data between distributed ECUs and software correctly, safely, performance improvement in in-vehicle network is essentially needed [2]. Controller Area Network (CAN) is an asynchronous serial communication protocol which follows ISO 11898 standards and is widely accepted in automobiles due to its real time performance, reliability and compatibility with wide range of devices. The main features of CAN protocol are high-speed data transmission up to 1 Mbps, bus access control depending on a multi-master principle, and bus off function in the event of transmission abnormalities [3]. However, if a critical fault occurs in CAN (e.g. disconnection), ECU must transfer the data to another ECU by replacement of CAN protocol.

In this paper, we present the new method for improvement reliability of CAN for in-vehicle network. To improve reliability, CAN is replaced with ZigBee when fault is generated by a disconnection of CAN bus. Also, in order to verify the proposed CAN system, the ESC operation is analyzed by using an EILS. The experimental set-up for EILS of ESC system consists of two 32-bit microcontroller boards which can be communicated with the CAN or the ZigBee protocol.

## 2 The High Reliability CAN System

Typically, CAN is used in a wide range of industrial automation and an important element in a protocol of distributed real-time control. The distributed real-time control functionalities have been studied for reliable communication network systems when separate ECUs are connected with each other through a CAN protocol [4]. To prevent a critical fault such as a disconnection, CAN protocol must be replaced with another protocol. Accordingly, we propose a fault-tolerant CAN controller system called the high reliability CAN system, which consists of CAN and ZigBee to avoid a fault such as a disconnection. This proposed CAN system for improvement of reliability is based on ZigBee to tolerate any single permanent fault in one CAN controller [5]. In this study, the proposed high reliability CAN system using ZigBee is as shown in Fig. 1.

When CAN is disconnected in this system, disconnected ECU transfers information on disconnected CAN bus to another ECUs by using ZigBee. Then, the high reliability CAN system operates as a normal system in fault. The process to prove the performance of this system is as follows [6, 7]. The transmission time  $C_m$

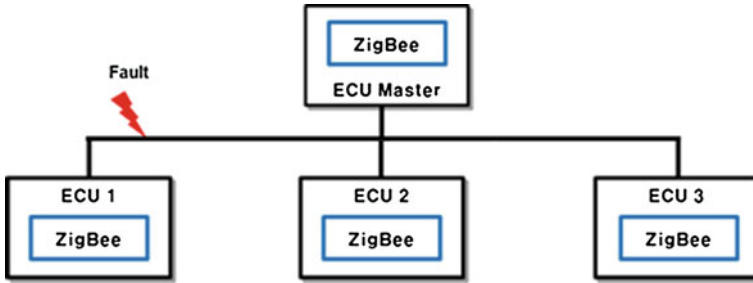


Fig. 1 The high reliability CAN network

of CAN messages containing  $s_m$  data byte, a 47 bit overhead per message and the transmission time for a single bit  $\tau_{bit}$  is given by:

$$C_m = (47 + 8s_m)\tau_{bit} \tag{1}$$

The transmission bit  $N_{CAN}$  for a single message of 8 byte CAN is given by:

$$N_{CAN} = (47 + 8s_m) = 111 \text{ bit} \tag{2}$$

If CAN transmits at 500 kbps, the number of transmission messages per second are 4505. Accordingly, the transmission period of CAN  $T_{CAN\_P}$  is given by:

$$T_{CAN\_P} = \frac{N_{CAN}}{4505} = 220 \text{ us} \tag{3}$$

In the same Eqs. (1) and (2), the transmission bit  $N_{ZigBee}$  for a single message of 1 byte ZigBee is given by:

$$N_{ZigBee} = (208 + 8s_m) = 216 \text{ bit} \tag{4}$$

At this time, ZigBee transmits at 250 kbps and the number of transmission messages per second is 1157. The transmission period of ZigBee  $T_{ZigBee\_P}$  is given by:

$$T_{ZigBee\_P} = \frac{N_{CAN}}{1157} = 864 \text{ us} \tag{5}$$

Considering the data length of 8 byte CAN, the control period multiplies  $T_{ZigBee\_P}$  by 8. And then, the control period  $T_{ZigBee\_P}$  is 6.91 ms.

We implement the high reliability CAN system in ESC for reliability verification. ESC is generally controlled by the control period at 10 ms because the bandwidth is considered for control, self-diagnosis and so on of a particular vehicle. The transmission period should not affect to the performance of control because the transmission messages are completed within 10 ms. In the two control period from Eqs. (3) and (5),  $T_{ZigBee\_P}$  and  $T_{CAN\_P}$  are included in the control period of ESC. Therefore, the high reliability CAN system is predicted to show a performance as well without a fault.

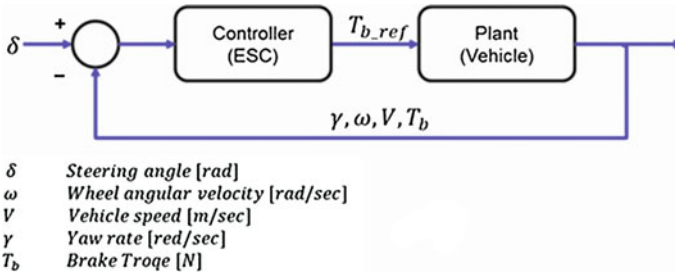


Fig. 2 The control scheme of ESC

### 3 Experimental Environment and Results

The experimental environment for EILS of ESC system consists of two 32-bit microcontroller connected by CAN protocol. ESC system consists of the controller and the plant as shown in Fig. 2. A 7-DOF vehicle model applied to Plant is developed to obtain the longitudinal, lateral and yaw motions of vehicle dynamics and the other four degrees of motion representing 4 wheel dynamics [8, 9].

In this paper, it is assumed that a fault occurs when CAN to transfer braking reference torque  $T_{b\_ref}$  from ESC to a vehicle model. In each microcontroller, a 7-DOF vehicle model and ESC algorithm are respectively implemented on ECU. Also, CAN monitoring device is used to confirm EILS as shown in Fig. 3.

When the vehicle drives at 80 km/h, the driver rapidly changes the lane as shown in Fig. 4. The behavior of vehicle by the steering input is shown in Figs. 5 and 6. The vehicle model without CAN fault similarly tracks the reference yaw rate. However, the vehicle model with CAN fault appears unstable condition at 2 s by compared with the reference yaw rate as shown in Fig. 5a and b.

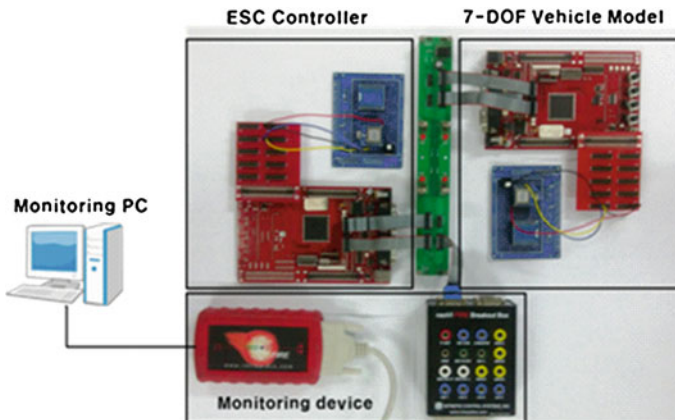


Fig. 3 EILS environment for ESC system

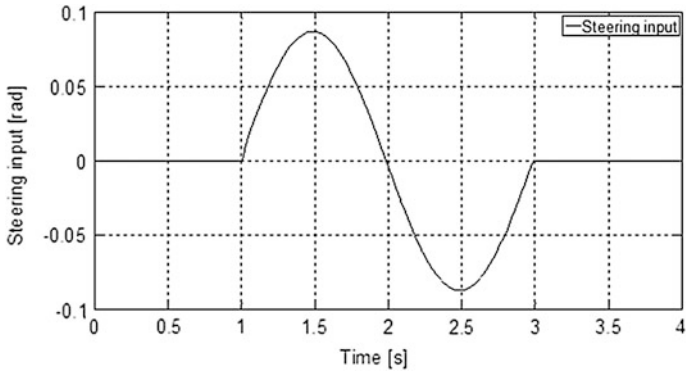


Fig. 4 The steering input for single lane change

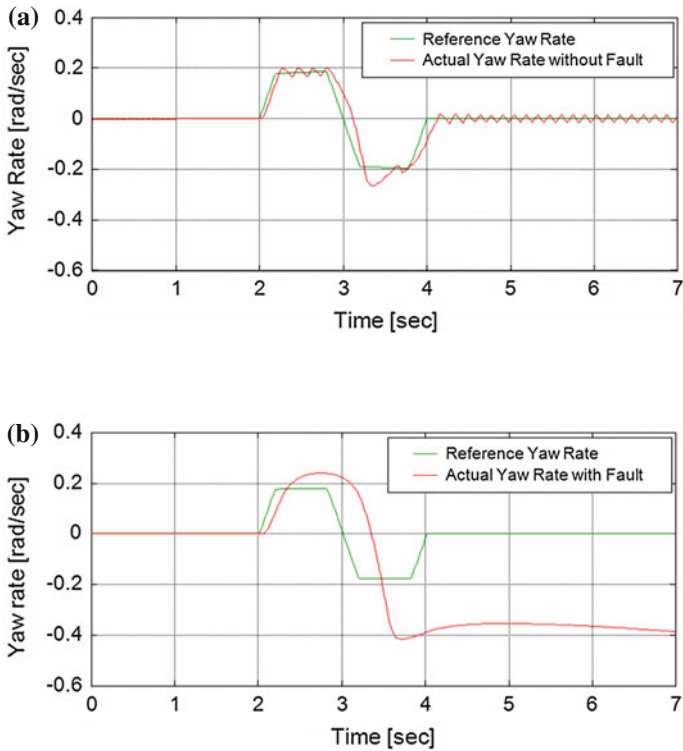
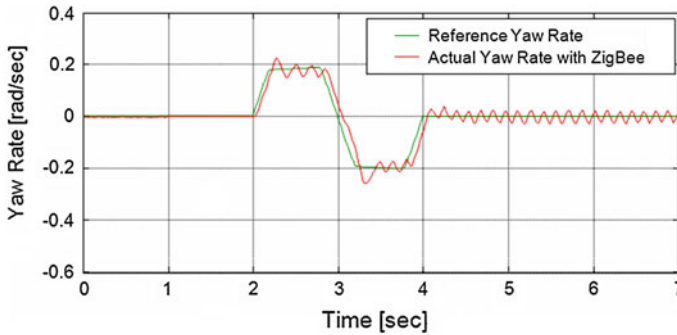


Fig. 5 EILS results with and without CAN fault. **a** Yaw rate in EILS without fault. **b** Yaw rate in EILS with fault

The experiment assumes that ESC system has a fault. The experimental result of the proposed high reliability CAN system using ZigBee is similar to the reference yaw rate as shown in Fig. 6.



**Fig. 6** EILS results of high reliability CAN system using ZigBee

Therefore, the performance of the high reliability CAN system to tolerate CAN fault is verified through the experimentation of EILS. Also, it is confirmed that all transmission messages are exactly processed by the calculated  $T_{CAN\_P}$  and  $T_{ZigBee\_P}$  within the control period of ESC.

## 4 Conclusion

In this paper, we proposed the high reliability CAN system based on ZigBee for tolerance of a CAN fault in ESC system. The main objects of this system were avoidance of a critical CAN fault and the performance as ESC system without a CAN fault. The efficiency of the high reliability CAN system was inferred from the transmission period calculation formula of CAN and ZigBee, and verified by the experiment of EILS of ESC system. EILS consists of a 7-DOF vehicle model and ESC algorithm. The experiment of EILS had been implemented on two 32-bit microcontroller and verified the performance of ESC system when ESC occurred a CAN fault. The experimental result of the proposed high reliability CAN system only using ZigBee was shown as a without CAN fault.

## References

1. Neuhaus D, Willms J (2005) Vehicle dynamics-continuous improvements in vehicle safety from abs to electronic stability control. SAE Technical paper 2005-26-065, pp 729–736
2. Chaavan K, Leserf P (2009) Simulation of a steer-by-wire system using flexray-based ecu network. International conference on ACTEA, pp 21–26
3. Chen H, Tian J (2009) Research on the controller area network. International conference on networking and digital society, vol 2, pp 251–254
4. Palai D (2012) Design methods to optimize the performance of controller area network. SAE Technical paper 2012-01-0194, pp 1–11



5. Guerrero C, Rodriguez-Navas G, Proenza J (2002) Hardware support for fault tolerance in triple redundant can controllers. *International conference on electronics, circuits and systems*, vol 2, pp 457–460
6. Mary GI, Alex ZC, Jenkins L (2012) Response time analysis of messages in controller area network: a review. *J Comput Netw Commun*, vol 2013, Article ID 14805
7. Johnstone MN, Jarvis JA (2011) Penetration of zigbee-based wireless sensor networks. *The 12th Australian information warfare and security conference*, pp 16–23
8. Zhao C, Xiang W, Richardson P (2006) Vehicle lateral control and yaw stability control through differential braking. *IEEE international symposium on industrial electronics*, vol 5, pp 384–389
9. Zhao C, Xiang W, Richardson P (2011) Monitoring system design for lateral vehicle motion. *J IEEE Trans Vehicular Technol* 6(4):1394–4103

# Wireless Positioning Techniques and Location-Based Services: A Literature Review

Pantea Keikhosrokiani, Norlia Mustafa, Nasriah Zakaria  
and Muhammad Imran Sarwar

**Abstract** With advent of satellite positioning system and availability of wireless communication network, it is possible for an end user to navigate even in the scarce location where there are fewer inhabitants. With affordable cost and vast coverage, millions of users can access location co-ordinates from any part of the world due to wireless positioning techniques. In this study, we highlight different positioning methods, location-based services and vast variety of applications benefited from these methods and services. This paper covers brief mathematical models used among all the wireless positioning systems along with their comparison. In today's fast pace information era, location-based services are not only used for hotspot navigation but, also used for marketing strategy and so on. In addition, this article includes location-based services that access mobile network, and utilized the current location of the mobile device appropriately. Finally we classify the location-based solutions that have been used in variety of models such health services, marketing, tourism, entertainment and advertisement, and so forth. The study concludes that with evolution of technological advancement, wireless positioning system will be more improved and will be used in every part of our daily life in an effective manner.

**Keywords** Location positioning methods • Wireless technologies • Location-based services • Global positioning system (GPS) • Cellular network

---

P. Keikhosrokiani (✉) · N. Mustafa · N. Zakaria  
School of Computer Sciences, Universiti Sains Malaysia, Minden 11800, Penang, Malaysia  
e-mail: pantea.kia@ieee.org

N. Mustafa  
e-mail: norlia@cs.usm.my

N. Zakaria  
e-mail: nasriah@cs.usm.my

M. I. Sarwar  
National Advance IPv6 Centre (NAv6), Universiti Sains Malaysia, Penang, Malaysia  
e-mail: imrans@live.com.my

## 1 Introduction

Rapid development of wireless communication technology and wide usage of wireless networks had a strong effect on the possibility of location-based services. One of the fundamental elements of many applications such as e-commerce, emergency and medical, advertisement, navigation and other location-based services (LBS) is object location positioning. Different functionalities and events are required to approximate the location of a node of interest. These functionalities involve coordinates in two or three dimensions as well as some information such as latitude, longitude, and altitude of the nodes. These information are available everywhere, outside environment, inside building, in the airplane, in the sea, etc. but different methods along with various mathematical principles are needed to track the location in each environment [1]. Location-based services (LBS) refer to services provided based geographical position of mobile device. Such services consist of commerce, emergency and medical, advertisement, navigation and routing, social networking, finding friends and so forth. LBS require location positioning techniques such as satellite and wireless technologies as well as geographical information system in order to map the location of a node of interest [2–4]. The main purpose of this paper is to overview the main mathematical principles of location positioning, different location positioning methods as well as emerging class of location-based services. Such information assists us to be familiar with various location-based services and recognize the right positioning methods and technologies for each service.

To understand how wireless positioning techniques and location-based services work, this paper first introduces basic mathematical positioning methods and their comparison. Furthermore, this paper presents an overview of existing positioning solutions using different wireless communication systems along with their weaknesses and strengths. Different location-based services and their technology is illustrated in next section followed by concluding remarks.

## 2 Positioning Technologies and Methods

**Mathematical Positioning Methods.** There are three basic mathematical principles that support every location positioning technique used today. Some positioning methods such as angle of arrival (AOA), time of arrival (TOA), time difference of arrival (TDOA), and received signal strength (RSS) are based on geometric principles to calculate the position of a node. These principles consist of Triangulation, multilateration and hyperbolic used lines and angles in order to calculate the position. Table 1 illustrates these basic mathematical principles in detail [1, 5].

Table 2 illustrates mathematical formulation of basic mathematical principles that are used in position location. The table shows three different columns for

**Table 1** Basic mathematical principles for location positioning [1, 5]

Method	Basic Approach	Figure
Triangulation	<ul style="list-style-type: none"> <li>– Angles of two sides, <math>\varphi_1</math> and <math>\varphi_2</math> are calculated to get desired location when the distance of points is unknown</li> <li>– Importantly the desired point has to be intersection of two lines from two sides</li> </ul>	
Multilateration	<ul style="list-style-type: none"> <li>– An extension of Triangulation with three reference points</li> <li>– Three point of intersection will give calculated distance value from reference point to object T</li> </ul>	
Hyperbolic principle	<ul style="list-style-type: none"> <li>– It is a set of points that have constant difference values from two fixed points</li> <li>– Hyperbola's focus is represented by each point where focus is an anchor node or reference point</li> <li>– Position can be calculated when the target resides between two foci of hyperbola curve</li> <li>– Curve's distance to each hyperbola focus are fixed</li> </ul>	

Triangulation, Multilateration and Hyperbolic principle. The purpose of the table is to explain the mathematical concept of getting final formulae used to calculate the location based positioning system.

**Types of Location positioning.** Location tracking of the handset devices becomes important among various fields such as advertisement, healthcare, tourism, navigation and routing, entertainment, observation and so forth. There are different ways and techniques in order to calculate the location of devices. Location positioning system consists of several components such as hardware, measuring unit and signal transmitter. Wireless location positioning systems can be classified based on the functionality of these components and their interactions. Hence, positioning systems are categorized into three groups: (1) handset-based positioning, (2) network-based positioning, and (3) hybrid-positioning system [1, 5]. In handset-based positioning, the handset calculates its own location while

**Table 2.** The formulation of three mathematical principle in position location

Triangulation	Multilateration	Hyperbolic
<p>The intersection point T, we get the 2 angles <math>\phi_1</math> and <math>\phi_2</math></p> $R = \frac{d}{\tan \phi_1 + \tan \phi_2}$ <p>R is line between reference point N1 and N2</p> $d = \frac{R \cdot \sin \phi_1 \sin \phi_2}{\sin(\phi_1 + \phi_2)}$ <p>d is perpendicular line between target point T and line R</p> $\phi_1 = \tan^{-1} \left( \frac{h-Y}{g-X} \right)$ $\phi_2 = \tan^{-1} \left( \frac{b-Y}{a-X} \right)$ <p>The coordinates of the target (X, Y) can be calculated as mentioned below</p> $Y = x \tan \phi_2 + (b - a \tan \phi_1)$ $X = \frac{b - h - a \tan \phi_2 + g \tan \phi_1}{\tan \phi_1 - \tan \phi_2}$ <p>distances between reference points N1 and N2, and the target point T</p> $d_1 = \ g - X\  = \sqrt{(g - X)^2 - (h - Y)^2}$ $d_2 = \ a - X\  = \sqrt{(a - X)^2 - (b - Y)^2}$	<p>To calculate the coordinates of the target T, first, the distances between reference nodes and the target T</p> $d_1 = (t_1 - t_0) \cdot c$ $d_2 = (t_2 - t_0) \cdot c$ $d_3 = (t_3 - t_0) \cdot c$ <p>where c is the speed of light, <math>t_0</math> is time of a signal sent from T, <math>d_1</math> is distance between <math>N_1</math> and T, <math>d_2</math> distance between <math>N_2</math> and T, <math>d_3</math> distance between <math>N_3</math> and T, <math>t_1</math> time of arrival of signal T to <math>N_1</math>, <math>t_2</math> time of arrival of signal T to <math>N_2</math>, <math>t_3</math> time of arrival of signal T to <math>N_3</math></p> <p>Equation of 3 intersecting circle with centers at the reference point</p> $d_1^2 = x^2 + y^2$ $d_2^2 = (x - x_2)^2 + y^2$ $d_3^2 = (x - x_3)^2 + (y - y_3)^2$ <p>Solving the above equations:</p> $x = \frac{2x_2}{x_2^2 + d_1^2 - d_2^2}$ $y = \frac{2x_3}{x_3^2 + y_3^2 + d_1^2 - d_3^2 - 2x_2x_3}$	<p>The equations of hyperbola are:</p> $1 = \frac{x^2}{a^2} - \frac{y^2}{b^2}$ $a^2 = \left(\frac{\Delta d}{2}\right)^2$ $b^2 = \left(\frac{D}{2}\right)^2 - a^2$ <p>Where a and b can be obtained from quantities d and D</p> $\Delta d = d_2 - d_1 = c(t_1 - t_1)$ <p>Where,</p> <p>c is the speed of light</p> <p><math>t_1</math> is the time of node <math>N_1</math></p> <p><math>t_2</math> is the time of node <math>N_2</math></p>

in network-based positioning the network calculates the handset's location. In hybrid-positioning method, there is collaboration between the network and handset in order to measure and calculate the device's position. GPS is one of the examples of handset-based positioning in which position estimation will be done by handset and GPS based on signals received from at least four satellites [6]. There are some examples for network-based positioning systems such as the cellular networks and Airborne Early Warning and Control System (AWACS) as stated in [1, 7]. Lastly, Assisted GPS(A-GPS) is a good example for hybrid-positioning system [8]. The most important concept of positioning technology is locating users in outdoor and indoor environment and it can be divided into three categories as mentioned above. The fundamental attributes of those approaches are reviewed in Table 3.

Each method illustrated in Table 3 has some weaknesses and strengths; thus, they must be used in the proper situation. For instance, from handset-based category, Global Positioning System (GPS) is appropriate to use for outdoor environment and it does not have indoor services. GPS coverage is poor in urban canyons, it has delay in calculating the location, and a modern handset along with power is required. On the other hand, GPS does not required new network infrastructures, and it is accurate with improved privacy for the user. The next method of handset-based positioning is Enhanced Observed Time Difference (E-OTD) that has enhanced privacy; whereas, some modification must be done in handset and network investment is needed. The next method is Forward Link Triangulation (FLT) that decrease complexity and cost for handset but same as E-OTD some handset modification and network investment is needed. FLT consists of two categories of Advanced Forward Link Trilateration (A-FLT) and Enhanced Forward Link Trilateration (E-FLT). These two categories have various accuracies as shown in Table 3. The first method from network-based category is Cell-ID, Cell of Origin (COO). COO is available now, no handset modification is required for this method and the cost is lower in compare with other methods. COO has lower accuracy in compare with other positioning methods especially in rural cells. Another weakness of COO is low privacy for users. Time of Arrival (TOA) and Uplink Time Difference of Arrival (U-TDOA) have better accuracy in compare with COO and in addition to position, it can determine velocity and heading. Moreover, TOA does not need any modification in handset while TDOA requires some handset modification. TDOA has lower accuracy for TDOA in analog and narrowband digital systems. One of the weaknesses of TOA and TDOA is new equipment is required for base stations. In addition, they have less privacy for users. Angle of Arrival (AOA), another network-based positioning method needs some special equipments for base stations such as special antennas and receivers. It has low privacy for users same as TOA, TDOA and COO but it does not require any handset modification. Received Signal Strength (RSS) is competitive in terms of simplicity and cost in compare with other methods. It is valuable to merge different positioning methods. For instance, combining AOA with RSS concludes better accuracy in compare with using one of the methods alone. Nevertheless, combination methods will increase cost of the network infrastructure. Fingerprint method overcomes many problems by using RSS at the

**Table 3** Positioning technologies for location-based services [1, 5]

Type	Positioning Method	Basic Approach	Technology	Accuracy
Handset-Based	Global Positioning System (GPS)	Triangulation method by using timing signals from at least 4 satellites	Satellite	50–100 m
	Enhanced Observed Time Difference (E-OTD)	Triangulation calculation is used to determine location	Cellular Network	60–200 m
	Advanced Forward Link Trilateration (A-FLT)	Measure the time difference of signals from nearby cellular base stations (BS) to triangulate location	Cellular Network	50–200 m
	Enhanced Forward Link Trilateration (E-FLT)	Existing pilot signal measurement message (PSMM) is used from mobile device to BS	Cellular Network	250–300 m
Network-Based	Cell-ID, Cell of Origin (COO)	Location of base station is used to illustrate subscribers location	Cellular Network	10–35 km
	Time of Arrival (TOA)	Uses timing of signals sent by mobile device to triangulate the location	Cellular Network	100–400 m
	Uplink Time Difference of Arrival (U-TDOA)	Uses differences in arrival time between the received signals to identify the location	Cellular Network	50–150 m
	Angle of Arrival (AOA)	It is based on the angle of the received signal of a mobile device into two or more base stations	Cellular Network, WLAN	50–150 m
Hybrid	Received Signal Strength (RSS)	The energy of the received signal at one end is used to estimate the distance between two nodes	Cellular Network, WLAN	
	(Multipath-) Fingerprint	Measured fingerprints at the existing position location of the nodes will be compared with the fingerprints of diverse positioning locations that are stored in a database	Cellular Network, WLAN	
	Timing Advance (TA)	The length of time a signal takes to reach the base station from a mobile phone	GSM	100–550 m
	Assisted Global Positioning System (A-GPS)	GPS receivers are embedded in the cellular network which assist a partial GPS receiver in the handset, reducing the calculation burden	GPS Satellite, Cellular Network	3–20 m

modeling location. Fingerprint method does not require any handset modification but some receiving equipment is needed for base stations. Furthermore, updating and development of database is required in fingerprint and users have less privacy. Assisted Global Positioning System (A-GPS) has some strength in compare with GPS. For example, it reduced the cost imposed by GPS handset and the handset can be smaller with better battery life. Moreover, A-GPS reduced delay in calculating the location. The only problem of A-GPS is new handset requirement as well as indoor positioning accuracy.

### **3 Location-Based Services**

Location-based services can be defined as services that can be accessed by mobile network, and utilized the current location of the mobile device appropriately. Many industries used GPS to enhance their products and services such as automotive industries that used navigation systems for their produced cars. Location-based services assist user to access to the information regarding to the current geographic area of the user [2–4]. Additionally, location-based services make possible two way communication and interaction between customers and businesses. In this way, users will get information according to their needs and requirements. Location based services are a combination of information and telecommunication technologies including Web GIS, Mobile GIS, Mobile Internet, Spatial Database, Internet and mobile devices [9]. Generally, location based services are composed of some components: a mobile device, a communication network, a positioning component, a service and application provider, and a data and content provider.

### **4 Classification of Location-Based Services**

Nowadays, a wide range of services has been offered by relying on user's location information. By accessing to different types of geographical information services (GIS), the location information can be provided simply. As mentioned before, there are several ways to exploit the location in order to provide new services. It will be more effective when the location information merged with other user profile information to offer new location-based services [10]. There are many classifications for location-based services. For instance, Levijoki (2000) categorized location-based services into billing, safety, information, tracking and proximity awareness [11]. Moreover, Kar and Bouwman (2001) grouped location-based services into different services such as information, entertainment, communication, transaction, mobile office and business process support [12]. Additionally, the classification that has been done by Steinfield et al. (2004) is: Emergency, Safety and Medical/Health, Information, Navigation/Routing,



Transactions and Billing, Asset Tracking and Fleet Management, Mobile Office, Entertainment, Proximity Services [13]. After reviewing different location-based services, we classified it into Emergency/Medical/Health Services, Tourism, Navigation/routing/Tracking, Proximity Services, E-Commerce, Vehicular Services, Entertainment, and Advertisement as shown in Table 4.

Rapid technological growth in mobile communications in the last decade has led to innovative and unique mobile healthcare systems. Adding location-based services into healthcare systems will assist patient in terms of searching nearby doctors and healthcare centers. Doctors can check patients remotely and current patient's location can be tracked in case of emergency. References [14–16] are examples of location-based Emergency/Medical/Health Services. Moreover, location-based services can provide the wide range of information related to the Points of Interest (POI) such as hotels, restaurant, tourism attraction and so on. This information can be offered based on the current location of the tourists who are looking for an appropriate POI. References [17–19] proposed tourism location-based services. In addition to using location-based services in healthcare and tourism areas, location-based services can guide users in order to find the best routes. Navigation/routing/Tracking services can be used for tracking the friends, patients, etc. It can be used in order to direct users to their destination and find a way with less traffic. References [20–22] offered navigation, localization and monitoring of patients, and pedestrian navigation respectively. Proximity is another type of the location-based services. Proximity services can notify users while they are within the certain distance of other people, businesses, and so forth. On the other hand, businesses can be informed while users are in their proximity; therefore, they can send advertisement to those users and attract them to their businesses. References [23–25] are good cases for proximity services. Rapid technological evaluation affects electronic commerce (e-commerce) by changing the way of shopping, booking and marketing. Location-aware shopping has been developed by [26] in order to provide information of the customer's preferred vendors that are in their neighborhood.

Furthermore, [27] provides a dynamic service discovery mechanism that enables mobile users in a given coverage area to easily access available services that are provided by suppliers. In addition, [28] designed an intelligent agent based hotel search and booking system. The system is agent-based to perform hotel-booking activities. The agent will check all of the hotels in terms of available facilities, price, customer experience, transportation etc. and forward this information back to the user's mobile phone. Another category of location-based services is vehicular services. It can be either related to traffic information sharing like [29], or it can be used for vehicle location prediction such as [30] and [31]. Entertainment is the next category of location-based services. There are many location-based services such as [32–34] that can be used for the purpose of language learning specially [32]. The last location-based services is advertisement. Businesses can detect near-by customers and send them their promotion and new product in order to attract them to their businesses. On the other hand, this service will benefit customers who are looking for their favorite product. While the

**Table 4** Classification of location-based services

Application Type	Years	Framework	Technology	Application Type	Years	Framework	Technology
Emergency/ Medical/ Health Services	2012	Location-Based Mobile Cardiac Emergency System (LMCES) [14]	GPS/GPRS	E-Commerce	2008	Location-aware recommender system for mobile shopping environments [26]	Internet, Cellular Network
	2010	Location Application for Healthcare System [15]	Internet, GPS		2007	Location-based M-Commerce [27]	Cellular Network
	2011	Android-based emergency alarm and healthcare management system [16]	GSM, GPS		2007	Hotel Search and Booking System [28]	Internet, Web Application
Tourism	2009	Personalized Tourism Information System in Mobile Commerce [17]	RFID	Vehicular Services	2006	Sharing Traffic Jam Information using Inter-Vehicle Communication [29]	GPS, WLAN
	2012	MyTourGuide.com [18]	GPS		2012	Vehicular location prediction based on mobility patterns for routing in urban VANET [30]	VANET
Navigation/ routing/ Tracking	2012	A Trajectory-Based Recommender System for Tourism [19]	GPS		2009	Wireless LAN-Based Vehicular Location Information Processing [31]	WLAN
	2012	RFID Assisted Navigation Systems for VANETs [20]	RFID	Entertainment	2009	Location-based Game for Supporting Effective English Learning [32]	WLAN
	2010	Localization and monitoring of patients [21]	Zigbee		2009	Mobile Game Based Learning [33]	GPS
	2010	Pedestrian Navigation System [22]	GPS		2011	iDetective [34]	GPS

(continued)

**Table 4** (continued)

Application Type	Years	Framework	Technology	Application Type	Years	Framework	Technology
Proximity Services	2011	Framework for quantifying the system performance of proximity-based services (PBS) [23]	Cellular Network	Advertisement	2005	Location Based Information/Advertising [35]	Bluetooth
	2008	Proximity-based peer selection [24]	GPS, Cellular Network		2011	e-Brochure [36]	GPS
	2000	Context-aware Electronic Tourist Guide [25]	Cell-based wireless communications		2012	Targeted mobile advertising system (TMAS) [37]	Internet

customer is confused between different products, sending promotion and product information can draw their attention to the special business close to their current location. References [35–37] are proposed for location-based advertisement services.

## 5 Concluding Remarks

Accuracy is one of the important parameters of positioning methods used in location-based services. Accuracy of different positioning methods along with their pros and cons are discussed in this paper. As mentioned in this paper, the accuracy of different positioning methods varies about 10 m–35 km. Therefore the developer must be careful to decide which method is best fitting, depending on the needs of the location based service. Another important parameter of selecting positioning method is the environment of positioning method. It is important to select appropriate method for indoor and outdoor environment. For instance, GPS is a reliable outdoor positioning method while it cannot be used for indoor positioning. TOA, Cell ID, and E-OTD are suitable for indoor positioning; whereas, GPS and A-GPS are not recommended to locate the indoor position of any node. Location positioning is the key features of location-based services and existing location-based services prove the significance of location positioning method. Although LBS represent promising services for user, privacy, concerns, quality of service problems, fair access to location information, and the lack of standards for technology and service providers may hinder market development and represent critical policy issues to be resolved.

## References

1. Khalel AMH (2010) Position location techniques in wireless communication systems. Master electrical engineering emphasis on telecommunications. Blekinge Institute of Technology Karlskrona, Sweden
2. Schiller J, Voisard A (2004) Location-based services. Morgan Kaufmann Publishers-Elsevier, San Francisco
3. Virrantaus K et al (2001) Developing GIS-supported location-based services. In: Web information systems engineering, pp 66–75
4. Espinoza F et al (2001) GeoNotes: social and navigational aspects of location-based information systems. In: Abowd GD, Brumitt B, Shafer S (eds) UbiComp 2001. LNCS. vol 2201. Springer, Heidelberg, pp 2–17
5. Willaredt J (2010) WiFi and Cell-ID based positioning-protocols, standards and solutions. Presented at the 2nd international conference on Computer and Network Technology
6. Rappaport TS et al (1996) Position location using wireless communications on highways of the future. *Commun Mag IEEE* 34:33–41
7. Kayton M, Fried WR (1997) Avionics navigation systems, 2nd edn. Wiley-Interscience, New York

8. Ficco M, Russo S (2009) A hybrid positioning system for technology-independent location-aware computing. *Softw Pract Experience* 39:1095–1125
9. Brimicombe AJ (2002) GIS-Where are the frontiers now? In: *Proceedings GIS 2002, Bahrain*
10. Searby S (2003) Personalisation-an overview of its use and potential. *BT Technol J* 21:13–19
11. S. Levijoki “Title”, unpublished
12. van de Kar E, Bouwman H (2001) The development of location based mobile services. *Edispuut conference*
13. Steinfield C et al (2004) The development of location based services in mobile commerce. In: Preissl B, Bouwman H, Steinfield C (eds) *Elife after the dot.com bust*. Springer, Berlin, pp 177–197
14. Keikhosrokiani P et al (2012) A proposal to design a location-based mobile cardiac emergency system (LMCES). *Stud Health Technol Inform* 182:83–92
15. Sobh T et al (2010) Mobile application for healthcare system-location based. In: Sobh T (ed) *Innovations and advances in computer sciences and engineering*. Springer, Netherlands, pp 297–302
16. Yuanyuan D et al (2011) An android-based emergency alarm and healthcare management system. In: *International Symposium on IT in Medicine and Education (ITME), 2011*, pp 375–379
17. Zheng W (2009) Personalized tourism information system in mobile commerce. In: *International conference on management of e-Commerce and e-Government, 2009. ICMECG '09*, pp 387–391
18. Husain W et al (2012) MyTourGuide.com: a framework of a location-based services for tourism industry. In: *International conference on Computer and Information Science (ICIS), 2012*, pp 184–189
19. Huang R et al (2012) A trajectory-based recommender system for Tourism, in *Active Media Technology*. Springer Berlin Heidelberg, pp 196–205
20. Wei C et al (2012) On the design and deployment of RFID assisted navigation systems for VANETs. *Parallel Distrib Syst IEEE Trans* 23:1267–1274
21. Redondi A et al (2010) LAURA-LocAlization and Ubiquitous monitoRing of pAtients for health care support. In: *IEEE 21st international symposium on personal, indoor and Mobile Radio Communications Workshops (PIMRC Workshops), 2010*, pp 218–222
22. Popa M et al (2010) Car finding with a pedestrian navigation system. In: *3rd conference on Human System Interactions (HSI), 2010*, pp 406–411
23. Günes Karabulut K (2011) On the performance of proximity-based services. *Wirel Commun Mobile Comput*
24. El-Nahas A, Helmy D (2008) Proximity-based peer selection for service lookup in areas of sudden dense population. In: *IET 4th international conference on intelligent environments, 2008*, pp 1–7
25. Cheverst K et al (2000) Developing a context-aware electronic tourist guide: some issues and experiences. Presented at the proceedings of the SIGCHI conference on Human Factors in computing systems, The Hague, The Netherlands
26. Yang W-S et al (2008) A location-aware recommender system for mobile shopping environments. *Expert Syst Appl* 34:437–445
27. Mzila PD et al (2007) Service supplier infrastructure for location-based M-commerce. In: *Second international conference on Internet Monitoring and Protection, ICIMP 2007*, pp 35–35
28. McTavish C, Sankaranarayanan S (2010) Intelligent agent based hotel search and booking system. In: *IEEE international conference on Electro/Information Technology (EIT), 2010*, pp 1–6
29. Shibata N et al (2006) A method for sharing traffic jam information using inter-vehicle communication. *IEEE*
30. Xue G et al (2012) A novel vehicular location prediction based on mobility patterns for routing in urban vanet. *EURASIP J Wirel Commun Netw* C7–222(2012):1–14

31. Takeda K et al (2009) Wireless lan-based vehicular location information processing. In: Takeda K et al (eds) *In-Vehicle corpus and signal processing for driver behavior*. Springer, US, pp 69–82
32. Chih-Ming C, Yen-Nung T (2009) Interactive location-based game for supporting effective english learning. In: *Environmental Science and Information Application Technology, 2009, ESIAT 2009. International Conference*, pp 523–526
33. Schadenbauer S (2009) Mobile game based learning: designing a mobile location based game. In: Bruck P (ed) *Multimedia and e-content trends*. Vieweg Teubner, Wiesbaden, pp 73–88
34. Yoshii A et al (2011) iDetective: a location based game to persuade users unconsciously. In: *Embedded and Real-Time Computing Systems and Applications (RTCSA), 2011 IEEE 17th International Conference*, pp 115–120
35. Rashid O et al (2005) Implementing location based information/advertising for existing mobile phone users in indoor/urban environments. In: *Mobile Business, 2005. ICMB 2005. International Conference*, pp 377–383
36. Keikhosrokiani P et al (2011) A study towards proposing GPS-based mobile advertisement service. *Commun Comput Inf Sci* 252:527–544
37. Li K, Du TC (2012) Building a targeted mobile advertising system for location-based services. *Decis Support Syst* 54:1–8

**Part XIV**  
**Green and Human Information**  
**Technology**

# Performance Analysis of Digital Retrodirective Array Antenna System in Presence of Frequency Offset

Junyeong Bok and Heung-Gyoon Ryu

**Abstract** In this paper, we design and analyze a digital retrodirective array antenna (RDA) system based on bandpass sampling for wireless communication. The proposed system has low power consumption thanks to increased signal to interference noise ratio (SINR) because digital RDA can automatically make beam toward source with no information about the direction of incoming signal. Also, this paper presents a robust communications system to frequency offset due to digital PLL. Digital PLL can automatically compensate for frequency offset which occurs because of different frequency between transceivers. Simulation results show that the proposed scheme has better BER performance about 5 dB than that of without phase conjugation when the array elements are three.

**Keywords** Retrodirective array antenna · Phase detection · Phase conjugation · Phase lock loop

## 1 Introduction

Retrodirective array technique is able to transmit signal toward source without a priori knowledge of the arrival direction [1]. Retrodirective array has more simple structure than smart antenna technique and it is possible to do automatically beam-tracking. Also, retrodirective system has merit such as high link gain, easy interference elimination, and high energy efficiency. The design of phase conjugation is important in order to develop a retro-directive system. Various schemes

---

J. Bok · H.-G. Ryu (✉)

Department of Electronic Engineering, Chungbuk National University, Cheongju, Korea  
e-mail: ecomm@cbu.ac.kr

J. Bok

e-mail: bji84@nate.com



have been studied to design efficient phase conjugation. Above all, Corner reflector is well known as analog phase conjugation scheme [2]. The Corner reflector scheme is contributed by placing two intersecting flat reflectors perpendicular to each other. Passive retrodirective arrays such as these methods are easy to be implemented. But, it is cannot be update or modify of whole system. Another method heterodyne mixing is proposed to design phase conjugation as another schemes. The phase conjugation is achieved by mixing the received signal of know frequency with double frequency of RF frequency. It is very big disadvantage to need double frequency of RF.

Retrodirective array system using direct down-conversion scheme is proposed for resolving these problem [3]. Direct down conversion method is sensitivity to DC offset and frequency offset. Recently, retrodirective array system using under-sampling (bandpass sampling) based on SDR (software-defined-radio) is studied to solve the problem of direct conversation methods [4, 5].

In this paper, we analyze the BER performance of digital retro-directive array system based on bandpass sampling considering noise at retro-directive array antenna system.

## 2 Digital Retrodirective Array Antenna

The each received signal has different phase lags ( $0, \Delta\varphi, 2\Delta\varphi, 3\Delta\varphi$ ) when incident wave is  $\theta$  in shown Fig. 1. Frequency offset is defined as  $\Delta\gamma$ . For example, first array element has phase lag ( $0$ ), and adjacent second array element has phase lag ( $\Delta\varphi$ ) when frequency offset effect does not exist. In presence of frequency offset,

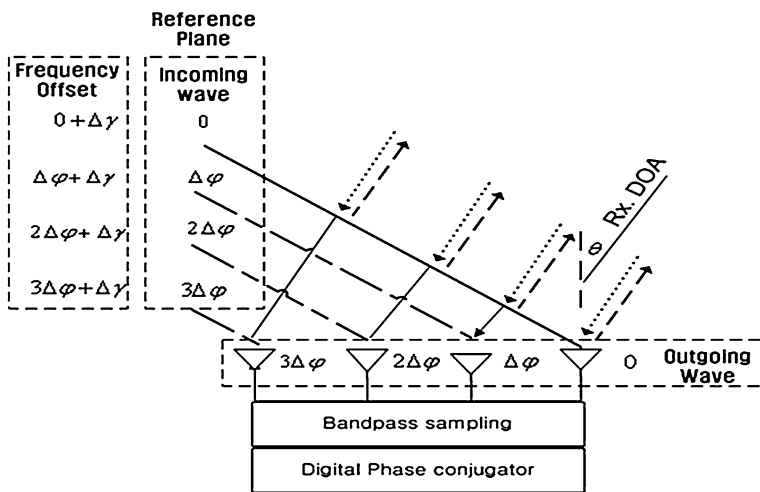
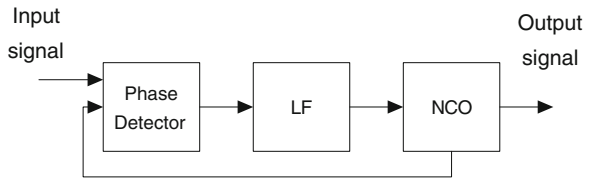


Fig. 1 Digital RDA consider frequency offset

**Fig. 2** Block diagram of digital PLL



first array element and second array element have phase lag such as  $0 + \Delta\gamma$  and  $\Delta\phi + \Delta\gamma$ . In order to compensate for the frequency offset, we design the digital PLL.

### 2.1 Digital Phase Lock Loop (PLL)

Figure 2 shows the block diagram of digital PLL. Digital PLL is a control system that generates an output signal whose phase is related to the phase of an input “Reference” signal. “Reference” signal is hard decision value of received signal. The proposed scheme has three step process to make that reference phase is locked as 0 degrees. Firstly, different phase between input signal and feedback signal was detected using the phase detector block. Secondly, output signal of phase detector is passing through loop filter (LP) for stabilizing the signal. Lastly, the output signal of LF is passing through numerical control oscillator (NCO). NCO is a digital signal generator which creates a signal with amplitude 1 and with conjugation about the phase of output of phase detector.

Transfer function of NCO and LF given by

$$H_{NCO}(z) = \frac{z^{-1}}{1 - z^{-1}} \tag{1}$$

$$H_{LP}(z) = \frac{(G_2 - G_1)z^{-1} + G_1}{1 - z^{-1}} \tag{2}$$

### 2.2 Phase Detector

In the case of QPSK modulation, phase difference is calculated by using equation (5). The phase difference can be expressed as follows equation.

$$\begin{aligned} e^{j\phi} &= e^{(\phi_a - \phi_b)} = \cos(\phi_a - \phi_b) + j \sin(\phi_a - \phi_b) \\ &= \frac{I_a}{\sqrt{I_a^2 + Q_a^2}} \frac{I_b}{\sqrt{I_b^2 + Q_b^2}} + \frac{Q_a}{\sqrt{I_a^2 + Q_a^2}} \frac{Q_b}{\sqrt{I_b^2 + Q_b^2}} \\ &\quad + j \left( \frac{Q_a}{\sqrt{I_a^2 + Q_a^2}} \frac{I_b}{\sqrt{I_b^2 + Q_b^2}} - \frac{Q_b}{\sqrt{I_b^2 + Q_b^2}} \frac{I_a}{\sqrt{I_a^2 + Q_a^2}} \right) \end{aligned} \tag{3}$$

where,  $Q_a, I_a$  are the quadrature and in phase component of received signal.  $Q_b, I_b$  are the quadrature and in phase component of hard decision signal.

The amplitude of QPSK signal has  $\sqrt{I_b^2 + Q_b^2} = \sqrt{2}$ , we assume ( $|\varphi| < 20$ ), phase difference  $\varphi$  can be approximated as ( $\varphi \simeq \sin \varphi$ )

(3) can be reformed as

$$\varphi \simeq \sin \varphi = \frac{1}{\sqrt{2(I_a^2 + Q_a^2)}} \cdot (I_b Q_a - I_a Q_b) \tag{4}$$

Finally, only phase information is given by

$$\varphi = (I_b Q_a - I_a Q_b) \tag{5}$$

We can detect the phase difference by using Eq. (5).

### 3 Simulation Results

Table 1 shows simulation parameters. We assume that transmitter has one antenna, receiver is array antenna (the number of array is 1, 2, and 3). Each received signal has different phase delay. The phase of received signal of first array element is fixed to 0 degrees by using digital PLL.

Figure 3 is constellation of received signal at receiver. Figure 3a–c show that phase delays are 10, 35, and 55 degrees respectively when phase offset is 10.

The phase delay of the received signal is 0 by digital PLL as shown in Fig. 4a. We ensure that digital PLL can compensate for phase delay (10 degrees) by frequency offset. The phase delay of 1st element is 0 degrees after passing through the digital PLL. Phase delays of 2nd element and 3rd element are 20 and 40 after passing through the digital PLL as shown in Fig. 4b, c.

Figure 5 shows comparison of bit error ratio (BER) performance according to the number of array elements at digital RDA. The BER performance of receiver is improved by array gain in the case of increasing the number of array elements. The proposed system can efficiently retransmit the data signal using digital RDA. When the number of array elements is 2, the propose system need 3 dB lower SNR compare to that array element with 1.

Figure 6 shows comparison of BER performance by with and without phase conjugation schemes. Simulation results show that the propose system with phase

**Table 1** Simulation parameters

Parameters	Values
Symbol rate	1 Mbps
Modulation	QPSK
Channel	AWGN
# of array elements	1, 2, 3

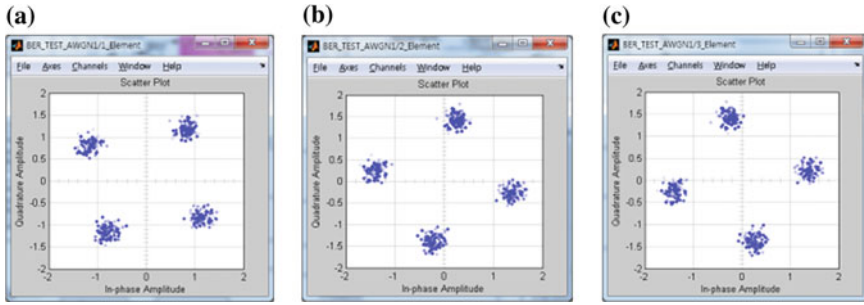


Fig. 3 Constellation of received signal of digital RDA. a 1st element, b 2nd element, c 3rd element

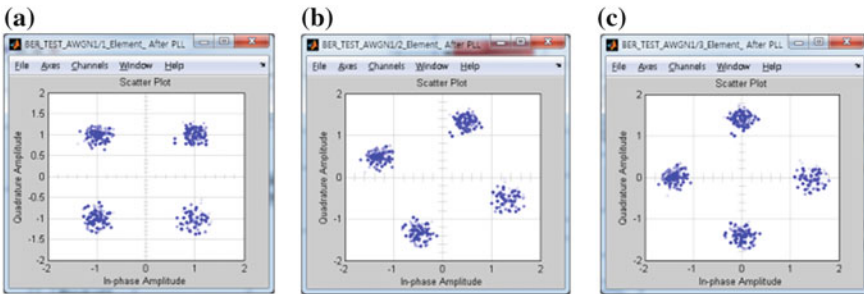


Fig. 4 Constellation of received signal by digital PLL. a 1st element, b 2nd element, c 3rd element

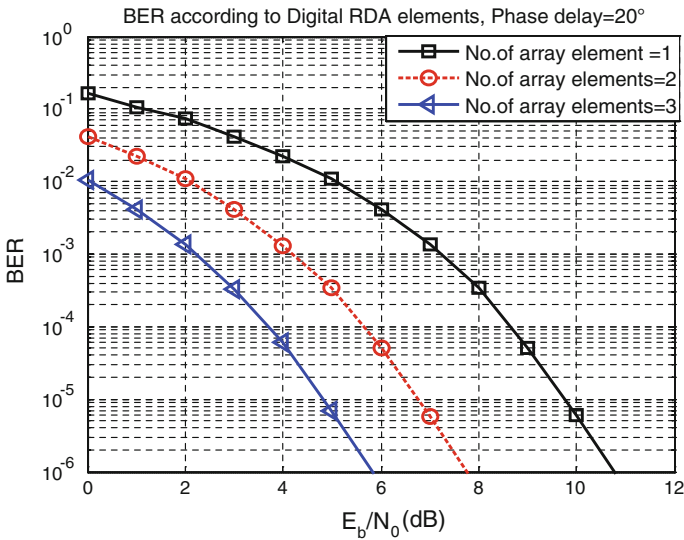


Fig. 5 Comparison of performance by the number of elements

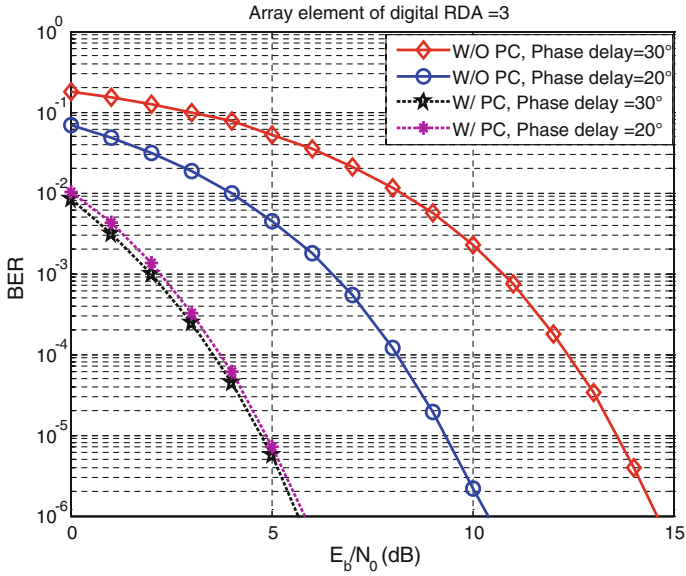


Fig. 6 Comparison of BER performance by w/ and w/o phase conjugation (PC)

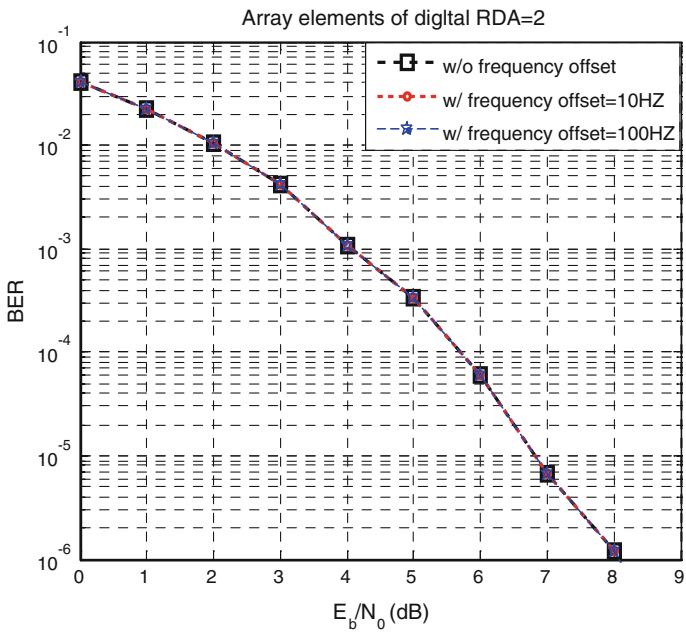


Fig. 7 Comparison of BER performance in presence of frequency offset

conjugation has better BER performance than that without phase conjugation. When we do not use phase conjugation technique, BER performance at receiver is poor because transmitter send data signal toward different direction at source. Simulation results show that the proposed scheme has better BER performance about 5 dB than that of without phase conjugation when the array elements are three.

Figure 7 shows comparison of BER performance by frequency offset. The received signal is shifted by frequency offset. The proposed system has not changing BER performance when same frequency offset occurs at receiver. The proposed system is robustness to frequency offset by digital PLL.

## 4 Conclusion

We study wireless communication using digital retrodirective array antenna for low power consumption and high quality. The proposed system can efficiently communicate compare to using omni-directional antenna because digital RDA can make automatically beam without prior information about source position. Simulation results that the proposed scheme has better BER performance about 5 dB than that of without phase conjugation when the array elements are three. We ensure that the proposed scheme is an energy efficiency system and robustness to frequency offset through designing the digital PLL.

**Acknowledgments** This research was supported by the Korea Communications Commission (KCC), Korea, under the R&D program supervised by the Korea Communications Agency (KCA) (KCA-2012-11-921-04-001).

## References

1. Sharp ED, Diab MA (1960) Van Atta reflector array. *IEEE Trans Antennas Propag* 8:436–438
2. Pon C (1964) Retrodirective array using the heterodyne technique. *IEEE Trans Antennas Propag* 2:176–180
3. Miyamoto RY, Qian Y, Itoh T (2001) A reconfigurable active retrodirective direct conversion receiver array for wireless sensor systems. In: *Proceedings of IEEE MTT-S international microwave symposium, Phoenix*, pp 1119–1122
4. Sun J (2007) A bandpass sampling retrodirective antenna array for time division duplex communications. M.A.Sc. thesis, Dalhousie University, Halifax, NS, Canada
5. Sun J, Zeng X, Chen Z (2008) A direct RF-undersampling retrodirective array system. In: *Proceedings of IEEE radio and wireless symposium, Orlando*, pp 631–634

# A Novel Low Profile Multi-Band Antenna for LTE Handset

Bao Ngoc Nguyen, Dinh Uyen Nguyen, Tran Van Su,  
Binh Duong Nguyen and Mai Linh

**Abstract** A low profile antenna, using coupled-fed, meandered, and folded Planar Inverted-F antenna (PIFA), is proposed to cover multi-band Wireless Wide Area Network (WWAN) operations. The proposed antenna, which is suitable for modern 4G mobile phones, requires only a small foot print of  $44 \times 20 \text{ mm}^2$ . The antenna covers band 14 of LTE-700, GSM-850, GSM-900, DCS-1800, PCS-1900, WCDMA-2100, and band 41 of LTE-2500. In addition to the common WWAN frequencies for mobile communication, the antenna also covers the IEEE802.11b band. The proposed antenna has three simple structures comprising a coupled-fed strip, a meandered shorted-patch, and a folded-patch. These elements are capacitively coupled to each other to form resonant regions in the low bands and the high bands. Modifications to the ground plane are added to achieve a good operation in LTE low band. In the scope of this paper, simulation results, using CST software, are presented to show the effectiveness of the proposed antenna.

**Keywords** PIFA · Low profile · Broadband · Folded patch · Meandered-line · LTE · Modified ground

## 1 Introduction

Currently, mobile communication devices require antennas to have the ability to operate in multi-frequency bands, such as GSM 850/900, DSC 1800, PCS 1900, UMTS 2100, and new LTE bands. In addition, mobile devices are becoming

---

B. N. Nguyen (✉) · D. U. Nguyen · T. Van Su

B. D. Nguyen · M. Linh

School of Electrical Engineering, International University—Vietnam National University,  
Hochiminh City, Vietnam

e-mail: baongocvt1@gmail.com

D. U. Nguyen

e-mail: nduyen@hcmiu.edu.vn

smaller, slimmer and multi-function integrated, forcing the size of the internal antenna to be smaller. Thus, many researches in designing compact and broad band handset antenna have been published recently [1–9].

In the near future the demanding of the 4th generation of communications is expected to grow. Therefore, the mobile communication antenna will have to incorporate the LTE bands as well. Although LTE standard can be used with many different frequency bands, LTE 700 is of interest because of its availability and low cost deployment. Designing an antenna for a low frequency, specifically the LTE-700 band, with a low profile poses a difficulty on the limited dimensions of typical modern mobile phones. The typical size of the antenna at the low frequency band is usually much larger than the typical size of the mobile devices.

Recently, planar inverted-F antenna (PIFA) has been employed widely as internal antenna for mobile handsets thanks to its low profile, light weight, low cost, versatile characteristics. Nonetheless, PIFA in its original shape has disadvantage of narrow bandwidth. Thus, some techniques to make PIFA broadband and compact such as inserting slots, matching network, meandering and folding must be applied to PIFA [10, 11].

In this paper, we will propose a novel low profile multiband antenna which has the potential to cover 8 frequency bands of Long Term Evolution US band 14 (LTE 758–798 MHz), Global System for Mobile communications (GSM850/900 824–960 MHz), Digital Communication System (DCS 1710–1880 MHz), Personal Communication Services (PCS 1850–1990 MHz), Universal Mobile Telecommunications System (UMTS 1920–2170 MHz), IEEE802.11b (2400–2495 MHz), and LTE band 41 (2496–2690 MHz).

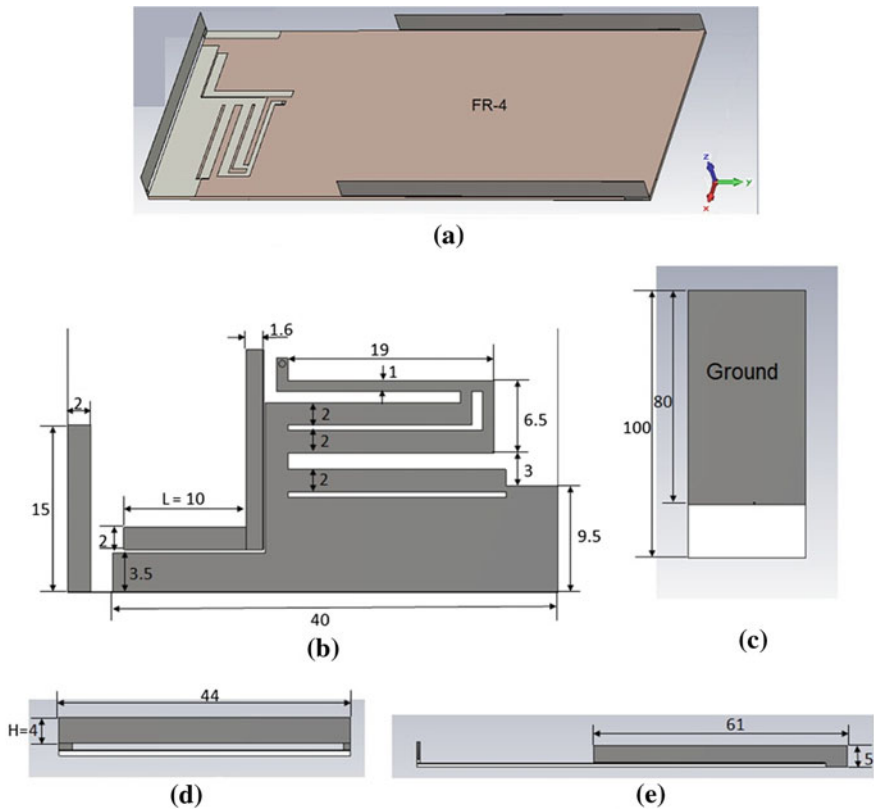
The proposed antenna is a coupled-fed, meandered, and folded PIFA which is printable on a FR4 substrate circuit board of the mobile devices, making it easy to fabricate at low cost and attractive for slim mobile phone applications.

## 2 Proposed Antenna Geometry

The proposed coupled-fed folded PIFA antenna with modified ground for eight-band LTE/GSM/UMTS/WLAN operation in the mobile phone is shown in Fig. 1a. The antenna is printed on a no-ground space of  $20 \times 44 \text{ mm}^2$  and occupies the bottom of the system circuit board which is a 0.8-mm thick FR4 substrate of relative permittivity 4.3, loss tangent 0.025, length 100 mm and width 44 mm.

The detailed dimensions of the antenna are described, in various types of view, in Fig. 1b, c, d, e, respectively. The proposed antenna is composed of a monopole antenna acting as a coupled-fed strip and a shorted folded-patch antenna. The monopole antenna is a quarter-wavelength inverted-L shape antenna centered around 2436 MHz. The monopole antenna is used as the coupled-fed structure to the shorted folded-patch antenna. The coupled-fed structure was proven to enhance considerably impedance bandwidth of the antenna in both high bands and low bands [2, 4, 6, 7]. To obtain the low bands for the antenna, the folded antenna is





**Fig. 1** Geometry and detailed dimensions of the proposed antenna; **a** The overall antenna on the circuit board, **b** Front view, **c** Back view, **d** Bottom view, **e** Right side view

meandered with two strip lines and a shorting pin. The shorting pin here plays important roles not only in the making the antenna physically smaller but also improving impedance matching [10]. To extend the covering of the low band, the radiator is lengthened with a perpendicular patch (size  $4 \times 44 \text{ mm}^2$ ) to the main radiator. The perpendicular patch is connected to the radiator by two connectors; one is a metal piece (size  $1 \times 1 \text{ mm}^2$ ) connected at one end and the other is a strip line (size  $2 \times 15 \text{ mm}^2$ ) connected at other end.

The shorted meandered strip lines and the perpendicular patch together (hence the shorted folded-patch antenna) attribute in generating a resonant mode to form the antenna’s lower band to cover LTE band 14, GSM850 and GSM900. Moreover, two slots are created in order to gain more impedance matching at both low band (758–1107 MHz) and high band (1.68–2.18 GHz).

Even the antenna is designed to resonate at the low frequency bands; the impedance matching is still a problem due to the limitation of the ground length [3]. To achieve resonance at LTE band 14 which has the longest wavelength, the

ground is at least 100 mm in length, thus the total circuit board must be 120 mm; which is considerable larger than the typical length of the mobile phone. To make the ground larger without increasing the antenna size, folded arms are attached to the ground [3]. Our proposed antenna has two arms with optimized length of 61 mm are attached to the ground (Fig. 1e).

### 3 Simulation Results and Discussions

Figure 2 shows the simulated return loss, S11 parameter, of the proposed antenna. The simulated results were obtained using Microwave Studio software from Computer Simulation Technology (CST). The proposed antenna has three simple structures, the coupled-fed strip, a meandered shorted-patch, and a folded-patch. By varying each of the structure individually, the optimum combination was used to achieve the desired bandwidth at the selected frequency bands with better return loss less than  $-6$  dB. The  $-6$  dB threshold is generally the acceptable level for broadband internal mobile phone antenna. The optimum combination includes the coupled-fed strip length  $L = 10$  mm, the folded-patch with the height of  $H = 4$  mm, and the two meandered lines with the total length of 46.5 mm. In Fig. 2, there are two frequency areas of interest that satisfy the  $-6$  dB recommendation; ranging from 0.756–1.042 GHz and 1.709–2.725 GHz. Within these two areas of interest, all the desired frequency bands have the return loss below  $-6$  dB threshold. In other words, the proposed antenna can operate in LTE 700 MHz bands 14 758–798 MHz, GSM 824–960 MHz, DCS 1710–1880 MHz, PCS 1850–1990 MHz, UMTS 1920–2170 MHz, and LTE band 41 2496–2690 MHz. In addition to the mobile communication bands, the result shows the covering of the WLAN as well, specifically the IEEE802.11b 2400–2495 MHz frequency band.

The maximum simulated radiation gains and efficiency at the centered frequencies are shown in Table 1.

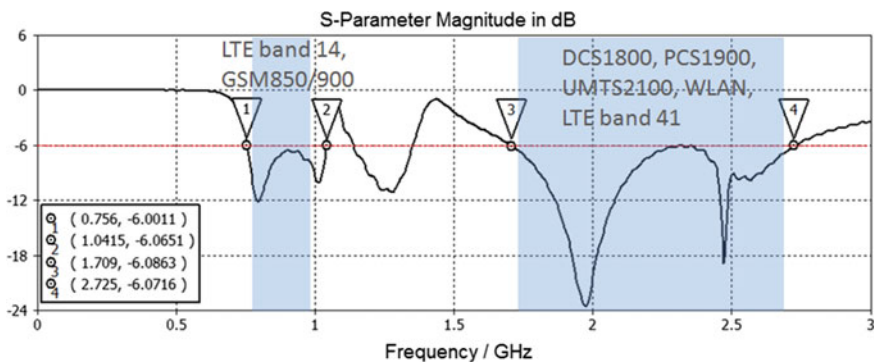
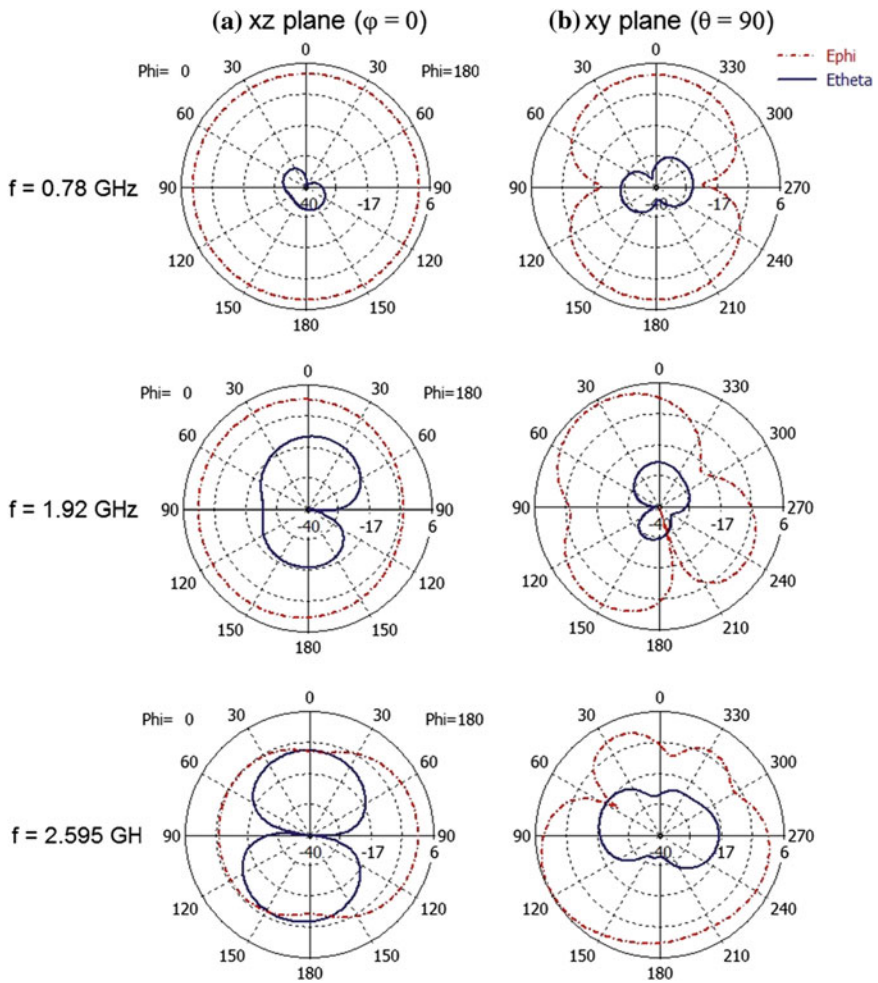


Fig. 2 Return loss simulated result of the proposed antenna

**Table 1** Maximum simulation gain of the proposed antenna

Application	Centre frequencies (MHz)	Peak gain (dBi)	Total efficiency (%)
LTE band 14 (758–798 MHz)	780	1.9	90.4
GSM850 (824–894 MHz)	860	2.0	82.5
GSM900 (880–960 MHz)	920	1.8	76.6
DCS (1710–1880 MHz)	1795	4.0	84.2



**Fig. 3** Polar view of  $E_\phi$  and  $E_\theta$  radiation pattern of the proposed antenna in xz-plane and xy-plane

The polar views of the radiation patterns of the proposed antenna are shown in Fig. 3. At low bands from 758–960 MHz, antenna's patterns are similar to a dipole's pattern. On the other hand, at the high bands, the radiation patterns do not follow any specific patterns.

## 4 Conclusion

The novel wide bands LTE-700 band 14, GSM-850, GSM-900, DCS-1800, PCS-1900, WCDMA-2100, LTE-2500 band 41 and WLAN 2400 has been proposed for mobile phone application. Coupling the feeding strip with the microstrip antenna was proven to enhance considerably the bandwidth of the antenna in both high bands and low bands. Adding the two meandered strip lines with the PIFA was proven to improve the impedance matching at the lower band. The shorted meandered strip line and the perpendicular patch together attribute in generating a resonant mode at the lower band to cover LTE band 14, GSM850 and GSM900. Moreover, to improve the antenna operating in LTE low band, the ground is lengthened by attached two arms. The proposed antenna, simulated on a FR4 substrate, has shown the simulation results with acceptable of return loss (S11), radiation patterns, gains and efficiencies.

## References

1. Lee WY, Jeong YS, Lee SH, Oh JR, Hwang KS, Yoon YJ (2010) Internal mobile antenna for LTE/DCN/US-PCS. In: IEEE microwave Conference Proceedings (APMC), 2010 Asia-Pacific, pp 2240–2243, 7–10 Dec 2010
2. Wong KL, Chen WY, Kang TW (2011) On-board printed coupled-fed loop antenna in close proximity to the surrounding ground plane for penta-band WWAN mobile phone. *IEEE Trans Antennas Propag* 59(3):751–757
3. Jeong YS, Lee SH, Yoon JH, Lee WY, Choi WY, Yoon YJ (2010) Internal mobile antenna for LTE, GSM850, GSM900, PCS1900, WiMAX, WLAN. In: IEEE conference publications on Radio and Wireless Symposium (RWS) 2010, pp 559–562 (Jan 2010)
4. Ying LJ, Ban YL, Chen JH (2011) Low-profile coupled-fed printed PIFA for internal seven-band LTE, GSM, UMTS mobile phone antenna. In: IEEE conference publications on cross strait quad-regional radio science and wireless technology conference (CSQRWC), Vol. 1, pp 418–421 (July 2011)
5. Ying Z (2012) Antennas in cellular phones for mobile communications. *IEEE J Mag, Proc IEEE* 100(7):2286–2296
6. Yang CW, Jung YB, Jung CW (2011) Octaband internal antenna for 4G mobile handset. *IEEE J Mag, Antennas Wirel Propag Lett* 10:817–819
7. Kim MH, Lee WS, Yoon YJ (2011) Wideband antenna for mobile terminals using a coupled feeding structure. In: IEEE international symposium on antennas and propagation (APSURSI) 2011, pp 1910–1913, July 2011
8. Zheng M, Wang H, Hao Y (2012) Internal hexa-band folded monopole, dipole, loop antenna with four resonances for mobile device. *IEEE Trans Antennas Propag* 60(6):2880–2885

9. Tsai PC, Lin DB, Lin HP, Chen PS, Tang IT (2011) Printed inverted-f monopole antenna for internal multi-band mobile phone antenna. In: Vehicular technology conference (VTC Spring), 2011 IEEE 73rd, 15–18 May 2011
10. Wong KL (2002) Compact and broadband microstrip antennas. John Wiley & Sons, Inc., New York
11. Chen ZN, Chia MYN (2006) Broadband planar antennas design and applications. John Wiley & Sons, Inc., New York

# Digital Signature Schemes from Two Hard Problems

Binh V. Do, Minh H. Nguyen and Nikolay A. Moldovyan

**Abstract** In this paper, we propose two new signature schemes and a novel short signature scheme from two hard problems. The proposed schemes have two prominent advantages. Firstly, they are developed from some signature schemes where the security and efficiency have been proven. Therefore, they inherit these properties from the previous schemes. Secondly, the security of the proposed schemes is based on two hard problems. Therefore, they are still safe even when cryptanalysis has an effective algorithm to solve one of these problems, but not both. Moreover, we also propose a method for reducing signatures and this is the first attempt to reduce signatures based on two hard problems. Therefore, our proposed schemes are suitable for the applications requiring long-term security in resource limited systems.

**Keywords** Cryptographic protocol • Digital signature • Factorization problem • Discrete logarithm problem • Short signature scheme

## 1 Introduction

One of the vital objectives of a information security systems is providing authentication of the electronic documents and messages. Usually this problem is solved with digital signature schemes (DSSes) [1]. There were many proposals for

---

B. V. Do (✉)

Military Information Technology Institute, Hanoi, Vietnam

e-mail: binhdv@gmail.com

M. H. Nguyen

Le Qui Don Technical University, Hanoi, Vietnam

e-mail: hieuminhmta@ymail.com

N. A. Moldovyan

St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, 14

Liniya, 39, St., Petersburg, Russia199178,

signature schemes published based on a single hard problem such as factoring (FAC), discrete logarithm (DL) or elliptic curve discrete logarithm (ECDL) problems [1, 2]. However, these schemes only guarantee short-term security. In order to enhance the security of signature schemes, it is desirable that the signature schemes are developed based on multiple hard problems. This makes it much harder to attack these schemes since it requires solving multiple problems simultaneously. Some schemes based on two problems, FAC and DL, have been published [3–5]. However, designing these schemes is not easy. Moreover, most of them have been proven that they are not secure [6–8]. Therefore, it is necessary to develop new safe signature schemes based on two hard problems.

In bandwidth and resource limited systems, it is important that the signature schemes have a short signature length. So far, the problem of signature reducing is only investigated for the schemes with single hard problem [9, 10]. We can easily implement a combination of two or more hard mathematical problems in a unified DSS. Breaking such schemes requires simultaneously solving all hard problems. Such implementations require increasing signature length, because the signatures must be present elements belonging to different mathematical problems. It is therefore of interest to develop DSSes, that provide an acceptable signature length. The rest of this paper is organized as follows. In Sect. 2, describes the DSSes based on two hard problems (FAC and DL). Section 3, presents the design of two new DSSes, which requires the simultaneous breaking of FAC and DL problems. Section 4 proposes a novel and efficient short signature scheme. Section 5, describes the security analysis of our schemes. Section 6, describes the performance analysis of our schemes. In the last section, the conclusion of our research is presented.

## 2 Signature Schemes Based on Factoring and Discrete Logarithms

Previously, DSSes were proposed based on the difficulty in solving the factorization and discrete logarithm problems. For example, the scheme in [11] used a prime modulo  $p$  with a special structure  $p = 2n + 1$ , where  $n = q'q$ ,  $q'$  and  $q$  are large prime numbers with at least 512 bits. We use the following notations to describe these signature schemes.  $H$  is a hash value computed from the signed document  $M$ .  $F$  is a one-way function, for which can be used to calculate the value of  $H = F_H(M)$ .  $\alpha$  is a primitive element in  $Z_p^*$  with order  $q$  satisfying  $\alpha^q \equiv 1 \pmod{p}$ . The value of  $\lambda$  is a bit length of  $q$ , where  $q$  is a prime divisor value of  $n$ .

The public key is a triple of  $(p, \alpha, \lambda)$ . The private key is  $q$ .

*Signature generation procedure:*

- (1) Compute  $r = F_H(\alpha^k \pmod{p})$ , where  $k$  is a secret random number,  $1 < k \leq q - 1$ .
- (2) The equation generating the parameter  $S$  is given by the following equation:  

$$S = k(Hr)^{-1} \pmod{q}.$$

The signature is a pair of values  $(r, S)$ , in which the length of the second value is equal to  $|S| \leq \lambda$ ;

When using 1024-bit prime  $p$  and a compression function  $F$  whose output is a  $t$ -bit length and assuming  $t = 160$  bits, the length of the digital signature is  $|F| + |q| \approx 160 + 512 \approx 672$  bits.

*Signature verification procedure:*

The verification equation is as follow:  $r = F_H(\alpha^{HSr} \bmod p)$ .

An important part of the verification procedure is to verify the authenticity of a digital signature with the condition  $|S| \leq \lambda$ , because signature  $(r, S')$  with second element of which has the size  $|S'| \approx 1023$  bits (if  $|p| \approx 1024$  bits) can be easily generated without knowing of the secret parameter  $q$ . Such signature  $(r, S')$  will satisfy the verification equation. However the signatures  $(r, S')$  do not satisfy the condition  $|S'| \leq \lambda$ . Computing the forged signature  $(r, S')$  satisfying both the verification equation and the condition  $|S'| \leq \lambda$  without knowing the private key  $q$  is not easier than factoring the number  $n = (p - 1)/2$  [11]. Security of the considered DSS is based on the difficulty of solving any of the following two problems, factorization and discrete logarithm. Indeed, it is easy to show that solving the factorization problem or solving the discrete logarithm problem allows one to compute the private key and to forge the signature.

In the Schnorr signature in [1], we can use a prime module with the structure of  $p = 2n + 1$ . This leads to the DSS with public key in the form of four values  $(p, \alpha, \lambda, y)$ , where the first three parameters are defined as in the scheme [11] and  $y$  is calculated by the formula  $y = \alpha^x \bmod p$ , where  $x$  is one element of the secret key.

*Signature generation procedure:*

- (1) Compute  $R = \alpha^k \bmod p$ , where  $k$  is a secret random number,  $1 < k \leq q - 1$ .
- (2) Compute  $E = F_H(M||R)$ .
- (3) Compute  $S = k - xE \bmod q$ , such that  $R = \alpha^S y^E \bmod p$ .

The signature is the pair  $(R, S)$ .

*Signature verification procedure:*

- (1) If  $|S| \leq \lambda$ , then calculating the value of  $R^* = \alpha^S y^E \bmod p$ . Otherwise, the signature is rejected as invalid.
- (2) Compute  $E^* = F_H(M||R^*)$ .
- (3) Compare the values  $E^*$  and  $E$ . If  $E^* = E$ , then signature is valid.

Breaking the last signature scheme can be done by simultaneously solving the discrete logarithm problem, which allows to find the secret key  $x$  and the factorization problem, which allows to find the value of  $q$ , required to compute the value of signature  $S$ , whose size will not exceed the value of  $\lambda|q|$ .

However, the simultaneously solving of these two independent hard problems is not necessary to break this scheme. Indeed, the secret parameters of the scheme can be calculated by solving only the discrete logarithm problem.

*This can be done as follow:*

We choose an arbitrary number  $t$ , the bit length does not exceed the value  $\lambda - 1$ . Then calculate the value of  $Z = \alpha^t \bmod p$ . After that we find the logarithm of  $Z$  on



the basis of  $\alpha$ , using the index calculus algorithm [1]. This gives a value of  $T$ , calculated modulo  $n = (p - 1)/2$ . With a probability close to 1, the size of this value is equal to  $|T| \approx |n| > |t|$ . Because  $\alpha$  is number with order  $q$  over  $Z_p^*$  then we have  $t = T \bmod q$ , so  $q$  evenly divides the difference between  $T - t$ . This means that by following the factorization of  $T - t$ , we can find the secret parameter  $q$ . The probability that a factorization of  $T - t$  will have a relatively low complexity is quite high. This means that following the above procedure several times, we will find the value of  $T - t$ , which can be easily factored.

Thus, for breaking of the two DSSes in this section, we only need to solve discrete logarithm problem modulo a prime. In order to design the DSS, which requires simultaneous solving both the factorization problem and the discrete logarithm problem to break, the last signature scheme should be modified. For example, one can use the value  $\alpha$  having order equal to  $n$  and introduce a new mechanism for calculating the value  $S$ , which will require knowledge of the factors of  $n$  while computing  $S$ .

### 3 New Signature Schemes Based on Difficulty of Solving Simultaneously Two Hard Problems

In this section, we propose two new signature schemes from two hard problems. Breaking the modified signature schemes described below requires simultaneous solving two different hard problems, computing discrete logarithm in the ground field  $GF(p)$  and factoring  $n$ .

#### 3.1 The First Scheme

The following modifications have been introduced in the first signature scheme: (i) as parameter  $\alpha$  it is used a value having order equal to  $n$  modulo  $p$ ; (ii) instead of the value  $S$  in the signature verification equation it is introduced the value  $S^2$ .

*Key generation:*

- (1) Choose large distinct primes  $q'$  and  $q$  in the form  $4r + 3$ , and compute  $n = q'q$ .
- (2) Choose randomly a secret key  $x$  with  $x \in Z_p^*$ .
- (3) Compute  $y = \alpha^x \bmod p$ .

The public key is  $(p, \alpha, y)$ . The secret key is  $(x, q', q)$ .

*Signature generation procedure:*

- (1) Compute  $R = \alpha^k \bmod p$ , where  $k$  is a secret random number,  $1 < k \leq n - 1$ .
- (2) Compute  $E = F_H(M||R)$ .

(3) Calculate the value  $S$ , such that  $S^2 = k - xE \pmod n$ .

The signature is the pair  $(E, S)$ .

*Signature verification procedure:*

- (1) Compute  $R^* = \alpha^{S^2} y^E \pmod p$
- (2) Compute  $E^* = F_H(M || R^*)$ .
- (3) Compare the values  $E^*$  and  $E$ . If  $E^* = E$ , then signature is valid.

It is easy to see that, the advantage of using this exponent 2 (calculate the value  $S$ ) is computational load smaller compared to larger exponents. The disadvantage is if  $S^2 = k - xE \pmod n$  has no solution, the signature cannot be directly generated [1].

### 3.2 The Second Scheme

The following modifications have been introduced in the second signature scheme: (i) as parameter  $\alpha$  it is used a value having order equal to  $n$  modulo  $p$ ; (ii) it is used one additional element  $e$  of the public key; (iii) it is used one additional element  $d$  of the private key; (iv) instead of the value  $S$  in the signature verification equation it is introduced the value  $S^e$ . The values  $e$  and  $d$  are generated like in the RSA cryptosystem [1].

*Key generation:*

- (1) Choose randomly an integer  $e \in Z_n$  such that  $\gcd(e, n) = 1$ .
- (2) Calculate a secret  $d$  such that  $ed \equiv 1 \pmod{\phi(n)}$ .
- (3) Choose randomly a secret key  $x$  with  $x \in Z_p^*$ .
- (4) Compute  $y = \alpha^x \pmod p$ .

The public key is  $(e, \alpha, y)$ . The secret key is  $(x, d)$ .

*Signature generation procedure:*

- (1) Compute  $R = \alpha^k \pmod p$ , where  $k$  is a secret random number.
- (2) Compute  $E = F_H(M || R)$ .
- (3) Calculate the value  $S$ , such that  $S^e = k - xE \pmod n$ , i.e.  $S = (k - xE)^d \pmod n$  such that  $R = \alpha^{S^e} y^E \pmod p$ .

The signature is the pair  $(E, S)$ . It is easy to see that the length of signature is  $|E| + |S| \geq 1184$  bits.

*Signature verification procedure:*

- (1) Compute  $R^* = \alpha^{S^e} y^E \pmod p$ .
- (2) Compute  $E^* = F_H(M || R^*)$ .
- (3) Compare the values  $E^*$  and  $E$ . If  $E^* = E$ , then signature is valid.

### 4 Novel Short Signature Scheme

One of important problems is developing digital signature schemes with short signature length [9]. To reduce the signature length in the case of DSSes from two hard problems we use signature formation mechanism, which is based on solving a system of equations [10].

We use the signature formation mechanism that can be applied while developing DSSes with three-element signature denoted as  $(k, g, v)$ .

The mechanism is characterized in using a three element public key with the structure  $(y, \alpha, \beta)$ , where  $y = \alpha^x \text{ mod } p$ ;  $\alpha$  is the  $\delta$  order element modulo  $p$ , i.e.  $\alpha^\delta \text{ mod } p = 1$ ;  $\beta$  is the  $\gamma$  order element modulo  $n$ , i.e.  $\beta^\gamma \text{ mod } n = 1$  ( $p = 2n + 1$ , where  $n = q'q$ ) and in solving a system of three equations while generating signature. The secret key is  $\gamma$ .

In this scheme,  $q$  and  $q'$  are strong primes and easy to generate using Gordon's algorithm [1]. The prime  $q$  and  $q'$  are supposed to be of large size  $|q| \approx |q'| \geq 512$  bits. Gordon's algorithm allows to generate strong primes  $q$  and  $q'$  for which the numbers  $q - 1$  and  $q' - 1$  contain different prime devisors  $\gamma'$  and  $\gamma''$ , respectively.

Some internal relation between the  $\beta$  and  $n$  values provides potentially some additional possibilities to factorize modulus  $n$ . This defines special requirements to the  $\beta$  element of the public key [10]. One should use composite  $\gamma$ , i.e.  $\gamma = \gamma'\gamma''$ , where  $\gamma' | q - 1$ ,  $\gamma'' | q' - 1$ ,  $\gamma' \nmid q' - 1$  and  $\gamma'' \nmid q - 1$ . To choose the size of the  $\gamma$  value we should take into account that the  $\beta$  value can be used to factorize the  $n$  modulus calculating  $\text{gcd}(\beta^i \text{ mod } n - 1, n)$  for  $i = 1, 2, \dots, \min\{\gamma', \gamma''\}$ . Therefore we should use the 80-bit values  $\gamma'$  and  $\gamma''$ . Thus, for  $\gamma$  we get the following required length:  $|\gamma| = 160$  bits.

A secure variant of the DSS with the 480-bit signature length is described by the following verification equation:  $k = (y^k \alpha^{gH} \text{ mod } p + \beta^{kgv+H} \text{ mod } n) \text{ mod } \delta$ , where  $\delta$  is a specified prime number and  $H$  is the hash value of the signed message.

*The signature generation is performed as follows:*

- (1) Generate two random number  $u_1$  and  $u_2$  calculate  $z_1 = \alpha^{u_1} \text{ mod } p$  and  $z_2 = \beta^{u_2} \text{ mod } n$ .
- (2) Solve simultaneously three equations:

$$k = (z_1 + z_2) \text{ mod } \delta; g = (u_1 - kx)H^{-1} \text{ mod } \delta; v = (u_2 - H)k^{-1}g^{-1} \text{ mod } \gamma.$$

Breaking this scheme requires the simultaneously solving of the factorization the modulus  $n$  and the discrete logarithm modulo  $p$ .

In this scheme the signature length is compared for different DSSes in the case of minimum security level that can be estimated at present as  $2^{80}$  operations [1]. The minimum level of security provided under the following size parameters:  $|p| \geq 1024$  bits,  $|n| \geq 1024$  bits,  $|\delta| \geq 160$  bits and  $|\gamma| \geq 160$  bits. It is easy to see that the size of a digital signature is  $|k| + |g| + |v| \geq 480$  bits.

## 5 Security Analysis

This section presents an analysis on the security of the proposed signature schemes. The results show that the new schemes are only broken when two hard problems, DL and FAC, are solved simultaneously.

*The first scheme:* In this scheme, solving the DL problem in  $GF(p)$  is not sufficient for breaking the modified scheme. The solution of the DL problem leads to the computation of the secret key  $x$  and to the possibility to calculate the value  $S^* = (k - xE) \bmod n$ . However, calculating the signature  $S$  requires to extract the square root modulo  $n$  from the value  $S^*$ . The last represent a hard problem until the value  $n$  is factorized.

*The second scheme:* Similar to the first scheme, solving the DL problem in  $GF(p)$  is not sufficient for breaking the modified scheme. To break this signature scheme it is required to know the factorization of  $n$ . The solution of the DL problem leads to the computation of the secret key  $x$  and to the possibility to calculate the value  $S^* = (k - xE) \bmod n$ . However, to calculate the signature  $S$ , it is required to extract the  $e$ th root modulo  $n$  from the value  $S^*$ . This requires factoring the modulus  $n$ .

**Theorem 1** *If an ORACLE  $O$  can solve DL and FAC problems, then it can break the proposed schemes.*

In other words, if an ORACLE  $O$  has the prime factors  $(q', q)$  of  $n$  and  $(x, k)$  by solving FAC and DL problems, then  $(E, S)$  will be the eligible sign of document  $M$  generated by the proposed methods.

We indicate that the following attacks can be used to break the proposed schemes.

- Attack 1: In order to break these schemes, the adversary needs to calculate all secrete elements in the systems. In this case, the adversary needs to solve DL problem to calculate values  $(x, k)$ . Moreover, the adversary also have to solve FAC problem. It means that the adversary have to solve both DL and FAC problems in order to break the proposed schemes.
- Attack 2: The adversary may receive values  $(R, E, S)$ . By selecting  $S$  arbitrarily and computing  $E = F_H(M||R)$ , the adversary try to find  $S$  satisfying equation  $R = \alpha^{S^e} y^E \bmod p$ . In order to solve this equation, the adversary also needs to solve both DL and FAC problems.
- Attack 3: All attacks on RSA, Rabin, Schnorr [1] can not be successful on the proposed schemes, because these schemes are the combination of two fundamental algorithms.

**Table 1** Time complexity comparison of the proposed schemes and the scheme of [5]

	Time complexity (our first scheme)	Time complexity (our second scheme)	Time complexity [5]
Key generation	$T_{EXP}$	$T_{EXP} + T_{INV}$	$T_{EXP} + T_{INV}$
Signature generation	$T_{EXP} + T_{MUL} + T_{SR} + T_H$	$2T_{EXP} + T_{MUL} + T_H$	$3T_{EXP} + 3T_{MUL} + 2T_{SR} + T_H$
Signature verification	$3T_{EXP} + T_{MUL} + T_H$	$3T_{EXP} + T_{MUL} + T_H$	$4T_{EXP} + 2T_{MUL} + T_H$

## 6 Performance Analysis

The performance of the proposed algorithms is evaluated based on the complexity of the following procedures: key generation, signing generation and verification. For the sack of comparison, we use the following notations.  $T_{EXP}$  denotes Time complexity for executing the modular exponentiation.  $T_{MUL}$  denotes Time complexity for executing the modular multiplication.  $T_H$  denotes Time complexity for performing hash function.  $T_{SR}$  denotes Time complexity for executing the modular square root computation.  $T_{INV}$  denotes Time complexity for executing the modular inverse computation.

The results in Table 1 show that the proposed scheme have better performance than the previous scheme in [5].

## 7 Conclusion

This paper presents the ability to efficiently develop signature schemes based on the widely used fundamental schemes. Based on some well-know schemes, RSA, Rabin and Schnorr, we proposed two new signature schemes. The proposed schemes possess the higher security than well-know schemes because they are based on two independently difficult problems.

The paper also introduces a new method for reducing signature length. This leads to the proposed signature schemes have the shortest signature length in comparison with the other schemes based on two hard problems.

## References

1. Menezes AJ, van Oorschot PC, Vanstone SA (1996) Handbook of applied cryptography. CRC Press, Boca Raton
2. Pieprzyk J, Hardjono T, Seberry J (2003) Fundamentals of computer security. Springer, New York
3. Harn L (1994) Public-key cryptosystem design based on factoring and discrete logarithms. IEEE Proc Comput Digit Tech 141(3):193–195

4. Tzeng SF, Yang CY, Hwang MS (2004) A new digital signature scheme based on factoring and discrete logarithms. *Int J Comput Math* 81(1):9–14
5. Ismail ES, Tahat NMF (2011) A new signature scheme based on multiple hard number theoretic problems. *ISRN Commun Netw*
6. Li J, Xiao G (1998) Remarks on new signature scheme based on two hard problems. *Electron Lett* 34(25):2401–2402
7. Chen T-H, Lee W-B, Horng G (2005) Remarks on some signature schemes based on factoring and discrete logarithms. *Appl Math Comput* 169:1070–1075
8. Buchmann J, May A, Vollmer U (2006) Perspectives for cryptographic long term security. *Commun ACM* 49(9):50–55
9. Boneh D, Lynn B, Shacham H (2001) Short signatures from the Weil pairing. In: *ASIACRYPT '01*, vol 2248. LNCS, pp 514–532
10. Moldovyan NA (2009) Short signatures from difficulty of factorization problem. *Int J Netw Secur* 8(1):90–95
11. Dernova ES (2009) Information authentication protocols based on two hard problems. PhD Dissertation, St.Petersburg State Electrotechnical University. St. Petersburg, Russia

# Performance Improvements Using Upgrading Precedences in MIL-STD-188-220 Standard

Sewon Han and Byung-Seo Kim

**Abstract** Deterministic Adaptable Priority Net Access Delay, one of the channel access methods defined in MIL-STD-188-220 standard, is designed focusing on the reliable transmissions of the highest priority packets. Therefore, it degrades the performances of the low priority traffics. This paper proposes a method to improve the performances of the low priority traffics while maintaining the performances of the highest priority traffics. The method upgrades intentionally the priority of the traffics that stays in a queue for a certain time period, so that it gives an opportunity for even lower priority traffic to be transmitted. The proposed method is simulated and the results show it improves the network performances.

**Keywords** MIL-STD-188-220 • Tactical networks • DAP-NAD • Precedence

## 1 Introduction

MIL-STD-188-220 standard specifies physical, data link, and intranet protocol layers for narrowband-based tactical Digital Message Transfer Devices (DMTD) [1] used for remote access to automated Command, Control, Communication, Computer and Intelligence (C4I) systems and to other DMTDs. DMTD is a portable tactical communication devices having limited data generation and processing capability. The standard has been versioned up from MIL-STD-188-220A to MIL-STD-188-220D with Change 1. MIL-STD-188-220 standard defines 6

---

S. Han

Korea Radio Promotion Association, Seoul, Korea  
e-mail: swan@rapa.or.kr

B.-S. Kim

Department of Computer and Information Communications Engineering, Hongik University, 2639, Sejong-ro, Sejong-si, Korea

different protocols for data link layer which are Random-Network Access Delay (R-NAD), Prioritized-NAD (P-NAD), Hybrid-NAD (H-NAD), Radio Embedded-NAD (RE-NAD), Deterministic Adaptable Priority NAD (DAP-NAD), and Data And Voice NAD (DAV-NAD). Moreover, MIL-STD-188-220 standard defines three precedences for the traffics: Urgent, Priority, and Routine. Urgent precedence is the highest precedence and Routine precedence has the lowest precedence. P-NAD, DAP-NAD, and DAV-NAD provide methods for priority-based transmissions.

While conventional tactical communication devices process mainly voice traffic and short messages over the narrowband systems, recent tactical communication devices have been designed to deal with various types of traffics including voice, short/long messages, pictures and video over wideband systems as shown in [2] because sufficient information on the battle field leads to the better command and control. As a part of the movements on the tactical communications, MIL-STD-188-220 standard also needs to be modified to deal with various traffic types over wideband systems. Even though there are some studies as shown in [3–11], all studies are based on the narrowband systems. Studies in [3–9] have focused on the old version of MIL-STD-188-220 standard. Recently, MIL-STD-188-220 standard has been evaluated over wideband channel environments in [12–14]. In [12], R-NAD and DAP-NAD are evaluated comparing with IEEE802.11-based Wireless Local Area Networks (WLAN) over wideband network scenarios and the study concludes using the standard over wideband systems is feasible. In [13], the utility of bump-slot, that is one of time slot used in DAP-NAD, is evaluated over wideband systems and it is concluded that the bump-slot is not useful over wideband systems. In [14], a method to provide enhancements on voice packet transmission over MIL-STD-188-220-based standard system is proposed. The method utilizes the bump slots to give higher opportunity to voice traffics.

In this paper, a method to improve the performances of the lower precedence traffics using DAP-NAD is proposed. While as shown in [12], DAP-NAD gives the best performances on Urgent traffics. Routine traffic may not have any transmission opportunity. The proposed method upgrades the precedences of the frames if the lower frame stays more than certain time period, so that it give an opportunity of transmission to even the lower precedence frames.

In this paper, Sect. 2 introduces the specifications on DAP-NAD specified in MIL-STD-188-220 standard. In Sect. 3, the proposed method is introduced and the extensive simulation results are provided in Sect. 4. Finally, the conclusions and future works are made in Sect. 5.

## 2 DAP-NAD

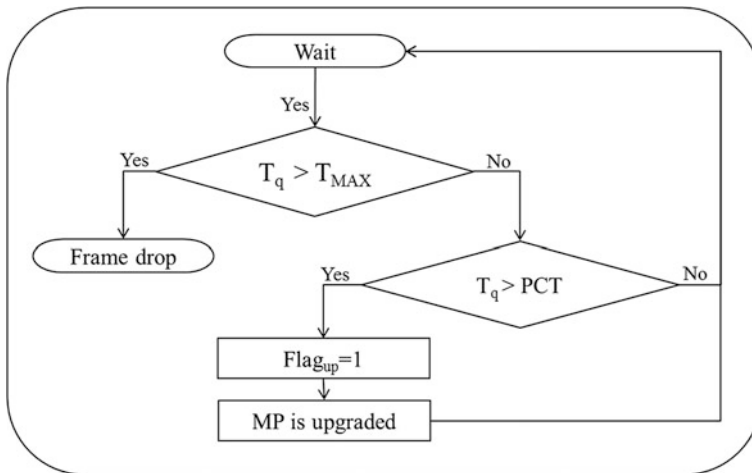
DAP-NAD is Time-Division Multiple Access (TDMA)-based medium access protocol and the transmissions sequence among the participating nodes is pre-scheduled based on the assigned unique node number and the sequence is repeated



whenever the last ordered node's turn is completed. Each node transmits its own frames during its own time slot whose length is variable depending on the length of frame if the Network Precedence (NP) has to be same as Message Precedence (MP) of a node's frame. NP indicates the precedence that is allowed to be transmitted during one transmission sequence. NP might be changed after one transmission sequence is completed. Each transmission sequence is set to one of three precedences. The first NP is set to Urgent, so that during a whole one-sequence, only a node with Urgent pending frame is allowed to transmit. If a node does not have Urgent pending frame when the NP is Urgent, it give up its turn and the next node have opportunity to transmit. If no node transmits during one sequence, then the NP is downgraded to Priority which means nodes having higher precedence frame than Priority are allowed to transmit. If any node transmits a frame during this sequence with Priority NP, the NP is upgraded to Urgent. Otherwise, the NP is downgraded to Routine which means nodes having higher precedence frame than Routine are allowed to transmit. That is, only a node having a frame whose MP is same as NP can transmit. If a node transmits its frame no matter what NP is, the next sequence automatically upgraded to Urgent and start over again. Each node has to keep tracking the NP by itself. For synchronizing NP over all participating nodes, the header of frame contains the MP, so that nodes enable to adjust their NPs by overhearing the header.

### 3 Proposed Method

As mentioned in Sect. 2, DAP-NAD provides the highest transmission opportunity on the Urgent traffics. While it gives the high reliability on the Urgent traffic, the performances of the lower priority traffics are degraded. In certain case, the transmission opportunities for the lower priority traffics are totally prohibited even though there are some opportunities for the lower priority traffics. For example, If a node has periodic Urgent traffic and others have lower traffics, the node with Urgent traffic keep transmitting and the time slots for other nodes is passed without any transmission. This is because only one transmission make NP upgrade to Urgent. In this scenario, some transmissions of the lower MP frames may not degrade the performances of Urgent MP traffic. To resolve this problem, we proposes a method to intentionally upgrade the higher MP of the lower MP to provides transmission opportunities even for the nodes with the lower MP frames. The process of the proposed method is shown in Fig. 1. The proposed method is reclusively performed for the frames in a queue. When a frame is arrived in the queue, the queuing time of the frame,  $T_q$ , is recorded. In every a unit time, the  $T_p$  is evaluated with  $T_{MAX}$  and Precedence-Critical-Time (PCT).  $T_{MAX}$  is the life time of the frame and PCT is time threshold to upgrade the frame's precedence. In every  $T_u$ , if  $T_p$  is larger than  $T_{MAX}$ , the frame is dropped. Otherwise,  $T_p$  of the frame is compared with PCT. If the  $T_q$  is larger than PCT and  $Flag_{up}$  is not set, then the precedence of the frame is upgraded to next higher level.  $Flag_{up}$  indicates



**Fig. 1** Flow chart of the proposed method

if the frame's precedence has been upgraded. Upgrading the precedence is done by changing the values in the *Data Link Precedence* subfield in the transmission header. '1 0' in *T-Bits* indicates the network uses DAP-NAD method and *First Station Number* indicate the node number that is owner of the first time slot in the new sequence. Based on the standard, the values for Urgent, Priority, and Routine are 00, 10, and 11, respectively. Therefore, 10 and 11 in the current value in the *Data Link Precedence* subfield is changed to 10 and 11, respectively. In order to prevent from over-crowded in Urgent traffics, the precedence level is upgraded once. That is, the precedence is upgraded one level higher, not more than two level. As the standard defines, when the next time slot for the node is arrived, the current Network Precedence (NP) is compared with the Frame Precedence (FP) of the pending frame in queue and if both precedences are equal, the pending frame is transmitted.

## 4 Performance Evaluations

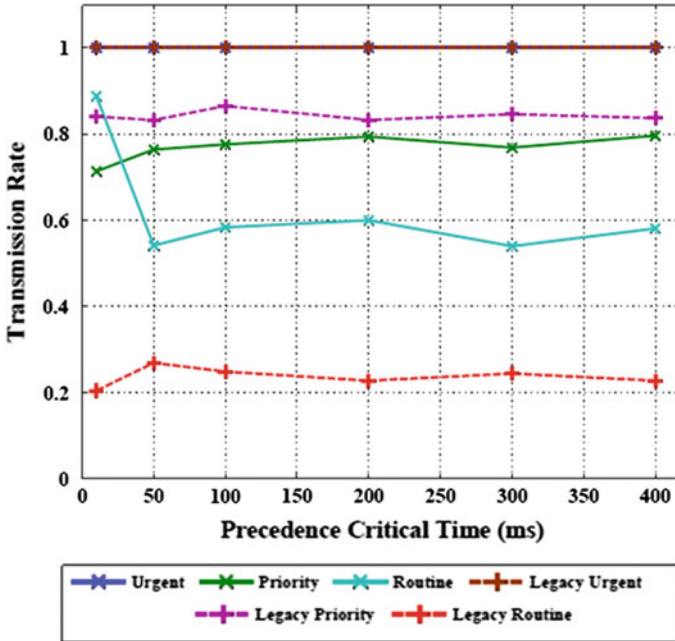
The proposed method is simulated using the simulator used in [13, 14]. Because we are targeting to the wideband systems and the latest commercial tactical communication system in [2] adopts Orthogonal Frequency Division Multiplexing (OFDM)-based system, IEEE802.11a-based OFDM system [15] is used as a physical layer for the evaluation of the proposed method. Because based on [2], the channel bandwidth is 10 MHz, the system parameters of IEEE802.11a is redefined as Table 1 in order to corresponding to 10 MHz channel bandwidth. The traffic type for this simulation is Type 3 connectionless and coupled acknowledgement operation mode defined in [1] which is similar to unicast transmission in

**Table 1** Simulation parameters

Parameter	Value
Data rate	3 Mbps
Preamble	32 us
Physical layer header	8 us
MAC header	272 bits
Default slot time	13 us
ACK packet	88 us
SIFS	32 us

IEEE802.11-based system. The channel error is not considered. In each simulation case, the number of nodes with Urgent traffic is fixed to 6 and the nodes with Priority and Routine traffics are randomly chosen between 1 and 10. The frame sizes of all traffics are fixed to 256-byte and the packet inter-arrival times of Urgent, Priority, Routine traffics are set to 0.005, 0.01, 0.05 s, respectively. As the part of evaluations, the proposed method is applied to only Routine traffics. That is, only Routine traffic is upgraded to Priority if the waiting time in queue is larger than PCT. This is to evaluate if there are some opportunities for Routine traffic even though only upgrading Routine traffic to Priority traffic.

Figure 2 shows the transmission rates as a function of PCT and the precedences of traffics. The transmission rate is defined as the ratio of successfully transmitted



**Fig. 2** Average transmission rates as a function of PCT

frames to totally generated frames. In the figure, Legacy means conventional MIL-STD-188-220-standard-based system. As shown in Fig. 3, the performances of Routine traffic using the proposed method is improved up to 3 times comparing to that using conventional DAP-NAD. On the other hand, the performances of Priority traffic is degraded up to 9 %.

## 5 Conclusion and Future Works

In this paper, a method to improve the performances of the lowest precedence traffic is proposed while minimizing the impacts on the performances of the higher precedence traffic. The method upgrades MPs of Priority and Routine frames if the waiting-time of the frames in queue is larger than a certain time, and as a consequence, the lower MP frames have the more opportunities for their transmissions. Through the simulation studies, it is proved that the proposed method improves the performances of Routine frames up to three times while the performance of Priority frame is degraded 9 % and the performance of Urgent frames is not degraded. However, the scenario might be different with different environments. Therefore, as the future works, the proposed method will be evaluated with the actual tactical traffic model and various network environments such as erroneous channel.

**Acknowledgments** This research was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (2012-0003609) and in part by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-(H0301-12-2003)).

## References

1. MIL-STD-188-220D with Change1 (2008) Digital message transfer device subsystem. 23 June 2008
2. Abacus Programming Corporation. <http://www.abacuscop.com/se.htm>
3. Thuente DJ, Borchelt TE (1997) Simulation studies of MAC algorithms for combat net radio. In: 16th military communications conference, pp. 193–199. IEEE Press, New York
4. Thuente DJ, Borchelt TE (1998) Simulation model and studies of MIL-STD-188-220A. In: 17th military communications conference, pp. 198–204. IEEE Press, New York
5. Thuente DJ, Borchelt TE (2000) Efficient data and voice media access control algorithm for MIL-STD-188-220B. In: 19th military communications conference, pp. 115–121. IEEE Press, New York
6. Thuente DJ, Whiteman JK (2001) Modified CSMA/implicit token passing algorithm for MIL-STD-188-220B. In: 20th military communication conference, pp. 838–844. IEEE Press, New York
7. Thuente DJ (2002) Improving quality of service for MIL-STD-188-220C. In: 21st military communication conference, pp. 1194–1200. IEEE Press, New York

8. Yang J, Liu Y (2006) An improved implicit token passing algorithm for DAP-NAD in MIL-STD-188-220C. In: 2nd international conference on wireless communications, pp. 1–4
9. Liu Y, An J, Liu H (2008) The modified DAP-NAD-CJ algorithm for multicast applications. In: 4th international conference on wireless communications, networking and mobile computing, pp. 1–4
10. You J, Baek I, Kang H, Choi J (2010) Effective traffic control for military tactical wireless mobile ad-hoc network. In: 6th IEEE international conference on wireless and mobile computing, networking and communications, pp. 1–8. IEEE Press, New York
11. Kim J, Kim D, Lim J, Choi J, Kim H (2011) Effective packet transmission scheme for real-time situational awareness based on MIL-STD-188-220 tactical ad-hoc networks. In: Military communication conference, pp. 956–960. IEEE Press, New York
12. Kim B-S (2010) Comparative study of MIL-STD-188-220D standard over IEEE802.11 standard. *SK Telecommun Review* 20:256–264
13. Han S, Kim B-S (2011) Evaluations on effectiveness of bump-slot over DAP-NAD-based tactical wideband wireless networks. In: Altma E, Shi W NPC 2011. LNCS, vol. 6985, pp. 341–350. Springer, Heidelberg
14. Han S, Kim B-S (2012) Efficient voice transmissions for MIL-STD-188-220-based wideband tactical systems. *IEICE Trans Commun.* E95-B, 264–2967
15. Part 11 (2007) Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, IEEE Std. 802.11, 12 June 2007

# Blind Beamforming Using the MCMA and SAG-MCMA Algorithm with MUSIC Algorithm

Yongguk Kim and Heung-Gyoon Ryu

**Abstract** Satellite communication system does not use training sequence because the satellite communication channel is similar to the additive white gaussian noise (AWGN). But, in the mobile satellite communication environment, inter-symbol-interference (ISI) seriously occurs due to movement of receiver. We must use the blind equalization for remove the ISI in mobile satellite communication. Blind equalization techniques such as MCMA and SAG-MCMA is suitable for channel equalization in the mobile satellite environment. But equalization performance of blind equalizer were not as satisfactory as expected. In this paper, we propose a blind equalization technique based on coordinate change and beamforming method in order to improve the BER performance of receiver in mobile satellite communication. The simulation results show that the proposed scheme with coordinate change need to less SNR about 1 dB to satisfy BER performance (10<sup>-5</sup>).

**Keywords** Blind equalizer · Coordinate change · MUSIC algorithm · MCMA · SAG-MCMA

## 1 Introduction

In digital communication system, it's important to transmit more information data. According to the given power, the amount of information is limited based on information theory. Channel noise and inter symbol interference (ISI) are main factors to limit amount of information. Conventional adaptive equalizations are

---

Y. Kim · H.-G. Ryu (✉)

Department of Electronic Engineering, Chungbuk National University, Cheongju, Korea  
e-mail: ecomm@cbu.ac.kr

Y. Kim

e-mail: coolfeelyg@naver.com

using the training sequence to estimate the channel characteristic. Through the channel characteristic, we estimate the characteristic coefficient of reverse channel. After that, transmit signals are passed, have a characteristic coefficient of reverse channel, the filter. Using this method, we reduce the ISI and random phase rotation influence. Therefore, the communication system can improve overall performance. Training sequence is promised signal between transmitter and receiver. In other words, training sequence is additional information. So, Bandwidth efficiency is decreased.

In the blind equalization, using the cumulative rate of received signal and using the modulus constant modulus algorithm (CMA) is represented in a way [1, 2]. Inter symbol interference (ISI) and the phase rotation can be restored at the cumulative rate method. However, it requires high-level operation. So, high speed transmission may have a problem as equalization. In the CMA, ISI and phase rotation compensate is impossible at a time. However, this method has the advantage of reduces the amount of computation. CMA equalization method for updating the equalizer coefficients, using the LMS adaptive filtering algorithm the actual implementation is very simple. LMS method the Eigen value distribution of the correlation matrix of the input signal is large; the rate of convergence is slow. CMA blind equalization algorithm is one of the most used techniques [3, 4].

MCMA(modified CMA) can compensate phase rotation problem. The MCMA accomplishes the correction of phase error and frequency offset with the modified cost functions. But, the MCMA does not judge whether the adjustment of tap coefficients is correct or not. Picchi and Prati was define the SAG (stop and go) algorithm. SAG algorithm is comparing the received signal decision error with the Sato algorithm error. If two error sign is equal, tap coefficient is updated (go). Another case, tap coefficient is not updated (stop) [5].

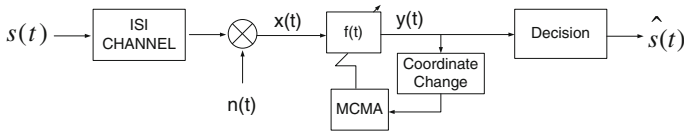
In this paper, MCMA blind equalization system using the beamforming, MUSIC algorithm and coordinate change method. Receive SNR is increased through the beamforming and the MUSIC algorithm. The propose method improve BER performance because through the coordinate change reduces the modulus and error function.

## 2 Coordinate Change for MCMA and SAG-MCMA

### 2.1 Proposed MCMA Algorithm

Figure 1 shows block diagram of the MCMA with coordinate change. We explain the coordinate change scheme in 16APSK for improving the performance of equalizer. The ratio of the inner circle and outer circle is expressed as follows.

$$\gamma = \frac{R_2}{R_1} \quad (1)$$



**Fig. 1** Block diagram of the proposed scheme in MCMA

**Table 1** Coordinate change for 16-APSK

Original coordinates	New coordinates	Original coordinates	New coordinates
$1 + i$	$1 + i$	$1 - i$	$1 - i$
$2.0153 + 2.0153i$	$1 + i$	$2.0153 - 2.0153i$	$1 - i$
$2.7529 + 0.7376i$	$1 - i$	$2.7529 - 0.7376i$	$1 + i$
$0.7376 + 2.7529i$	$-1 + i$	$0.7376 - 2.7529i$	$-1 - i$
$-1 + i$	$-1 + i$	$-1 - i$	$-1 - i$
$-2.0153 + 2.0153i$	$-1 + i$	$-2.0153 - 2.0153i$	$-1 - i$
$-2.7529 + 0.7376i$	$-i - i$	$-2.7529 - 0.7376i$	$-1 + i$
$-0.7376 + 2.7529i$	$1 + i$	$-0.7376 - 2.7529i$	$1 - i$

$\gamma$  of 16-APSK signal has a value of 2.85, each symbol has a value of  $\{\pm 1 \pm i, \pm 2.0153 \pm 2.0153i, \pm 2.7529 \pm 0.7376i, \pm 0.7376 \pm 2.7529i\}$ . Coordinate Change can be seen in Table 1.

We can get the new coordinate value using Table 1. Coordinate change of  $R'_2$  is defined as follows.

$$R'_{2,R} = \frac{E[|a'_R(t)|^4]}{E[|a'_R(t)|^2]}, R'_{2,I} = \frac{E[|a'_I(t)|^4]}{E[|a'_I(t)|^2]} \tag{2}$$

In the case of coordinate change, constant modulus values are calculated by using new coordinate values in shown Table 1.

The tap coefficients are updated through the following equation.

$$f(t + 1) = f(t) - mu(fr_R(t)e'_r(t) + jfi(t)e'_i(t))x(t) \tag{3}$$

where  $m$  is the step size.

## 2.2 Proposed SAG MCMA Algorithm

Figure 2 shows block diagram of the SAG-MCMA with coordinate change.

Coordinate change of  $R'^2$  of SAG-MCMA defined as follows.

$$R'_{2,R} = \frac{E[|s'_R(t)|^4]}{E[|s'_R(t)|^2]}, R'_{2,I} = \frac{E[|s'_I(t)|^4]}{E[|s'_I(t)|^2]} \tag{4}$$



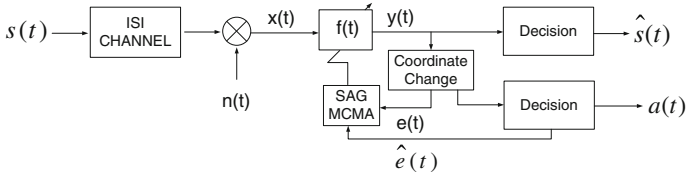


Fig. 2 Block diagram of the propose scheme in SAG-MCMA

The output signal of the equalizer is

$$y(t) = f^T(t)x(t) \tag{5}$$

The proposed error function is

$$\begin{aligned} \widehat{e}_R(t) &= y_R'(t)(y_R'(t)^2 - a_R(t)^2) \\ \widehat{e}_I(t) &= y_I'(t)(y_I'(t)^2 - a_I(t)^2) \end{aligned} \tag{6}$$

Cost function of the proposed CMA is as follows.

$$J'_{CMA}(f) = E[\{e'(t)\}^2] \tag{7}$$

The tap coefficients are updated through the following equation.

$$f(t + 1) = f(t) - mu(fr_R(t)e'_r(t) + jfi(t)e'_i(t))x(t) \tag{8}$$

### 3 Blind Beamforming System

This Fig. 3 shows the block diagram of blind beamforming system. Each element has weighting factor for beamforming. Transmit signal is passed ISI channel. After then, we detect direction of Passed signal using MUSIC algorithm. Through the detected direction, receiver elements have weighting factor for receive beamforming. Received signals are coordinate changed. Finally, signals are equalized using the SAG MCMA and MCMA algorithm.

### 4 Simulation Results

In this paper, we like to compare the proposed method with the conventional MCMA. We can find a better BER performance by blind beamforming system. We consider Table 2 for analyzing the improvement of BER. In the simulation, the ISI channel was used. SNR is 30 dB. Equalizer has 21 tabs.

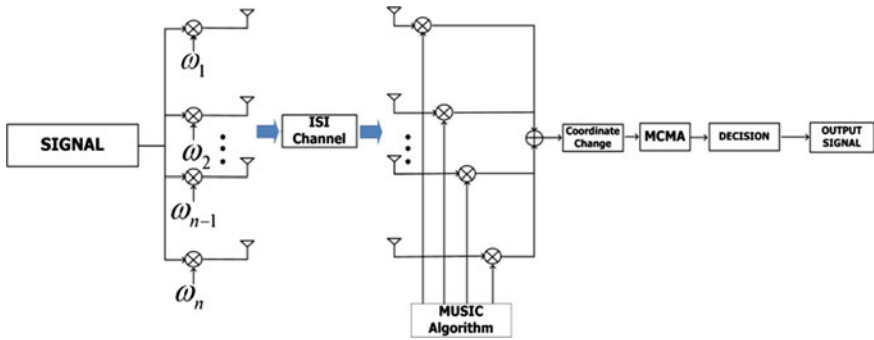


Fig. 3 Block diagram of propose system

Table 2 Simulation parameters

Modulation	16-APSK
Channel	ISI Channel [0.8, 0.3, 0, 0.2 + j0.2, 0, 0]

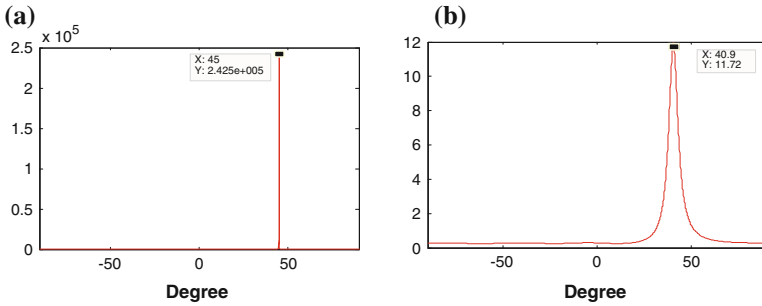
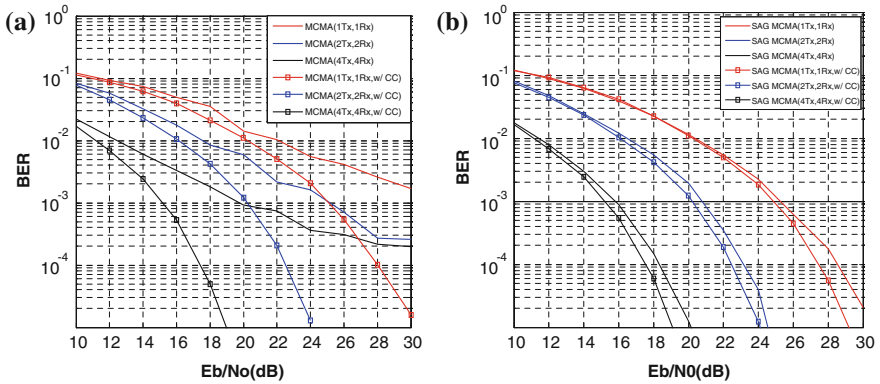


Fig. 4 MUSIC spectrum(desired angle = 45°). (a) AWGN channel (b) ISI channel

Figure 4 shows the results the result of tracking the direction of the signal in the AWGN and ISI channels. In the ISI channel, detection result has lower accuracy than in the AWGN channel. But, MUSIC algorithm can accurately detect arrival direction of receiver.

Figure 5 is shows the BER performance of the SAG MCMA and MCMA. Both MCMA and SAG MCMA, we confirm that the BER performance is improved approximate 1 dB when using coordinate change scheme. We can see that the BER performance change in MCMA is greater than in the SAG MCMA.



**Fig. 5** BER performance of the SAG MCMA and MCMA using coordinate change. (a) BER performance of MCMA (b) BER performance of SAG MCMA

### 5 Conclusion

In this paper, we propose the stop-and-go MCMA algorithm and MCMA based on the coordinate change using MUSIC algorithm for beamforming. The proposed scheme has better BER performance than that of general system in 16-APSK in shown the Fig. 5. To improve receive performance, we using the beamforming and the MUSIC algorithm. The case of the MUSIC algorithm in AWGN channel environment, detecting very accurately the direction of the signal. However, In ISI channel environment, less accurate than in the AWGN channel environment in shown Fig. 4.

**Acknowledgments** This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(No. 2012017339).

### References

1. Rao W, Yuan K-M, Guo Y-C, Yang C (2008) A simple constant modulus algorithm for blind equalization suitable for 16-QAM signal. In: The 9th international conference on signal processing, vol. 2, pp 1963–1966
2. Godard D (1980) Self-recovering equalization and carrier tracking in two dimensional data communication systems. In: IEEE Trans Commun COM-28: 1867–1875
3. Johnson CR Jr, Schniter P, Endres JT et al (1998) Blind equalization using the constant modulus criterion: a review. Proc IEEE 86(10):1927–1949
4. Rao W, Guo, Y-C (2006) New constant modulus blind equalization algorithm based on variable segment error function. J Syst Simul 19(12): 2686–2689
5. Suzuki Y, Hashimoto A, Kojima M, Sujikai H, Tanaka S, Kimura T, Shogen K (2009) A study of adaptive equalizer for APSK in the advanced satellite broadcasting system. In: Global telecommunications conference. GLOBECOM 2009. IEEE, pp 1–6

# Performance Evaluation of EPON-Based Communication Network Architectures for Large-Scale Offshore Wind Power Farms

Mohamed A. Ahmed, Won-Hyuk Yang and Young-Chon Kim

**Abstract** In order to meet the growing demand of large-scale wind power farms (WPF), integration of high reliability, high speed, cost effectiveness and secure communication networks are needed. This paper proposes the Ethernet passive optical network (EPON) as one of promising candidates for next generation WPF. Critical communication network characteristics such as reliability, mean down-time, optical power budget, path loss and network cost are evaluated and compared with conventional switched-based architectures. The results show that our proposed EPON-based network architectures are superior to conventional switched-based architectures.

**Keywords** Ethernet network · SCADA · Communication network · EPON · Reliability · Power budget · Path loss · Cost · Wind power farm

## 1 Introduction

There is a rapid development in wind farm industry around the world. Many large-scale projects are scheduled for construction in the coming years with a huge number of wind turbines. The communication networks are considered a fundamental infrastructure that enable transmission of measured information and control signals

---

M. A. Ahmed · W.-H. Yang · Y.-C. Kim  
Department of Computer Engineering, Chonbuk National University, Jeonju, Korea  
e-mail: mohamed@jbnu.ac.kr

W.-H. Yang  
e-mail: whyang@jbnu.ac.kr

Y.-C. Kim (✉)  
Smart Grid Research Center, Jeonju 561-756, Korea  
e-mail: yckim@jbnu.ac.kr

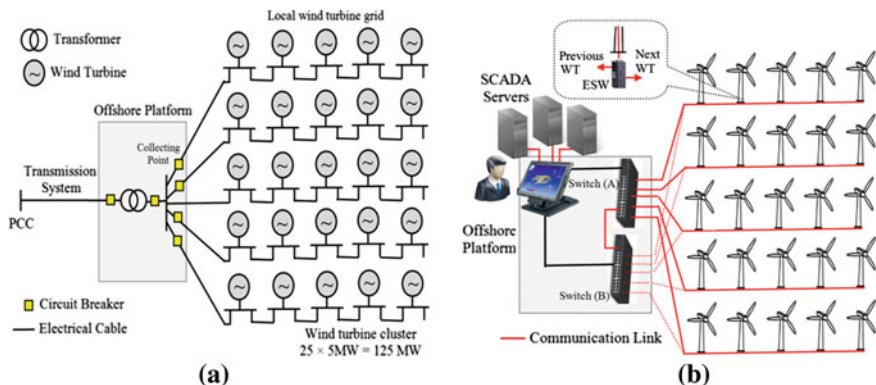
between the wind turbines and control center. Traditional communication infrastructures for monitoring the wind power farms (WPF) are based on Ethernet communication, and consist of an independent set of network switches and communication links in every wind turbine [1]. In case of network failure, serious problems may interrupt the operation, generation and control of the whole wind farm. To consider a new infrastructure for WPF communication network, critical characteristics need to be evaluated such as reliability, scalability and network cost [2].

EPON technology provides high performance data communication with a high bandwidth, flexibility, high reliability, low maintenance costs and compatibility with existing Ethernet networks. EPON could be configured with different topologies includes star, bus, tree and ring. It consists of optical line terminal (OLT), passive optical splitter (POS) and optical network unit (ONU) [3]. This paper proposes the Ethernet passive optical network (EPON) as one of promising candidates for next generation WPF.

## 2 Related Work

### 2.1 Offshore Wind Farm Layout

The offshore wind power farm consists of wind turbines, local wind turbine grid, collecting point and transmission system. The electrical layout can be designed with different configurations depending on the wind farm size and redundancy, such as radial, ring and star. Figure 1a shows an offshore wind farm consists of five radials, each of them with five turbines, connected to an offshore platform via a circuit breaker and switch. Cables with different cross section areas are used to connect the turbines. The voltage is stepped up using an offshore transformer and the transmission system transmits the total output power from the 25 turbines to shore at the point of common coupling (PCC) [4].



**Fig. 1** a Layout of wind power farm. b Conventional communication network of wind farm

## 2.2 Conventional Wind Farm Communication Network

The communication network for wind power farm defines the SCADA communication between the control center and wind turbines. This configuration usually follows the electrical topologies because the optical fibers are integrated with the submarine medium voltage cables. Furthermore, a wireless network or radio link can be incorporated into the design to increase the reliability [5].

## 3 Proposed EPON-Based Architectures

This section describes the proposed EPON-based communication network architectures for offshore wind power farms based on electrical topologies of Ref. [4]. There are four different cases designed with different number of feeders. The spacing between individual turbines is equivalent to 1.12 km within a row and between rows.

The proposed network model consists of an ONU deployed on the wind turbine side. All ONUs from different wind turbines are connected to a central OLT, placed in the control center. We considered each wind turbine have two devices; one ONU device and one POS ( $1 \times 2$ ). The ONU collects data from different devices (turbine controller, video cameras, internet telephones, etc.). The POS has two output ports as shown in Fig. 2, one port is connected to the ONU unit and the other is connected to next wind turbine. At offshore platform side, the OLT unit is installed, and connected to WTs-ONUs using feeder fiber (FF), distributed fiber (DF) and POS.

Architecture of case (1) begins with an OLT unit located at offshore platform, connected with five cables with different length of feeder fiber, (FF1  $\rightarrow$  FF5 to WT-A01  $\rightarrow$  WT-E01). Cascade splitters are used to reduce the amount of deployed fiber in the network. For all wind turbines (from WT-A01  $\rightarrow$  WT-E05), each WT has only one POS ( $1 \times 2$ ); one port is connected using DF to the next WT, while the other port is connected to the WT ONU unit. All DFs in all architectures are of 1.12 km length (the distance between WTs). Note that, POS ( $1 \times 2$ ) is used at the end of the feeder in order to help extending the network, in case of installing new WTs (one port is used, while the other is left free).

Architectures of case (2) and case (3) differ with configuration (A) with respect to number of feeder fibers, based on the electric system layout. All POS are ( $1 \times 2$ ), the same like case (1) with some interconnection between turbines such as WT-B03 to WT-C03 and WT-D01 to WT-E01. In Architecture of case (4), the WT-B01 and WT-D01 have two identical POS (primary and secondary); where POS is ( $1 \times 2$ ), with different insertion loss. Table 1 shows the details of network elements for the proposed EPON-based network architectures.

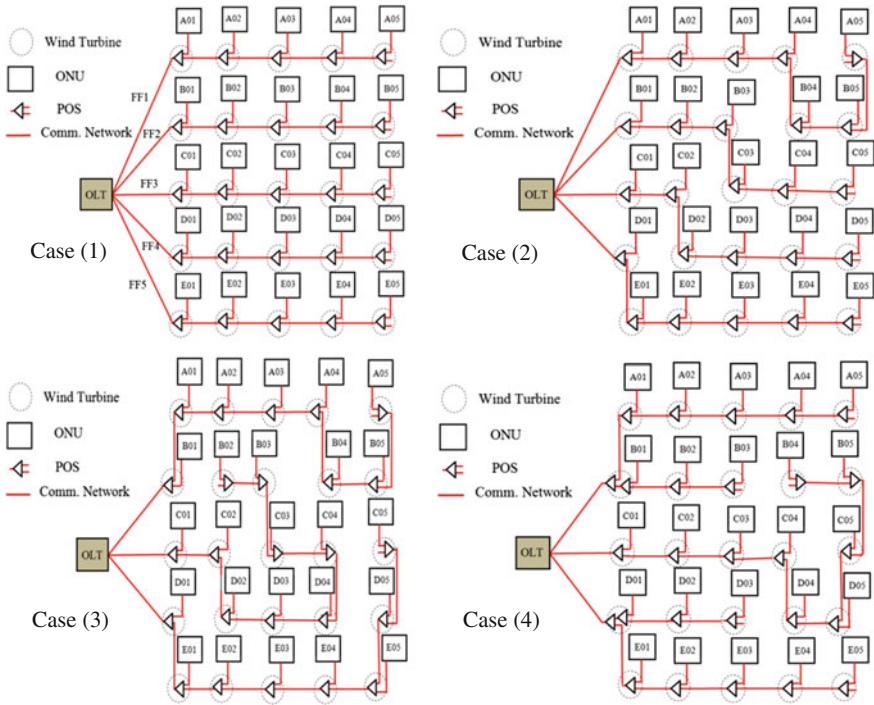


Fig. 2 Proposed wind farm communication network based on EPON

Table 1 Network elements of EPON-based large-scale WPF (unprotected)

Architecture	# OLT	FF (Km)	# POS	FF (Km)	# ONU
Case (1)	1	9.28	25 (1 × 2)	22.4	25
Case (2)	1	6.78	25 (1 × 2)	23.52	25
Case (3)	1	4.28	25 (1 × 2)	24.64	25
Case (4)	1	4.28	27 (1 × 2)	24.64	25

## 4 Performance Evaluation

### 4.1 Reliability

We studied the connection availability between the OLT located at offshore platform and each wind turbine. We consider TDM-PON architecture defined by ITU-T, with unprotected architecture in [6]. Using the failure rate of a communication network component, unavailability of a component ( $U_x$ ) is derived from its failure rate in FIT (1 FIT = 1 failure/10E09 h) and the mean repair time (MTTR) in hours. We considered that MTTR is 24 h, as all network elements located offshore and the only way

**Table 2** Component reliability of EPON-based WPF

Components	Failure rate (FIT)	MTTR (h)	Unavailability
OLT	256	24	6.144E-06
ONU	256	24	6.144E-06
Splitter (1 × 2)	50	24	1.20E-06
ESW	1250	24	3.00E-05
Fibe (/Km)	570/Km	24	1.368E-05

to access is by boat or helicopter. The expression for the connection unavailability ( $U_{EPON}$ ) for EPON-based architecture is given as follow:

$$U_{EPON} = U_{OLT} + U_{FF} + KU_{POS} + U_{DF} + U_{ONU} \tag{1}$$

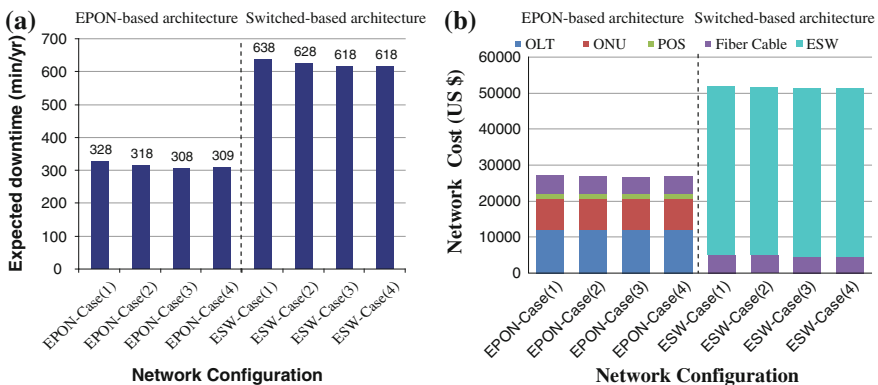
where,  $U_{OLT}$ ,  $U_{FF}$ ,  $U_{POS}$ ,  $U_{DF}$  and  $U_{ONU}$  are the unavailabilities of OLT, FF, POS, DF and ONU, respectively.  $K$  is the number of passive optical splitters. The calculations of unavailability for different network elements are shown in Table 2.

Figure 3a shows the expected downtime of EPON-based and switched-based architectures. As we can see, switched-based architectures have the highest downtime. The lowest MDT is 309 min/year for EPON-based architectures compared with 618 min/year for switched-based architectures (Table 3).

### 4.2 Network Cost

The communication network cost can be divided into active devices cost (OLT, ONU and Ethernet switch) and passive components cost (POS and fiber). Table 4 detailed the components cost used in our network model [6]. The total network cost for EPON-based and switched-based architectures can be represented as follow:

$$Cost_{EPON} = C_{OLT} + C_{FF} + C_{POS} + C_{DF} + N_{ONU} \cdot C_{ONU} \tag{2}$$



**Fig. 3** a Mean down time for WPF architectures. b Network cost for WPF architectures



**Table 3** Component cost (US \$)

OLT	ONU	Ethernet switch	Splitter (1 × 2)	Splitter (1 × 16)	Fiber (/Km)
12100	350	1800	50	800	160

**Table 4** Component insertion loss

Component	Fiber	Connector	Splitter
Attenuation	0.4 dB/Km	0.2 dB	1 × 2 (5 %:95 %) 0.4 dB 1 × 2 (50 %:50 %) 0.4 dB

$$\text{Cost}_{\text{Ethernet}} = C_{\text{ESW}} + C_{\text{FF}} + C_{\text{DF}} \quad (3)$$

where  $C_{\text{OLT}}$ ,  $C_{\text{POS}}$ ,  $C_{\text{ESW}}$  and  $C_{\text{ONU}}$  represent the component cost of OLT, POS, Ethernet switch and ONU, respectively.  $C_{\text{FF}}$  and  $C_{\text{DF}}$  represent the costs of optical fiber cable of feeder fiber and distributed fiber, respectively.  $N_{\text{ONU}}$  represents the number of WTs-ONUs.

Figure 3b shows the network cost for EPON-based and switched-based architectures in US\$. EPON-based architectures have the lowest costs by about 52 % which represent the most economic deployment solution compared with switched-based architectures which have the dominant cost of Ethernet switches.

### 4.3 Power Budget

The optical power budget is analyzed to ensure that received signal power is enough to maintain acceptable performance. The power budget for EPON specified in IEEE 802.3ah standard is 26.0 dB in case of 1000Base-PX20 for both upstream and downstream traffic. The optical budget [dB] is defined as the difference between the minimum transmitter launch power ( $P_{\text{Tx}}$ , dBm) at the input of the optical link, and the minimum sensitivity of the receiver ( $P_{\text{Rx}}$ , dBm) at the output of optical links [7].

$$\text{Power Budget} = P_{\text{Tx}} - P_{\text{Rx}} \quad (4)$$

There are many sources of attenuation including splitters ( $\text{Loss}_{\text{POS}}$ ), connections ( $\text{Loss}_{\text{conn}}$ ) and the fiber cable itself ( $\text{Loss}_{\text{fiber}}$ ). The optical path loss is equal to:

$$\text{Loss}_{\text{epon}} = \sum \text{Loss}_{\text{POS}} + \sum \text{Loss}_{\text{conn}} + \sum \text{Loss}_{\text{fiber}} \quad (5)$$

The total insertion loss must be less than the value of power budget. Table 3 shows the network elements insertion loss. A safety margin should be considered in total optical power loss calculation due to factors of aging the Tx/Rx elements and the effect of temperature. The network component insertion loss is shown in Table 3.

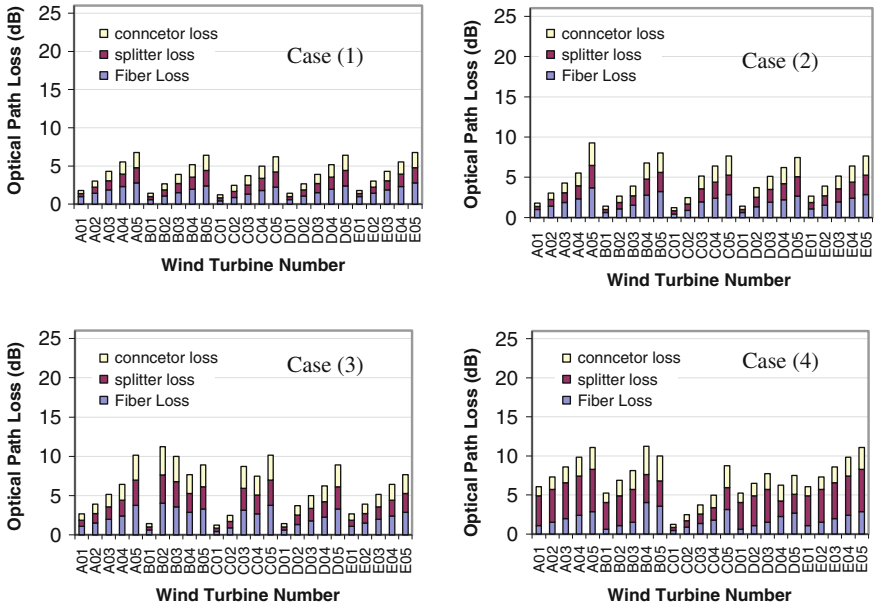


Fig. 4 Total optical path loss for WPF architectures

Figure 4 shows the total optical path loss calculation for four different configurations. The highest optical path loss value represents the farthest turbine, while the lowest value represents the nearest turbine. For example, the highest optical path loss value for WT-B02 in case (3) is about 11.23 dB, while the lowest value for WT-C01 is about 1.25 dB. Considering the IEEE 802.3 std. requirements, all EPON-based architectures satisfy the standard requirements.

### 5 Conclusion

In this paper, we proposed EPON-based network architectures for large-scale wind power farm. We evaluated the network performance in view of reliability, path loss, optical power budget and network cost. The results show that EPON-based architectures have superior performance to conventional switched-based networks. This work proves the applicability and robustness of EPON-based communication network architectures for next generation WPF.

**Acknowledgments** This work was supported by the National Research Foundation of Korea (NRF) funded by the Korea government (MEST) (2012-0009152).

## References

1. Carolsfeld R (2011) Practical experience from design and implementation of IEC 61850 based communications network in large offshore wind installation. In: Technical Meeting, CIGRE-AORC, Thailand
2. Yu R, Zhang P, Xiao W, Choudhury P (2011) Communication systems for grid integration of renewable energy resources. *IEEE Netw* 25(5):22–29
3. Kramer G (2005) Ethernet passive optical networks. McGraw-Hill, New Jersey
4. Shin J-S, Cha S-T, Wu Q, Kim J-O (2012) Reliability evaluation considering structures of a large scale wind farm. In: European power electronics (EPE) wind energy and T&D chapter seminar, Aalborg
5. Ahmed MA, Kim Y-C (2012) Network modeling and simulation of wind power farm with switched gigabit ethernet. In: 12th international symposium on communications and information technologies, Australia
6. Wosinska L, Chen J, Larsen CP (2009) Fiber access networks: reliability analysis and Swedish broadband market. *IEICE Trans Commun* E92-B(10):3006–3014
7. Huang H, Zhang H (2011) Application and analysis of long-distance EPON in transmission lines monitoring system. *Adv Mater Res* 317–319:1583–1589

# A User-Data Division Multiple Access Scheme

P. Niroopan, K. Bandara and Yeon-ho Chung

**Abstract** The conventional Interleave Division Multiple Access (IDMA) employs interleavers to separate users, while the conventional Code Division Multiple Access (CDMA) uses user specific spreading sequences for user separation. In this paper, we propose a User-Data Division Multiple Access (UDMA) scheme that employs user data as the spreading sequence for user separation with chip-by-chip iterative multiuser detection strategy. As such, this spreading sequence is not only as random as user data and independent of current symbols, but also dynamically changes from one symbol to another according to the user data. Therefore, this spreading sequence makes unwanted detection of the data by unintended receivers practically impossible. Also, in UDMA, identical interleavers are used and thus do not require to store all interleaving patterns. The simulation results show that the proposed scheme is superior to the bit error rate (BER) performance of the system in flat fading channel.

**Keywords** IDMA · Interleaver · Multiple access · Spreading

---

P. Niroopan · K. Bandara · Y. Chung (✉)  
Department of Information and Communications Engineering, Pukyong National  
University, Busan, Korea  
e-mail: yhchung@pknu.ac.kr

P. Niroopan  
e-mail: niroopan86@gmail.com

K. Bandara  
e-mail: kassae6@gmail.com

## 1 Introduction

Wireless technology has been gaining rapid popularity for some years. Multiple access schemes are the major concern for researchers that can support high data rate and high reliability for next generation wireless communications systems.

Interleave Division Multiple Access (IDMA) is a new multiple access technique [1], where interleaver that gives the name to IDMA has an important role in the system architecture. Every single user has its own interleaver which differs from others. IDMA is a special case of the CDMA system [2]. A user specific spreading sequence is used in the CDMA system where the spreading sequence must be orthogonal and also needs to maintain synchronization. It is shown that the IDMA system has the edge over the CDMA system [3].

In this paper, we propose User-Data Division Multiple Access (UDMA) systems that use user data for user separation instead of interleavers in the IDMA system. As the randomly generated user data are used as a spreading sequence, it will be more secure and less probable to be intercepted [4, 5]. In fact, this dynamically changing spreading sequence makes unwanted detection of the data by unintended receivers practically impossible. Also, we remove the repetition code that serves as a spreading sequence in the IDMA system. This repetition code is not only unsophisticated but also bandwidth inefficient. Thus, the UDMA system provides more secure and efficient communications than CDMA and IDMA systems. For the error checking in the despreading sequence at the receiver of the UDMA system, we use genetic search algorithm and Markov chain analysis [6].

## 2 IDMA System

The transmitter and receiver structures of an IDMA system with  $K$ -simultaneous users are shown in Fig. 1. At the transmitter, the block size of  $N$ -length information bits from each user- $k$  is denoted as  $d_k = [d_k(0), \dots, d_k(N-1)]^T$ ,  $k = 1, 2, \dots, K$ . The data sequence is encoded using a convolutional code into  $b_k = [b_k(0), \dots, b_k(N_C-1)]^T$ . That is, the code rate is defined as  $R_1 = N/N_C$ . Then each bit of  $b_k$  is again encoded using a low rate code such as a spread encoder with a rate of  $R_2 = 1/S_k$ , where  $S_k$  is a spreading factor. Thus, the overall code rate is  $R_1R_2$ , which produces a chip signal. The second encoder output is fed into the user specific interleaver  $(\pi_1, \pi_2, \dots, \pi_K)$  for user separation, which generates  $x_k(j)$ ,  $j = 1, 2, \dots, J$ , where  $J$  is the user frame length. The resultant signal is then transmitted through the multiple access channel. In the receiver, the received signal is given by

$$r(j) = \sum_{k=1}^K h_k x_k(j) + n(j), \quad j = 1, 2, \dots, J \quad (1)$$

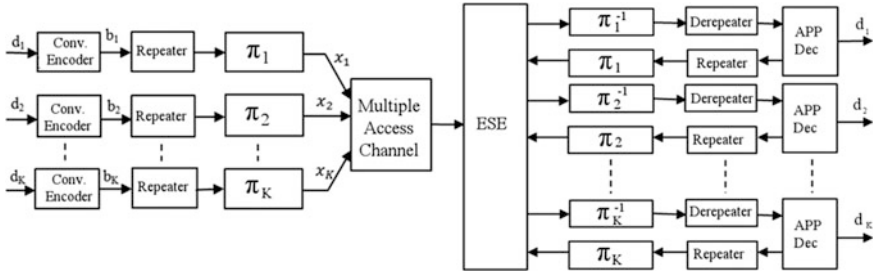


Fig. 1 The transmitter and receiver structure of the IDMA system

where  $h_k$  is the channel gain for user- $k$ ,  $x_k$  is the corresponding transmitted signal and  $n$  is the additive white Gaussian noise (AWGN) process with zero mean and variance,  $\sigma^2 = N_0/2$ . It is assumed that the channel coefficients  $\{h_k\}$  are known a priori at the receiver.

This received signal is passed to a multi-user detection (MUD) receiver that consists of an elementary signal estimator (ESE) and  $K$  a posteriori probability (APP) decoders (DECs), one for each user. The ESE performs chip-by-chip detection to roughly remove the interference among users. The outputs of the ESE and DECs are extrinsic log-likelihood ratios (LLRs) about  $\{x_k\}$  defined as

$$e(x_k(j)) = \log \left[ \frac{p(y|x_k(j) = +1)}{p(y|x_k(j) = -1)} \right] \tag{2}$$

Those LLRs are further distinguished by  $e_{ESE}(x_k(j))$  and  $e_{DEC}(x_k(j))$ , depending on whether they are generated by the ESE or DECs. For the ESE section,  $y$  in (2) denotes the received channel output while for the DECs,  $y$  in (2) is formed by the deinterleaved version of the outputs of the ESE block. These results are then combined using a turbo-type iterative process for a pre-defined number of iterations. Finally the DECs produce hard decisions on information bits for each user.

### 3 UDMA Scheme

We have developed a practical method to generate spreading sequences from the user data and regenerate them for data detection in the intended receiver. This user-data based spreader not only spreads the information bits but also separates from individual users. Note that the user data based spreading sequences are not only random but also changing dynamically from symbol to symbol.

The system model is shown in Fig. 2. In the transmitter, the user data are first encoded using the convolutional code. Then, this coded data enter the spreading sequence generator. We use shift registers for the spreading sequence. Initially, all shift registers are initialized with '1'. The all '1' sequence is multiplied by the first coded data. After outputting the first sequence from the generator, the shift

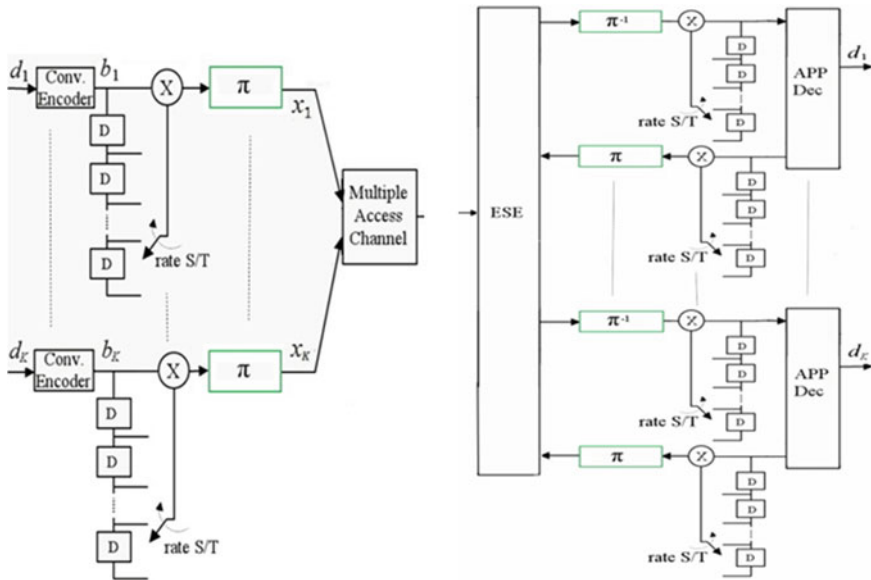


Fig. 2 The transmitter and receiver structure of the UDMA scheme

registers shift one bit right. This process continues for all data. Thus, the spreading sequences are changing dynamically according to the user data. This spread data is fed into the interleaver ( $\pi$ ) that is identical to all users, instead of user specific interleavers. The resultant signal is then transmitted through the multiple access channel.

In IDMA, user specific interleavers are used for user separation and thus the specific interleaving pattern needs to be known to the receiver. User specific spreading sequences are employed in CDMA and the spreading sequences must be orthogonal between users and have to be synchronous. In UDMA, however, all identical interleavers are used and thus do not require to store all interleaving patterns at the uplink transmission. In addition, the UDMA sequences do not require orthogonality and synchronization.

The UDMA scheme employs a chip by chip detection. Initially, we assign the same spreading sequences for spreading and despreading in the transmitter and the receiver. In the detection process, despreading sequences are provided by decoded data for each user. Severe multiple access interference and error propagation in the receiver may cause the despreading sequences to be mismatched with the spreading sequences. We consider the recovery of the spreading sequence at the receiver without a priori knowledge. The received signal strength would determine the integrity of the recovered spreading sequence and would thus affect bit decisions subsequently. For the error checking in the despreading sequence, we use genetic search algorithm and Markov chain analysis. Those algorithms help to refine despreading sequences and updates with optimization efficiently.

**Table 1** Parameters used in the simulation over flat fading channel

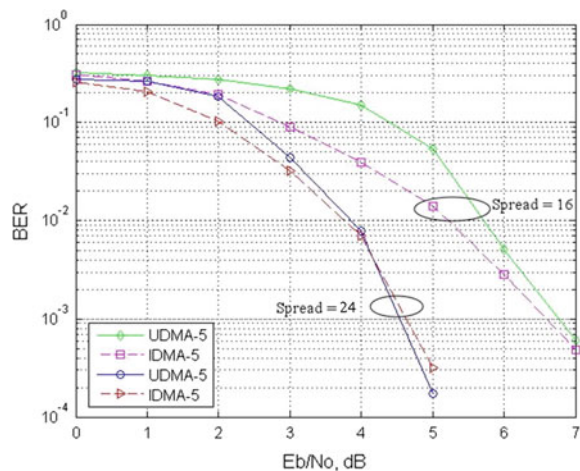
Parameters	Specifications
Number of users	5, 10
Data length	1024 bits
Encoder	Convolutional code (23, 35) <sub>8</sub>
Spreader length	16, 24
Modulation	BPSK
Iteration	10
Interleaver	Random interleaver

### 4 Simulation Results

To verify the performance of the UDMA scheme, we have conducted performance evaluation and comparative study. For a comparison, we have used the IDMA system. It is assumed that all users use the same energy level. The simulation parameters used in the simulation are given in the Table 1.

Figure 3 shows the BER performance comparison between UDMA and IDMA with 5 simultaneous users. The spreading sequence lengths used for the simulation are 16 and 24. When the spreading length is 16, the performance of the UDMA scheme shows comparable performance to the IDMA system. However, when the spreading length increases, the UDMA scheme gives better performance than the IDMA system. In Fig. 4, we further performed the evaluation with 10 simultaneous users with the spreading lengths remained unchanged. Compared with IDMA, the BER performance of the UDMA scheme is better than the IDMA system. It is important to note that the performance gain of the UDMA over the IDMA increases as the number of the user increases. Likewise, when the spreading sequence length increases, the BER performance further improves. This performance improvement stems from more sophisticated spreading and efficient detection process.

**Fig. 3** Performance comparison between UDMA and IDMA with K = 5





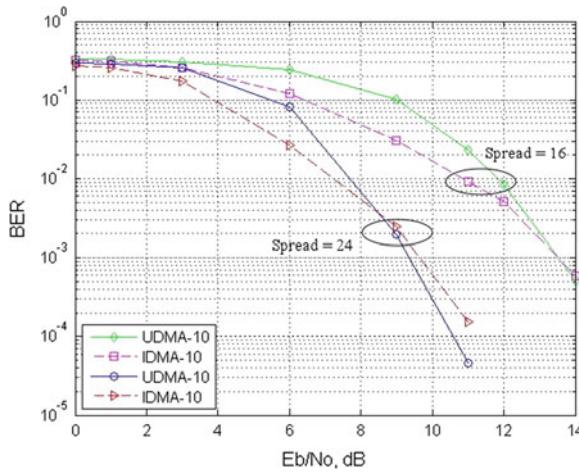


Fig. 4 Performance comparison between UDMA and IDMA with  $K = 10$

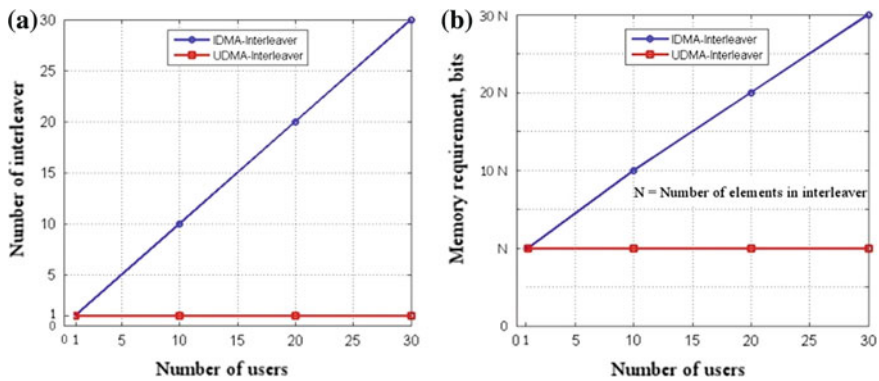


Fig. 5 Comparison of interleaver between UDMA and IDMA (a) Number of interleaver (b) Memory requirement of interleaver

Figure 5 shows comparison of interleaver between UDMA and IDMA. When the number of simultaneous user increases, number of interleaver also increases linearly in the IDMA system which requires the storage of the entire interleaving pattern for each user. This can be expensive or infeasible for applications that have limited storages when the number of users is large. But, the UDMA scheme uses identical interleaver. Thus, it uses a same interleaving pattern for all users.

## 5 Conclusion

We have proposed a User-Data Division Multiple Access (UDMA) scheme with user data based spreading sequences. Unlike IDMA and CDMA, the proposed scheme does not require the storage of all interleaving patterns and orthogonality and synchronization of spreading sequences. In addition, these spreading sequences vary dynamically from symbol to symbol according to the user data. As a result, the UDMA scheme provides enhanced security and privacy. For a BER performance comparison, the proposed scheme improves the BER performance as the spreading sequence length and the number of users increase.

## References

1. Li P, Liu L, Wu KY, Leung WK (2006) Interleave-division multiple-access. *IEEE Trans Wirel Commun* 5:938–947
2. Prasad R, Ojanpera T (1998) A survey on CDMA: evolution towards wideband CDMA. In: *Proceedings, IEEE 5th international symposium on spread spectrum techniques and applications*, vol. 1, pp 323–331 (IEEE)
3. Li P, Liu L, Wu KY, Leung WK (2003) Interleave division multiple access (IDMA) communication systems. In: *Proceedings of 3rd international symposium on turbo codes & related topics*, pp 173–180
4. Kim YS, Jang WM, Nguyen L (2006) Self-encoded TH-PPM UWB system with iterative detection. In: *The 8th international conference on advanced communication technology*, pp 710–714 (ICACT)
5. Jang WM, Nguyen L (2012) Distributed and centralized iterative detection of self-encoded spread spectrum in multi-channel communication. *J Commun Netw* 14(3):280–285 (IEEE)
6. Nguyen L, Jang WM (2008) Self-encoded spread spectrum synchronization with genetic algorithm and Markov chain analysis. In: *42nd annual conference on information sciences and systems*, pp 324–329 (CISS)

# On Channel Capacity of Two-Way Multiple-hop MIMO Relay System with Specific Access Control

Pham Thanh Hiep, Nguyen Huy Hoang and Ryuji Kohno

**Abstract** For the high end-to-end channel capacity, the amplify-and-forward (AF) scheme multiple-hop MIMO relays system is considered. The distance between each transceiver and the transmit power of each relay node are optimized to prevent some relays from being the bottleneck and guarantee the high end-to-end channel capacity. However, when the system has no control on Mac layer, the interference signal should be taken in account and then the performance of system is deteriorated. Therefore, the specific access control on MAC layer is proposed to obtain the higher end-to-end channel capacity. The optimum number of relays for the highest channel capacity is obtained for each access method. However, there is the trade-off of channel capacity and delay time.

**Keywords** Multiple-hop MIMO relays system • MAC-PHY cross layer • Optimization distance • Optimization transmit power • Specific access control • Channel capacity-delay time tradeoff • Outdated channel state information

## 1 Introduction

In order to achieve the high performance, the multiple-hop relays system is considered. [1–3]. However, in these papers the SNR at receiver(s) is assumed to be fixed and the location as well as the transmit power of each transmitter(s) are not dealt. In the multiple-hop MIMO relay system, when the distance between the source (Tx) and the destination (Rx) is fixed, the distance between the Tx to a relay

---

P. T. Hiep (✉) · R. Kohno  
Graduate School of Engineering, Yokohama National University, Yokohama, Japan  
e-mail: hiep@kohnolab.dnj.ynu.ac.jp

N. H. Hoang  
Le Quy Don Technical University, Ha Noi, Viet Nam

(RS), RS to RS, RS to the Rx called the distances between transceivers, is shorten. Consequently, according to the number of relay and the location of the relay, the SNR and the capacity are changed. Hence, to achieve the high end-to-end channel capacity, the location of each relay meaning the distance between each transceiver needs to be optimized. We have analyzed the one-way AF scheme multiple-hop MIMO relay system (MMRS) in case the interference is taken in account and optimized distance between each transceiver to obtain the high end-to-end channel capacity [4]. However, in order to achieve the higher end-to-end channel capacity when the interference is taken in account, the specific access control on Mac layer for multiple-hop MIMO relay system needs to be analyzed. In this paper, we propose the specific access control on MAC layer for one-way multiple-hop relay system and apply this method into two-way multiple-hop relay system. The proposed access control is compared to the existing method using network coding technology [5, 6]. The end-to-end channel capacity, the delay time and the relation of them is analyzed. Note that the channel capacity which is analyzed in this paper is the ergodic channel capacity. The rest of the paper is organized as follows. We introduce the concept of MMRS in Sect. 2. Section 3 shows specific access control on MAC layer. The two-way MMRS is described in Sect. 4. Finally, Sect. 5 concludes the paper.

## 2 Multiple-hop MIMO Relays System

The MMRS is described in details in [4]. However we choose some important parts to help the reader understand easier.

### 2.1 Channel Model

Let  $M$ ,  $N$  and  $K_i$  ( $i = 1, \dots, m$ ) denote the number of the antenna at the  $Tx$ ,  $Rx$  and  $RS_i$ , respectively. The distance between each transceivers is denoted by  $d_i$  ( $i = 0, \dots, m$ ). The distance between the  $Tx$  and the  $Rx$  is fixed as  $d$ . The  $Tx$  and all the relays employ amplify-and-forward strategy. Mathematical notations used in this paper are as follows.  $x$  and  $X$  are scalar variable,  $\mathbf{x}$  and  $\mathbf{X}$  are vector variable or matrix variable  $(\bullet)^H$  is conjugate transpose. In order to easily describe, the  $Tx$ ,  $Rx$  are also be denoted as the  $RS_0$  and  $RS_{m+1}$ , respectively. Since the path loss is taken into consideration, channel matrix is a composite matrix and we model as  $\sqrt{l_i}H_i, i = 0, \dots, m$ , of which  $l_i$  and  $H_i$  represent the path loss and the channel matrix between the  $RS_i$  and the  $RS_{i+1}$ , respectively.  $H_i$  is a matrix with independent and identical distribution (i.i.d.), zero mean, unit variance, circularly symmetric complex Gaussian entries. We assume that the transmit power of the  $Tx$  ( $E_{Tx}$ ), the  $Rx$  ( $E_{Rx}$ ) and the total transmit power of relays ( $E_{rs}$ ) are fixed and are not affected

by the change in the number of relays and antennas at each relay. In order to simplify the composition of relay and demonstrate the effect of optimizing the distance and the transmit power of each relay, we assume that the transmit power of each relay is equally divided into each antenna and the number of antenna in each relay is the same. Moreover, the perfect channel state information is assumed to be available and the zero forcing algorithms is applied to both the transmitter and the receiver.

### 3 Specific Access Control on MAC Layer

#### 3.1 Multiple-Phases Transmission

The transmission of each relay in the system can be divided into the multiple-phases. The relays in the same phases transmit the signal in the same time and the allocation time ( $t_i$ ). In the other phases, the relay keeps the silence or receives the signal. Since the neighbor relay transmits the signal in different phases, the interference signal is weaker than that of the system without control.

Figure 1 shows 2 phases and 3 phases transmission protocol. The 2 phases transmission protocol is explained as follows. The even-number relays and the odd-number relays transmit the signal in phase 1 and phase 2, respectively. The system has no control on MAC layer can be seem as the system with 1 phase transmission protocol. Therefore, the end-to-end channel capacity of the system with  $n$  phases can be written as

$$C = \log_2 \left( \det \left( I_M + \frac{HH^H \left( \sum_{i=0}^m l_i p_i \right)}{\sigma^2 + \sum_{i=1}^{m+1-n} l_{i-1+i} p_{i-1} + \sum_{i=0}^{m-1-n} l_{i+1+n} p_{i+1+n}} \right) \right) \quad (1)$$

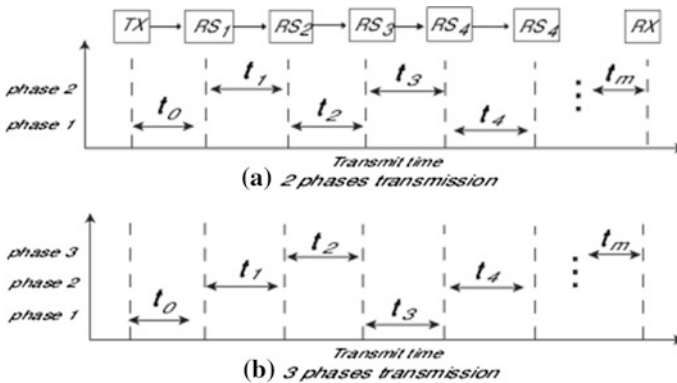


Fig. 1 2 phases and 3 phases transmission protocol

Compare the interference component of system has no control to that of the system with n phases transmission protocol (1), the distance from interference relay is longer and the number of interference relay is also larger. Hence, we can say that according to the control on MAC layer, the power of interference is decreased, thus the end-to-end channel capacity is expected to be higher.

### 3.2 Comparing to the existing method

The access control for one-way was proposed in Sec. 3.1. For two-way transmission, the transmission of downlink and uplink is assumed to alternate. Therefore, although the delay time increases 2 times, this transmission protocol can be extended for two-way transmission. The uplink end-to-end channel capacity is the same as the channel capacity of downlink in (1). We compare the proposed access method with existing once. The access method for two-way have being considered. There are some methods using the network coding technologies [5, 6]. In case of interference from 2d (2 times of distance), the transmission of all transmitters is divided into 3 phases. It means the delay time in this case is 3 s if we assume that the transmission time on each phase is 1 s. Moreover, in the proposed access method, the 1 phase method with MIMO beamforming to cancel the interference from uplink has the interference from 2d. However, it needs only 2 phases for two-way. It means the delay time is 2 s, smaller than the delay time of network coding method. Similarly, in case of interference from 3d, the delay time is 4 s for network coding method. In the proposed method, the 2 phases has the same distance of interference and the same delay time for two-way.

### 3.3 Numerical Evaluation for Proposed Access Method

In order to obtain the high end-to-end channel capacity, the distance and the transmit power should be optimized. The mathematical optimization method is explained in [4]. However, the mathematical method is complicated in case the channel model between each transceiver is different. Hence, the particle filter method is applied to optimize the distance and the transmit power simultaneously. The system parameter is summarized in Table 1.

**Table 1** Numerical parameters

Antennas at TX, RX, RS	4
Transmit power of TX (mW)	100
Transmit power of RX (mW)	10
Total transmit power of RS [mW]	100
Noise power (mW)	6.12e-011
Reflection factor	0.38
Distance between TX-RX (m)	3000

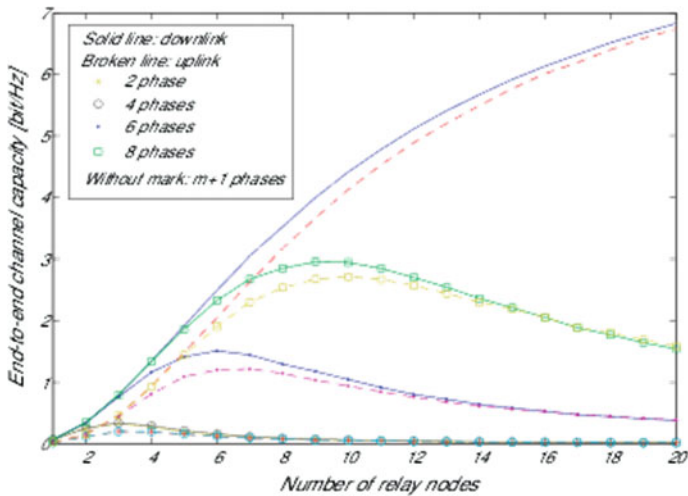


Fig. 2 The end-to-end channel capacity of two-way transmission under access control on MAC layer, the interference from both uplink and downlink

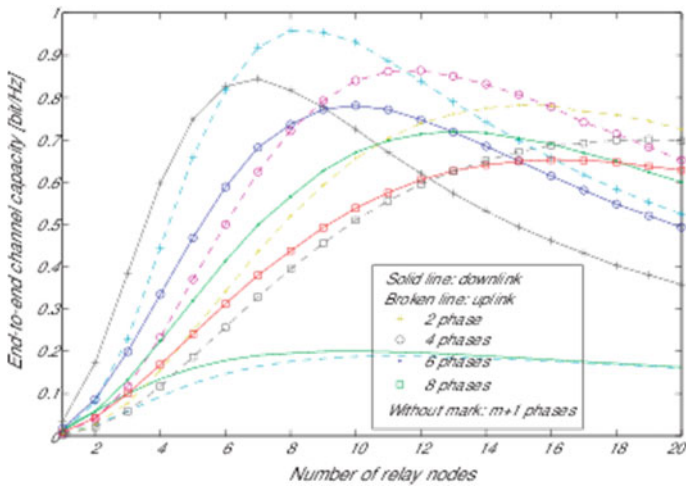


Fig. 3 The end-to-end channel capacity in case the transmission time is normalized

The end-to-end channel capacity of two-way under control on MAC layer is shown in Fig. 2. There is the optimum number of relays that has the maximum end-to-end channel capacity. In addition, the end-to-end channel capacity of the high number of phases is higher than that of the low number of phases. However, according to the transmission environment and the access method, the optimum number of relays is changed. Moreover, the allocation time for each phase was assumed as 1 s. Thus, the delay time of each access method increases when the

number of phases increases. It means that there is the trade-off between channel capacity and delay time. In case the transmission time of the system is normalized meaning the transmission time from the  $T_x$  to the  $R_x$  is 1 s, the end-to-end channel capacity is shown in Fig. 3. According to the channel model (the transmission environment, the transmit power and so on), the optimum number of relays and the number of phases is changed for the highest end-to-end channel capacity

## 4 Conclusion

The access control method on MAC layer for two-way MMRS is proposed based on the access method of one-way and compared to the existing method. There are the trade-off of channel capacity-delay time and the optimum number of relays for highest end-to-end channel capacity. According to the channel model and the number of phases, the optimum number of relays is different. In this paper, we have optimized the distance and the transmit power for each transmission protocol on MAC layer to obtain the highest end-to-end channel capacity. However, the combination of physical layer and MAC layer is not optimized. Additionally, the perfect channel state information is assumed and the ergodic channel capacity is analyzed. In the future, the system with the imperfect channel state information and the instantaneous channel capacity will be analyzed.

## References

1. Gastpar M, Vetterli M (2005) On the capacity of large Gaussian relay networks. *IEEE Trans Inf Theor* 51(3):765–779
2. Levin G, Loyka S (2010) On the outage capacity distribution of correlated keyhole MIMO channels. *IEEE Trans Inf Theor* 54(7):3232–3245
3. Peyman R, Yu W (2009) Parity forwarding for multiple-relay networks. *IEEE Trans Inf Theor* 55(1):158–173
4. Hiep PT, Kohno R (2010) Optimizing position of repeaters in distributed MIMO repeater system for large capacity. *IEICE Trans Commun* E93-B(12):3616–3623
5. Katti S, Rahul H, Hu W, Katabi D, Mdard M, Crowcroft J (2006) Xors in the air: practical wireless network coding. *Proceedings of the 2006 conference on applications, technologies, architectures, and protocols for computer communications*, vol 36, no 4, pp 243–254
6. Popovski P, Yomo H (2007) Physical network coding in two-way wireless relay channels. *IEEE International Conference on Communications (ICC07)*, pp 707–712
7. Kita N, Yamada W, Sato A (2006) Path loss prediction model for the over-rooftop propagation environment of microwave band in suburban areas (in Japanese) *IEICE Trans Commun* J89-B(2):115–125
8. Edwards HM (1997) *Graduate texts in mathematics: Galois theory*, Springer, New York



# Single-Feed Wideband Circularly Polarized Antenna for UHF RFID Reader

Pham HuuTo, B. D. Nguyen, Van-Su Tran, Tram Van  
and Kien T. Pham

**Abstract** In this paper, a single-feed wideband circularly polarized antenna has been proposed for UHF RFID reader. This antenna is designed to cover the frequency range from 860 to 960 MHz. In this antenna, the main patch is a modified form of the conventional E-shaped patch to obtain a circular polarization. A parasite patch is placed at the same layer of main patch to enhance axial ratio bandwidth. A short-circuited cylinder is also added in main patch to broaden the impedance bandwidth. The 3 dB axial ratio bandwidth is over 11 %, from 850 to 960 MHz. The impedance bandwidth is of 17 % (850–1000 MHz). Thus, The impedance bandwidth and 3 dB axial ratio bandwidth totally covers the universal UHF RFID band (860–960 MHz). The simulated and measured results indicate that the proposed antenna will be a good candidate for UHF RFID reader system.

**Keywords** Wideband · UHF · RFID · Circularly polarized antenna · LHCP

## 1 Introduction

Basic RFID systems are based on wireless communication between a reader and a tag. Antenna is one of the most important components; it will affect the performance of the whole RFID system. The operating frequencies authorized for UHF RFID

---

P. HuuTo (✉) · B. D. Nguyen · V.-S. Tran · T. Van · K. T. Pham  
School of Electrical Engineering, International University, Ho chi minh, Vietnam  
e-mail: phto@hcmiu.edu.vn

B. D. Nguyen  
e-mail: nbduong@hcmiu.edu.vn

V.-S. Tran  
e-mail: tvsu@hcmiu.edu.vn

K. T. Pham  
e-mail: ptkien@hcmiu.edu.vn

applications are varied in different countries and regions (866–869 MHz in Europe, 902–928 MHz band in North and South of America, 866–869 and 920–925 MHz in Singapore, and 952–955 MHz in Japan...). Hence, a universal UHF RFID antenna for reader is necessary to cover all UHF RFID frequency range. Since the RFID tags are always arbitrarily oriented, circularly polarized (CP) antenna is the best solution for RFID system to ensure the reliability of communications between readers and tags. It can increase orientation diversity and reduce the loss caused by the multipath effects between the reader antenna and the tag antenna. In the literature, there are several configurations proposed to create the circular polarization. The commonly used methods are rectangular patch with truncated corners and selecting suitable feed position [1], or a power splitting network to excite two orthogonal patch modes in phase quadrature [2]. However, these antennas have inherent narrow impedance bandwidth and axial ratio (AR) bandwidth (typically 1–4 %). Some designed structures allow a wide axial ratio (AR) bandwidth covering the entire ultra-high frequency (UHF) band (860–960 MHz) [3–6], while some of them show complex feeding networks.

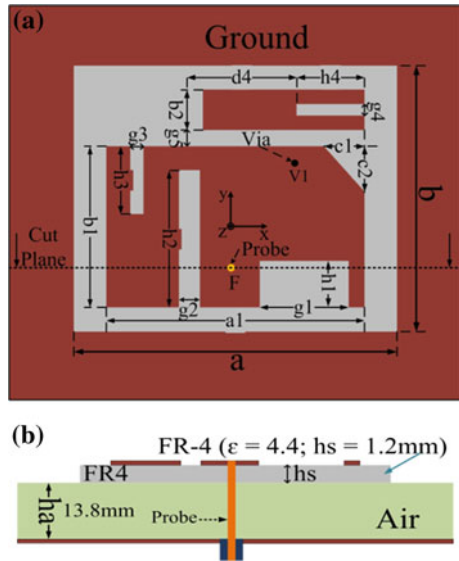
In this paper, we propose a new geometry of antenna to achieve a wide band of circular polarization. The antenna which is built on the low-cost FR-4 substrate, is a single layer with main patch and parasite patch, and fed directly by a single coaxial probe. The main patch is a modification from conventional E-antenna [7] to generate the circular polarization. A parasite patch is added on the same layer with main patch to broaden the circular polarization band. A slot and short-circuited cylinder is also added to the main patch antenna in order to widen impedance bandwidth. After this model has been designed, the return loss was measured to validate the simulation results.

## 2 RFID Antenna Design

In order to obtain a circular polarization, two conditions must be satisfied. One of them is to have two degenerated orthogonal modes. The other is that the difference of phase of two orthogonal modes is  $90^\circ$ . By this way, a conventional E-patch antenna is modified to become dissymmetric to provide a circular polarization. Fig. 1 shows the geometry of the proposed antenna for RFID. This antenna is printed on the FR-4 substrate with relative permittivity  $\epsilon_r = 4.4$ , loss tangent  $\tan \delta = 0.02$  and the substrate thickness is 1.2 mm. The medium between the FR-4 substrate and the ground is an air-layer with height of 13.8 mm. The designed antenna structure is simulated and optimized by using HFSS simulation software. The optimized geometric parameters are listed in Table 1.

The proposed geometry of this antenna is shown in Fig. 1. The main patch antenna is modified from a rectangular patch antenna (198 mm, 138 mm). The antenna is fed by a coaxial probe at position F (–37 mm, 0 mm) at the middle line of the patch. Basically, the dissymmetric dimensions of slot No. 1 ( $g_1, h_1$ ) and slot No. 2 ( $g_2, h_2$ ) leads to two orthogonal currents on the patch; hence, circularly

**Fig. 1** Geometry of the proposed UHF-RFID antenna  
**a** Top view of the antenna.  
**b** Cut-plane view of the antenna



**Table 1** Optimized parameters of the antenna

Main patch	Unit: mm	Sub-patch	Unit: mm
a1	198	g4	5
b1	138	h4	65
c1	28.5	b2	35
c2	28.5	g5	3
g1	71	d4	58
h1	36	<b>Ground plan</b>	<b>Unit: mm</b>
g2	14	a	270
h2	120	b	270
g3	8		
h3	56		

polarized fields are excited. The resonant lengths of the x and y orthogonal currents are thus dependent on the width (g) and length (h) of slots. The patch is also connected to the ground via a short-circuited cylinder with diameter of 2 mm to improve the impedance bandwidth; it is placed at the position V1 (53 mm, -37 mm). This antenna is fabricated as shown in Fig. 2 and measured by E5071CAgilent Network Analyzer.

In order to widen and tune the axial ratio band, a parasitic patch is added on the same layer with main patch. On the surface of parasite patch, a slot is also created to improve and tune the axial ratio band. The antenna is truncated at one corner with equal side lengths. The purpose of truncating is to enhance the 3 dB axial ratio band.

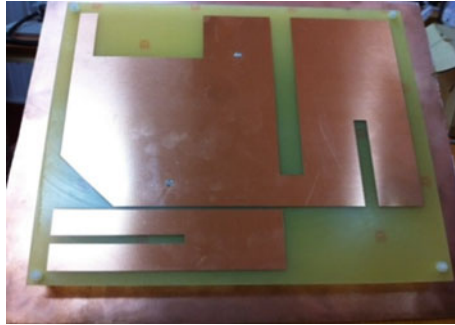


Fig. 2 The fabricated UHF-RFID antenna

Figure 3 shows the current-vector distribution on the antenna at 860 MHz for different phase states. As shown in Fig. 3, at time instant of  $0^\circ$ , the current at left side of the main patch flows in the negative x-axis, while current at right side of

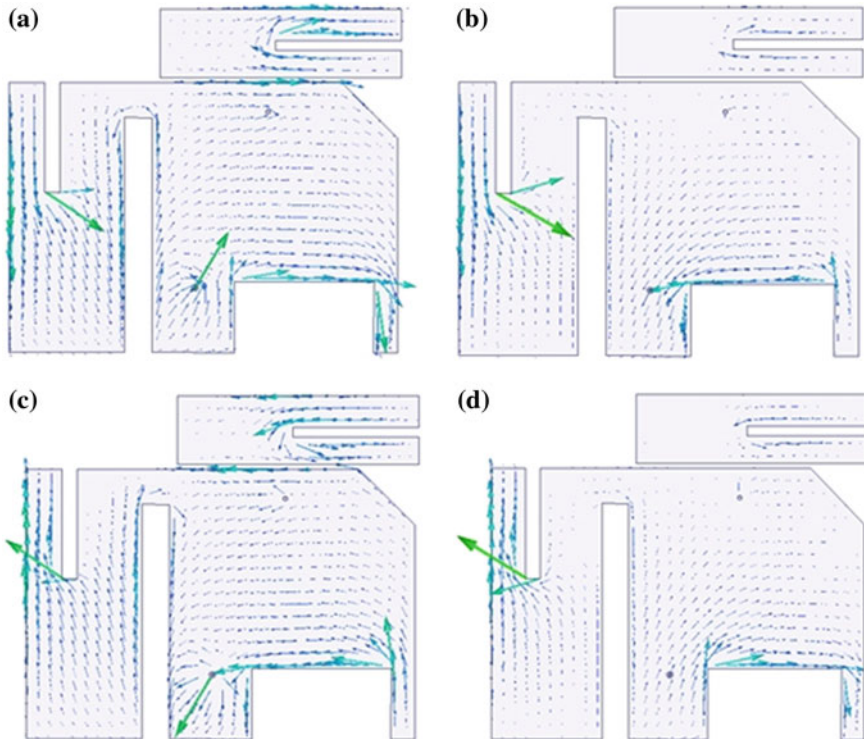
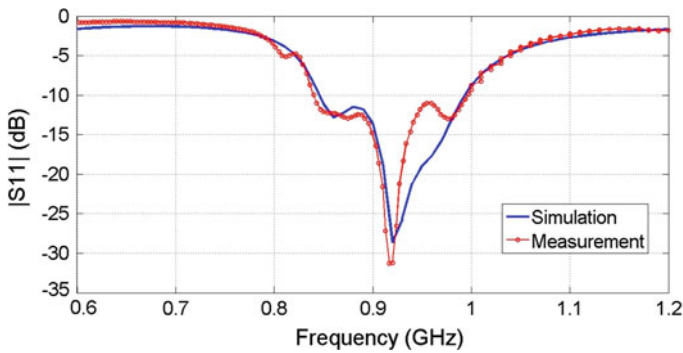
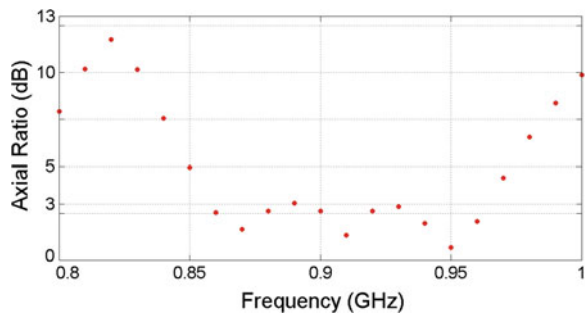


Fig. 3 The surface current distribution and orientation on the patches for the designed UHF-RFID antenna at frequency of 860 MHz at four time instants; a Time instant of  $0^\circ$ . b Time instant of  $90^\circ$ . c Time instant of  $180^\circ$ . d Time instant of  $270^\circ$



**Fig. 4** The simulated and measured return loss of the proposed antenna

**Fig. 5** The simulation axial ratio with respect to frequency



that flows in the negative  $y$ -axis. Then, at time instant of  $90^\circ$ , the current direction at right side of the patch changes to  $+y$  axis while at left side of the patch still remains its own direction. Similarly, at  $180^\circ$  and  $270^\circ$ , both currents are in opposite directions with that in case of  $0^\circ$  and  $90^\circ$  respectively. This implies a quadrature phase between the  $x$ - and  $y$ -directed currents. As a result, the current flows turning the  $x$ -axis into  $y$ -axis like a left-handed circularly polarized (LHCP). Note that the surface currents at other frequencies within the 3 dB axial ratio band are varied as functions of time in a similar manner.

### 3 Results and Discussion

Figure 4 shows the simulated and measured return loss of antenna. It is observed that the measured return loss is less than  $-10$  dB over the frequency range of 850–1000 MHz (17 %), which can easily cover the entire universal UHF RFID frequency band of 860–960 MHz.

Figure 5 illustrates the simulated axial ratio of antenna. As can be seen from it, the simulated 3 dB AR bandwidth is of 860–960 MHz or 11 %. It is able to cover the entire universal UHF RFID frequency band of 860–960 MHz.

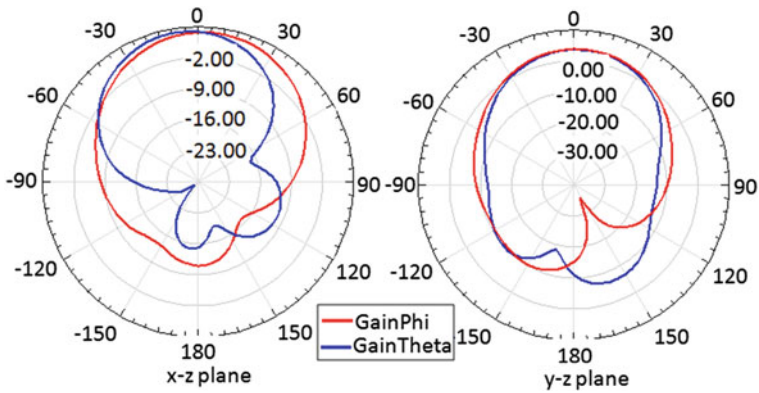


Fig. 6 Simulation radiation pattern of the UHF-RFID antenna at 860 MHz

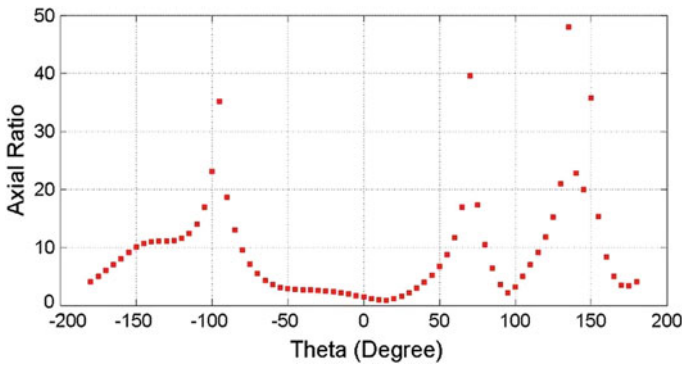
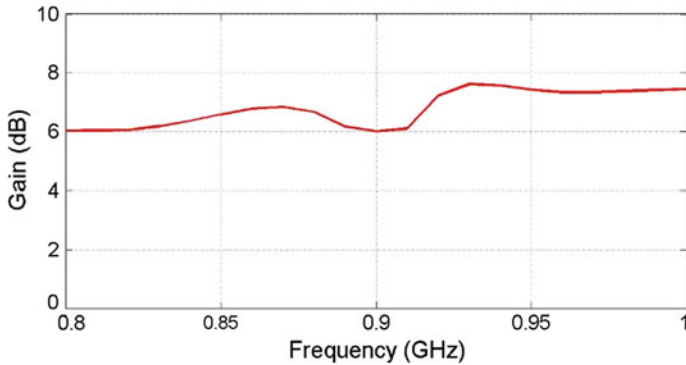


Fig. 7 The simulated axial ratio with respect to theta of the implemented antenna simulated at 860 MHz

The radiation patterns at 860 MHz are illustrated in Fig. 6 in x-z and y-z planes. In both planes, wide-angle axial ratio characteristics have been examined.

As shown in Fig. 7, the simulated 3 dB AR beam width of the implemented antenna is about 90° (from -50° to 40°).

The performance of simulation gain with respect to frequency is illustrated in Fig. 8. The value of gain varies from 6.0 to 7.6 dBic along the universal UHF RFID bandwidth, with a peak gain of 7.6 dBic at 930 MHz.



**Fig. 8** Simulation gain of the implemented UHF antenna

## 4 Conclusion

In this paper, a broadband circularly polarized antenna has been presented for UHF RFID applications. The antenna structure is fed by a single coaxial probe. By combining several techniques, the implemented antenna has achieved the desired performance over the UHF band: the return loss of 17 % or 850–1000 MHz, 3 dB axial ratio bandwidth of 11 % or 860–960 MHz, the gain of more than 6dBic. A prototype is fabricated to validate the simulation results. The measured results show that the designed antenna can provide broad impedance bandwidth of 17 % (850–1000 MHz) and prove that it is a good candidate for UHF RFID reader system.

**Acknowledgments** The authors would like to thank Advantech Jsc Company for their support of Ansoft Designer HFSS used as simulation tool to obtain these results in this paper.

## References

1. Chen W-S, Wu CK, Wong K-L (1998) Single feed square-ring microstrip antenna with truncated corner for compact circular polarization operation. *Electron Lett* 34:1045–1047
2. Targonski SD, Pozar DM (1993) Design of wideband circularly polarized aperture-coupled microstrip antenna. *IEEE Trans Antennas Propag* 41:214–220
3. Chen ZN, Qing X, Chung HL (2009) A universal UHF RFID reader antenna. *IEEE Trans Microw Theory Tech* 57(5):1275–1282
4. Lau P-Y, Yung KK-O, Yung EK-N (2010) A low-cost printed CP patch antenna for RFID smart bookshelf in library. *IEEE Trans Industr Electron* 57(5):1583–1589
5. Wang P, Wen G, Li J, Huang Y, Yang L, Zhang Q (2012) Wideband circularly polarized UHF RFID reader antenna with high gain and wide axial ration beam widths. *Prog Electromagn Res* 129:365–385

6. Kwa HW, Qing X, Chen ZN (2008) Broadband single-fed single-patch circularly polarized antenna for UHF RFID applications. In: IEEE AP-S International Symposium on Antennas and Propagation, San Diego, pp 1–4
7. Yang F, Zhang XX, Ye X, Ramat-Samii Y (2001) Wide band E-shaped patch antenna for wireless communications. *IEEE Trans Antennas Propag* 49(7):1094–1100



# Experimental Evaluation of WBAN Antenna Performance for FCC Common Frequency Band with Human Body

Musleemin Noitubtim, Chairak Deepunya and Sathaporn Promwong

**Abstract** Wireless communication systems have become important in daily life such as wireless body area network (WBAN). An ultra-wideband (UWB) technology is chosen to be used for short-range communication scenarios, low-power and high data rate technology which accommodates appropriate technology in WBAN. In this paper, we design a printed circular monopole with coplanar waveguide (CPW) fed for WBAN with common frequency band following FCC common band (7.25–8.5 GHz). The antenna structure is simple, using FR4 circuit board (PCB) with overall size of  $18 \times 20 \times 1.6 \text{ mm}^3$ . The simulation and experiment results show that the proposed antenna achieves good impedance matching and stable radiation patterns over operating bandwidth of 6.8–12.34 GHz. Moreover, the authors evaluate UWB antenna in two scenarios are free-space and with human-body to consider the antenna performance in time domain.

**Keywords** WBAN · BAN · UWB · Wideband antenna · Human body

---

M. Noitubtim (✉) · S. Promwong  
Department of Telecommunication Engineering, King Mongkut's Institute of Technology  
Ladkrabang, Bangkok, Thailand  
e-mail: m\_leemin@hotmail.com

S. Promwong  
e-mail: kpsathap@kmitl.ac.th

C. Deepunya  
Department of Electrical Engineering, King Mongkut's Institute of Technology Ladkrabang,  
Bangkok, Thailand  
e-mail: kdchaira@kmitl.ac.th

## 1 Introduction

There are increasing interests in ultra wideband (UWB) communication systems because of a radio technology with high data rate, anti-multipath interference and simple transceiver structures that is possible to make it as potentially powerful technology for low complexity, low cost communications. The power density of the UWB signal is considered to be noise for other communication systems because its power spectrum is below the noise level or part 15 limited. Therefore, UWB technology can exist with other RF technologies and can use any applications. The Federal Communications Commission (FCC) [1] in the United States allocated the fractional bandwidth  $\geq 0.2$  and having occupied bandwidth  $\geq 500$  MHz.

The regulation for UWB indoor devices defined the frequency of high band UWB and the common band is ranged from 7.25 to 8.5 GHz with power spectral density (PSD) is  $-41.3$  dBm/MHz for the UWB applications [2] indoor communication systems such as wireless body area network (WBAN). WBAN is wireless communication system that enable communications between electronic devices, that place on and/or into the human body. The systems are of great interest for various applications such as sport, multimedia, health care, and military applications [3].

The antenna is an important component for ultra wide band system. The coplanar waveguide (CPW) is very suited for patch antenna design and has been widely used [4, 5]. It also has wide bandwidth characteristic and can be easily integrated with microwave monolithic integrated circuits (MMICs). The printed circular monopole with CPW fed antenna used in WBAN fundamental requirements such as optimized characteristics in frequency and time domains small size, low profile and good on-body propagation. The proposed antenna has good return loss frequency range 6.88–12.23 GHz. It cover FCC common band (7.25–8.5 GHz).

In this paper, printed circular monopole with coplanar waveguide (CPW) fed is proposed for WBAN applications. The common parameters of the proposed antenna such as return loss, radiation patterns are shown. Moreover, the authors evaluate UWB antenna in two scenarios are free-space and with human-body to consider the antenna performance in time domain.

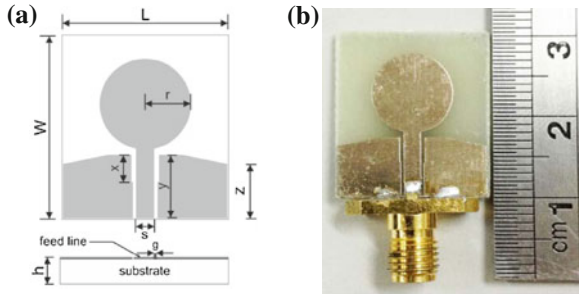
## 2 Common Parameter of the Proposed Antenna

All the simulations were carried out with CST Microwave Studio. The proposed antenna optimal parameter values are listed in Table 1. The printed circular monopole with CPW fed antenna verify in this paper depicted in Fig. 1a. The proposed antenna fabricated was show in Fig. 1b. The antenna was printed on one side of a FR4-Epoxy (PCB) substrate which has dielectric constant of 4.3, thickness of 1.6 mm and size  $18 \times 20$  mm<sup>2</sup>. The CPW transmission line is designed with 50 ohm and terminated with SMA connector for measurement purpose in this paper. In the Fig. 2 is shown simulated compare with measured return loss ( $S_{11}$ ).

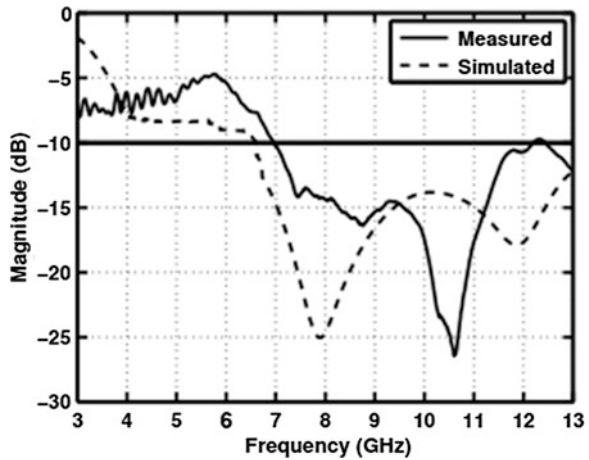
**Table 1** The optimized parameters of the printed circular monopole with CPW fed antenna

Parameter	Value (mm)	Parameter	Value (mm)
W	20	L	18
x	3	y	7
z	6	g	0.35
s	2	r	5

**Fig. 1** The printed circular monopole with CPW fed antenna. **a** Geometric and dimensions. **b** Phototype



**Fig. 2** Characteristic of the printed circular monopole with CPW fed antenna IS111



### 3 UWB-IR Transmission

#### 3.1 System Transfer Function

Friis’ transmission formula cannot be directly applied to the UWB radio as the bandwidth of the pulse is extremely wide. The complex form Friis’ Transmission Formula is extended for estimating the link budget of UWB transmission System [6–9]. Transfer function can be defined as the ratio between the voltages received

(Rx-antenna) and the voltage at the Transmitter antenna (Tx-antenna), the transfer function  $H(f)$  can simply expressed

$$H(f) = H_{tx}(f)H_f(f)H_{rx}(f) \quad (1)$$

$H_{tx}$  is transfer function of the transmitter antenna,  $H_{rx}$  is transfer function of the receiver antenna, and  $H(f)$  is free space transfer function, can be written as

$$H_f(f, d) = \frac{c}{4\pi fd} \exp(-j \frac{2\pi f}{c} d) \quad (2)$$

The transfer function  $H(f)$  can be directly obtained from measurement, and the system response can be completely determined when the transfer function is known. In this paper we used the rectangular passband transmitted waveform, which is in time domain and its spectral density as the model which is given by

$$v_t(t) = \frac{A}{f_b} [f_H \sin c(2f_H t) - f_L \sin c(2f_L t)] \quad (3)$$

$$V_t(f) = \begin{cases} \frac{A}{2f_b}, & ||f| - f_c| \leq \frac{f_b}{2} \\ 0, & ||f| - f_c| \geq \frac{f_b}{2} \end{cases} \quad (4)$$

where is the maximum amplitude,  $f_b$  is the occupied bandwidth,  $f_c$  is the center frequency,  $f_L = f_c - f_b / 2$  and  $f_H = f_c + f_b / 2$  are the minimum and maximum frequencies. Investigate the waveform occupying the entire UWB band,  $f_L = 7.25$  GHz and  $f_H = 8.5$  GHz. After knowing the channel transfer function and determining the transmitted waveforms, the receiver antenna output waveform  $v_r(t)$  is given by

$$v_r(t) = \int_{-\infty}^{\infty} H_c(f)V_t(f) \exp(j2\pi ft)df \quad (5)$$

And in case of using isotropic antennas on both sides, the receiver output waveform  $v_{r-iso}(t)$  can be written

$$v_{r-iso}(t) = \int_{-\infty}^{\infty} H_r(f)V_t(f) \exp(j2\pi ft)df \quad (6)$$

### 3.2 Power Delay Profile

The mean relative power of the taps are specified by the power delay profile (PDP) of the channel, defined as the variation of mean power in the channel and  $h(\tau)$  is the channel impulse response.

$$PDP_\tau = |h(\tau)|^2 \quad (7)$$

### 4 Experimental Setup

In Fig. 3 shows the sketch of the experimental setup. The UWB radio channel transfer function was measured as  $S_{21}$  in frequency domain by using the vector network analyzer. Measurement in frequency range from 7–11 GHz, number of frequency points are 801 and dynamic range is 80 dB. In this study the printed coplanar waveguide (CPW) antennas was used as Tx and Rx antennas. Tx-antenna was rotate start from  $0^\circ$  to  $350^\circ$  for  $10^\circ$  step. Two scenarios shows in this experiment, first take data in azimuth plane without body and second in this case takes  $S_{21}$  when port 1 or Tx-antenna place on body. The Tx-antenna was fixed on human bodies (chest) and Rx-antenna for receive signal height 1.3 m. The distance between Tx-antenna and Rx-antenna are 1 m.

### 5 Measured Results and Discussion

The radiation patterns of the E-plane and H-plane which is obtained at 7.25 GHz, 8.5 GHz and 10.25 GHz was shown Fig. 4. From the figure, note that the antenna has Omni-directional radiation pattern. The return loss from measurement is less than  $-10$  dB from 6.8–12.34 GHz covering the entire common band 7.25–8.5 GHz. The power delay profiles in free-space and with human-body are shown in Fig. 5. However, power delay profile has worse case when antenna place on-body in angle from  $75^\circ$  to  $260^\circ$ .

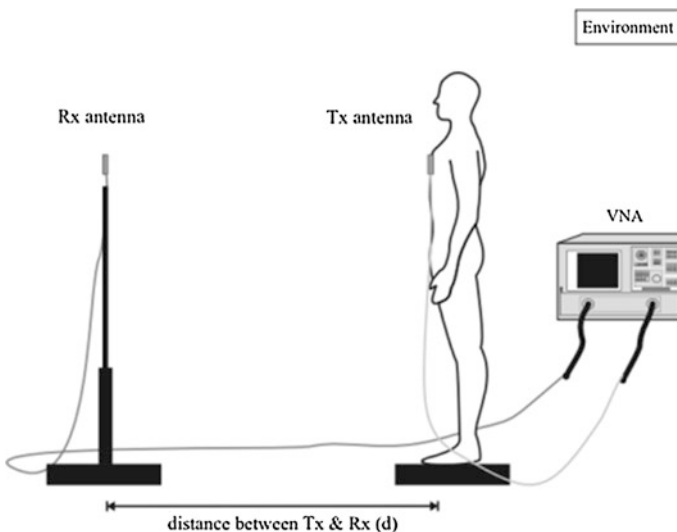
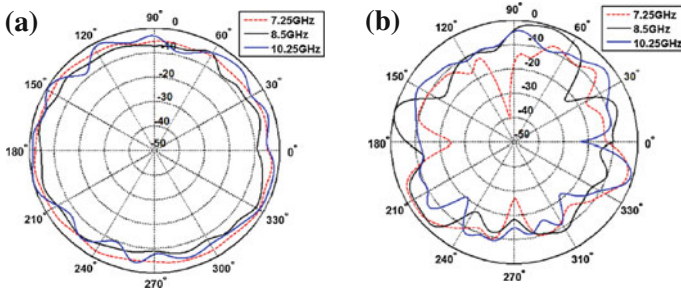
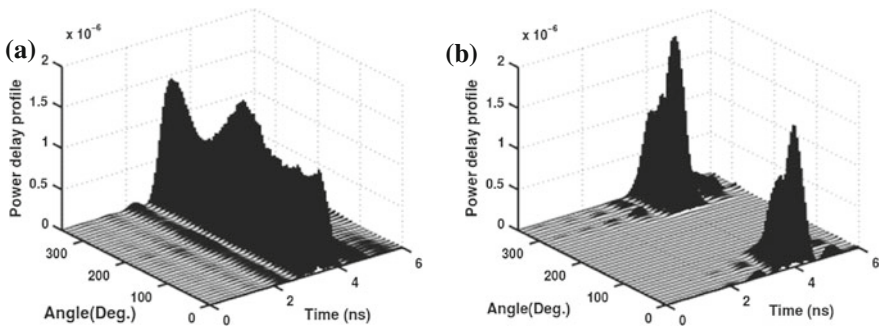


Fig. 3 Experimental setup



**Fig. 4** Radiation pattern of the printed circular monopole with CPW fed antenna. **a** E-plane. **b** H-plane



**Fig. 5** Power delay profile. **a** Free-space. **b** With human-body

## 6 Conclusion

The printed circular monopole with CPW fed antenna for common frequency band with human body is proposed antenna. The proposed antenna has structure simple and compact size covering bandwidth from 6.8 to 12.34 GHz. The result from simulations and measured show the proposed antenna has good return loss has achieved good impedance matching, Omni-directional for common frequency band with human body (7.25–8.5 GHz) WBAN applications. However, transmission performance for on-body shows not good quality in case NLOS.

## References

1. Federal communications commission: revision of part 15 of the commission's rules regarding ultra-wideband transmission systems, First report, FCC 0248, Apr (2002)
2. Ian Oppermann (2004) *UWB Theory and Applications*, Ian Oppermann, Matti Hamalainen and Jari Iinatti
3. Chen ZN (2006) *Broadband planar antennas design and applications*. Wiley, London
4. IEEE P802.15 (2010) *Wireless personal area Networks (WPANs)*. Channel model for body area network (BAN), 10 Nov
5. Natarajamani S(2009) CPW-fed octagon shape slot antenna for UWB application, International conference on microwaves. Antenna propagation and remote sensing
6. Promwong S, Supanakoon P, Takada J (2010) Waveform distortion and transmission gain due to antennas on ultra wideband impulse radio. IEICE Trans Commun E93-B:2644–2650
7. Promwong S (2008) Experimental evaluation of complex form Friis transmission formula with indoor/outdoor for ultra wideband impulse radio. Computer and Communication Engineering ICCCE 13–15 May 2008
8. Promwong S, Hachitani W, Ching GS, Takada J (2004) Characterization of ultra-wideband antenna with human body. In: International symposium on communication and information technologies, Sapporo, Japan, 21–24 Oct 2004
9. Promwong S (2005) Experimental study of UWB transmission antennas for short range wireless system. In: International symposium communications and information technology, 12–14 Oct 2005

# Performance Evaluation of UWB-BAN with Friis's Formula and CLEAN Algorithm

Krisada Koonchiang, Dissakan Arpasilp and Sathaporn Promwong

**Abstract** An ultra wideband impulse radio (UWB-IR) are developing to use in communication system and medical application because it has been an increase interest in using on body for health monitoring and body area networks (BAN). This research want to improving UWB channel propagation on body by using CLEAN algorithm for eliminate noise in channel propagation. In addition to, we use result from previous work for easier to compare performance before system. Moreover, in this paper us analysis performance of system when using CLEAN algorithm and compare without use CLEAN Algorithm by us will show BER in each position on body for analysis performance of CLEAN algorithm can be reduce noise or effect on body when without CLEAN Algorithm.

**Keywords** BAN · UWB · Impulse radio · Friis's transmission formula · CLEAN Algorithm

## 1 Introduction

Current demand in the connection network electronics device with other electronics device for convenience of use has increased. Whether the network connections within buildings. Or a network connection to the entertainment

---

K. Koonchiang (✉) · D. Arpasilp · S. Promwong  
Department of Telecommunication Engineering, Faculty of Engineering,  
King Mongkut's Institute of Technology Ladkrabang, Chalongkrung Rd,  
Ladkrabang, Bangkok 10520, Thailand  
e-mail: Boatwi@hotmail.com

D. Arpasilp  
e-mail: pangdumjungka@hotmail.com

S. Promwong  
e-mail: kpsathap@kmitl.ac.th



within the housing. The popular wireless technology used to connect such devices include WiFi, Bluetooth and shortwave technology, however, present a wide range of interesting trends about technology, Ultra wideband (UWB). Which is expected to change the data communication system with high efficacy over a WiFi or Bluetooth technology was an obvious, Ultra wideband, or UWB is likely to increase and become a standard wireless network in indoor environment. Thus, this re-research sees a critical need to consider and study the technology, Ultra wideband, which this project was to study the impact of the human body that affect the signal propagation radio impulse for Ultra wide band technology. There are a variety of modeling to analyze and compare.

The CLEAN algorithm (de-convolution algorithm) was first used to enhance the radio astronomical imaging of the sky and microwave communication. Whereat, it has been widely use in both narrowband [1] and UWB [1, 2] communication in localization and UWB biomedical imaging applications [1]. In general, the algorithm processes data by serially cancelling (i.e., cleans) the similarity between a dirty map (e.g., the measurement) and the a priori information (e.g., the template), and reconstructs the clean map (i.e., CIR) based on these detected similarities. However, CLEAN inherently assumes the channel to be non-dispersive, and that the resultant CIR is simply a summation of amplitude scaled and time-shifted versions of the a priori information. For time domain UWB channel sounding, this assumption must be considered with care when it involves probing the channel with sub-nanosecond impulses. Because of the wide spectral occupancy of these pulses, and a significant number of objects in the channel, the received signal is always severely distorted due to the frequency selectivity of the propagation phenomena, which often arise due to object's material, orientation and shape, especially for non line-of-sight (NLOS) and long-range line-of-sight (LOS) measurements.

The purpose of this paper is first to using CLEAN algorithm for improving UWB channel propagation on body and compare without use CLEAN Algorithm. So, this research has been using CLEAN algorithm from UWB-localization to UWB-BANs by implement apply CLEAN algorithm for application on body because communication system or medical want to performance of channel propagation more than directional of signal or estimate channel propagation. Where fore, we must be using CLEAN algorithm for reduce noise in UWB channel propagation (in time domain) by building new channel impulse response (CIR) or CLEAN map of the original. Specifically, we have used the results of previous research [3] but in this research change to use CLEAN for improving CIR for easier to compare in previous work. Finally, we use type of CLEAN algorithm is single-template [1] because it basic CLEAN algorithm and easy to analysis.

The rest of paper is organized as follow: in [Sect. 2](#), describes about measurement setup from previous work, [Sect. 3](#) we present some background of CLEAN algorithm, theory and analysis in channel propagation, [Sect. 4](#) result and discussion we give some show BER and receive signal wave form from match filter, [Sect. 5](#) conclusion.

## 2 Measurement Setup

This research will be implementing to experiment the UWB channel propagation on body by using vector network analyzer (VNA) in indoor-environment and we install Tx and Rx Antenna on body as well as standing position in room of experiment and antenna under test (AUT), It has been shown in previous work [3]. However, this research uses correlator in receiver side for signal distortion analysis, it shown in Fig 1.

In UWB channel propagation characterization, when we get data in frequency-domain (UWB channel) from VNA by we sent frequency range between 3–11 GHz (sampling frequency 801 point) and we processing data in matlab program for converse to time-domain by use Inverse Fourier Transform (IFT) because CLEAN technique must be process in time domain [2] or channel impulse response (CIR). In this research we generate pulse wave form from programming by generate sinc function (rectangular wave form in frequency-domain) and sent thought UWB channel, so it can be analysis power delay profile and bit error rate (BER) of channel on body.

## 3 Theory and Analysis

### 3.1 UWB Channel Impulse Response (UWB-CIR)

In this research we get data from VNA in frequency-domain (UWB channel) and we multiply by transmit signal for we want to get receive signal, the after that must be use Inverse Fourier Transform (IFT) for converse from frequency-domain to time-domain includes receive signal in time-domain show in equation.

$$V_r(f) = V_t(f)H_c(f) \tag{1}$$

$$v_r(t) = F^{-1}\{V_r(f)\} \tag{2}$$

Where is receive signal in frequency-domain, is UWB channel propagation in frequency (UWB-CIR) and is receive signal in time-domain from inverse fourier transform technique.

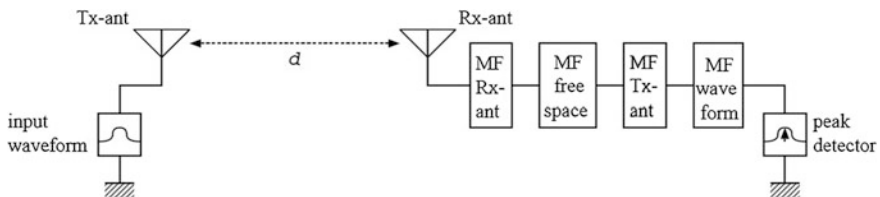


Fig. 1 Block diagram of extention Friis’s transmission formula for UWB system [4]

### 3.2 CLEAN Algorithm

In this topic, a CLEAN algorithm for UWB-CIR characterization is proposed. Although, CLEAN algorithm were used in localization [1, 2] for estimate signal or decrease of signal distortion but this paper present CLEAN algorithm for UWB-BANs application, so object of this algorithm was Improved can be reduce noise and decrease signal distortion.

The basic algorithm to process the narrowband channel was introduced in [2] from image processing to estimate details of the time of arrival (ToA) but This paper implementation for UWB-BAN by involves the computation of the correlation coefficient function and the removal and the reconstruction of detected Fig. 2.

Similarity on both the dirty and the clean maps, respectively, for each iteration, as follows:

1. Initialize normalized cross-correlation between  $v_r(t)$  and  $v_t(t)$  normalized autocorrelation of  $v_t(t)$  as  $C_{cc}(\tau) = v_r(t) \odot v_t(t)$  and  $C_{au}(\tau)$  respectively, and define the dirty and clean maps as  $d_o(\tau) = C_{cc}(\tau)$  and  $c_o(\tau) = 0$ .
2. Compute  $a_k = \max|C_{cc}(\tau)|$
3. If all  $a_k < \text{threshold}$ , go to step 7.
4. Clean the dirty map by  $d_t = d_{t-1} - (a_k \times C_{au}(\tau))$ .
5. Update the clean map by  $c_t = c_{t-1} - (a_k \times \delta(t_0))$ .
6. Go to step 2.
7. The CIR is then  $c_t = h_{\text{clean}}(t)$ .

The above algorithms assume to be independent of the generator output, the measurement system, and the antennas used. Despite accurate estimation, the aforementioned algorithms are still based on a modeled approach. Therefore, their outputs (i.e., the CIR) must be carefully interpreted.

### 3.3 Bit Error Rate

This parameter will show performance when using CLEAN algorithm for improve channel impulse response on body that show in the equation by this paper using correlation coefficient for analysis about performance of signal when sent on body, so this research will show equation as:

$$H_{\text{clean}}(f) = F\{H_{\text{clean}}(t)\} \quad (3)$$

$$V_{\text{rc}}(f) = V_t(f)H_{\text{clean}}(f) \quad (4)$$

$$v_{\text{rc}}(t) = F^{-1}\{V_{\text{rc}}(f)\} \quad (5)$$

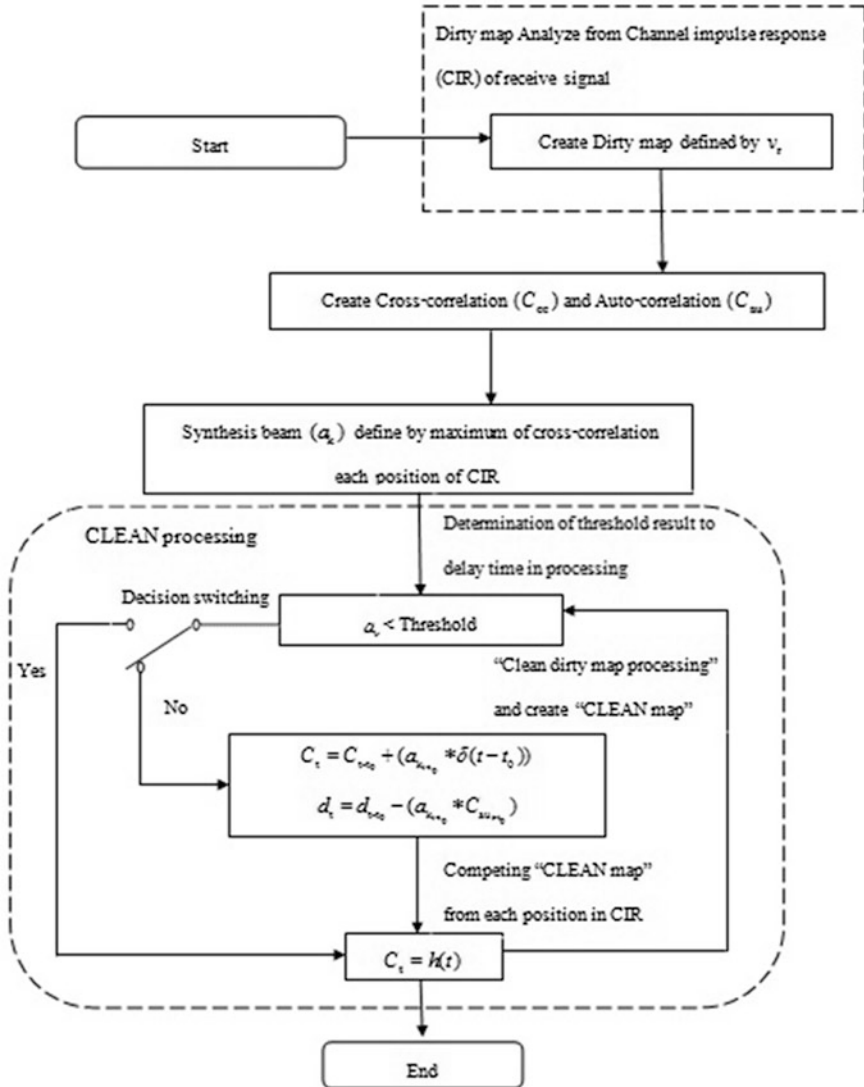


Fig. 2 Flow chart of CLEAN Algorithm for UWB system

The equation above shows  $V_{rc}(f)$  is receive signal in frequency-domain when through UWB channel from processing by using CLEAN algorithm or “CLEAN map” and  $v_{rc}(t)$  is receive signal in frequency-domain by using inverse fourier transform technique for analysis in correlation coefficient and that can be analysis bit error rate by use correlation coefficient in error function [4].

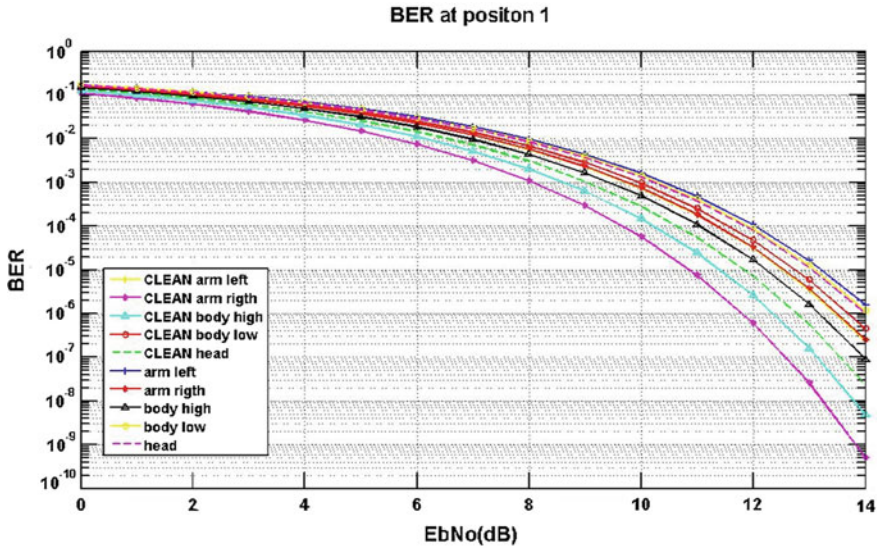


Fig. 3 Example result of bit error rate at position 1

### 4 Result and Discussion

This result will show the performance of the system by using the CLEAN technique to improve channel propagation. Figure 3 shows the BER at standing position 1, observing the BER of the signal when it passes through channel propagation created by the CLEAN technique or "CLEAN map" has a BER better than in [1]. Since the performance of the CLEAN technique can decrease signal distortion from the effect of the body or multipath, this figure shows that the position on the body has the lowest BER at the arm right and the position has the worst at the stomach area when observed. Figure 4 shows the BER at standing position 4, so farthest between Tx and Rx antenna. The best of position on the body (lowest BER) is at the head and arm right position, while the previous work [3] the position of the worst is at the head because we know CLEAN can improve the channel to reduce signal distortion.

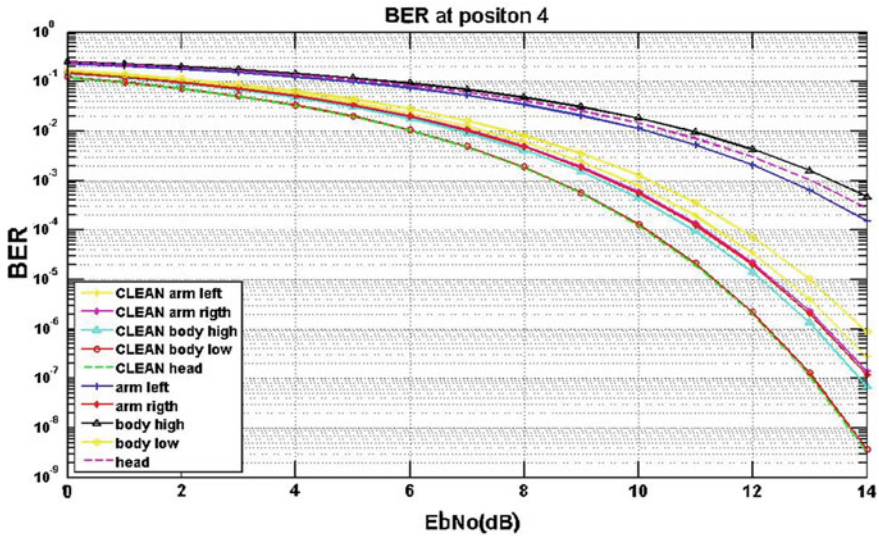


Fig. 4 Example result of bit error rate at position 4

### 5 Conclusion

This paper present improves UWB channel propagation by using CLEAN technique for solve MPC, signal distortion, noise from free space and effect on body. The resulted shows that CLEAN algorithm can be reduce noise and effect of multipath by show in PDP and solve problem of signal distortion in receiver side by show in BER when compare [3]. Specifically, CLEAN algorithm can be implementing about time of arrival (ToA) by time of receive has reduce in receiver side. However issue of CLEAN is amplitude of signal very low when compare using Extension Friis's formula and this algorithm modified from single-template CLEAN technique may be considered further.

Finally, for solve problem about amplitude of receive signal. That will present new receiver by using RAKE receiver for increase amplitude or power of signal at receiver side. Although, CLEAN algorithm has problem about amplitude of signal as very low but CLEAN technique still benefit for using on body and waiting develop.

### References

1. Liu TCK, Kim DI, Vaughan RG (2007) A high-resolution, multi-template deconvolution algorithm for time-domain UWB channel characterization. *Can J Elect Comput Eng* 32(4):207-213
2. Yang W, Naitong Z (2006) A new multi-template CLEAN algorithm for UWB channel impulse response characterization, This project was supported by the key program of national natural science foundation of China, 27-30 Nov 2006

3. Arpasin D, Narongsak M, Promwong S (2011) Experimental characterization of UWB channel model for body area networks. In: ISPACs, 7–9 Dec 2011
4. Takada J, Promwong S, Hachitani W (2003) Extension of Friis' transmission formula for ultra-wideband systems. IEICE Tech Rep WBS2003-8/MW2003-20
5. Molisch AF (2005) Ultrawideband propagation channels: theory measurement and modeling. *IEEE Trans Veh Technol* 54(5):1528–1545
6. Xia L, Redfield S, Chiang P (2011) Experimental characterization of a UWB channel for body area networks. *EURASIP J Wirel Commun Netw* 2011(703239):11 (Hindawi publishing corporation)

# A Study of Algorithm Comparison Simulator for Energy Consumption Prediction in Indoor Space

Do-Hyeun Kim and Nan Chen

**Abstract** In last couple of years many research have been done to develop the technology for minimizing the energy consumption, security and maintaining a comfortable living environment in smart buildings. In this paper, we propose comparison simulator to analyze algorithms such as averaging method, moving averaging, Low-pass filter, Kalman filter and Gray model for predicting energy consumption in indoor space. Additionally, we evaluate energy prediction algorithms in order to facilitate the testing. Our propose comparison simulator support to verify the performance of the prediction algorithms and effective estimation of energy usage in indoor environment.

**Keywords** Indoor space · Energy consumption prediction · Simulator

## 1 Introduction

Recent patterns of economic growth, worldwide energy consumption of resources, its conservation and agreements to minimize the global carbon emission and environments friendly consumption patterns are converged. Research in this connection still continues to develop the efficient technology for minimizing the energy consumption, security and maintaining a comfortable living environment in smart buildings. Such kind of building energy management systems provides energy saving effects through optimal operation of different equipment's by getting energy consumption information and comprehensively analyzing operation information of various equipment's connected to building automation system.

---

D.-H. Kim (✉) · N. Chen

Department of Computer Engineering, Jeju National University, Jeju, Korea  
e-mail: kimdh@jejunu.ac.kr

N. Chen

e-mail: xuehu001@gmail.com



Energy management and consumption in future buildings is predicted to analyze space–time form of energy consumption patterns or configures converted meaningful information in step on simple storing energy collected base of the interior or extracting statistical data. In particular, by drawing relation between energy data collected by importing ontology concept to building energy management system and space–time, user, subject etc., it is predicted on reflecting decision-making or policy for indoor energy savings or efficiency. The interior space of the building consists of floors, rooms, and hallways. Here we focus on displaying real-time energy data and demand on the map by showing room and the object in the room centrally. Demand energy expected value calculated by using predicted model and real-time energy data.

In this paper, we present algorithm comparison simulator for energy consumption prediction in indoor space. This is a part of energy information collector of indoor energy monitoring system and we focus on short-term energy consumption information using prediction algorithm for providing meaningful information based on real-time energy data.

The rest of this paper is structured as follows. In [Sect. 2](#), we will describe prediction algorithms of indoor energy consumption in detail. In [Sect. 3](#), we describe our proposed simulator and show how our design addresses the problems. Finally we conclude in [Sect. 4](#).

## 2 Prediction Algorithms of Indoor Energy Consumption

In order to show the characteristics of the set of measured values usually we use the averaging method. Averaging method is the numerical summation observations. Therefore, the averaging method considers the mid value of all the observations. All the other values of averaging method are fluctuating around this mid-point. We apply the averaging method to predict the indoor energy consumption. Equation 1 below can be used to calculate the mid value using averaging method. In this equation “ $x_k$ ” is the measurement value and “ $i$ ” is the accumulative value from 1 to “ $k$ ”. The accumulated value is divided by the number “ $k$ ”, the result is the predicted value in time “ $k + 1$ ”, “ $k$ ” is the current time and “ $x_i$ ” is the energy consumption value at time “ $i$ ”.

$$X_{k+1} = \frac{1}{k} \sum_{i=1}^k X_i \quad (1)$$

Figure 1 shows the sequence diagram of averaging method for energy prediction in indoor space. Initially import the simulated energy consumption value and then we get first stage prediction values and error values through the averaging method, prediction module and finally we correct the predicted value of the first stage through the Kalman filter to get the second stage prediction value and error values.

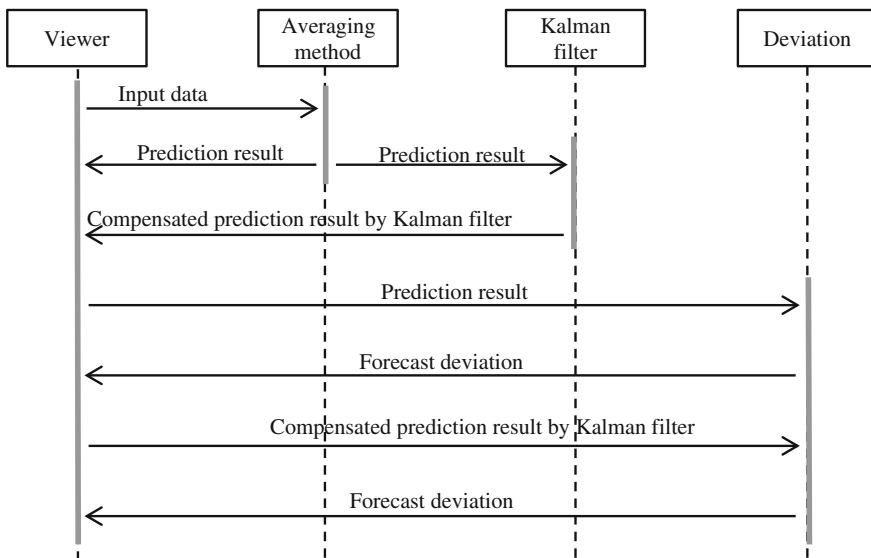


Fig. 1 Sequence diagram of energy prediction using averaging method

Disadvantage of prediction using averaging method is that, it is unable to reflect the dynamic changes in the system. In order to solve this problem the moving averaging method is introduced. The moving averaging algorithm is the modification of the averaging algorithm. The characteristic of the moving averaging method is that, it does not take all the measured value as the object and just select couple of values as objects. When it receives the latest data, the previous data will be replaced with this new data automatically. This method can be maintained at a certain number of sample data.

We calculate the moving averaging method for prediction of indoor energy consumption. At first we set “S” as move set. Then obtain a measured value “ $x_k$ ” at the present stage. Second we accumulate values from the time of k to S. ‘k’ is the sum value and it is divided by move set value “S”.

$$x_{k+1} = \frac{1}{n} \sum_{i=k-n}^k x_i \tag{2}$$

Equation 2 is the moving averaging method for prediction. Here “k” represents the current stage. The value of “ $x_i$ ” is the energy consumption value at time “i”. The value of “n” is move set of the moving averaging method. First, moving averaging method energy prediction import the simulated energy consumption value and we get first stage prediction values and error values through the moving averaging method prediction module and correct the predicted value of the first stage through the Kalman filter to get the second stage prediction value and error values. Moving average method algorithm disadvantage is that, the new and old

values use the equivalent specific gravity. Therefore, the Low-pass filter algorithm can adjust the proportion of value came into existence. High-frequency signals will be filtered out when the signal go through the low-frequency pass filter. The moving set value “ $S$ ” and weight value “ $W$ ” are adjusted. Then input the load value “ $x_k$ ” of the current time. Accumulate all values from “ $k-s$ ”, moment measured value to a measured value of the present time  $k$ . Then we get moving average method value by using the result to divide moving set value and obtain the prediction value from the proportion of adjusting this value and the current time.

$$x_{k+1} = \alpha \bar{x}_{k-1} + (1 - \alpha)x_k \quad (3)$$

Equation 3 is the Low-pass filter prediction of the prediction formula. “ $x_k$ ” is the energy consumption value at current time and “ $\alpha$ ” is the weight value.

First, Low-pass filter prediction algorithm import the simulate energy consumption value and then get first stage prediction values and error values through the Low-pass filter prediction module and correct the predicted value of the first stage through the Kalman filter to get the second stage prediction value and error values.

Kalman filter correct prediction error to get the right estimated value. Therefore, it reduces the error using the correction function of Kalman filter. As shown in Fig. 2, the Kalman filter comprises five computation phases. In Kalman model, input has the measurement value “ $Z_k$ ” and after taking a series of internal calculations automatically we get the predicted value “ $x_k$ ” as a output. In addition to the first time we run the initialization steps, while the remaining four are calculation steps.

The short-term indoor energy consumption predictions also use gray prediction algorithm. The gray model is proposed by Deng Julong who is from China Huazhong University of Science and Engineering in 1982. This model is used to predict the unknown information by the known part of the information. The Gray model can find out the law of development of the data by the collation of the raw data. The advantage of this model is that it does not require much of specimen data that is needed in case of moving averaging method. The model generates small specimen of data. Then new sequence is generated by accumulating and near averaging method sequence, generated by calculating the averaging method of adjacent elements. Figure 3 is prediction model of energy consumption using Gray model.

### 3 Simulator for Prediction Algorithms Comparison in Indoor

In order to predict the energy, load data must analyze the raw energy consumption data. We assume that the data in second interval determined by the room where energy consumption sensors are located.

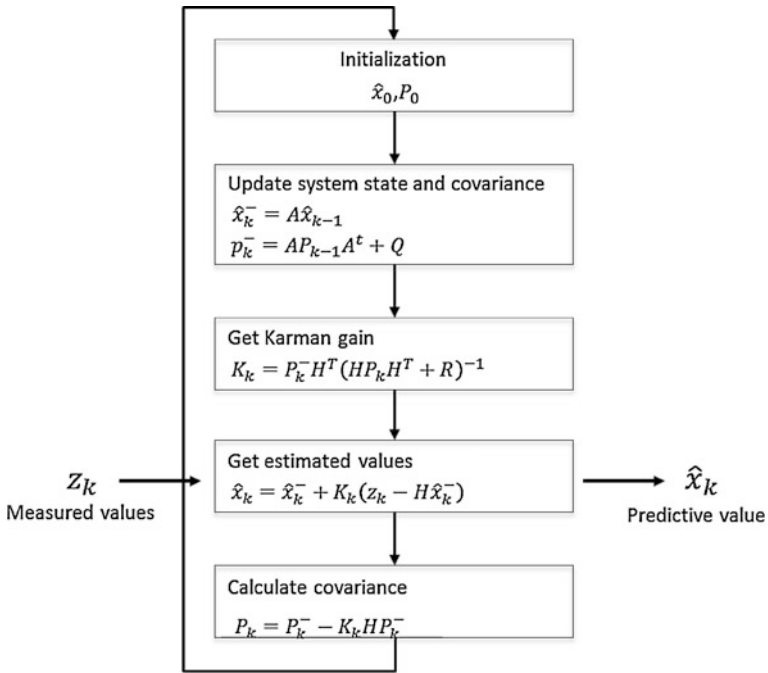


Fig. 2 Kalman filter prediction model

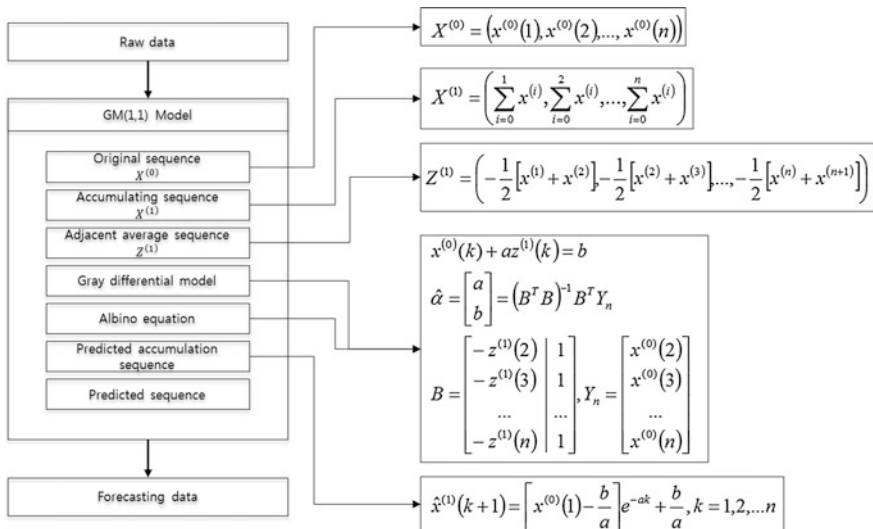


Fig. 3 Prediction model of energy consumption using Gray model

A room for the object and hour interval record load of energy 3600 s an hour, we must generate 3600 simulation energy consumption values. Room energy consumption is 5474 kW. So every second load will be 5474 kW centered fluctuated. We assume that this fluctuation is to meet the normal distribution.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

Equation 4 is a normal distribution equation.  $f(x)$  is the probability of occurrence of the “ $x$ ” events.  $\mu$  is the overall averaging method of the event and  $\sigma$  is the offset value of the specimen. The following procedure describes how the program automatically generates simulated load data.

Random energy consumption data for simulation used various prediction algorithms (averaging method, moving average method, Low-pass filtering, Kalman filter and Gray model, compensated methods by Kalman filter) and it results in a linear graph display. Following picture shows a linear plot about 5474 kW as center for 3600 data that is generated by simulating data generator.

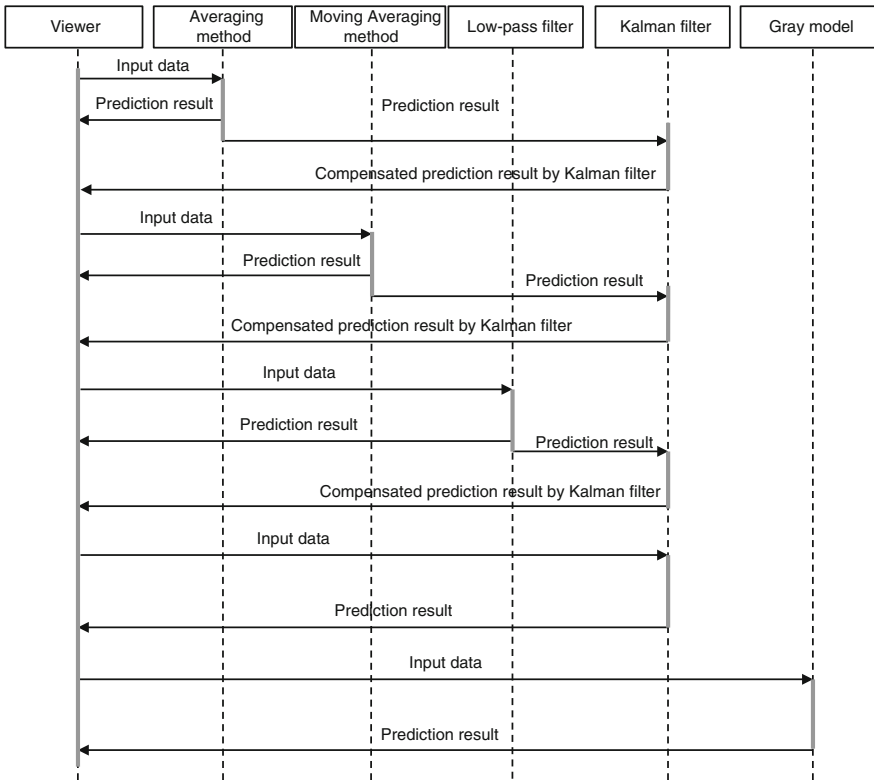


Fig. 4 Sequence diagram of the simulation of energy consumption prediction

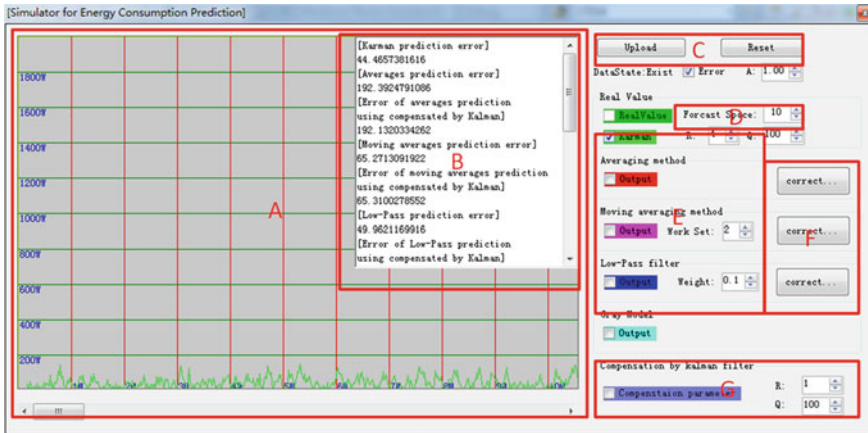


Fig. 5 Simulator for evaluating energy consumption prediction in indoor space

Figure 4 shows the data processing sequence diagram of the simulator. The simulator imports simulated data and display it in the form of a linear graph. It predicts indoor energy consumption values through the averaging method, moving average method, Low-pass filter, Kalman filter, Gray model prediction module. As the .Net environment does not have class library for dealing with the matrix. So we need to implement a matrix calculation module for matrix addition, subtraction, division and multiplication.

Figure 5 shows the indoor energy consumption performance evaluation simulator. This simulator is divided into seven areas. ‘A’ area is used to show linear diagram. Simulated data, prediction data, prediction error values are shown in an intuitive way through converting the values to a linear plot. ‘B’ area is used to display the error value of the predicted results of the prediction module. ‘C’ area is used to control the import simulation data through the “Upload” button to import a text file to get simulated data and clear it by “Reset” button. ‘D’ area is used to adjust the moving set value of the moving averaging method. ‘E’ area is used to set the parameters of each prediction model. ‘F’ area is used to correct the various modules of the prediction results through the Kalman filter. ‘G’ area is used to set the parameter “R” and “Q” of the Kalman filter.

## 4 Conclusions

In this paper, we present the simulator for comparing prediction methods of energy consumption in indoor space. Additionally, we evaluate energy prediction algorithms in order to facilitate the testing. We compare statistical prediction algorithms such as the averaging method, moving averaging method, Low-pass filter, Kalman filter, Gray model. Then we verified that prediction by using Low-pass

filter and Kalman filter demonstrate the best performance. Additionally, we have developed energy prediction simulator and energy data generation tools based on normal distribution in order to facilitate the testing.

**Acknowledgments** This work was supported by the Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy (MKE, Korea). [10038653, Development of Semantic based Open USN Service Platform]. (No. 2011-0015009). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0015009).

## References

1. Jia J, Niu D (2008) Application of improved gray markov model in power load forecasting. *Electr Util Deregul Restruct Power Technol* 1:1488–1492
2. Crossley F (2007) Advanced metering for energy supply in Australia. Energy Future Australia Pty Ltd, Australia July
3. Gu T, Pung HK, Zhang DQ (2005) Service-oriented middleware for building context-aware services. *J Netw Comput Appl* 28:1–18

# Energy Efficient Wireless Sensor Network Design and Simulation for Water Environment Monitoring

Nguyen Thi Hong Doanh and Nguyen Tuan Duc

**Abstract** In recent years, flooding in Ho Chi Minh city, Vietnam has become more and more serious, it affects much the lives and economics of the citizens. There are some reasons which caused this situation; among them the unfavorable natural conditions like rainfall, tide, are the crucial one. It is the reason why to build an intelligent network which is suitable for data collection precisely, rapidly and timely for water environment monitoring is essential. While traditional networks spend much money and effort to ensure continuous information in specific regions, Wireless Sensor Network is best known its advantages such as low cost, reliable and accurate over long term and required no real maintenance. Our contribution is that to build a network topology with suitable hardware and Low Energy Adaptive Clustering Hierarchy (LEACH) routing protocol in order to reduce energy consumption of whole network and prolong network life time. Fact has shown that LEACH protocol achieves a factor of 8–15 % energy improvement compared to direct transmissions.

**Keywords** Wireless sensor network • LEACH protocol • Energy consumption • Network lifetime

## 1 Introduction

A wireless sensor network consists of sensor nodes deployed over a geographical area for monitoring physical phenomena like water level, temperature. For a wireless sensor network design, to select a network topology and hardware are two

---

N. T. H. Doanh (✉) · N. T. Duc  
School of Electrical Engineering, Vietnam National University—International University,  
Vietnam, China  
e-mail: doanhnth04@hcmiu.edu.vn

N. T. Duc  
e-mail: ntduc@hcmiu.edu.vn



main points need to be considered for purpose of network energy efficient [1]. In traditional design, a network is designed with distance between nodes about 50–100 m. This means we have to use hundreds even thousands nodes to monitor a stage of a river. This, actually, is challenge because it required much money and time to maintain a huge system. Hence, firstly, a suitable hardware is chosen to guarantee transmission range at least from 2000 to 2500 m, saving energy and increasing network lifetime. After a long time study and test existed hardware in cutting edge technology, a table below makes a comparison some hardware [2].

	MRF24J40 MB (Microchip ZigBee)	CC1110 (Texas Instrument)	CC1120 (Texas Instrument)
Standard	IEEE 802.15.4	TI proprietary	IEEE 802.15.4 g
Data rate	250 Kbps	1.2–500 Kbps	200 Kbps
Output power	+20 dBm	10–12 dBm	+16 dBm
Received sensitivity	−102 dBm (at 250 Kbps)	−110 dBm (at 1.2 kBaud) −94 dBm (at 250 kBaud)	−123 dBm (at 1.2 kbps) −110 dBm (at 50 kbps)
Frequency band	ISM band (industrial, scientific and medical band)	868/915 MHz ISM band	868 MHz ISM band
Max. current consumption (RX/TX)	25 mA/130 mA	20.4 mA/ 36.2 mA	22 mA/45 mA

It can be concluded that the SmartRF Transceiver CC1120 evaluation board for Low Power RF transceiver devices from Texas Instrument is the best choice for environment applications.

### ***1.1 LEACH for Energy Constrained Wireless Sensor Network Hardware***

LEACH is traditional topology that helps to reduce energy consumption of whole network because of its energy efficient and simplicity. LEACH divides nodes into clusters with one node from each cluster serving as a cluster-head (CH). It randomly selects some predetermined number of nodes as cluster heads. CHs then advertise themselves and other nodes join one of those cluster heads whose signal they found strongest (i.e. the CH which is nearest to them). In this way, a cluster is formed. The CHs collect the data from their clusters and aggregate it before sending it to the other CHs or base station (BS) [3].

### 1.2 Network Topology for Water Environment Monitoring

In reliability, to build a wireless sensor network for water environment monitoring meets many difficult because of its complex geography. In addition, HCMC is locating in the lower basin of the Saigon-Dongnai river, HCMC has an area of 2093.7 km<sup>2</sup> and at 0.5–32.0 m above mean sea level. In our simulation, we assumed that there are 31 sensor nodes organized into 3 clusters. Distance between each nodes is about from 500 to 2500 m (Fig. 1).

## 2 Simulation Materials

### 2.1 Simulation Environment

I integrated LEACH code by making all of the changes as specified in the uAMPS changes package to the ns-allinone-2.34.tar.gz package, which latest version because the MIT uAMPS created by Massachusetts Institute of Technology on ns-2.e1b5 release, a very old version of the program. In order to change some parameters that are suitable for TI CC1120 hardware, we make some changes in physical layer. In MAC protocol, it is a combination of a carrier-sense multiple access (CSMA), Time division multiple access (TDMA), and a simple model of direct-sequence spread spectrum (DS-SS) to send data messages to avoid inter-cluster interference.

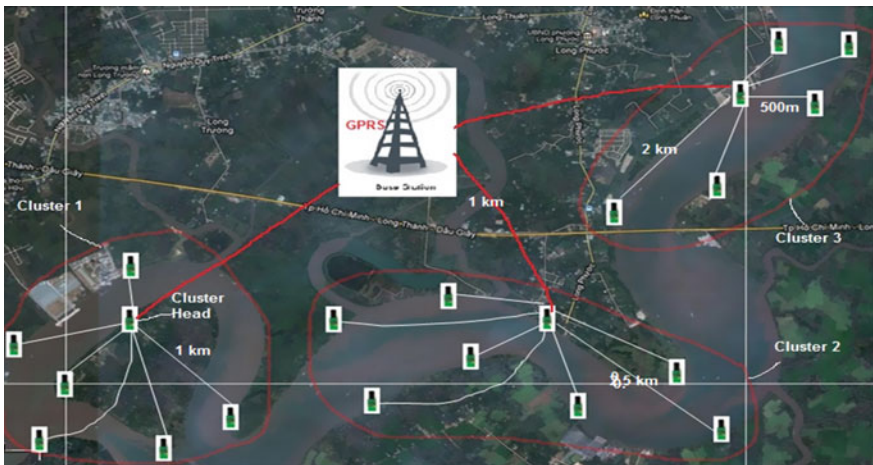


Fig. 1 Network topology for HCMC water monitoring

## 2.2 Channel Propagation Model

There are many propagation models that can describe comprehensively environment between transmitter and receiver, but we use two-ray ground propagation model is for simple observation [5]. Received power is calculated as:

$$P_r = \frac{P_t * G_t * G_r * h_t^2 * h_r^2}{d^4}$$

where  $P_r$  is the received power at distance  $d$ ,  $P_t$  is transmitted power,  $G_t$  is gain of the transmitting antenna and  $G_r$  is gain of the receiving antenna.

In order the packet is successfully detected and to avoid the collision, there are three factors should to be considered. They are: Receive threshold value (RXthresh), Carrier Sense Threshold (CStresh) and Capture Threshold (CPthresh). The CC1120 transceiver is built to be very sensitive. It is expected to decode signals with power as low as  $-123$  dBm.

$$10 \times \log\left(\frac{P_{RXthresh}}{1 \text{ mW}}\right) = -123$$

Hence,  $P_{RXthresh} = 5.0118e - 16 \text{ W}$  and  $CPthresh > = 10$ .

## 3 Simulation Analysis

In order to evaluate performance of this system, we make a comparison between direct transmission and LEACH in some parameters like energy consumption average, number of alive nodes and total end to end delay.

### 3.1 Simulation Parameters

Parameters	Value
Number of nodes	31
Number of clusters	3
Size of network	10000 m × 10000 m
BS location	6000 m × 8000 m
Initial energy of node	10 J
Propagation model	Tworayground
Time simulation	1000 s
Mac	CSMA/CA
Distance between nodes	500–2500 m

(continued)

(continued)

Parameters	Value
Frequency	868 MHz/139 MHz
CSThresh	1e-15 W (-12 dBm)
RXThresh	5.1594e-15 W (-12.3 dBm)

### 3.2 Simulation Performance and Evaluation

#### 3.2.1 Total Energy Consumption

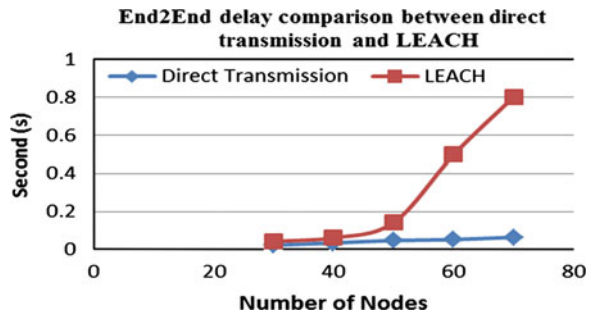
Energy consumption is the most importance factor that considered in our research. In order to dominate LEACH’s advantage, I simulate scenarios that consisted 31 nodes (Fig. 2).

As can be seen, when the time increases, the energy consumption rises up significantly. Energy consumption in direct transmission is always higher than LEACH from 0 to nearly 800 s. After that, energy consumption in former is constant because all nodes are dead while the later is continuously goes up due to LEACH’s nodes still alive. In this situation, LEACH protocol actually saved energy of whole network than other routing protocol.

#### 3.2.2 Lifetime of Network

There is a reduction in number of alive nodes in both routing protocol due to the far distance from nodes to base station. In LEACH routing protocol, sensor nodes witnessed a slight decrease while other protocol goes down strongly and ends at nearly 800 s (Fig. 3).

Fig. 2 Energy consumption average between direct transmission and LEACH



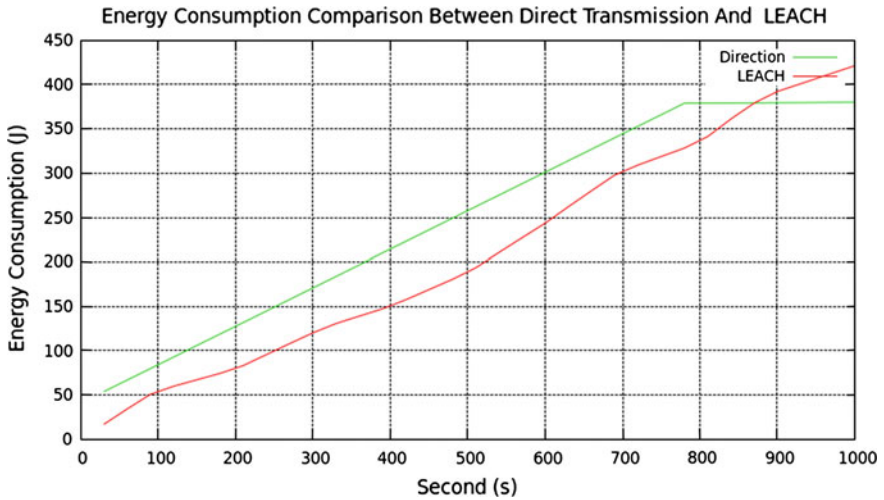


Fig. 3 Number of alive nodes comparison between direct transmission and LEACH

### 3.2.3 End to End Delay

Although LEACH protocol gives us the optimal energy consumption, end2end is not minimum in comparison with direct transmission. This is explained because LEACH has to undergo many steps before sending data to base station while direct transmission sees opposite trend. We cannot fix cluster head position because in case of far cluster head, it cannot send message to base station successfully (Fig. 4).

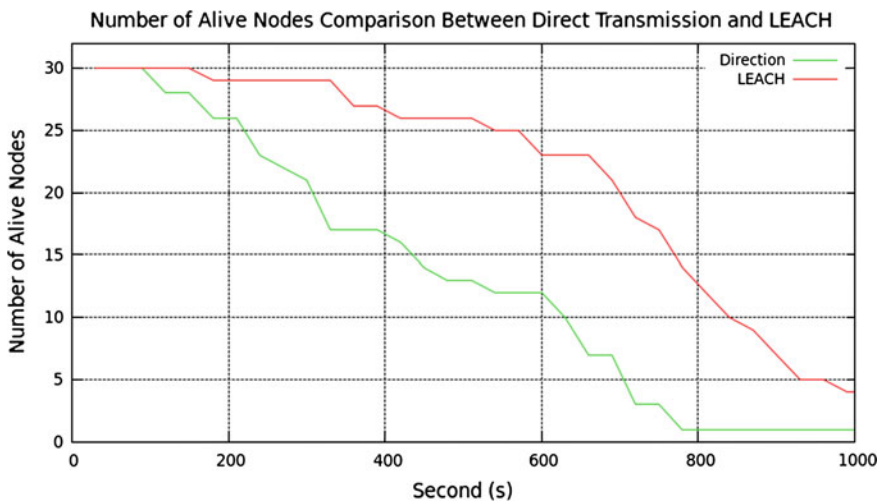


Fig. 4 End to end delay comparison between direct transmission and LEACH

## 4 Conclusion

In this paper, we integrated the wireless sensor network hardware for water environment monitoring in a large geography. For this model, LEACH protocol is also studied in order to reduce the energy consumption and prolong network lifetime of whole network.

## References

1. Padmavathy TV, Gayathri V, Indhumathi V, Karthiga G (2012) Energy constrained reliable routing optimized cluster head protocol for multihop under water acoustic sensor networks. *Int J Netw Secur Appl (IJNSA)* 4(3):57
2. Texas Instruments. High performance RF transceiver for narrowband systems, CC1110, CC1120. Datasheet
3. Tong H, Zheng J (2011) An energy and distance based clustering protocol for wireless sensor networks. 978-1-61284-307-0/11/\$26.00 ©2011IEEE
4. The LEACH code is only compatible with ns-2.1b5. It was originally developed by WendiB.Heinzelman and is no longer being updated. <http://www.mtl.mit.edu/uamps/research/cad.shtmlresearchgroups/icsystems/>
5. Rao GS, Vallikumari G (2012) A beneficial analysis of node deployment schemes for wireless sensor networks. *Int J Advanced Smart Sensor Netw Syst (IJASSN)* 2(2):33–43

# An Energy Efficient Reliability Scheme for Event Driven Service in Wireless Sensor Actuator Networks

Seungcheon Kim

**Abstract** Wireless sensor network has been evolving to wireless sensor actuator networks (WSAN), which is supposed to provide more dynamic services based on events from sensor nodes. WSANs require event reliability to support more accurate services. The optimum number of event notification messages from sensor nodes is the prerequisite for the further service reliability in WSANs. In this paper, we provide the analysis about how the number of event notification affects the energy consumption of actuator node in wireless sensor actuator networks.

**Keywords** WSAN · Event reliability · Service reliability · Energy consumption · Actuator

## 1 Introduction

Sensor network usually refers to the network that provides information services by using sensing information from sensor nodes [1]. To provide sensing information, it should have an infrastructure of sensor nodes and sink node, which is dealing with how to deliver and process the sensing information. Those efforts resulted in making Zigbee, IEEE 802.15.4, as PHY and MAC of wireless sensor network (WSN). And lots of works are concentrated in providing Internet services in WSNs with IPv6 [2].

But now WSN is evolving toward the wireless sensor actuator network (WSAN) that can react promptly to the event of an interest based on the sensing information from sensor nodes. In WSANs, actuator nodes are required to do specific actions or services after exchanging query or response messages with

---

S. Kim (✉)

Department of Information and Communication Engineering, Hansung University,  
Seoul, Korea

e-mail: kimsc@hansung.ac.kr

sensor nodes [3]. The reliability matters a lot especially in WSN, which can be categorized into the network reliability, data transmission reliability and the service reliability. Among those, service reliability means how the sink or actuator node can verify if the wanted event happens. Usually the number of event notification message is used to verify and confirm that the wanted event happens [4].

As time goes, QoS of WSN has been issued depend on the services of WSN. One of the main concerns in WSN is how to reduce the energy consumption when it exchanges sensing information. Since the energy consumption is directly connected with the life time of sensor network, nearly all the searches have been done from MAC to Application Layers.

When it comes to WSN, the energy consumption in WSN is also a crucially important matter especially in aspect of network life time. Usually actuator node uses more energy than normal sensor node since it has to perform a specific action as a reaction to the events. The more the actuator node reacts to the unwanted event, therefore, the shorter the network life time becomes [5]. We need to find out the ways to reduce the energy consumption in WSN.

This paper is focused in investigating the proper number of event notification messages that satisfies the service reliability with minimum energy consumption of actuator node in WSN and provides the scheme to reduce the energy consumption in WSN.

## 2 Wireless Sensor Actuator Network

### 2.1 Reliability in WSN

Reliability in WSN/WSAN can be categorized as Network reliability, Data transmission reliability and Service reliability.

First, the network reliability of WSN means the network stability itself. WSN/WSAN can change its shape or configurations when its purpose of service has been changed or modified. Even though those changes happen, WSN/WSAN should be operated stably.

Second, the data transmission means to deliver the sensing data safely and reliably from sensor node to sink node. For this, we need to reconsider the way the data is transmitted in wireless sensor network environments. Typically there are three methods for data delivery: End-to-End Delivery, Hop-to-Hop Delivery and Cache Mode Delivery. Those are probably compared in aspect of reliability and efficiency. And also it should be revised to deal with the transmission errors for the data transmission reliability. Considerations on the use of ACK, NAK or Timer should be done for better transmission reliability.

Lastly, the service reliability or event reliability is very far from the data transmission reliability in WSN. The event is the most important interest in WSN, where the specific reactions are related to the specific events. The service



reliability, therefore, means that the actuator node can verify whether it receives proper events from sensor nodes. Usually WSANs are composed of lots of sensor nodes. The event reliability depends on the number of event notification messages and confirming messages from sensor nodes. Sometimes sensor nodes can send a notification message on a wrong event. Therefore actuator node is recommended to react when it receives a certain number of notification messages from sensor node to identify the event.

## ***2.2 Sensor–Actuator Coordination***

The first thing we should think about in WSAN is how to assign an actuator node to the corresponding sensor nodes considering the reacting pattern of actuator node over the event. To determine the mutual relationship between sensor node and actuator node, we should look into how many actuator nodes are required to respond to the event notification and which actuator node are to be selected in the end. And also we need to categorize the event according to the task that is assigned to the actuator in WSANs.

### **2.2.1 Single-Actuator Task Versus Multi-Actuator Task**

The most general example of WSANs is the single actuator task (SAT) that is performed by single actuator on an event. In this case, the actuator normally performs a reaction towards the event after exchanging some messages with sensor nodes about the event.

On the contrary, multi actuator task (MAT) requires actuators to do a more complicated task than SAT. In this case, task for actuators might be able to be assigned according to the mutual communication between sensor node and actuator nodes and the proper reaction plan for a specific event can be changed after communication among actuator nodes.

### **2.2.2 Centralized Decision Versus Distributed Decision**

The decision making method in WSANs can be divided into centralized decision (CD) and distributed decision (DD). In CD, the event is notified to the server in the information center through sink node and the server determines how the actuator reacts to the event as shown in Fig. 1.

In DD, the actuator has the right to decide whether it will perform a reaction or not when it receives the event notification from sensor nodes. The result of reaction is reported to the server through sink node by the actuator node.

In DD, the reaction is much prompter than in CD but the actuator node takes the responsibility of wrong decision of the event, which result in shortening the life

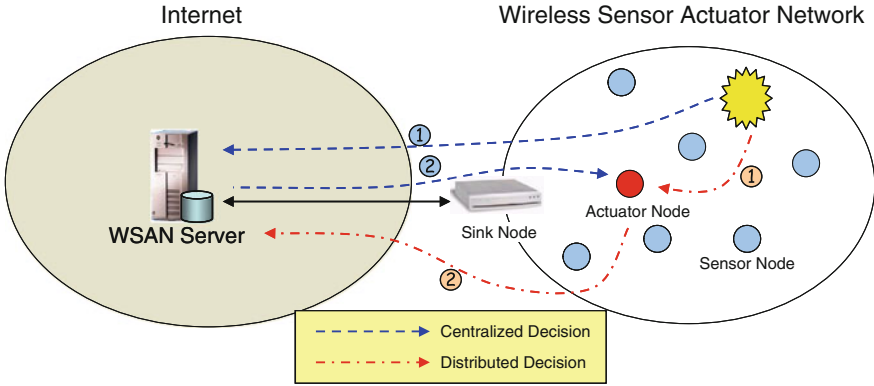


Fig. 1 Decision making in WSAN

time of actuator and WSAN since the energy consumption of reaction is much larger than the energy consumption needed for communication.

### 3 Analysis of the Event Reliability

In WSAN, the number of notification messages for an event is definitely related to the energy consumption of actuator, which can affect the whole WSAN. Here we are going to investigate the relationship between the number of notification message and the energy consumption of actuator node.

#### 3.1 Probability of Receiving $k$ Messages During $\tau$ in Actuator

Assuming that every sensor node can send only one notification message when it detects an event, the probability of receiving  $k$ -messages from  $n$ -sensor nodes is possibly calculated as follows.

$$P_e(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \text{ where } p \text{ is the event detection probability in sensor node.}$$

If we assume that the average event rate is  $\lambda$  and the interval time of each event is independent, we can use exponential distribution for the probability density function of the interval between two events in WSANs. Then, the probability of event during a specific time  $\tau$  can be described like this.

$$P_\tau = \int_0^\tau \frac{1}{\lambda} e^{-\lambda t} dt = 1 - e^{-\lambda\tau} \quad (1)$$

Therefore, the probability that the actuator could receive  $k$  messages from  $n$ -sensor nodes during  $\tau$  is finally described like this.

$$P_e(k) = \binom{n}{k} p_\tau^k (1 - p_\tau)^{n-k}, \text{ where } P_\tau = 1 - e^{-\lambda\tau} \quad (2)$$

### 3.2 Power Consumption in Actuator

The power consumption in actuator ( $E_{av}$ ) is composed of the energy ( $E_{act}$ ) that is needed for action for specific service and the energy ( $E_c$ ) for exchanging data with sensor nodes.

$$E_{av} = E_{act} + E_c$$

$E_T$  requires actuator node to receive more than the threshold ( $T$ ) number of messages from sensor nodes to verify the event in WSANs. If the number of messages in actuator node is less than  $T$ , actuator is supposed not to react on the event notification messages from sensor node and thus it consumes only communication energies needed to exchange only  $k$  messages with sensor nodes. Usually the energy for actuator to perform a service in WSANs is more than 10 times as big as the one needed for communication. Therefore, the average energy consumption in actuator is possibly described like this.

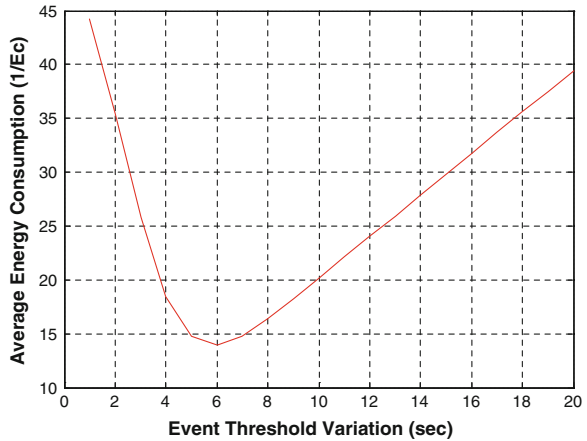
$$\begin{aligned} E_{av} &= Prob[k < T] \times E_c + Prob[k \geq T] \times (E_{av} + E_{act}) \\ &= \sum_{i=1}^{T-1} p_e(i) \times E_c + \sum_{i=T}^n p_e(i) \times (E_c + E_{act}) \end{aligned} \quad (3)$$

When the number of sensor nodes is set to 20 and the average arrival rate of notification message is set to 0.01/s and the duration time of accepting notification messages in actuator is set to 2 s, the result is shown in Fig. 2.

## 4 Proposed Scheme

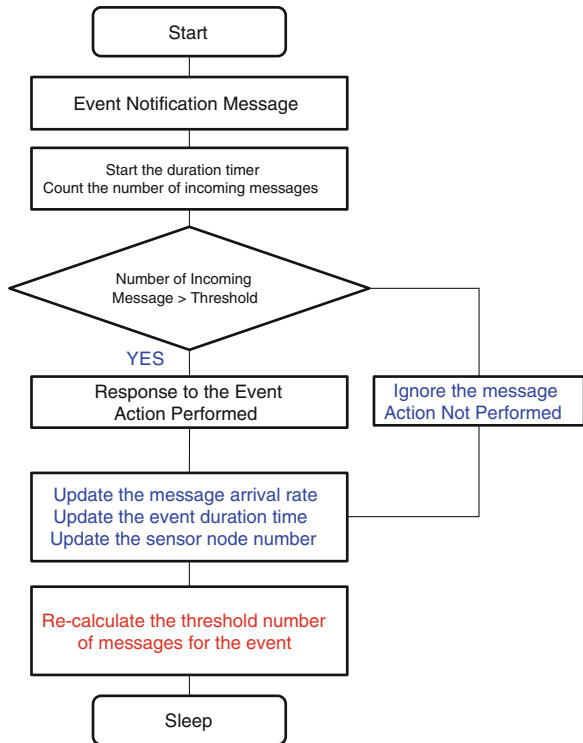
Consequently we found that the number of the notification messages is connected to the energy consumption of the actuator in WSAN. Based on the information we found between the number of the notification messages and the energy consumption of actuator in WSAN, we propose a method that can regulate the threshold number of notification messages for the event to minimize the energy

**Fig. 2** Variation of energy consumption



consumption of the actuator. The idea is described in Fig. 3. As shown in the Fig. 3, after receiving the notification messages from sensor nodes, actuator needs to start the timer for event checking and count the number of event notification message to see if it can be considered as a real event. Upon seeing the number of messages going over the threshold, it would perform the scheduled action for the

**Fig. 3** Proposed algorithm



corresponding event. Otherwise, it would ignore the notification messages from sensor nodes and continue to count the number of the messages and finally update the information required to calculate the threshold of notification messages for an event.

With the updated information about arrival rate of event notification message, event checking duration and the number of sensor nodes, the proper number of event notification messages for minimum energy consumption is calculated with the Eq. (3). This is also updated in an actuator for later detection of an event.

## 5 Conclusions

In WSAAN, the number of notification messages for an event is definitely related to the energy consumption of actuator, which can affect the whole WSAAN. Here we investigated the relationship between the number of notification message and the energy consumption of actuator node. And we propose a new scheme that can regulate the threshold of event notification messages for an event and reduce the energy consumption of an actuator in WSAAN. The proposed scheme is only used in an actuator and does not affect the other sensor nodes in WSAAN. And also it is considered as a self-adaptive scheme for better performance. With the minor improvements, the proposed scheme is expected to contribute in alleviate the burden of energy consumption problem in WSAAN/WSN.

**Acknowledgments** This Research was financially supported by Hansung University.

## References

1. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) A survey on sensor networks. *IEEE Commun Mag*, August 2002
2. Akyildiz IF, Kasimoglu IH (2004) Wireless sensor and actor networks: research challenges. *Ad Hoc Netw* 2(4):351–367
3. Yick J, Mukherjee B, Ghosal D (2008) Wireless sensor network survey. *Comput Netw* 52(12):2292–2330
4. Xi F (2008) QoS challenges and opportunities in wireless sensor/actuator networks. *Sensors* 8:1099–1110
5. Xia F, Tian Y-C, Li Y, Sun Y (2007) Wireless sensor/actuator network design for mobile control applications. *Sensors* 7:2157–2173

# Efficient and Reliable GPS-Based Wireless Ad Hoc for Marine Search Rescue System

Ta Duc-Tuyen, Tran Duc-Tan and Do Duc Dung

**Abstract** Based on work in wireless ad hoc network for Marine Search and Rescue Sys-tem (MSnR) system, this paper presents an improving GPS-based wireless ad hoc network capable of providing location and emergency service to small fishing boats by improving the weak sea-to-land wireless radio link from small fishing boats to the central in-land stations. The proposed approach provides continuous report and monitoring of all boats and its locations for searching and rescuing process during emergencies. The message priority assignment allows the system to operate more efficient and reliable when one or several boats boat in distress. System model, communication mechanism and network simulation results are presented.

**Keywords** Ad hoc network · Global positioning system (GPS) · Medium access control (MAC)

## 1 Introduction

According to our recent study in real-time location monitoring of small fishing [1], we propose an improve model of GPS-based mobile ad hoc network which will pro-vide more efficient and reliable sea-to-land communication link from small fishing boats to central base-stations. The proposed network combines the Global Positioning System (GPS) service (positioning) with a wireless ad hoc network (wireless communication). In addition, the message priority assignment allows the system to operate more efficient and reliable when one or several boats boat in distress. For the land-to-sea communication link, the proposed network simply

---

T. Duc-Tuyen (✉) · T. Duc-Tan  
VNU University of Engineering and Technology, 144 Xuan Thuy,  
Cau GiayHa Noi, Viet Nam  
e-mail: tuyentd@vnu.edu.vn

D. Duc Dung  
Samsung Mobile R&D Center, Samsung Viet Nam, Yen Phong, Bac Ninh, Viet Nam

utilizes the existing coastal radio network; hereby greatly reduces cost and simplifies network design.

The paper is organized as follows: In [Sect. 2](#) we introduce model of Marine Search and Rescue System and benefits of the proposed system. [Section 3](#) present the communication mechanism of proposed system, included packet structure, medium access control and routing protocol. The simulation scenarios and results is shown in [Sect. 4](#). Finally, in [Sect. 5](#) we will conclude the paper and discuss future possibilities for this research.

## 2 Monitoring, Searching and Rescuing System

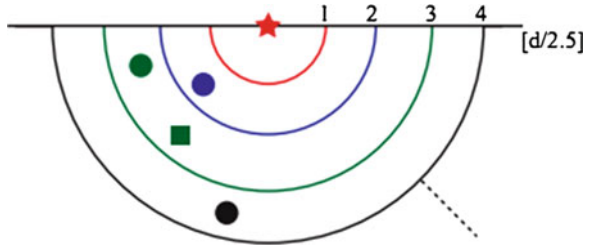
Operates of the proposed GPS-based wireless ad hoc network for marine monitoring, searching and rescuing system as follows. The forward link from land to sea uses the current coastal radio system with high-power base-stations along the coastal line to transmit weather and relevant information to all fishing. The forward link can reach all of fishing boats based on the high power transmission. In contrast to the forward link, the return link from sea to land is limited by a low transmit power and a low antenna height mounted on small fishing boats. This problem is remedied by the proposed wireless ad hoc network, which enables the return link to be first established between fishing boats and then finally connected from the closest boats to the central base-station. The routing protocol of the proposed system is location-based routing. A data packet being routed in these links contain a boats identification number (boat ID), its current GPS location, and a short message. In order to improve the reliability and efficiently of proposed system in case of emergency mode, a message priority assignment is added. It means that messages are divided into three kinds: emergency message, the node's message and normal forwarded message with descending priority. In addition, to identify boats position and establish a wireless connection, a commercially off-the-shelf integrated GPS receiver (location determination) and a programmable digital signal processing (DSP) board (wireless routing and medium access control) are added to the existing low-power radio on fishing boats. Finally, packets are transmitted over existing radio system. Compare with the previous system, the priority mode ensure the probability of successful reception of emergency message from the boat in distress to the base station.

## 3 Communication Mechanism

### 3.1 Medium Access Control (MAC)

In this work, we propose a hybrid MAC solution. The core idea of the hybrid MAC protocol is a combination and smart selection of time division medium access (TDMA) and carrier sense multiple access (CSMA). The operating principle is as

**Fig. 1** A graphical representation of a coverage area

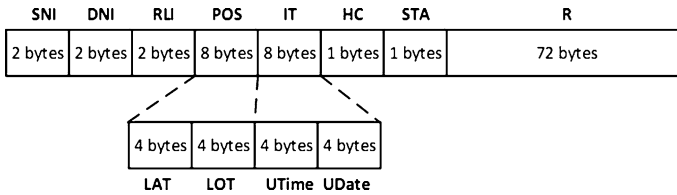


follows. A coverage area is divided into an equal number of concentric circles (red, blue, green, black in Fig. 1, which share the same central point at the local base station (red star). These circles have a radius equal to a multiple of 2.5 km ( $r = n \times 2.5$ ) where  $n$  is an integer number ( $n = 1; 2; 3\dots$ ). Nodes residing on different concentric circles [i.e. circles with  $n = 2$  (blue) and  $n = 4$  (black)] are given two different time slots to transmit/receive. This is similar to TDMA. When there are two or more nodes within the same concentric circle [i.e. circle with  $n = 3$  (green)], these nodes employ CSMA protocol with a carrier detection capability to avoid transmit collision. In addition, handshaking is implemented in conjunction with the CSMA/CA scheme in order to reduce the hidden node problem in the same concentric circle. When node A wants to transmit data, it will broadcast the Ready-to-Send (RTS) message to check the medium idle or not. The received node forward this message to all node in its radio cover-age. If no other nodes in the same concentric circle with A makes a transmission at-tempt, the medium is idle. A Clear-to-Send (CLS) message is replied to node A. Else, the medium is busy and A must be wait for a random time and then transmits again.

### 3.2 Data Packet Format

Figure 2 shows a format of each data packet used in the proposed network. A SNI is the source node identifier or the ID of a transmit node. A DNI is the destination node identifier or the ID of a receive node. When DNI is set to all 0s, the data are sent to base station nodes (default mode). A RNI is the relay node identifier or the ID of a message forwarded node. It set same the SNI field at message’s source. At immediately node, which the message will be, relay, the RNI is the node identifier. A source position (POS) is the position of node at the time a packet is sent. It includes latitude (SLA) and longitude (SLO) of the source nodes at that time. An initial time (IT) field is the original time when a packet is first transmitted. It contains two sub-fields: UDate and UTime, which are Coordinated Universal Time (UTC) date and time of day of the GPS signal. A hop count (HC) indicates the number of nodes that a packet has been traversed. It is set to 0 at the





**Fig. 2** Format of a data packet used in the proposed GPS-based wireless ad hoc network

source node and incremented by 1 at each subsequent forwarding node. A status (STA) field indicates the status of the source node. It is set to all 0s when the source node is in its normal mode of operation, and set to all 1s if it is in an emergency mode, i.e., when the source node seeks help from other ships or the base-station. Reserved field a reserved area for other purposes if any.

### 3.3 Routing Protocol

The main use of the proposed ad hoc network is monitoring and reporting ships location, therefore it does not require a high data rate and can support a large network delay tolerance [2, 3]. Based on these requirements, we adopt a modified hybrid proactive-passive, location-based routing protocol similar to the DREAM protocol first proposed by Basagni in [4]. In the proposed hybrid routing protocol, a lookup table (LUT) at each node is updated when a node receives a data packet. Unlike DREAM or other passive protocol, the LUT only contains locations of one-hop neighboring nodes. As a result, the LUT is significantly smaller and hence less time is required updating the LUT entries.

In this protocol, each node in our ad hoc network broadcasts its packet *m* to all of its one-hop neighbors regardless of their directions and locations. Default that the emergency message will be broadcast immediately at receiving node. It means the emergency packet is served with the highest priority. In other hand, when the neighbor receives a normal message, it decides to relay or drop the packet based on the relative position of the neighbors to the base station. In cases the neighbor decides to relay the packet, it first switches from receive to transmit mode, then re-transmits the packet *m* and updates its forward table (FT). Other node repeats the same procedure, until the final destination (base station) is eventually reached. The routing algorithm is described next. Algorithm 1 shows operating of a node in it is receive mode when Algorithm 2 shows two different transmit scenarios for a given node.

```

Data: receive package
Result: decide to drop or forward package
initialization;
while Receive Mode do
  if receive package then
    detect neighbor;
    update LUT;
    compare location of receive node and source
    node;
    if receive node is in the higher concentric circle
    than source node then
      Drop packet;
    else
      if Packet is sent before then
        Drop packet;
      else
        Update Forward Table (FT);
        if Packet is a emergency packet then
          storage in emergency (FIFO)buffer;
        else
          storage in normal (FIFO)buffer;
        end
      end
    end
  end
else
  if no message is received after 10 consecutive
  time intervals then
    Alert mode;
  else
    wait for next receive mode
  end
end
end

```

Algorithm 1: The operation of a node in its receive mode

```

initialization;
while Transmit Mode do
  if In emergency mode then
    transmit emergency package;
  else
    if (forward)emergency buffer is NULL then
      transmit its owner message;
      transmit (forwarded) normal message;
    else
      transmit (forward) emergency message;
      transmit its owner message;
      transmit (forwarded) normal message;
    end
  end
end
end

```

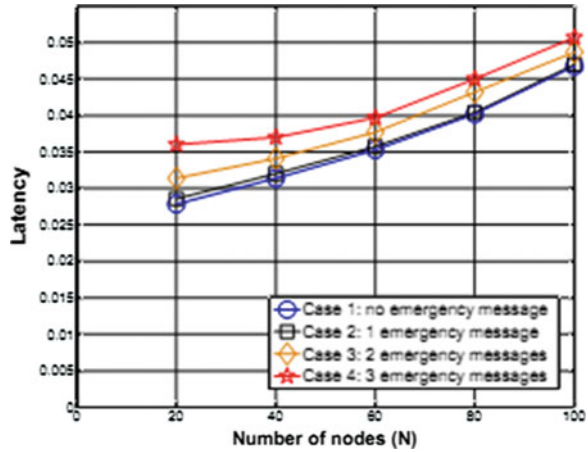
Algorithm 2: The operation of a node in its transmit mode

## 4 Network Simulation and Results

The proposed routing algorithm and MAC are implemented and tested in OMNET++ tool with INETMANET framework, a discrete event simulation tool for Mobile Ad Hoc Networks. The simulation environment is a Mobile Ad Hoc Network consists of 20–100 nodes in a 1000 × 1000 area. We assume that each node will be active by a 8.1 MHz radio frequency with bandwidth of 100 kHz and support data rate of 200 Kbps. The correspondence convergence circle area of each node with radio radius of 100. The simulated area is considered as a two dimensional square and nodes movement freely throughout the area. The movement of node has been simulated according to Random Way-point model with maximum movement speed is 10 m per second (36 km per hour or 19.4 nautical mile per hour).

In order to evaluate the performance and the efficiently of the proposed system, a set of simulation were operated with duration of 2000 s. We select a set of parameters to show the efficiency of our algorithm in two case: (1) no emergency message is generated and (2) at least one emergency message is generated in network. These parameters include:

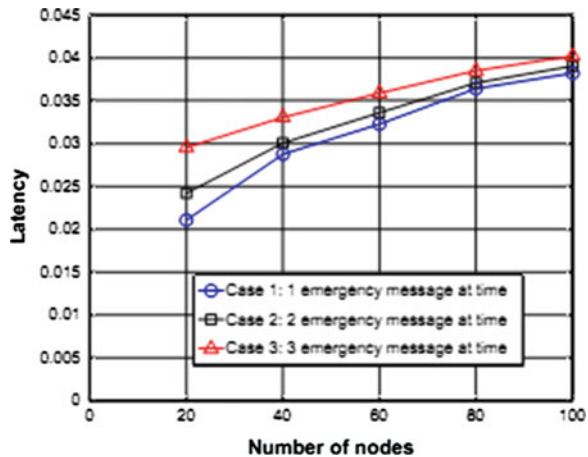
**Fig. 3** Average packet latency for different number of nodes (N) and different number of emergency message



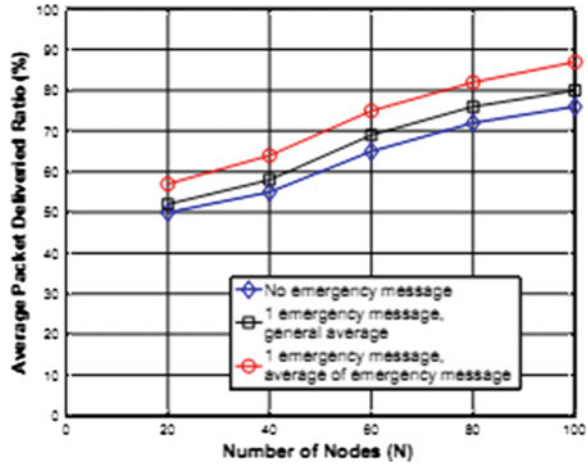
### 4.1 Average Packet Latency

Figure 3 is a plot of the average packet latency versus the number of network nodes. That is the average time taken by any node to receive the message. In our case, this is the time taken by base station to receive the message from boat. In case of no emergency message is generated, the packet latency increases proportionally with the number of network nodes. This is expected since a large network necessary causes a long delay because packets must be relayed through more number of nodes. In other hand, when at least one emergency packet is generated, the average packet latency has slightly decreased. However, when the number of emergency message at the same time is increased, the average packet latency has increased. This is due to fact that the crashed node only broadcast the emergency message and does not participate in forwarding the message from other

**Fig. 4** Average latency of emergency packet for different number of emergency packets at the same time



**Fig. 5** The average packet delivery ratio (PDR) for different number of nodes



node in network. In worst case, the network will be collapse when the number of network nodes in an emergency increased.

The average latency of emergency packet is shown in Fig. 4. We can see that latency increases faster in small-scale network than it in large-scale network. This can be explained by the increase in emergency nodes reduce the ability to forward messages on the network, especially for small networks.

### 4.2 Average Delivery Ratio

The average packet latency increases when increase size of network. In addition, a larger ad hoc network increases the possibility of packet delivery, too. Figure 5 indicates the average packet delivery ratio for normal message in case of normal mode and emergency and normal message in case of emergency mode with 1 emergency message is generated.

## 5 Conclusions

A wireless ad hoc network capable of providing location service to small fishing boats and improving the weak sea-to-land wireless radio link from small fishing boats to the central in-land stations has been proposed. The proposed location-based routing protocol (DREAM) and hybrid MAC have been verified using OMNET++ tool with INETMANET Framework. The simulation results show that the message priority assignment will be help to improve the reliability and the great potential of the proposed concept for marine monitoring, searching, and rescuing applications.

**Acknowledgments** The authors want to thank the project CN 12.04 of VNU University of Engineering and Technology for the financial support of this work.

## References

1. Do DD, Nguyen HV, Tran NX, Ta TD, Tran TD, Vu YV (2011) Wireless ad hoc network based on global positioning system for marine monitoring, searching and rescuing (MSnR). In: APMC 2011: Asia-Pacific microwave conference, Melbourne, pp 1510–1513
2. Karl H, Willig A (2005) Protocols and architectures for wireless sensor networks. Wiley
3. Zhang Z (2009) Routing protocols in intermittently connected mobile ad hoc networks and delay-tolerant networks. In: Boukerche a (ed) Algorithms and protocols for wireless and mobile ad hoc networks. Wiley
4. Basagni S, Chlamtac I, Syrotiuk VR (1998) A distance routing effect algorithm for mobility (DREAM). In: Proceeding of the annual international conference on mobile computing and networking, USA

# Improved Relay Selection for MIMO-SDM Cooperative Communications

Duc Hiep Vu, Quoc Trinh Do, Xuan Nam Tran  
and Vo Nguyen Quoc Bao

**Abstract** In this paper, we propose two relay selection algorithms based on the signal-to-noise ratio (SNR) and the eigenvalue which achieve improved bit error rate (BER) performance compared with the previous one based on the mean square error (MSE) at the same complexity order.

**Keywords** Cooperative communication · Relay selection · MIMO · SDM

## 1 Introduction

The modern wireless communication is developing very fast in order to meet the human demand for high-speed data access. The last decade has witnessed various successful developments in the air interface technology. The most important development is probably the multiple-input multiple-output (MIMO) transmission [1]. MIMO transmission systems can be implemented in the form of either the transmit diversity [2] or spatial division multiplexing (SDM) [3]. The aim of the transmit diversity is to achieve diversity gain in order to reduce the bit error rate (BER) and thus increasing the link reliability. This transmit diversity scheme is also known as the space–time block code [2]. The MIMO-SDM systems, on the other hand, aim at achieving multiplexing gain in order to increase the spectral efficiency. For a centralized MIMO system where multiple antennas are placed at the transmitter and the receiver, it was shown in [4] that there is a trade-off between the diversity and multiplexing gain. This means that the centralized

---

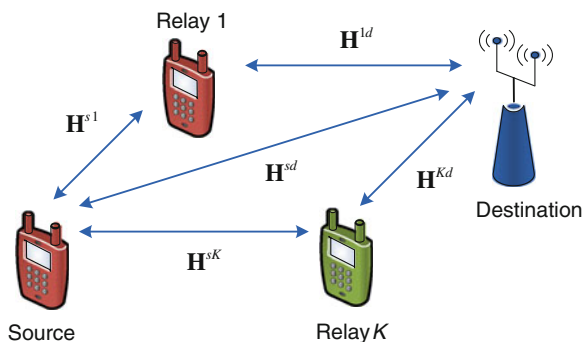
D. H. Vu (✉) · Q. T. Do · X. N. Tran  
Le Quy Don Technical University, 236 Hoang Quoc Viet, Cau Giay, Ha Noi, Vietnam  
V. N. Q. Bao  
Post and Telecommunications Institute of Technology, 11 Nguyen Dinh Chieu Str.,  
District 1, Ho Chi Minh City, Vietnam

MIMO systems do not achieve full diversity and multiplexing gain at the same time. In order to achieve both diversity and multiplexing gain, a so-called MIMO-SDM cooperative communication system was proposed in [5]. In this work the authors proposed three distributed relay selection schemes and a linear minimum mean square error (MMSE) combining scheme which achieve full diversity and full multiplexing gain at the same time. Among the three proposed selection algorithms based on maximum channel matrix norm, maximum channel harmonic mean, and minimum MSE, the MSE-based algorithm was shown to achieve the best BER performance [5]. In this paper, based on the idea of [6] for the case of MIMO-SDM, we developed two relay selection algorithms based on the signal to noise ratio (SNR) and eigenvalue. The two proposed algorithms have improved BER performance over the MSE-based algorithm while requiring the same complexity order.

## 2 System Model

We consider a MIMO-SDM cooperative communication network similar to [5] as illustrated in Fig. 1. The network consists of a source and a destination communicating with each other with the help of a relay node via a relaying path. Without loss of generality, we assume that all nodes (including source, destination and intermediate) are equipped with  $N = 2$  antennas for both transmission and reception. There are  $K$  capable intermediate nodes  $k = 1, 2, \dots, K$  between the source and the destination. Based on a distributed relay selection protocol [5, 7] the  $K$  intermediate nodes will interact with one another to select the best one to act as the relay (denoted by the index  $r$ ). The channels between nodes are assumed flat uncorrelated Rayleigh fading and unvarying during a transmission period. We denote  $\mathbf{H}^{sd}$ ,  $\mathbf{H}^{sk}$ ,  $\mathbf{H}^{kd}$  the channel matrices between the source and the destination, the source and the intermediate node  $k$  and the destination, respectively. The channel between a node  $a$  and a node  $b$  is denoted as the matrix  $\mathbf{H}^{ab} =$

**Fig. 1** System model of the MIMO cooperative communications



$[h_{11}^{ab}, h_{12}^{ab}, h_{21}^{ab}, h_{22}^{ab}]$  where  $h_{mn}^{ab}$  is the channel between the  $m$ th antenna of node  $b$  to the  $n$ th antenna of  $a$ .

The communication between the source and the destination involves two phases: relay selection and signal transmission. The relay selection is done using the distributed protocol as mentioned above while the signal transmission uses two time slots. In the first slot, the source transmits a signal vector  $\mathbf{s} = [s_1, s_2]^T$  consisting of two symbols  $s_1$  and  $s_2$  from the two antennas to both the destination and the relay. Here the superscript  $T$  denotes the matrix transpose. The received signal vector at the destination and the relay is given by  $\mathbf{y}_1 = \mathbf{H}^{sd}\mathbf{s} + \mathbf{z}_1, \mathbf{x}_r = \mathbf{H}^{sr}\mathbf{s} + \mathbf{z}_r$ , where  $\mathbf{z}_1$  and  $\mathbf{z}_r$  are the noise vector at the destination and relay  $r$ , respectively. In the second time slot, the relay performs amplifying-and-forwarding (AF) the received signal  $\mathbf{x}_r$  to the destination. The amplification matrix  $\mathbf{G}_r$  is a diagonal matrix with the amplification factor used for the  $i$ th branch given by [5]:  $g_i^r = \sqrt{E_s/N(E_s/N\|\mathbf{h}_i^{sr}\|^2 + 1)}$ , where  $E_s$  is the transmit symbol energy and  $\mathbf{h}_i$  is the  $i$ th row of the appropriate channel matrix. The received signal at the destination during the second time slot is given by [5]

$$\mathbf{y}_2 = \mathbf{H}^{rd}\mathbf{G}_r\mathbf{x}_r + \mathbf{z}_2 = \mathbf{H}^{rd}\mathbf{G}_r\mathbf{H}^{sr}\mathbf{s} + \mathbf{H}^{rd}\mathbf{G}_r\mathbf{z}_r + \mathbf{z}_2 \tag{1}$$

where  $\mathbf{z}_2$  is the noise vector at the destination in the second time slot. The destination will combine the received signal vectors  $\mathbf{y}_1$  and  $\mathbf{y}_2$  to obtain the received signal vector  $\mathbf{y}$ . Define  $\mathbf{y} = [\mathbf{y}_1^T, \mathbf{y}_2^T]^T$ ,  $\mathbf{H} = [(\mathbf{H}^{sd})^T, (\mathbf{H}^{sr}\mathbf{H}^{rd})^T]^T$  and  $\mathbf{z} = [\mathbf{z}_1^T, (\mathbf{H}^{rd}\mathbf{G}_r\mathbf{z}_r + \mathbf{z}_2)^T]^T$ , the system equation is given by

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{z}. \tag{2}$$

### 3 Proposed Relay Selection Algorithms

The proposed selection algorithms are performed in a distributed manner as described in [5, 7]. The intermediate nodes  $k = 1, 2, \dots, K$  are assumed to know the forward channel from itself to the destination  $\mathbf{H}^{kd}$  and the backward channel from it back to the source  $\mathbf{H}^{ks}$ . Due to reciprocity the channel  $\mathbf{H}^{sk}$  is assumed to be the same as  $\mathbf{H}^{ks}$ . Each node  $k$  will calculate the channel quality index (CQI) of the relaying path via itself. The node with the largest CQI, denoted by  $\kappa$  be selected as the relay.



**Table 1** Eigenvalue-based relay selection algorithm

---

Input: $K, \mathbf{H}^{sk}, \mathbf{H}^{kd}$
For $k = 1$ to $K$
Calculate $\lambda_{1,2}^{sk}, \lambda_{1,2}^{kd}$ using (3)
Select $\lambda^k = \min\{\lambda_{1,2}^{sk}, \lambda_{1,2}^{kd}\}$
$\text{CQI}_k = \lambda^k$
$\kappa = \arg \max_k \{\text{CQI}_k\}$
End
Output: node $\kappa$ as relay $r$

---

### 3.1 Eigenvalue-Based Relay Selection

The idea of selecting relay based on eigenvalues comes from the fact that in the MIMO systems eigenvalue of the channel matrix is considered the power gain of the channel [8]. As a result, the channel which has larger eigenvalues will have better power gain. The eigenvalues of the channel  $\mathbf{H}^{ab}$  is the solutions to the following characteristic equation  $\det(\mathbf{H}^{ab} - \lambda \mathbf{I}) = 0$ , where  $\mathbf{I}$  is an  $N \times N$  identity matrix and  $\det(\cdot)$  represents the determinant of the matrix formed by  $(\mathbf{H}^{ab} - \lambda \mathbf{I})$ . For an  $N \times N$  complex matrix  $\mathbf{H}^{ab}$  there are at most  $N$  distinct eigenvalues. For the  $2 \times 2$  channel matrix  $\mathbf{H}^{ab}$  considered here there are two eigenvalues given by

$$\lambda_{1,2}^{ab} = \frac{(h_{11}^{ab} + h_{22}^{ab}) \pm \sqrt{(h_{11}^{ab} + h_{22}^{ab})^2 - 4(h_{11}^{ab}h_{22}^{ab} - h_{12}^{ab}h_{21}^{ab})}}{2} \tag{3}$$

In order to obtain the associated CQI each intermediate node  $k$  will first select  $\lambda^k = \min\{\lambda_{1,2}^{sk}, \lambda_{1,2}^{kd}\}$  and then calculate  $\text{CQI}^k = \lambda^k$ . The max-min selection algorithm based on eigenvalues is summarized as pseudocodes in Table 1. It is worth noting that for the case of using a larger number of antennas, i.e.  $N > 2$ , the calculation of the eigenvalues as used in (3) is not straightforward and a more complicated calculation algorithm should be used.

### 3.2 SNR-Based Selection Algorithm

As SNR is inversely proportional to BER, selecting a relaying path with better SNR promises lower BER. In order to perform SNR-based relay selection, we assume that the destination uses the linear MMSE detector proposed in [5]. Based on this assumption intermediate nodes will calculate the received SNR at the destination via its relaying path. The CQI associated with each path will be assigned based on the calculated SNR. From [5] we can write the combining

**Table 2** SNR-based algorithm

---

Input: $K, \mathbf{H}^{sk}, \mathbf{H}^{kd}, \mathbf{G}_k$
For $k = 1$ to $K$
Calculate $\text{SNR}_n^k$ using (6)
$\text{SNR}^k = \min\{\text{SNR}_1^k, \text{SNR}_2^k\}$
$\text{CQI}_k = \text{SNR}^k$
$\kappa = \arg \max_k \{\text{CQI}_k\}$
End
Output: node $\kappa$ as relay $r$

---

weight matrix that the destination would use to combine the relaying signal with that from the direct path as follows

$$\mathbf{W}_2^k = \left[ \frac{E_s}{N} \mathbf{H}^{skd} (\mathbf{H}^{skd})^H + \sigma_{z_k}^2 \mathbf{H}^{kd} \mathbf{G}_k^2 (\mathbf{H}^{kd})^H + \sigma_{z_2}^2 \mathbf{I}_2 \right]^{-1} \frac{E_s}{N} \mathbf{H}^{skd}, \quad (4)$$

where  $\mathbf{H}^{skd} = \mathbf{H}^{sk} \mathbf{G}_k \mathbf{H}^{kd}$ ,  $\sigma_{z_k}^2$  and  $\sigma_{z_2}^2$  are the variance of the noise induced at node  $k$  and at the destination during the second time slot, respectively. For simplicity, we assume that  $\sigma_{z_d}^2 = \sigma_{z_1}^2 = \sigma_{z_2}^2$  and that  $\sigma_{z_k}^2 = \sigma_{z_d}^2$ . Since we use the assumption that the source sends two parallel streams, the estimated symbol  $\tilde{s}_n$  of  $s_n$  if node  $k$  acts as the relay would be

$$\tilde{s}_n = (\mathbf{w}_{2,n}^k)^H \mathbf{y}_2 = (\mathbf{w}_{2,n}^k)^H \mathbf{H}^{kd} \mathbf{G}_k \mathbf{H}^{sk} \mathbf{s} + (\mathbf{w}_{2,n}^k)^H (\mathbf{G}_k \mathbf{H}^{kd} \mathbf{z}_k + \mathbf{z}_2) \quad (5)$$

where  $\mathbf{w}_{2,n}^k$  is the  $n$ th column of  $\mathbf{W}_2^k$  and  $(\bullet)^H$  denotes the Hermitian operation. The received SNR at the destination is defined as follows

$$\text{SNR}_n^k = \frac{E_s \| (\mathbf{H}^{sk})^H \mathbf{G}_k^H (\mathbf{H}^{kd})^H \mathbf{w}_{2,n}^k \|^2}{N \left( \sigma_{z_k}^2 \| \mathbf{G}_k^H (\mathbf{H}^{kd})^H \mathbf{w}_{2,n}^k \|^2 + \sigma_{z_2}^2 \| \mathbf{w}_{2,n}^k \|^2 \right)} \quad (6)$$

From this equation the SNR-based relay algorithm as summarized in Table 2.

## 4 Complexity Analysis

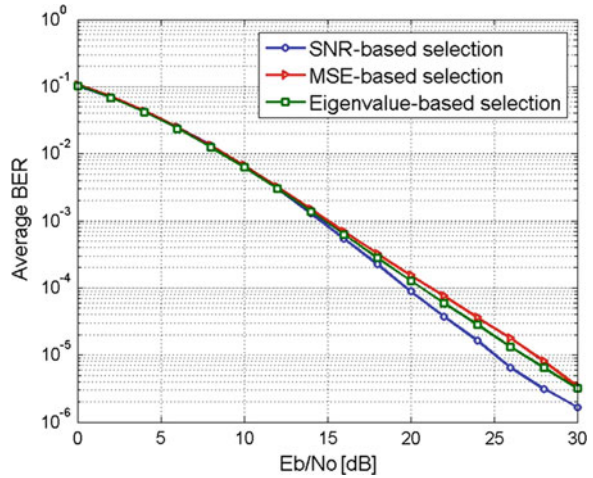
In order to compare the complexity of the proposed algorithms with that based on the MSE, we perform detailed calculation of the number of addition/subtraction, multiplication and division for all the case of complex–complex, complex–real, and real–real operations. These computational operations will be then converted into floating points (flops) for comparison. The complexity of the SNR-based algorithm involves mainly with calculating (5) and (7). The approximated complexity of the SNR-based algorithm is given by  $C_{\text{SNR}} = 36N^3 + 34N^2 + 28N + 5$  [flops]. The main complexity of the eigenvalue-based algorithm is used for computing the eigenvalues of the two square matrices  $\mathbf{H}^{sk}$ ,  $\mathbf{H}^{kd}$  both of the same

size  $N \times N$ . The complexity for calculating the eigenvalues using singular value decomposition is  $72N^3$  [flops]. The complexity of the MSE-based relay selection algorithm mainly involves with calculating equations (30), (34) and (35) in [5]. The number of computational operations required by the MSE-based algorithm is  $C_{\text{MSE}} = 20N^3 + 26N^2 + 4N + 3$  [flops]. Therefore, it is clear that all the algorithms have the same complexity order  $O(N^3)$ .

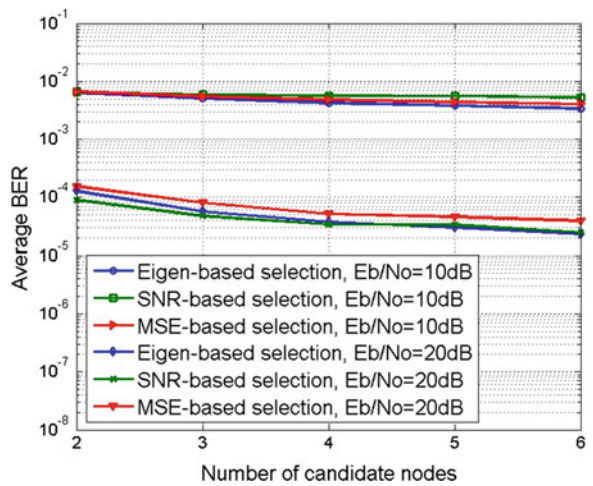
## 5 Simulation Results

In order to demonstrate the advantage of the proposed algorithms, we have performed Monte-Carlo simulations to obtain the average BER. In the first simulation, we use a simple model with three nodes, i.e., source, relay, and destination. In order to select the relay, we assume that there are two intermediate nodes within the coverage area of the source and destination. The proposed algorithms will be used to select the better node as the relay. The channels between the source to intermediate nodes and from the intermediate nodes to the destination are all assumed to undergo flat uncorrelated Rayleigh fading. All nodes are equipped with two antennas and transmit BPSK signal over the two parallel branches. The average symbol energy of each node is normalized to  $E_S$ . The destination employs the MMSE detector in [5] to estimate the transmit signal. In the second simulation, we use a similar model but the ratio  $E_b/N_0$  is fixed while the number of intermediate nodes is increased to analyze the effect of selecting a relay from a large number of nodes. In all simulations, BER of the MSE-based algorithm is also plotted for comparison. The average BER curves obtained using the proposed algorithms and the MSE based are shown in Fig. 2. It can be seen clearly from the figure that both the proposed algorithms have the same BER performance at the low  $E_b/N_0$  region but outperform the MSE-based for large  $E_b/N_0$ . Specifically, at  $\text{BER} = 10^{-5}$ , the proposed eigenvalue based algorithm has about 0.5 dB better  $E_b/N_0$  while the SNR based achieves up to 2.5 dB improvement. It is also clear that the gap between the SNR-based algorithm and the MSE-based is much larger than that of the eigenvalue-based. Figure 3 illustrates the BER performance of the three algorithms obtained at  $E_b/N_0 = 10, 20$  dB for the case of  $N = 2, 3, 4, 5, 6$ . It still can be seen that the two proposed algorithms achieve better BER performance than the MSE-based, particularly at high  $E_b/N_0$ . However, similar to [5] it is interesting to note that increasing the number of intermediate nodes does not achieve better improvement. This is the inherent property of the MIMO-SDM cooperative communication as explained in [5].

**Fig. 2** BER performance of different selection algorithm,  $N = 2$ ; 2 select 1



**Fig. 3** BER performance versus the number of candidate nodes



## 6 Conclusions

In this paper, we have proposed two relay selection algorithms based on SNR and eigenvalue for the MIMO-SDM cooperative communication networks. Both the proposed algorithms have better BER performance over the previous MSE-based algorithm. We have also carried out detailed complexity analysis to show that both the proposed algorithms and the MSE-based have the same complex order  $O(N^3)$ . The proposed SNR-based algorithm was shown to be the best candidate in terms of both BER performance and required complexity.

**Acknowledgments** This work is sponsored by National Foundation for Science and Technology Development (Nafosted) under project number 102.03-2012.18.

## References

1. Foschini GJ, Gans MJ (1998) On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Pers Commun* 6(3):311–335
2. Alamouti SM (1998) A simple transmit diversity technique for wireless communications. *IEEE J Sel Areas Commun* 16(8):1451–1458
3. Wolniansky P, Foschini GJ, Golden GD, Valenzuela RA (1998) V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel. In: *The URSI international symposium on signals, systems, and electronics, Italy*, pp 295–300
4. Zheng L, Tse D (2003) Diversity and multiplexing: a fundamental tradeoff in multiple antenna channels. *IEEE Trans Inf Theory* 49(5):1073–1096
5. Tran XN, Nguyen VH, Bui TT, Dinh TC (2012) Distributed relay selection for MIMO-SDM cooperative networks. In: *IEICE Trans Commun E95-B:1170–1179*
6. Paulraj A, Heath RW (2001) Antenna selection for spatial multiplexing systems with linear receivers. *IEEE Commun Lett* 5(4):142–144
7. Bletsas A (2003) A simple cooperative diversity method based on network path selection. *IEEE J Sel Areas Commun* 24(3):659–672
8. Andersen JB (2000) Array gain and capacity for known random channels with multiple element arrays at both ends. *IEEE J Sel Areas Commun* 18(11):2172–2178

# Freshness Preserving Hierarchical Key Agreement Protocol Over Hierarchical MANETs

Hyunsung Kim

**Abstract** Recently, Guo et al. proposed an efficient and non-interactive hierarchical key agreement protocol applicable to mobile ad-hoc networks, which is a try to solve the open question: How can secrets be established if an adversary can eavesdrop on every message exchange? However, their protocol does not support freshness of the established session key that key agreement protocols should have. Thereby, we propose a freshness preserving hierarchical key agreement protocol over the hierarchical MANETs. Compared with other existing protocols, the proposed protocol offers much better performance on the bandwidth consumption, the computational cost, and the storage cost.

**Keywords** Hierarchical key agreement · Security protocol · Mobile ad-hoc network · Information security · Cryptography

## 1 Introduction

Mobile Ad hoc Network (MANET) is a collection of mobile nodes that are dynamically and arbitrarily located in such a manner that communication between nodes does not rely on any fixed network infrastructure. The absence of static infrastructure and centralized administration makes MANETs to be self organized and relying on the cooperation of neighboring nodes in order to find the routes between the nodes for reliable communication. Hence, the performance of MANETs is highly dependent on collaboration of all the participating nodes. The more the number of nodes that participate in packet routing, greater aggregate bandwidth and shorter routing paths can be realized. It will further minimize network partition in the case of failures. The MANET has a wide range of applications in diverse fields

---

H. Kim (✉)

Department of Cyber Security, Kyungil University, Kyungsansi, Kyungpook, Korea  
e-mail: kim@kiu.ac.kr

ranging from low power military wireless sensor networks to large scale civilian applications, emergency search and rescue operations [1, 2].

The major challenges in providing secure authenticated communication for MANETs come from the following unique features of such networks [3]: lack of a fixed reliable public key infrastructure, dynamic network topology due to high mobility and joining/leaving devices, energy and resource constrained nodes with limited storage, communication and computation power, lack pre-distributed symmetric keys shared between nodes, high-level of self-organization, vulnerable multi-hop wireless links, etc.

Key agreement is a fundamental tool for secure communication, which lets two nodes in a MANET agree on a shared key that is known only to them, thus allowing them to use that key for secure communication [4–10]. In environments where bandwidth is at a premium, there is a significant advantage to non-interactive schemes, where two nodes can compute their shared key without any interaction. Diffie-Hellman key agreement protocol is an example of a non-interactive scheme [6]. But the nodes in the Diffie-Hellman protocol must still get each other's public keys, which require coordination. To minimize the required coordination, one may use identity-based key-agreement that provides each node with a secret key that corresponds to that node's name [7]. In this setting, the non-interactive identity-based scheme of Sakai et al. in [8] is based on bilinear maps. Recently, Guo et al. proposed an efficient and non-interactive hierarchical key agreement protocol, named as HNAKA, applicable to mobile ad-hoc networks [9]. However, the HNAKA could not support freshness for the established session key to support non-interactive. Thereby, the purpose of this paper is to remedy the HNAKA to preserve freshness. The proposed revision, named as HNAKA<sub>fresh</sub>, establishes a secure channel by setting up a fresh session key between any two nodes in the MANETs. The HNAKA<sub>fresh</sub> could support security and robustness over the hierarchical MANETs.

## 2 Related Works

This section reviews Guo et al.'s hierarchical non-interactive identity-based authenticated key agreement protocol (HNAKA) based on the pairing [8, 9]. Furthermore, we show that the HNAKA does not provide session freshness, which is one of important features for the key agreement protocol.

### 2.1 Guo et al.'s Hierarchical Non-interactive Authenticated Key Agreement

Similarly to other identity-based authenticated key agreement protocols, Guo et al.'s HNAKA requires a private key generator (PKG) and consists of three phases: system setup, private key generation, and key agreement [9]. Let  $k$  be the

security parameter,  $G$  and  $G_T$  be two cyclic groups of prime order  $q$ , and  $\hat{e}: G \times G \rightarrow G_T$  be a bilinear pairing. They denote by  $G^*$  the non-identity elements set of  $G$ . They assume that public keys (identities or IDs) at depth  $l$  are vectors of elements in  $(G^*)^l$ . The  $j$ th component corresponds to the identity at level  $j$ . They later extend the construction to public keys over  $\{0, 1\}^*$  by first hashing each component  $I_j$  using a collision resistant hash  $H_1: \{0, 1\}^* \rightarrow G$ .

**Setup** To generate system parameters for our scheme of maximum depth  $l$ ,

- Select a random generator  $P_0 \in G$  and choose master keys  $s_i$  ( $i = 1, \dots, l$ ) uniformly at random from  $Z_q$  and compute the PKG's public key as  $s_i P_0$

The resultant public parameters and the master key are  $params = \{q, G, G_T, P_0, \hat{e}, H_1, s_1 P_0, \dots, s_l P_0\}$  and master—key =  $\{s_1, \dots, s_l\}$ .

**KeyGen** For user  $A$  with the identity tuple  $(ID_1, \dots, ID_t)$ , given the master secret key  $\{s_1, \dots, s_l\}$  and the system public parameters, compute  $P_i = H_1(ID_i)$  and  $d_i = s_i P_i$  ( $i = 1, \dots, t$ ). Send  $D_A = (s_1 P_1, \dots, s_t P_t, s_{t+1}, \dots, s_l)$  to user  $A$  via an authenticated and private channel. Indeed, given user  $A$  with the identity tuple  $(ID_1, \dots, ID_t)$  and his parent  $C$  with the identity tuple  $(ID_1, \dots, ID_{t-1})$ ,  $C$  can compute the private key  $D_A$  for user  $A$  using his own private key  $D_C = (s_1 P_1, \dots, s_{t-1} P_{t-1}, s_t, \dots, s_l)$  as follows:

- $C$  computes  $P_t = H_1(ID_t)$  and  $s_t P_t$  using his private key.
- $C$  sends  $D_A = (s_1 P_1, \dots, s_t P_t, s_{t+1}, \dots, s_l)$  to user  $A$  via an authenticated and private channel.

Suppose user  $A$  and user  $B$  want to establish a shared session secret key:  $A$  has the identity  $(ID_1, \dots, ID_m)$  and  $B$  has the identity  $(ID'_1, \dots, ID'_n)$ , where  $m > n$ .

**Key agreement** To establish a shared session secret key,  $A$  and  $B$  conduct the following tasks:

- $A$  computes  $P'_i = H_1(ID'_i)$ , where  $i = 1, \dots, n$ , and  $B$  computes  $P'_j = H_1(ID'_j)$ , where  $j = 1, \dots, m$ .
- $A$  computes

$$sk_{AB} = \hat{e}(s_1 P_1, P'_1) \cdots \hat{e}(s_n P_n, P'_n) \cdot \hat{e}(s_{n+1} P_{n+1}, P'_n) \cdots \hat{e}(s_m P_m, P'_n) \quad (1)$$

- $B$  computes  $sk_{BA} = \hat{e}(P_1, s_1 P'_1) \cdots \hat{e}(P_n, s_n P'_n) \cdot \hat{e}(P_{n+1}, P'_n)^{s_{n+1}} \cdots \hat{e}(P_m, P'_n)^{s_m}$

## 2.2 No Key Freshness Support Problem in HNAKA

A key establishment/agreement process among the participants should guarantee that each shared session key is fresh, i.e. has not been reused by one of the



participants. This also means that a key used in one cryptographic association has not been used in another association. Thus, the session key needs to be changed over time since a key may be compromised during pre-deployment or operational phases of communication networks. In Guo et al.'s HNAKA, each party computes  $sk_{AB}$  via the Eq. (1), which depends on both of their own private key and identity tuples but not on the session dependent random value. Thereby, the HNAKA does not provide key freshness. No freshness support means that the established session keys in different sessions are always the same, which could provide some means or useful information to attacker. One of serious effects is traffic analysis attack, which is focused on traffic flow identification, traffic flow tracking, or disclosing application-level information.

### 3 Freshness Preserving Hierarchical Key Agreement Protocol

This section proposes a freshness preserving hierarchical key agreement protocol, named as HNAKA<sub>fresh</sub> over hierarchical MANETs, which is a remedy of Guo et al.'s HNAKA. This protocol falls into the same phases with the HNAKA: system setup, private key generation, and key agreement. The first two phases are the same with the HNAKA but the last one is different due to support key freshness to the protocol. Figure 1 shows the hierarchical key generation model for the HNAKA<sub>fresh</sub>.

The assumptions in the HNAKA<sub>fresh</sub> is the same as in Guo et al.'s protocol, which are user A with the identity tuple  $(ID_1, \dots, ID_m)$  and private key tuple  $D_A = (s_1P_1, \dots, s_mP_m, s_{m+1}, \dots, s_l)$  at level  $m$  and user B with the identity tuple

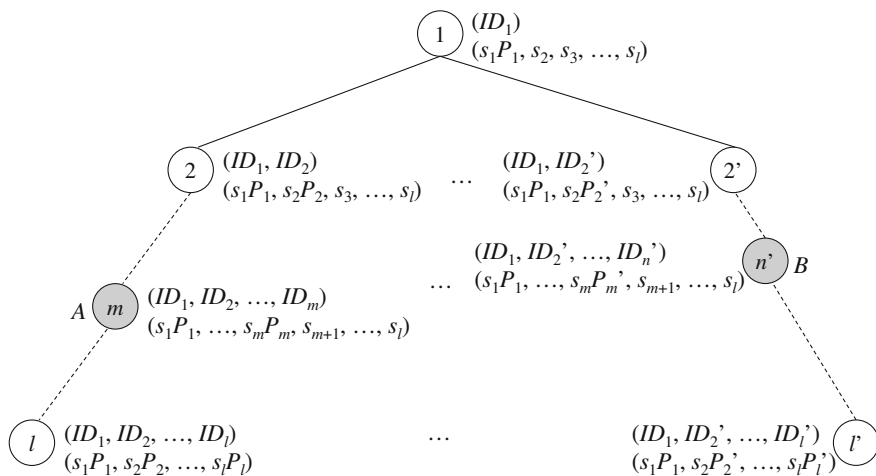


Fig. 1 Hierarchical key generation model

$(ID'_1, \dots, ID'_n)$  and private key tuple  $D_B = (s_1P'_1, \dots, s_nP'_n, s_{n+1}, \dots, s_l)$  at level  $n$ , where  $m > n$  and  $l$  is the depth of the hierarchy as shown in Fig. 1.

**Key agreement** To establish a shared session secret key,  $A$  and  $B$  conduct the following tasks:

- $A$  chooses a random number  $r_1$ , computes  $R_1 = r_1P_m$ ,  $P'_i = H_1(ID'_i)$ , where  $i = 1, \dots, n$ ,  $sk_{AB} = \hat{e}(s_1P_1, P'_1) \cdots \hat{e}(s_nP_n, P'_n) \cdot \hat{e}(s_{n+1}P_{n+1}, P'_n) \cdots \hat{e}(s_mP_m, P'_n)^{r_1}$  and  $MAC_1 = H_1(sk_{AB}, R_1)$ , and sends  $\{R_1, MAC_1\}$  to  $B$ .
- $B$  computes  $P'_j = H_1(ID'_j)$ , where  $j = 1, \dots, m$ , and  $sk_{BA} = \hat{e}(P_1, s_1P'_1) \cdots \hat{e}(P_n, s_nP'_n) \cdot \hat{e}(P_{n+1}, P'_n)^{s_{n+1}} \cdots \hat{e}(R_1, P'_n)^{s_m}$ .
- $B$  assures the correctness of the established session key only if the validity check of  $MAC_1$  is successful by comparing it with  $H_1(sk_{BA}, R_1)$ .

## 4 Security Analysis

This section provides security analysis of the  $HNAKA_{\text{fresh}}$ . Although it is important to provide a formal security proof on any cryptographic protocols, the formal security proof of the protocols remains one of the most challenging issues for cryptography research. Therefore, we follow the approaches used in [10].

### 4.1 Computational Problems

Bilinear map captures an important cryptographic problem, i.e., the Bilinear Diffie-Hellman (BDH) problem, which was introduced by Boneh and Franklin in [7]. The security of the  $HNAKA_{\text{fresh}}$  relies on a variant of the BDH assumption.

Let  $G$  and  $G_T$  be two groups of a prime order  $q$ . Suppose that there exists a bilinear map  $\hat{e}: G \times G \rightarrow G_T$ . We consider the following computational assumptions

- Bilinear Diffie-Hellman (BDH): For  $a, b$ , and  $c \in_{\mathbb{R}} Z_q^*$  and given  $aP, bP$  and  $cP$ , computing  $\hat{e}(P, P)^{abc}$  is hard
- Decisional Bilinear Diffie-Hellman (DBDH): For  $a, b, c$  and  $r \in_{\mathbb{R}} Z_q^*$ , differentiating  $(aP, bP, cP, \hat{e}(P, P)^{abc})$  and  $(aP, bP, cP, \hat{e}(P, P)^r)$  is hard

### 4.2 Security Analyses

Our security analysis is focused on verifying the overall security requirements for the  $HNAKA_{\text{fresh}}$  including passive and active attacks as follows.

**Proposition 1** *The  $HNAKA_{fresh}$  is secure against passive attack.*

*Proof* We assume that an adversary is success if the adversary could learn some useful information from the intercepted messages. We show that probability to succeed in learning them is negligible due to the difficulty of the underlying cryptosystem, the BDH problem and the DBDH problem.

1. A completeness of the key agreement protocol is already proven by describing the run of the protocol in Sect. 3.
2. If the adversary is passive adversary, all the adversary can gather are as follows: the identity tuples  $(ID_1, \dots, ID_m)$  of  $A$  and  $(ID'_1, \dots, ID'_n)$  of  $B$  and the message  $\{R_1, MAC_1\}$ . However, it is negligible to find the key related information from them due to the difficulty of the underlying cryptosystem, the keyed hash function.

Finally, we could say our protocol is secure against passive attack.

**Proposition 2** *The  $HNAKA_{fresh}$  is secure against active attack.*

*Proof* We assume that an adversary is success if the adversary finds the session key  $sk$  (from now on, we use  $sk$  instead of using  $sk_{AB}$  or  $sk_{BA}$  for simplicity) or the session key related private key information  $\{s_1, s_2, \dots, s_l\}$ . Therefore, we show that probability to succeed in finding them is negligible due to the difficulty of the underlying cryptosystem, the BDH problem and the DBDH problem.

1. The acceptance by all entities means that  $MAC_1$  in the corresponding message is successfully verified. That is,  $MAC_1$  is verified successfully by using the correct session key  $sk$  and the session dependent random related value  $R_1$ . We show that if it is the case that entities accept the message and continue the session, then the probability that the adversary have modified the message being transmitted is negligible. And the only way for the adversary to find the session key or security related information is to solve the difficulty of the underlying cryptosystem, the BDH problem and the DBDH problem.
2. Now, we consider the active adversary with following cases.
  - (a) There is no way that an adversary could get the secret information  $\{s_1, s_2, \dots, s_l\}$  due to the difficulty of the BDH problem and the DBDH problem.
  - (b) An adversary cannot impersonate  $A$  or  $B$  to cheat the others in the hierarchy. That is the attacker cannot generate valid message without deriving the correct session key  $sk$ , since the attacker cannot pass the verification of  $MAC_1$  in the protocol.
  - (c) An adversary cannot compute session key from any useful information from outside of the hierarchy nodes or gathered information from the network, which are the identity tuples  $(ID_1, \dots, ID_m)$  of  $A$  and  $(ID'_1, \dots, ID'_n)$  of  $B$  and the message  $\{R_1, MAC_1\}$  due to the difficulty of the underlying cryptosystem, the BDH problem and the DBDH problem.

Finally, we could say the  $HNAKA_{fresh}$  is secure against active attack.

## 5 Conclusion

Recently, Guo et al. proposed an efficient and non-interactive hierarchical key agreement protocol applicable to mobile ad-hoc networks, which is a try to solve the open question: How can secrets be established if an adversary can eavesdrop on every message exchange? However, their protocol could not support freshness of the session key. Thereby, we proposed a freshness preserving hierarchical key agreement protocol, named  $HNAKA_{fresh}$ , over the hierarchical MANETs. Compared with the other existing protocols, the  $HNAKA_{fresh}$  offers much better performance on the bandwidth consumption, the computational cost, and the storage cost.

**Acknowledgment** This work was supported by the Kyungil University Research Fund and was also partially supported by the National Research Foundation of Korea Grant funded by the Korean Government (MEST) (NRF-2010-0021575).

## References

1. Conti M, Giordano S (2007) Multihop ad hoc networking: the theory. *J IEEE Commun Mag* 45(4):78–86
2. Gopalakrishnan K, Uthariaraj VR (2012) Collaborative polling based routing security scheme to mitigate the colluding misbehaving nodes in mobile ad hoc networks. *Wireless Pers Commun* 67:829–857
3. Dutta R, Dowling T (2011) Provably secure hybrid key agreement protocols in cluster-based wireless ad hoc networks. *Ad Hoc Netw* 9:767–787
4. Yang H, Luo H, Ye F, Lu S, Zhang L (2004) Security in mobile ad hoc networks: challenges and solutions. *IEEE Wireless Commun* 11(1):38–47
5. Anjum F, Mouchtaris P (2007) Security for wireless ad hoc networks. Wiley, Hoboken
6. Diffie W, Hellman ME (1976) New directions in cryptography. *IEEE Trans Inf Theory* 22(6):644–654
7. Boneh D, Franklin M (2001) Identity-based encryption from the weil pairing. *Lect Notes Comput Sci* 2139:213–229
8. Sakai R, Ohgishi K, Kasahara M (2000) Cryptosystems based on Pairings. In: Proceedings of the symposium on cryptography and information security 2000
9. Guo H, Mu Y, Lin Z, Zhang X (2011) An efficient and non-interactive hierarchical key agreement protocol. *Comput Secur* 30:28–34
10. Kim H (2011) Location-based authentication protocol for first cognitive radio networking standard. *J Netw Comput Appl* 34:1160–1167

# A Deployment of RFID for Manufacturing and Logistic

Pacharaporn Choeksuwan and Somsak Choomchuay

**Abstract** Widespread adoption of Radio Frequency Identification System (RFID) has been widely used to develop and improve the Supply chain management. The Logistic control system has been focused in this work. In details, the system emphasizes at packaging, inventory and warehouse utilization that actually holds a good impact to final packing process and shipping process. The RFID hardware's specification are UHF frequency at 860–930 MHz on ISO/IEC 18000-6C Class 1 Gen 2 which well known to utilize in supply chain management. Data stored in the tag are the compressed version of all pieces inside the box. No indexing is further needed as required by the conventional barcode system. For the analysis of business deployment, an economic analysis tool is employed. The Net Present Value (NPV) of 5 year-period has been carried out for the purpose of medium-term to long-term investment.

**Keywords** Radio frequency identification system · Supply chain · Logistic Net present value

## 1 Introduction

RFID technology has grown rapidly for authentication applications on automatic identification and widely considered to represent the next generation beyond ubiquitous 1-D and 2-D barcode. One of the key emerging technologies is the

---

P. Choeksuwan (✉) · S. Choomchuay  
Department of Electronic Engineering, King Mongkut's Institute of Technology  
Ladkrabang, Bangkok, Thailand  
e-mail: patcharaporn@yahoo.com

S. Choomchuay  
e-mail: kchsomsa@kmitl.ac.th

opportunities and challenges related to the deployment of Electronic Product Code (EPC) and RFID as the unique identify object of each manufacturer. The relative impact on EPC/RFID attractiveness will vary based on each product's unique characteristic and each company's specific requirement.

EPC/RFID tag as well as barcode system are generally used for the same purpose as the product identification. To consider a tiny product or pallet packed in a small box with identification number either in a form of a barcode or an EPC tag. Many of these pallets are again packed in a rather bigger box. That box is of course identified unique number. Now, the real time awareness without going back to recheck the database can be achieved. Likewise manual checking can still be possible with some cost savings. These increase overall performance of supply chain.

Nowadays, RFID is rapidly becoming a cost-effective technology. One of the benefit potential is to reduce the cost of the system deployment that can be done in three major keys: hardware, software and services. Tag costs are one of the key considerations in RFID deployment. Tags are available in various design forms and memory size. Basically based on the application or where the tags are used.

In this work we concentrate on the hardware side that the user's memory size of the tag is a constraint. The main contributions are (1) the method to manage the big information to be able to store in a limited memory portion, and (2) the study whether such an effort can be deployed economically. In [Sect. 2](#), RFID system and components are given in brief. In [Sect. 3](#), production line manufacturing of the final packing process and shipping process are elaborated. Then in [Sect. 4](#), the details of RLE-like and lossless compression techniques are given. In [Sect. 5](#), the deployment of RFID system that the compressed information had been accommodated to the tag is detailed out. In [Sect. 6](#), the calculation of cost and investment consideration by using economic tool of NPV is studied. Finally, this work is concluded in [Sect. 7](#).

## 2 RFID Systems

### 2.1 Tags

An RFID system uses wireless radio communication technology to uniquely identify tagged objects. RFID tag usually holds an amount of memory for the both the system and the users. For example, the EPC Class 1 Gen 2 hold 4 memory banks; Reserve bank, EPC bank, Tag ID bank, and user memory bank. The user-defined data storage size is made vary from hundreds of byte to few kilobytes. The basic RFID system contains three major types of tags.

- *Passive tags*: There is no power source inside of the design. The power is provided by radio frequency wave sending from the reader to activate the tag. The most common RFID tags being used today are passive UHF RFID tags. The

ISO18000-6C defines the communication interfaces for UHF bands that work on 860–960 MHz frequency ranges. The EPC Class-1 Gen 2 standard is well known for UHF tags that also operate as much longer range. We also use this type of the tag in this experiment.

- *Active tag*: The internal power source is installed inside. The signal strength is higher than the passive one and can be read from the longer distance. The useful is applied to transportation truck system.
- *Semi-passive tags*: The design is a combined strength of each passive and active tag. This type of tags are commonly used in the car park system, smart shelf department store, and etc.

### 2.2 Reader/Interrogator

The components can transmit/receive the information from tags using the radio frequency waves via antenna for communication. The information can be read/written to tags based on the circuits and its associated protocol to protect the anticollision.

### 2.3 Middleware

The applications connect to hardware as reader to collect the unique number of each tag for processing the data by real-time acknowledgement and keep in back-end database for purpose analysis.

## 3 Production Line Manufacturing

Shown below in Fig. 1 is one of the final packing process; starting from the receiving of the parts-bundled plastic pack from the cleanroom. Then in the next process is the shipping process. This is shown in Fig. 2. One may notice some redundant procedures, in particular the barcode scanning. This regard is the major concern of our implementation that RFID can be utilized effectively.

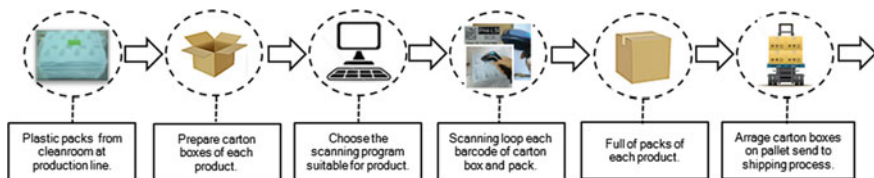


Fig. 1 Final packing process

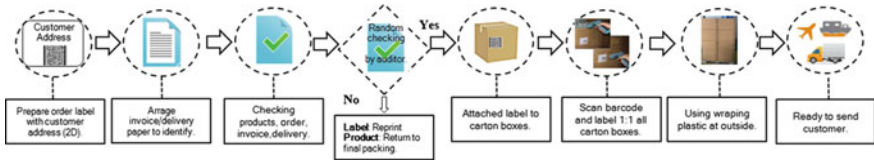


Fig. 2 Shipping process

## 4 Lossless Data Compression

### 4.1 Run Length-Like Coding

Upon the investigation of pattern characteristic of product serial that holds similar behavior of character value such as the same batch of product or the same data of built. Instead of compress the whole original serial shipment data [1], we can compress only those variable portions that hold only 5–8 character digits. Digit 5–12th is considered to lie in the variable concentrated portion. Digit 5–7th are found frequently changed while digit 8–9th are found more dynamic. and digit 10–12th are found rarely changed. For the pre-coding to be more efficient all input identifiers are ascending sorted and then compressed with the RLL-like algorithms. Let’s consider Fig. 3 where 15 identifiers are given as an example.

- X1: 7,5,B,5,Y,5,6,5,V,5,7,5,8,3,C,1,E,1,L,2,M,1,N,2,1,1,2,3,E,1
- X2: 276,5,2BV,5,2Y7,5,8,3,C,1,E,1,L,2,M,1,N,2,1,1,2,3,E,1
- X3: 3768,3,376C,1,376E,1,3BVL,2,3BVM,1,3BVN,2,3Y71,1,3Y72,3,3Y7E,1
- Y1: 1,10,3,5,0,5,4,5,2,5,1,5,9,5,5,5
- Y2: 1,10,3,5,201,5,249,5,225,5
- Y3: 3101,5,3149,5,3325,5
- Z1: digit 8th has less duplicated symbol. The code is *QQV2C3UPFGY0016*
- Z2: digit 9th also has less duplicated symbol. The code is *RVS2NAXHY6BRW07*.

Site Code	Year	Work week	Serial Number						Material Code			Revision	Code	Product Code			
			Frequent change		Mostly change		Rarely change		Same digit in same lot								
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
				Count by column			Z1	Z2	Count by column								
				X1					Y1								
				X2					Y2								
				X3					Y3								
No.	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17
1	T	0	3	0	7	6	8	Q	R	1	0	1	A	3	C	2	J
2	T	0	3	0	7	6	8	Q	V	1	0	1	A	3	C	2	J
3	T	0	3	0	7	6	8	V	S	1	0	1	A	3	C	2	J
4	T	0	3	0	7	6	C	2	2	1	0	1	A	3	C	2	J
5	T	0	3	0	7	6	E	C	N	1	0	1	A	3	C	2	J
6	T	0	3	0	B	V	L	3	A	1	4	9	A	3	C	2	J
7	T	0	3	0	B	V	L	U	X	1	4	9	A	3	C	2	J
8	T	0	3	0	B	V	M	P	H	1	4	9	A	3	C	2	J
9	T	0	3	0	B	V	N	F	Y	1	4	9	A	3	C	2	J
10	T	0	3	0	B	V	N	G	6	1	4	9	A	3	C	2	J
11	T	0	3	0	Y	7	1	V	B	3	2	5	A	3	C	2	J
12	T	0	3	0	Y	7	2	0	R	3	2	5	A	3	C	2	J
13	T	0	3	0	Y	7	2	0	W	3	2	5	A	3	C	2	J
14	T	0	3	0	Y	7	2	1	0	3	2	5	A	3	C	2	J
15	T	0	3	0	Y	7	E	6	7	3	2	5	A	3	C	2	J

Fig. 3 Arrangement of each character on 15 identifiers sampled parts



As a result of grouping, 9 patterns can be re-arranged. These are  $(X1 + Y1 + Z1 + Z2)$ ,  $(X1 + Y2 + Z1 + Z2)$ ,  $(X1 + Y3 + Z1 + Z2)$ ,  $(X2 + Y1 + Z1 + Z2)$ ,  $(X2 + Y2 + Z1 + Z2)$ ,  $(X2 + Y3 + Z1 + Z2)$ ,  $(X3 + Y1 + Z1 + Z2)$ ,  $(X3 + Y2 + Z1 + Z2)$ , and  $(X3 + Y3 + Z1 + Z2)$ . It is found that the group contains X1 or X2 shows similar compression ratio and better than that contains X3. With this RLL-like technique, the data of 12,144 bits can be reduced to 1,992 bits or 84 % after compression.

### 4.2 Huffman Coding and Arithmetic Coding

Two most common statistical compression methods are Huffman and Arithmetic coding. Traditional Huffman utilizes a static table to represent all the characters with their frequencies and then generates a probabilities code table to Huffman tree. In a Huffman tree where each node is the sum of its children have weighted sum of the leaves. For Adaptive Huffman coding, the tree and corresponding encoding scheme change accordingly base on technique of algorithm FGK. Arithmetic coding represents frequently characters using low bit and infrequently characters using high bit. The adaptive arithmetic model keeps the symbols, their counts frequencies of occurrence and their cumulative frequencies. The frequencies could be changed each time it is encoded and update the cumulative frequencies. We use the RLE-like compressed output to be the input of either Huffman or Arithmetic compression. As a result the data size can be further reduced by 40 %.

## 5 Deployment of the RFID System

The significant reduction of raw data provided by the double-stage compression detailed in the previous sections has enabled us to pack the necessary data input the limited RFID user memory. The deployment to an application is quite convinced. The final packing process and shipping process can be combined as shown below in Fig. 4. The tag is write-application by the store section where the typical barcode is attached to the carton box. Data can be revised in next process as read-application to all RFID tags information on pallet during wrapping that must be done before shipment. By doing so, we can save operation times by 2.61 min/unit.



Fig. 4 Flow chart of new process combines final packing process and shipping process

That means if we estimated the production of 100 K/day, we can save 47 min per 8 h rate working. Furthermore the labor force can be reduced from 10 to 7 persons as well as some other assets. Hence the hard cost can be able to quantify to financial statement or comparison. The soft cost such as customer satisfaction, productivity, increasing staff performance, etc. can be included in the key performance indicators (KPIs).

## 6 Cost and Investment Consideration

As the RFID deployment affects new operation processes and company’s financial statements, the cost analysis should be studied carefully. Economic analysis tool should be employed to estimate the value. The Net Present Value (NPV) [2] is the mechanism to understand the cash flow series, where  $V_t$  is the cash flow series at  $t$  time period,  $n$  is the number of analysis time period, and  $D$  is the discount rate at that time (i.e. time value of money).

$$NPV = \sum_{t=0}^n \frac{V_t}{(1 + D)^t} \tag{1}$$

In this work we consider 5-year of period for purpose of medium-term to long-term investment classified by fix cost/one-time cost and annual cost. In connection with the deployment of RFID-based system, all direct cost and effected cost are listed and/or estimated. As a result the total cost is 346,300.00 and estimated benefits and saving cost is 207,212.75 per year when the value assigned to  $D$  is 0.12 (12 %). NPV calculation for RFID implementation cost is detailed below.

$$NPV_1 = \frac{-346,300}{(1 + 0.12)^0} + \frac{-17,980}{(1 + 0.12)^1} + \frac{-17,980}{(1 + 0.12)^2} + \frac{-17,980}{(1 + 0.12)^3} + \frac{-17,980}{(1 + 0.12)^4} + \frac{-17,980}{(1 + 0.12)^5} = -411,113.88 \tag{2}$$

And NPV calculations for estimated benefits and saving cost,

$$NPV_B = 0 + \frac{207,212.75}{(1 + 0.12)^1} + \frac{414,425.50}{(1 + 0.12)^2} + \frac{621,638.25}{(1 + 0.12)^3} + \frac{828,851.00}{(1 + 0.12)^4} + \frac{1,036,063.75}{(1 + 0.12)^5} = 2,072,498.87 \tag{3}$$

The comparison between RFID implementation cost and estimated benefits and saving cost can be: loss of investment 43.12 % in the first year, in the 2nd year gain a small investment at 8.41 %, the 3rd year at 55.32 %, the 4th year at 98.19 % and the 5th year at 137.52 %. The value investment is double in the next year in this estimation. The long-term trend increases average gain investment 43 % per year.

## 7 Conclusion

The RFID deployment has been considered under medium-term to long-term project investment. The develop operation process have affected the financial statement, process reengineering, lean manufacturing and section labor (training performance, working hours, etc.). In the first year of investment cannot gain investment due to big component of an RFID installation, but in the next year can achieve the investment and increase by double in next year until time of period. The NPV calculation is the mechanism cash flow series that can be one tool for the investor decision on business.

**Acknowledgments** This work is supported by Industry/University Cooperative of Data Storage Technology and Applications Research Center (I/UCRC), King Mongkut's Institute of Technology Ladkrabang and National Electronic and Computer Technology Center (NECTEC), National Science and Technology Development Agency (NSTDA) under scholarship HDD-01-52-11 M. We also would like to thank Compart Precision (Thailand) company for their continue support to this research.

## References

1. Choeksuwan P, Choomchuay C (2010) A technique for label text compression applied to RFID passive tag. In: The 33rd electrical engineering conference, pp 1001–1004
2. Ozelkan EC, Sireli Y, Munoz MP, Mahadevan S (2006) A decision model to analyze costs and benefits of RFID for superior supply chain performance. In: Technology management for the global future, pp 610–617

# Real Time Video Implementation on FPGA

Pham Minh Luan Nguyen and Sang Bock Cho

**Abstract** Nowadays, real time video becomes popular in a lot of multimedia equipment, video cameras, tablets, camcorders. The more hardware improved the more application is used. Requesting a faster and cost-effective systems there are triggers a shift to Field Programmable Gate Arrays (FPGAs), where the inherent parallelism results in better performance. The implementation is based on efficient utilization of embedded multipliers and look up table (LUT) of target device to improve speed but also saves the general purpose resources of the target device. This paper proposes new hardware architecture for capture NTSC/PAL video stream. The whole system is implemented on a single low cost FPAG chip, capable of real time procession at frequency 60 MHz. In addition, to increase real-time performance, hardware architecture with streamlined data flow are developed.

**Keywords** FPGA · Video processing · Image processing · Real time processing

## 1 Introduction

In recent years, automated video surveillance system is developing and applying massive. That is enable when the progress in technology scaling more robust computationally intensive algorithms. The advantage of surveillance automation over traditional television based on system lies in the fact that it is a self contained

---

P. M. L. Nguyen (✉) · S. B. Cho  
School of Electrical Engineering, University of Ulsan, Ulsan, Korea  
e-mail: npmluan@gmail.com

S. B. Cho  
e-mail: sbcho@ulsan.ac.kr

system capable of automatic video processing. That is a request for a real-time video system. The implementation system on FPGA gets more advantages than the other hardware. The FPGA and SOC products improved fast. The traditional hardware implementation of image processing uses Digital Signal Processors (DSPs) or Application Specific Integrated Circuits (ASICs). An advanced system can process video stream real time, process image from video frames in time. As a logic capacity of FPGAs increases, they are being increasing used to implement large arithmetic-intensive applications. Since data-path circuits are designed to process multiple-bit-wide data, FPGAs implementing these circuits often have to transport a large amount of multiple-bit-wide signals from one computing element (such as a logic block, a DSP block, or a multi addressable memory cell) to another. With the advent of FPGAs having greater processing capability, it has been regarded as a useful means for implementing algorithms that massive parallelism is required.

In [9], the authors designed system based on Virtex-4 XC4VLX200-10 from Xilinx and two VCC-8350CL cameras there is more complicated in procession data. There also increase computationally in program. In [2], the system also implemented on Xilinx, Virtex-II pro vp30 FPGA. The camera in [2] is Kodak Kac-9648. In this case, the system has to reduce memory usage but they get more noise in results.

In this paper, we present hardware architecture capable of real time video processing with a screen resolution 30 frames per second. The paper is organized as follows. [Section 2](#) discusses the proposed implementation hardware and results. Finally, conclusions are covered in [Sect. 3](#).

## 2 Implementation FPGA

### 2.1 The Observation System

We use the system included: camera ST-400CD, kit Altera DE2\_115, LCD Monitor (Flatron Wide). There are some detailed descriptions of camera ST-400CD: Standard Camera Sonny 1/3" Super HAD Color CCD Digital Signal Processing, Video Auto Iris Lens.

The board we use in this paper is Altera Kit DE\_115 Development. It was implemented with a Altera FPGA chip, EP4CE115F29C7 (Cyclone IV E), that has 11,4480 LEs, 432 M9 K memory blocks, 3,888 embedded memory (Kbits), 4 PLLs, 528 maximum user I/Os, 230 maximum differential channels. The following hardware is provided on the DE2-115 board: Two 64 MB SDRAM, 2 MB SRAM, 8 MB, flash memory, VGA DAC with VGA out connector. The main improvement of our work include: the development of hardware and software components for a flexible, powerful and low-cost video processing engine, and the use of techniques such as run-time reconfiguration. The implementation has been placed

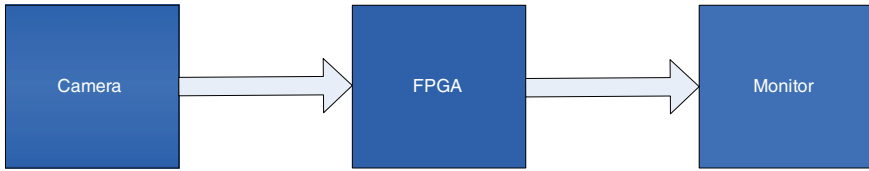


Fig. 1 Real-time video system

and routed using Quartus II v.11.1. The operating frequency of the design on DE2-115 is 50 MHz. Video stream is captured by camera. Video signal transmits to FPGA. Processing video stream on board FPGA, the output signal sends to VGA port. LCD Monitor gets video signal and capture it out screen. There is low cost and high performance system (Fig. 1).

Real-time image processing requires high computation power. For example, the NTSC video standard requires 30 frames second, with approximately 0.25 mega pixel per second. PAL video standard has a similar processing load with 25 frames per second, but the frame size is larger. The amount of processing required per pixel depends on the image processing algorithm. In our proposed hardware architecture, we design data flow from captured signal by camera to VGA output signal. In Fig. 2, there is data processing stream from TV decoder chip to VGA chip.

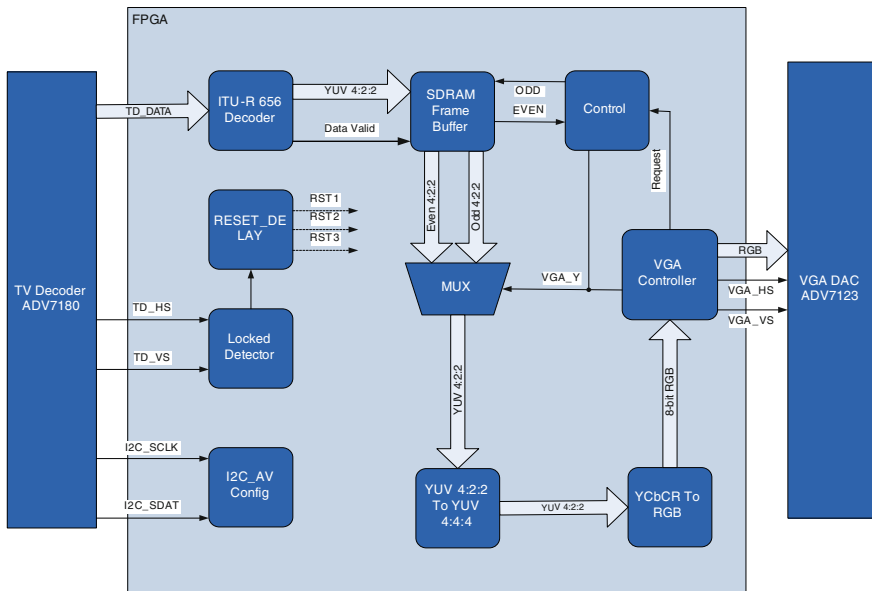


Fig. 2 Data flow in and out FPGA

## 2.2 Block Processing

- I2C Block (I2C\_AV Config): I2C block uses to connect FPGA and TV Decoder chip. Two signals I2C\_SClk, I2C\_SDATA connect to SCLK pin and SDA pin of Decoder chip. Data signal transmit in 3 bytes per frame.
- Locked detector block: This block check input conditions from VS and HS to know how the chip worked. When chip works, the TD\_Stable is high to make Reset\_Delay block work.
- ITU-R656 Block: ITU-R BT 656 describes a simple digital video protocol for streaming uncompressed PAL or NTSC Standard Definition TV (525 or 625 lines) signals. The protocol builds upon the 4:2:2 digital video encoding parameters defined in ITU-R Recommendation BT.601, which provides interlaced video data, streaming each filed separately, and uses the YCbCr Color space and a 13.5 MHz sampling frequency for pixels. The standard can be implemented to transmit either 8-bit values (the standard in consumer electronics) or 10-bit values (sometimes used in studio environments).
- SDRAM Frame buffer, MUX, VGA Controller and Control: We used the SDRAM Frame Buffer and a field selection multiplexer (MUX) which is controlled the VGA controller to perform the de-interlacing operation. Internally, the VGA Controller generates data request and odd/even selection signals to the SDRAM Frame Buffer and filed selection multiplexer (MUX).
- YUV 4:2:2 To YUV 4:4:4. The YUV422 to YUV444 block converts the selected YcrCb 4:2:2 (YUV 4:2:2) video data to the YcrCb 4:4:4 (YUV 4:4:4) video data format (Figs. 3, 4, 5).

## 2.3 Results

From our architecture, we do some experiments to get results. There are some data about our implementation hardware on Cyclone IV E in Table 1.

**Fig. 3** I2C config block connection



**Fig. 4** Locked detector block connection

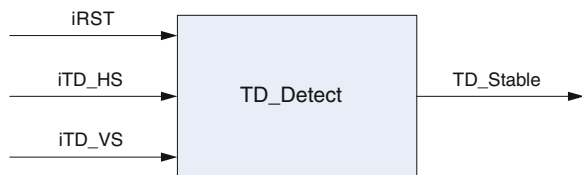




Fig. 5 ITU-R656 decoder block connection

Table 1 Usage percents on Cyclone IV E

	Origin	Usage	Usage percents (%)
Logic elements	114,800	1,695	1
Memory bits	3,981,312	45,028	1
Embedded multiplier 9-bit elements	432	18	4
Phase-locked loops (PLLs)	4	1	25

Fig. 6 Before download program on FPGA

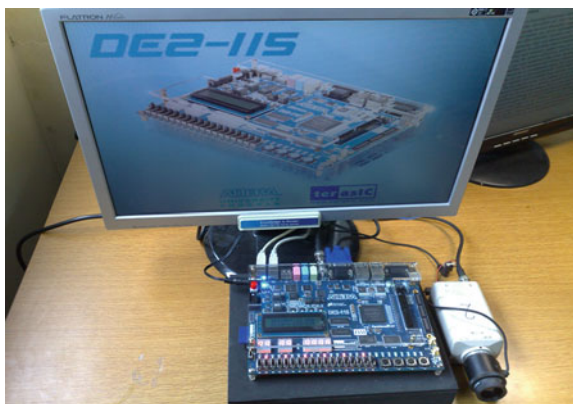


Fig. 7 After download program on FPGA





The time download program from PC to board about 20 s. The video displays on LCD response real time when we change the screen. In Fig. 6, when we connect LCD and DE2-115 board, the screen on LCD displays the image about DE2-115. That is original program on DE2-115. In Figs. 6 and 7, when we load the program to FPGA, the camera will capture the environment image, and video will display on LCD.

### 3 Conclusions

In this paper, we proposed new implementation hardware to get real video from camera by FPGA implementation. We get real video stream from camera that is very important detail. We apply some image processing application on real time in future.

We use DE2\_115 development kit for our study. This board gives more advance condition with video port, VGA port especial consumer chip Cyclone IV E.

### References

1. Abutaleb MM, Hamdy A, Saad EM (2008) FPGA-based real time video object segmentation with optimization schemes. *Int J Circ Syst Signal Process* 2:78–86
2. Jiang H, Owall V, Ardo H (2006) Real-time video segmentation with VGA resolution and memory bandwidth reduction. In: *Proceeding of the IEEE international conference on video and signal based surveillance*, Sydney
3. Lapalme FX, Amer A, Wang C (2006) FPGA architecture for real-time video noise estimation. In: *IEEE international conference on image processing*. Atlanta, pp 3257–3260
4. Chen PP, Ye A (2011) The effect of multi-bit correlation on the design of field-programmable gate array routing resources. *IEEE Trans VLSI Syst* 19:283–294
5. Cho J, Jin S, Kwon KH, Jeon JW (2010) A real-time histogram equalization system with automatic gain control using FPGA TIIS 633–654
6. DE2-115 user manual (2011)
7. Altera (2011) Introduction to the Altera SOPC builder using verilog designs
8. Altera (2011) Using the SDRAM on Altera's DE2-115 board with verilog designs
9. Jin S, Cho J, Pham XD, Lee KM, Park SK, Jeon JW (2010) FPGA design and implementation of a real-time stereo vision system. *IEEE Trans Circ Syst Video Technol* 20:15–26

# Recovery Algorithm for Compressive Image Sensing with Adaptive Hard Thresholding

Viet Anh Nguyen and Byeungwoo Jeon

**Abstract** Iterative hard thresholding (IHT) algorithm is one of the representative compressive sensing (CS) reconstruction algorithms. For applying to images, however, it has a problem of lacking in addressing human visual system (HVS) characteristics—its hard thresholding process treats all of coefficients in transform domain equally. To overcome the problem, this paper addresses an adaptive hard thresholding method accounting for the HVS characteristics. For this purpose, a suitable threshold level is adaptively selected for each coefficient in transform domain by utilizing the standard weighting matrix table used in JPEG together with the threshold value which is estimated over the noisy version of image. Experimental results show that the performance of the block compressive sensing with smooth projected Landweber (BCS-SPL) with the proposed adaptive hard thresholding algorithm remarkably outperforms that of the conventional BCS-SPL algorithm.

**Keywords** Compressive image sensing · Adaptive hard thresholding

## 1 Introduction

In conventional digital image acquisition and compression system, the encoding process is very time-consuming since all of transform coefficients had to be calculated even though most of them were discarded in quantization process. Obviously, the classical transform-coding procedure demands much of computational power and

---

V. A. Nguyen (✉) · B. Jeon  
School of Electrical and Computer Engineering Sungkyunkwan University,  
300 Chunchun-dong, Jangan-gu, Suwon, Korea  
e-mail: vietanh@skku.edu

B. Jeon  
e-mail: bjeon@skku.edu

memory storage. Recently, a novel sampling paradigm called compressive sensing (CS), which directly acquires compressible signal at a sub-Nyquist rate [1], has been being developed to overcome this problem. Due to the simplicity of signal acquisition, CS has drawn great attention. As a result, many efficient reconstruction algorithms have been proposed. Iterative hard thresholding (IHT) algorithm [2] is one of the promising compressive sensing reconstruction algorithms. IHT algorithm not only is very easy to implement in practical application and extremely fast but also has a strong performance guarantees in term of recovery error as shown in [3].

In applying to images, IHT algorithm has some problems. In fact, the hard thresholding process treats all coefficients of image in transform domain equally; any coefficients whose magnitude is lower than a threshold level is considered as noise and will be replaced by zero. However, the low frequency coefficients of image in transform domain are very important because of two reasons: energy in most of the natural images is mostly concentrated at the low frequency bands; the HVS is more sensitive to the loss of low frequency components than that of high frequency components. Moreover, the threshold value ( $\tau$ ) in the hard thresholding process is inaccurate since it is estimated from a noisy version of image. Therefore, applying the threshold level to all coefficients in transform domain may wrongly discard some important low frequency coefficients. As a result, the quality of reconstructed image can be degraded. Note that the iterative hard thresholding algorithm does not account for the HVS characteristics—the coefficients which are discarded in the hard thresholding process may contain visually significant information.

JPEG uses a standard weighting matrix table [4] for quantization which is designed based on human visual system (HVS) characteristics for compressing images without much visible artifacts. The quantization step size at frequency location  $(u, v)$ , denoted by  $Q_{u,v}$ , is obtained by multiplying a common factor (which is irrespective of frequency location) to a weight corresponding to the frequency position  $(u, v)$  in the matrix table. The weight is chosen as the perceptual threshold [5]. A higher quality factor leads to better image quality. In the quantization process, a coefficient whose magnitude is smaller than half of  $Q_{u,v}$  becomes zero [4]; it indirectly suggests that such a small value can be visually less significant. Motivated by this observation, we adaptively select a suitable threshold among  $Q_{u,v}/2$  and  $\tau$  which is estimated using a noisy version of image for applying to each coefficient to avoid wrongly discarding visually significant information in the hard thresholding process. In this way, the hard thresholding method can be made to account for the HVS characteristics.

In this paper, in order to evaluate the performance of the proposed method, we apply an adaptive hard thresholding method to the block compressive sensing with smoothed projected Landweber (BCS-SPL) algorithm—a prominent application of iterative hard thresholding algorithm to images [6, 7]. The objective quality (PSNR) of reconstructed image of BCS-SPL with the proposed adaptive hard thresholding method is compared to that of the conventional BCS-SPL algorithm.

The rest of the paper is organized as follows. [Section 2](#) reviews the fundamental of compressive sensing and the structure of BCS-SPL algorithm. Furthermore, [Sect. 3](#) presents the proposed method. Simulation results are illustrated in [Sect. 4](#). Finally, [Sect. 5](#) draws our conclusion.

## 2 Background

### 2.1 Compressive Sensing

Compressive sensing (CS) theory is built upon the work of Candès et al. [8], and of Donoho [9], who show that a finite-dimensional signal having a sparse (contains a lot of zero entries) or compressible representation in a selected domain (e.g., DCT, DWT) can be reconstructed from a small set of linear, non-adaptive measurements. Stated another ways, CS directly acquires signal in a compressed form by projecting it into a sensing matrix. In its study, designing a sensing matrix and designing a reconstruction algorithm are two main problems. The sensing matrix needs to satisfy some requirement conditions [1] (e.g., RIP condition, incoherent condition) to warrant a unique solution. The recovery algorithm should be fast and have a good performance guarantees in term of recovery error [3].

In being applied to images, a challenge for compressive sensing is to reduce the computational complexity in reconstruction process. The commonly used sensing matrix—i.i.d. Gaussian matrix is an unstructured random matrix, and its high complexity of matrix multiplication makes the reconstruction process slow, especially for a large-sized image. Furthermore, a huge memory is necessary to store the large-sized random matrix. In this context, due to the block-based processing, the BCS-SPL (Block Based CS with Smoothed Projected Landweber) algorithm seems to be suitable for applying to a large-sized image.

### 2.2 Block Based CS with Smoothed Projected Landweber Structure (BCS-SPL)

Figure 1a illustrates the structure of the BCS-SPL algorithm, which consists of block compressive sensing (BCS) at encoder side and the smoothed projected Landweber (SPL) reconstruction algorithm at decoder side. In BCS, an image is divided into small blocks. Then, each block is sampled by projecting it into a compact size of a random measurement matrix (e.g., i.i.d. Gaussian matrix). The measurement vectors of each block are transmitted to decoder. Due to the block-based processing, BCS reduces not only memory for storing the measurement matrix but also computational complexity in reconstruction process.

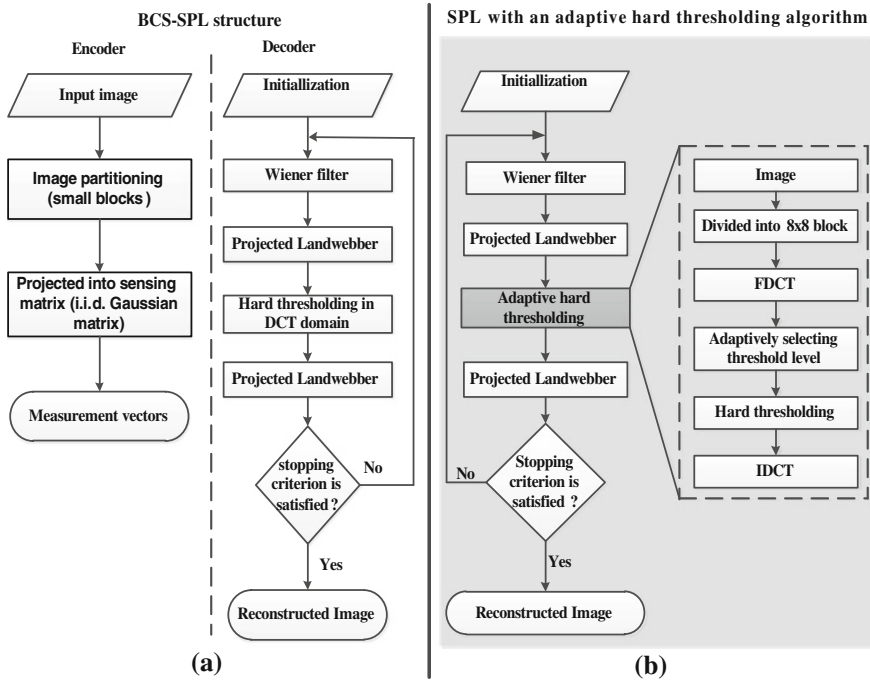


Fig. 1 a BCS-SPL structure; b SPL algorithm with an adaptive hard thresholding method

In the SPL reconstruction algorithm, an image is iteratively reconstructed. In each iteration, a procedure which incorporates Wiener filter and variant projected Landweber (i.e., namely hard thresholding and projected Landweber framework (PL) [7]) is performed using a noisy version of image. The wiener filter is applied in spatial domain of image signal for removing noise and blocking artifacts as well. The approximation of reconstructed image is calculated by projecting it into projected Landweber (PL) framework. The hard thresholding process imposes sparsity of image signal in transform domain (e.g., DCT) by eliminating any coefficient whose magnitude is less than a given threshold level. The level is calculated using the noisy version of image based on a universal threshold method [10]. In addition, if the stopping criterion is satisfied, SPL algorithm is terminated (see [7] for more details).

### 3 Proposed Method

As mentioned above, in order to improve the quality of reconstructed image, it is important to adaptively choose an appropriate threshold level for applying to each frequency band of image according to the HVS characteristics. The SPL algorithm with an adaptive hard thresholding method is illustrated in Fig. 1b.

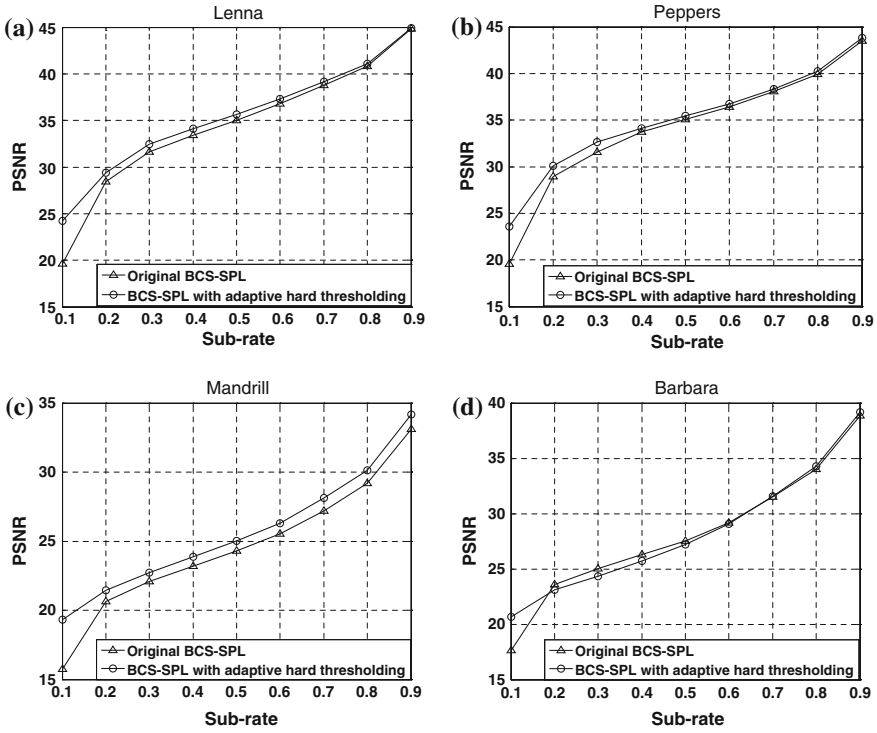
Firstly, the given image is divided into  $8 \times 8$  blocks each of which is put to the forward discrete cosine transform (FDCT). Before the hard thresholding process, we select a threshold value for each DCT coefficient between  $Q_{u,v}/2$  and  $\tau$ . The standard weighting table is derived from the psycho-visual experiments [5]; the weight values in the table are chosen as perceptual thresholds. If quantization step size  $Q_{u,v}$  is larger than the weight values in the standard weighting matrix, visible artifact will occur [4]. Therefore, we regard the weighting values as the quantization step size values and the resulting maximum quantization error is equal to  $Q_{u,v}/2$ . Thus, the coefficients whose magnitudes are larger than  $Q_{u,v}/2$  (i.e., the half of their corresponding quantization step size values) are considered as visually significant information and should be preserved even though their magnitudes are lower than  $\tau$ . Stated another way, for each coefficient, if  $Q_{u,v}/2 \leq \tau$ , the threshold level will be set as  $Q_{u,v}/2$ . Moreover, in the hard thresholding process,  $\tau$  is used to differentiate signals and noises; the coefficients whose magnitudes are larger than  $\tau$  are considered as signals and should be preserved. That is, for each coefficient, if  $Q_{u,v}/2 > \tau$ , the threshold level will be set as  $\tau$ . By adaptively applying the threshold level to each coefficient, we prevent the hard thresholding process from wrongly eliminating the visually significant information.

In following section, we evaluate the performance of our proposed method by comparing it with the conventional BCS-SPL algorithm.

## 4 Experimental Result

In our simulation, four  $512 \times 512$  gray-level benchmark images: Lena, Barbara, Mandrill, and Peppers, are used as input. Block size is set as  $8 \times 8$ . The objective quality (PSNR) is used to compare the performance of the proposed method with that of the conventional BCS-SPL algorithm [6].

Figure 2 shows that the PSNR performance of the proposed method remarkably outperforms that of the conventional BCS-SPL algorithm. Specially, at sub-rate 0.1, the PSNR of the proposed method is much better than that of the conventional BCS-SPL algorithm—PSNR gains of about 4.6, 4, 3.6, and 3 dB in case of Lena, Peppers, Mandrill, and Barbara, respectively. At higher sub-rates (from 0.2 to 0.9), the PSNR performance of the proposed method is also higher than that of the conventional BCS-SPL algorithm—average PSNR gains under sub-rates 0.2–0.9 are about 0.5, 0.5, and 0.8 dB in case of Lena, Peppers, Mandrill, respectively. Actually, at a sub-rate of 0.1, tested image contains a lot of noise, so the estimated threshold value from the noisy version of image is very large. Therefore, a lot of significant coefficients in the low frequency region are lost in the hard thresholding process. Moreover, except Barbara image, in the other images (Lena, Mandrill, and Peppers), most energy is concentrated at low frequency bands; therefore, by adaptively applying a suitable threshold level in each frequency band of transform



**Fig. 2** Performance comparison between the proposed method and the conventional BCS-SPL [6] on test image: **a** Lena; **b** Peppers; **c** Mandrill; and **d** Barbara

domain, we prevent a hard thresholding process from losing the visually significant information of image.

In case of Barbara image (an image with much detail), energy is mostly concentrated at high frequency region. After the filtering process, the high frequency components are distorted since the Wiener filter is used as a low pass filter [7]. Therefore, energy distribution of the image in DCT domain is changed (the energy in low frequency region is more than that in high frequency region), which leads the adaptive hard thresholding algorithm wrongly selects a suitable threshold level for applying to each coefficient. As a result, from sub-rates of 0.2–0.5, the averaged PSNR of the proposed method is less than that of the conventional algorithm (about 0.5 dB). At sub-rates from 0.6 to 0.9, the energy distribution of the image may not be affected much by the filtering process since not much noise is contained in the image. Therefore, the adaptive hard thresholding method may successfully select the appropriate threshold level for applying to each coefficient. As the result, the performance of our proposed method is slightly better than that of the conventional BCS-SPL algorithm.

## 5 Conclusion

In this paper, we propose an adaptive hard thresholding method accounting for the HVS characteristics. By adaptively applying the suitable threshold level by referring to HVS in each frequency band, the proposed method not only pursues the sparsity of signal but also preserves the visually significant information. Simulation results showed the proposed adaptive hard thresholding algorithm enhanced the BCS-SPL algorithm.

**Acknowledgments** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-001-7578).

## References

1. Baraniuk RG (2007) Compressive sensing. *IEEE Signal Process Mag* 24:118–121
2. Blumensath T, Davies ME (2008) Iterative thresholding for sparse approximation. *J Fourier Anal Appl* 14:629–654
3. Eldar YC, Kutyniok G (2012) Compressive sensing theory and applications. Cambridge University Press, Cambridge
4. Wallace GK (1992) The JPEG still picture compression standard. *IEEE Trans Consumer Electron* 38:18–34
5. Lohscheller H (1984) A subjectively adapted image communication system. *IEEE Trans Commun* 32:1316–1322
6. Fowler JE, Mun S, Tramel EW (2012) Block-based compressed sensing of images and video. *Found Trends Signal Process* 4:297–416
7. Gan L (2007) Block compressed sensing of natural images. In: Proceedings of international conference on digital signal processing. Cardiff, UK, pp 403–406
8. Candès EJ, Romberg J, Tao T (2006) Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans Inf Theory* 52:489–509
9. Donoho DL (2006) Compressive sensing. *IEEE Trans Inf Theory* 52:1289–1306
10. Donoho DL (1995) De-noising by soft-thresholding. *IEEE Trans Inf Theory* 41:613–627



# Estimation Value for Three Dimension Reconstruction

Tae-Eun Kim

**Abstract** This paper deals with a fundamental problem for 3D model acquisition after camera calibration [1]. We present an approach to estimate a robust fundamental matrix for camera calibration [2, 3]. Single axis motion can be described in terms of its fixed entities, those geometric objects in space or in the image that remain invariant throughout the sequence. In particular, corresponding epipolar lines between two images intersect at the image of the rotation axis. This constraint is then used to remove the outliers and provides new algorithms for the computing the fundamental matrix. In the simulation results, our method can be used to compute the fundamental matrix for camera calibration more efficiently.

**Keywords** 3D Reconstruction

## 1 Proposed Calibration Approach

In the past few years, the growing demand of realistic three-dimensional object models for graphic rendering, creation of nonconventional digital libraries, and population of virtual environments has renewed the interest in the reconstruction of the geometry of 3D objects from one or more camera images. One of the simple and robust methods to acquire 3D models from image sequences is using turntable motion. Therefore, it has been widely used by computer vision and graphics researchers. Turntable motion refers to the situation where the relative motion between a scene and a camera can be described as a rotation about a fixed axis. The motion is a practical case of the more general planar motion as all rotations are restricted to be around the same axis. The fundamental task for acquiring 3D

---

T.-E. Kim (✉)

Department of Multimedia, Namseoul University, Cheonan, Korea  
e-mail: tekim5@empas.com

models from single axis motion is to recover the camera parameters and the relative pose of the cameras. The estimation of the camera positions, or simply the rotation angles relative to a static camera, is the most important and difficult part of the modeling process [4, 5]. Traditionally, rotation angles are obtained by careful calibration. Fizgibbon extended the single axis approach to recover unknown rotation angles from uncalibrated image sequences based on a projective geometry approach. However, fundamental matrices and/or trifocal tensors have to be computed for each pair of images or each triplet of images. In the new algorithms, we try to improve the accuracy by estimating the robust fundamental matrices. Figure 1 shows the flow chart of the 3D reconstruction. From the epipolar geometry of circular motion, we can remove the outliers and estimate the robust fundamental matrix for camera calibration.

The fundamental matrix corresponding to a pair of cameras related by a rotation around a fixed axis has a very special parameterization, as shown in Fig. 2, which can be expressed explicitly in terms of fixed image features under circular motion (image of rotation axis, pole, and horizon, jointly holding 5 dof) Consider the pair of cameras  $P_1$  and  $P_2$ , given by (Fig. 3)

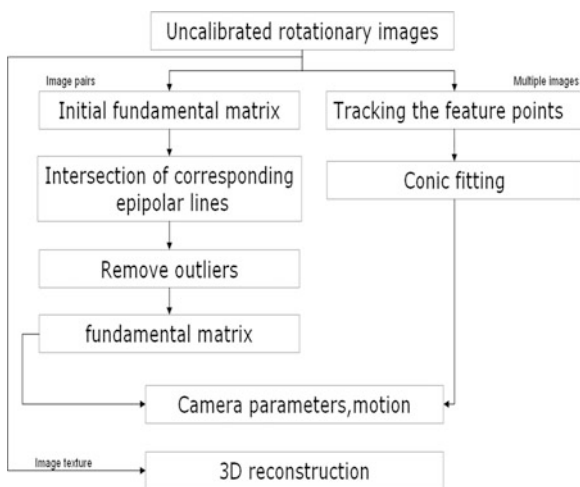
$$P_1 = K[I]t \tag{1}$$

$$P_2 = K [R_y(\theta)|t] \tag{2}$$

where

$$t = [001]^T \text{ and } R_y(\theta) = \begin{bmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{bmatrix}, K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

**Fig. 1** 3D Reconstruction algorithm



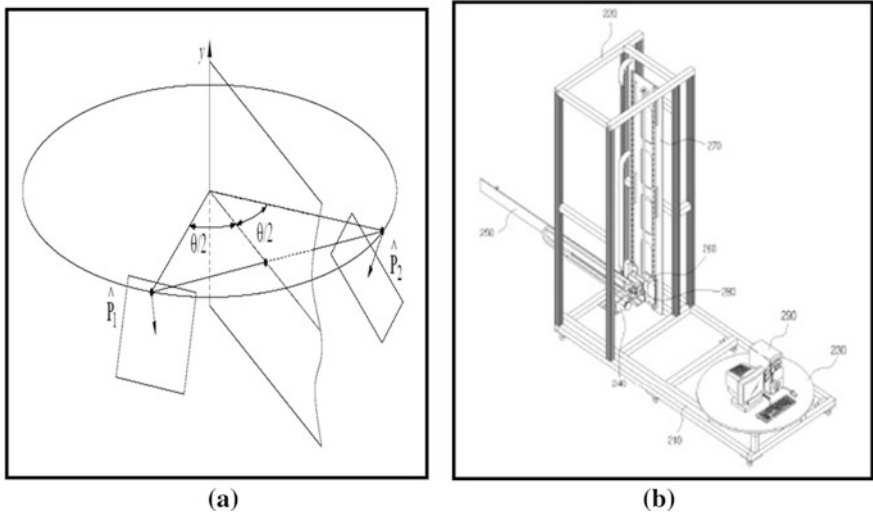
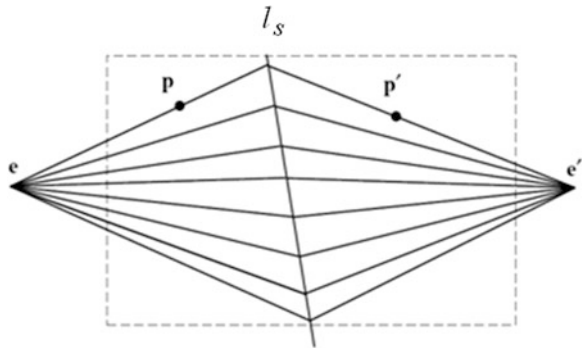


Fig. 2 a Geometry of turntable system, b Image acquisition system

Fig. 3 The epipolar geometry in circular motion. All corresponding epipolar lines must intersect at m (projection of the rotation axis)



Given a static camera and a generic object rotating on a turntable, single axis motion provides a sequence of different images of the object. Single axis motion can be described in terms of its fixed entities, those geometric objects in space or in the image that remain invariant throughout the sequence. In particular, corresponding epipolar lines between two images intersect at the image of the rotation axis. The epipolar constraint is then used to remove the outliers [6].

If the two views are taken by an uncalibrated moving camera, only the epipolar geometry between the two views was estimated. The epipolar geometry can be nicely coded by a  $3 \times 3$  rank 2 matrix  $F$ , called fundamental matrix. Solving fundamental matrix  $F$  is the algebraic representation of the epipolar constraint for the uncalibrated cameras. The epipolar constraint is described as follows: For each

point  $m$  in the 1st image plane, its corresponding point  $m'$  lies on its epipolar line  $l'm$  and similar for any point  $m'$  in the 2nd image plane. This relation can be given as

$$l'_m = Fm \tag{4}$$

$$l_m = F^T m' \tag{5}$$

Since  $m$  lies on  $l'_m$   $m'$  lies on  $l_m$ , following relations are obtained :

$$m'^T Fm = 0 \tag{6}$$

$$m^T F^T m' = 0 \tag{7}$$

where

$$m_i = [u_i, v_i, 1] \tag{8}$$

$$m'_i = [u'_i, v'_i, 1] \tag{9}$$

## 2 Experimental Results

**8-point algorithm.** If  $n$  corresponding points (at least 8) are given, a set of linear equations is obtained as:

$$Af = 0 \tag{10}$$

$$A = \begin{bmatrix} u'_1 u_1 & u'_1 v_1 & u'_1 & v'_1 u_1 & v'_1 v'_1 & v'_1 & u_1 & v_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u'_n u_n & u'_n v_n & u'_n & v'_n u_n & v'_n v'_n & v'_n & u_n & v_n & 1 \end{bmatrix}$$

$$f = \begin{bmatrix} f_{11} \\ f_{12} \\ f_{13} \\ f_{21} \\ f_{22} \\ f_{23} \\ f_{31} \\ f_{32} \\ f_{33} \end{bmatrix}$$

A robust solution of this equation is the eigenvector corresponding to the smallest singular value of  $A$ , that is, the column of  $V$  in the singular value decomposition(SVD) of  $A = UDV^T$ . In order to obtain a unique solution, the rank of  $A$  matrix must be equal to 8. Therefore, the closest singular  $F'$  matrix to  $F$  matrix can be obtained as:

$$F' = U \text{diag}(r,s,0) V^T \tag{11}$$

where  $D = \text{diag}(r,s,t)$  where  $r \geq s \geq t$ .

**8-point algorithm.** Hartly proposed a normalized 8-point algorithm to improve the performance. This normalization was performed by translating the center of corresponding points to the origin of the image reference frame and then scaling the corresponding points, so that the average distance from the origin becomes equal to. Finally, after the calculation of matrix using the 8-point algorithm, it is converted to F matrix of corresponding points before normalization as:

$$F = T_2^T \hat{F} T_1 \tag{12}$$

where T1 and T2 are transformation(normalization) matrices for the first and second images, respectively.

The corresponding epipolar lines between two images intersect at the image of the rotation axis. The epipolar constraint is then used to remove the outliers and estimate the fundamental matrix. Corresponding epipolar lines between two views and the epipolar lines meet at the rotation axis.

The algorithm to estimate the robust fundamental matrix can be summarized as follows:

1. Extract and match points across all images.
2. Compute the initial fundamental matrix using the normalized 8-point algorithm.
3. Compute the intersection of two corresponding epipolar lines.
4. Calculate the geometric error between rotation axis and intersection from the previous computations.
5. Estimate the fundamental matrix by minimizing the cost function over all correspondences.

Algorithm

1. Compute the initial F matrix using the normalized 8- point algorithm
2. Robustly estimate the F matrix

- a. Compute the intersection of two corresponding epipolar lines

$$m_i = l \times l', m_i(u_i, v_i)$$

- b. Calculate the geometric error between  $l_s$  and  $m_i$

$$d(l_s, m_i) = \frac{au_i + bv_i + c}{\sqrt{a^2 + b^2}}, l_s(a,b,c) \text{ image of rotation axis}$$

- c. Minimize the cost function over all correspondences

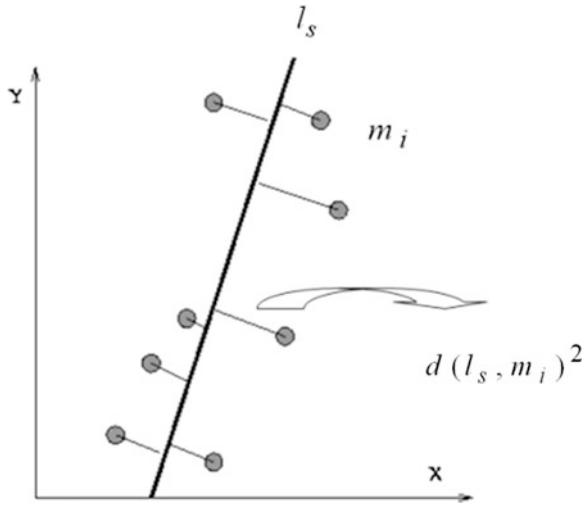


Fig. 4 Intersection of corresponding epipolar and lines geometric distance of rotation axis

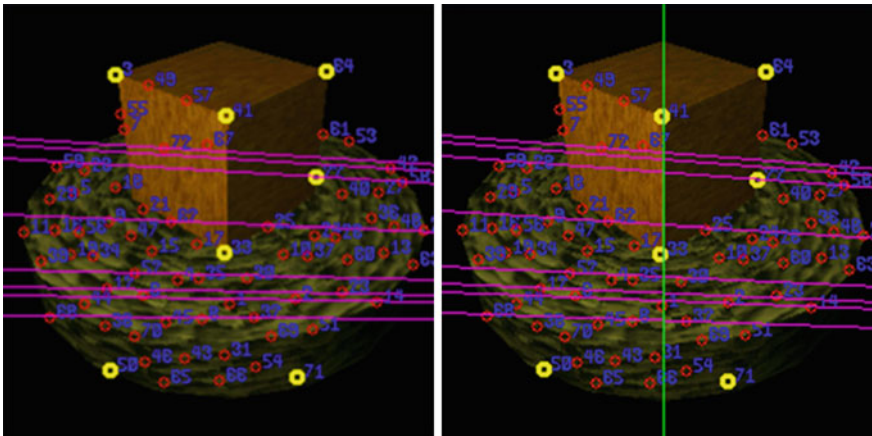


Fig. 5 a Initial epipolar lines and b Rotation axis of synthetic images

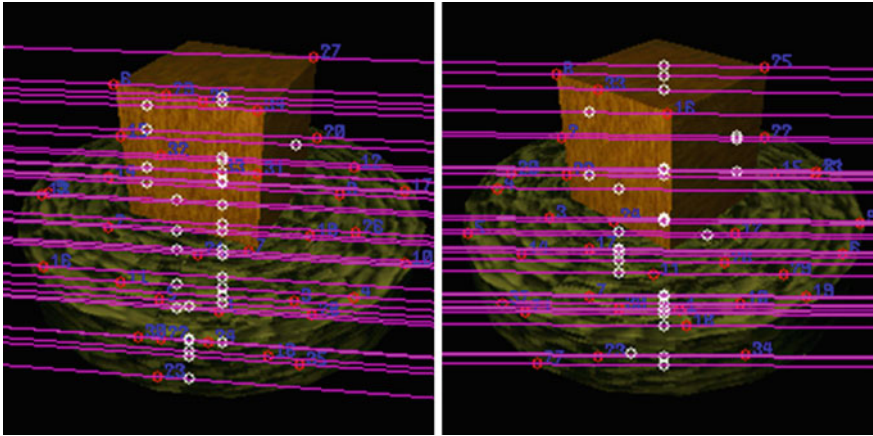


Fig. 6 White circles represent the intersection of two corresponding epipolar lines

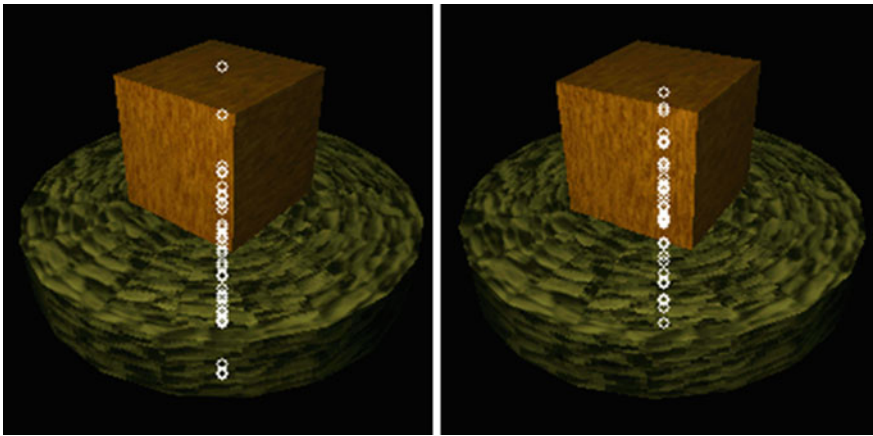


Fig. 7 Inliers of the intersection of epipolar lines where outliers are removed

### 3 Conclusion

In this paper, we have presented a simple and practical approach for computing the fundamental matrix to estimate the camera parameters. In this system, we need only uncalibrated images of a turntable sequence for input. The strong epipolar constraint can be used for estimating the robust fundamental matrix. The experiments on real and synthetic images demonstrate the usability of proposed algorithm (Figs. 4, 5, 6, 7, 8, 9, 10, 11).

**Acknowledgments** Funding of this paper was provided by Namseoul University.

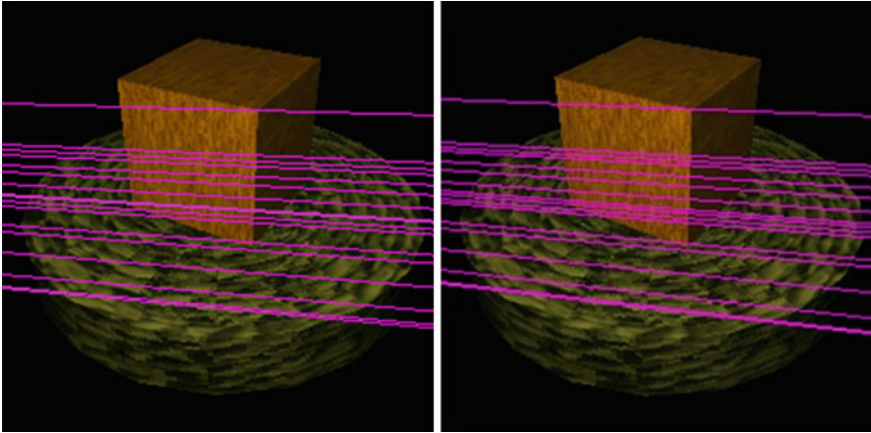


Fig. 8 b Recalculated epipolar lines

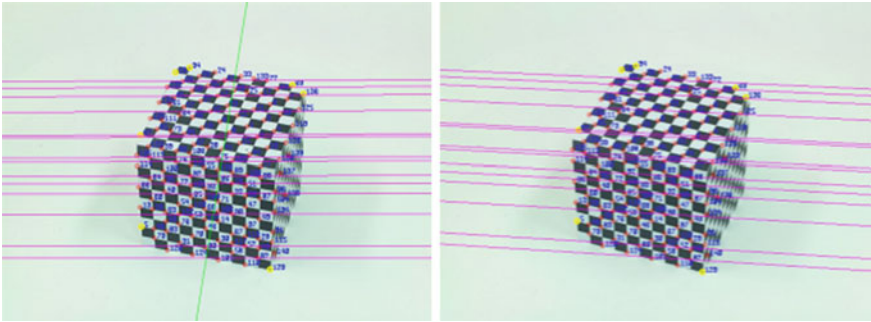


Fig. 9 Recalculated epipolar lines of the real image pair

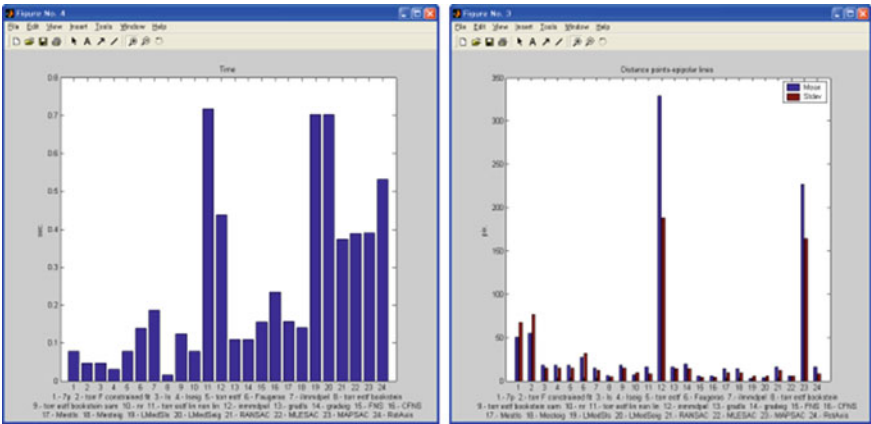
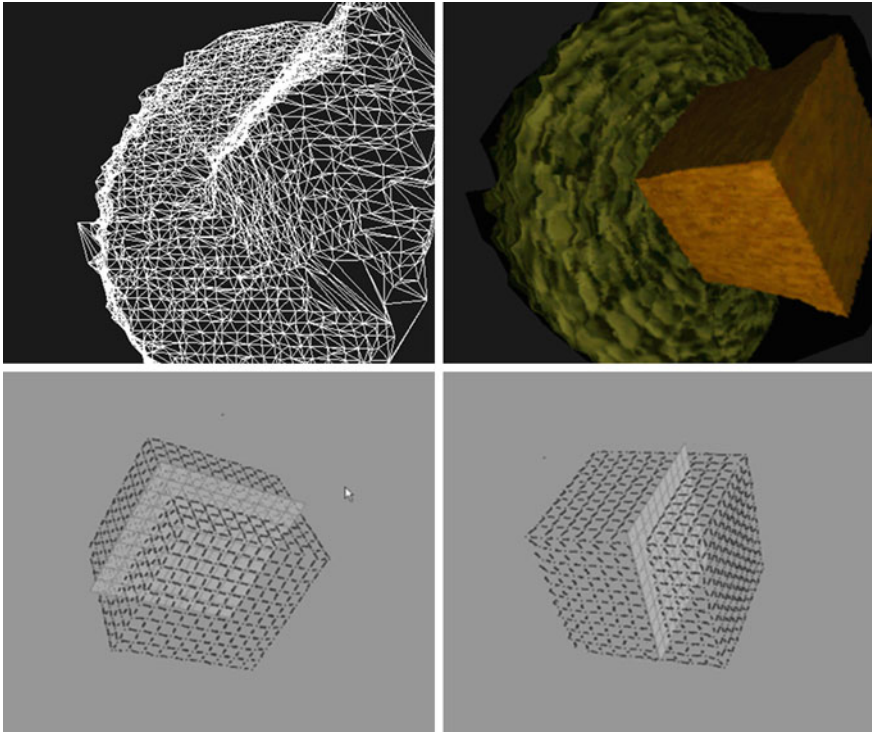


Fig. 10 Fundamental matrix error





**Fig. 11** Results of 3D reconstruction

**References**

1. Jang G, Tusi HT, Quan L, Zisserman A (2003) Single axis geometry by fitting conics. *IEEE Trans Pattern Anal Mach Intell* 25(10):1343–1348
2. Zhang Z (1998) Determining the epipolar geometry and its uncertainty: a review. *Int J Comput Vis* 27(2):161–195
3. Serra J (1982) *Image analysis and mathematical morphology*, vol 1. Academic Press, New York
4. Canny J (1986) A computation approach to edge detection. *IEEE Trans PAMI* 8(6):679–698
5. Rao K (1993) Extracting salient contours for target recognition: algorithm and performance evaluation. *Opt Eng* 32(11):2690–2697
6. Hartly R, Zisserman A (2000) *Multiple view geometry in computer vision*. Oxford university press, Oxford

# Gesture Recognition Algorithm using Morphological Analysis

Tae-Eun Kim

**Abstract** Recently, research into computer, vision-based methods to recognize gestures are widely conducted as means to communicate the volition of humans to a computer. The most important problem of gesture recognition is a reduction in the simplification treatment time for algorithms by means of real-time treatment. In order to resolve this problem, this research applies mathematical morphology, which is based on geometric set theory. The orientations for the primitive shape elements of hand signal shapes acquired from the application of morphological shape analysis include important information on hand signals. Utilizing such a characteristic, this research is aimed at suggesting a feature vector-based, morphological gesture recognition algorithm from a straight line which connects central dots of major primitive shape elements and minor primitive shape elements. It will also demonstrate the usefulness of the algorithm by means of experimentation.

**Keywords** Human motion • Computer vision

## 1 Introduction

In recent days, research is widely conducted to use gestures as a means to communicate between humans and computers. If an interface which understands gestures used during everyday dialogue is developed, subtle expressions and gestures for communicating ideas, emotions or sentiments can be widely used as natural inputs for gesture recognition.

---

T.-E. Kim (✉)

Department of Multimedia, Namseoul University, Cheonan, Korea  
e-mail: tekim5@empas.com

Broadly, gesture recognition techniques may be divided into an appliance attachment method which attaches many types of sensors, like gloves, to the body of humans and a vision-based image treatment method, which treats images acquired from a video camera. A method for attaching a sensor has some problems in that movements are limited, motions are unnatural, and handling is very inconvenient as discomfort and psychological burden is felt when it is worn. In this situation, currently, research is being actively conducted into fields which desire to use vision-based image treatment to recognize gestures in a non-touch manner, except for in special application fields.

The strength of a vision-based image treatment method is that a sensor need not be attached, but encounters problems as well. For example, the resolution limit of video data may be low, making hardware materialization and real-time treatment impossible. This is attributed to the fact that the recognition algorithm for hidden Markov model (HMM), neural network model, and others, which are mainly applied in form-based approach methods (and, also, major objects of research into vision-based image treatment) is complex and cannot be treated real-time. However, the core of gesture recognition is related to the real-time control of hardware as it concerns the interface between humans and machines. Therefore, big obstacles in putting it to practical use are that the algorithm is complex, and that it is difficult to treat the recognition in real-time and convert it into hardware [1–3].

So, this research suggests a morphological hand signal recognition algorithm applied with mathematical morphology in which hardware materialization is easy and high-speed operation is possible. In mathematical morphology, based on logical operations between pixels, various types of useful image treatment technologies, composed of morphological logic operations, are developed. Shape regions extracted from hand signal images are dismantled into primitive shape elements. This is because humans' visual recognition is a basic step for disintegrating complex shapes of objects contained in 2-dimensional images into simple primitive shape elements and expressing them in a stratum-wise manner. Based on this, a method is suggested to use the positional relationship of primitive shape elements and an experiment is conducted to demonstrate the usefulness of the suggested theory and confirm that hand signals can be recognized in search for video contents.

## 2 Disintegration of Morphological Shapes

**Morphological operation.** Morphological image treatment is a nonlinear treatment method used to interpret geometric characteristics of images; images in mathematical morphology, based on set theory, are a set of dots that can conduct set operations, such as translation, union and intersection. When  $A$ ,  $B$ ,  $C$  and  $K$  are described as open sets defined in  $Z^2$ , which is a 2-dimensional Euclidean space, and  $O$  is expressed as the starting point of  $Z^2$ , such operations as are widely used in the morphological image treatment field are shown in the following manner [4–6].

$$\text{Dilation } (\backslash) : \mathbf{A} \oplus \mathbf{K} \tag{1}$$

$$\text{Erosion } (\backslash) : \mathbf{A} \ominus \mathbf{K} \tag{2}$$

$$\text{Open } (\backslash) : \mathbf{A} \circ \mathbf{K} = (\mathbf{A} \ominus \mathbf{K}) \oplus \mathbf{K} \tag{3}$$

$$\text{Close } (\backslash) : \mathbf{A} \bullet \mathbf{K} = (\mathbf{A} \oplus \mathbf{K}) \ominus \mathbf{K} \tag{4}$$

**B**, which is mentioned here as a structuring element, is an image pattern used to convert images. An opening operation dilates the results of an erosion operation, and a closing operation erodes the results of a dilation operation. An opening operation has the nature of a filter that softens sharp corners of an object and eliminates smaller objects that are not contained in the structuring element. Meanwhile, a closing operation tends to fill up small holes on gorge-shaped objects. From these facts, it may be deduced that an opening operation and a closing operation can be used as a filter to remove positive noise components and negative noise components, respectively.

**Disintegration of the shape.** Disintegrating complex shapes of objects in 2-dimensional images into the elements of a simple primitive shape and stratum-wise expressions is a treatment which corresponds to a basic step of humans' visual recognition. This research utilizes primitive shape elements that are acquired through disintegrating gesture shapes into morphological shapes, so as to express gesture shapes.

The following is a morphological expression of shape disintegration which is intended to use morphological operations disintegrate 2-dimensional shape  $\mathbf{X} \in \mathbf{Z}^2$  into many sets or  $\{\mathbf{X}_i\}$ , which are primitive shape elements.

$$\mathbf{X} = \bigcup_{i=1}^n X_i, \quad X_i \in G(\mathbf{Z}^2) \tag{5}$$

Here,  $G(\mathbf{Z}^2)$  is a 2-dimensional open set defined in  $\mathbf{Z}$ , which is a 2-dimensional Euclidean space. When a primitive shape element for generating  $\mathbf{X}_i$  is expressed as  $\mathbf{Y}_i$  and a structuring element corresponding to  $\mathbf{Y}_i$  is expressed as  $\mathbf{B}$ , the morphological shape disintegration algorithm is expressed in the following manner. Also, the simplest example of primitive shape element  $\mathbf{Y}_i$  is  $n_i\mathbf{B}$ , which is the scalar multiple of structuring element  $\mathbf{B}$  selected as the original plate or a square having a unit area.

$$X_i = X_{n_i\mathbf{B}} = (\mathbf{X} \ominus n_i\mathbf{B}^S) \oplus n_i\mathbf{B} \tag{6}$$

Here,  $\mathbf{B}$  is a structuring element,  $n_i$  is the size of a structuring element, and  $\mathbf{B}^S$  is a reflection of  $\mathbf{B}$  against the starting point.

Formula (6) implies that a primitive shape element can be acquired through a dilation operation conducted many times because an erosion operation is conducted of the results of an erosion operation of shape  $\mathbf{X}$  against  $\mathbf{B}$  until the shape is reduced to dots or lines. Such a treatment is repeatedly conducted against  $\mathbf{X} - X_i$ .

The following expresses the treatment process mentioned above, in the form of regression [7, 8].

$$X_i = (X - X'_{i-1})n_iB, X'_i = \bigcup_{j=1}^i X_j, X'_0 = \phi \tag{7}$$

$$\text{Stopping condition: } (X - X'_k) \ominus B^S = \phi$$

Here, the stopping condition is a condition for disintegrating all and every region of the shape. And, k is the total number of disintegrated primitive shape elements. Each primitive shape element, disintegrated through applying the shape disintegration algorithm is a region generated as, is moved in parallel to track expressed as dot or line. Track of the maximum structuring element inscribed in a primitive shape element may be expressed in the following manner:

$$L_i = \left( X - \bigcup_{0 \leq j \leq i-1} (L_j \oplus n_iB) \right) \ominus n_iB^S \tag{8}$$

The relationship between the primitive shape element and track can be understood through combining the above formula (7) with formula (8).

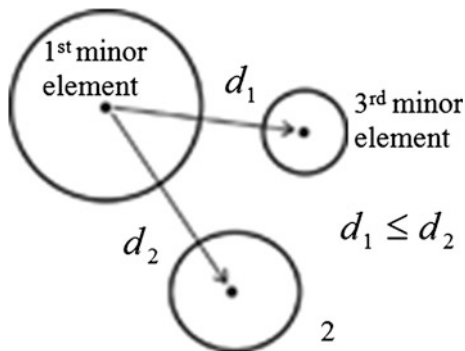
$$X_i = L_i \oplus n_iB \tag{9}$$

### 3 Recognition of the Shape

**Expression of the shape.** In the process of morphological shape disintegration, primitive shape elements are extracted starting from larger ones to the smaller ones. Again, these primitive shape elements are stratum-wise expressed in the following manner:

1. The largest primitive shape elements extracted first during shape disintegration set  $X_i$  as their main element.
2.  $X_i (i = 2, 3, \dots, n)$ , which are extracted during shape disintegration, are set as 1, 2, 3, ..., nth minor elements, according to their order.
3. As is seen in Fig. 1, primitive shape elements positioned between the higher two minor elements are removed from shape expression since they are not required for the recognition of hand signal orientation. The primitive shape elements positioned between the two elements in such a process are primitive shape elements—the distance from which to major elements is shorter compared to the immediately higher elements.
4. The said process is conducted for  $i = 2$  through  $n$ . Here,  $n$  is the frequency of shape disintegration, conducted until such a feature vector that recognizes hand signals can be acquired.

**Fig. 1** Primitive shape elements positioned between higher two minor elements



The next step is extracting the central dot of primitive shape elements. The central dot is acquired through repeatedly conducting an erosion operation until primitive shape elements get smaller than structuring elements.

**Extraction and recognition feature vectors.** A feature vector is referred to as the angle of such a line that connects the central dots of major primitive shape elements positioned in a 2-dimensional space via shape disintegration and shape expression processes, with the central dots of minor primitive shape elements.

$$\mathbf{x} = \{\theta_1, \theta_2, \theta_3, \dots, \theta_{n-1}\} \tag{10}$$

Here,  $n$  is the number of primitive shape elements acquired in the process of shape disintegration and  $\theta_{n-1}$  is central dots  $(x, y)$  of a major primitive shape element. If the central dot of  $n - 1$ st major primitive shape element is  $(x_{n-1}, y_{n-1})$ , its value is calculated in the following manner:

$$\theta_{n-1} = \tan^{-1}(|y - y_{n-1}|/|x - x_{n-1}|) \tag{11}$$

The mean value of feature vectors, which are composed of such angles that connect the central dots of major primitive shape elements and minor primitive shape elements, can be calculated in the following way:

$$x = (\theta_1 + \theta_2 + \theta_3 + \dots + \theta_{n-1})/(n - 1) \tag{12}$$

The values of feature vectors calculated through applying Formulas (10) and (11) are based on the thumb and the orientation which have important meanings for hand signals.

**Recognition of hand signals.** As is seen in Fig. 2, if the mean value of the feature vectors is  $224^\circ \sim 43^\circ$ ,  $45^\circ \sim 134^\circ$  and  $135^\circ \sim 225^\circ$ , feature vectors are recognized as region 2, 4 and 3, respectively, and minor primitive shape elements of region 1 are positioned around major primitive shape elements, then it is recognized as a case where only major primitive shape elements exist and where minor primitive shape elements are removed. This is under the assumption that only major primitive shape elements exist.

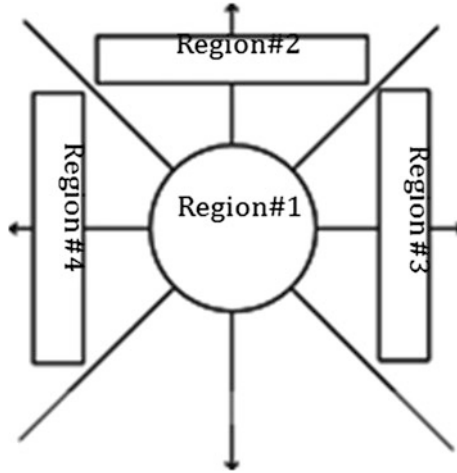


Fig. 2 Regions where feature vectors are recognized

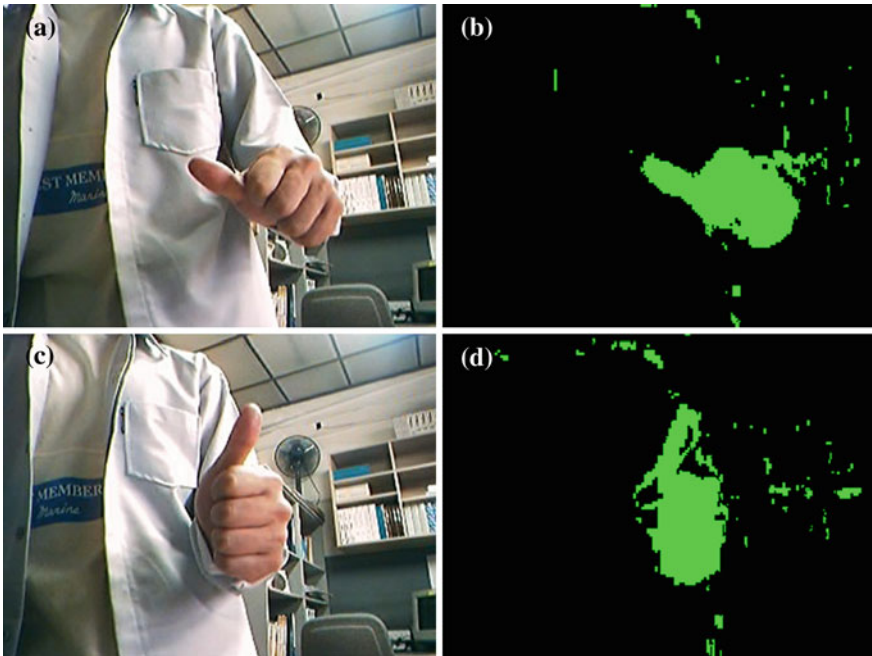


Fig. 3 a, c Original images, b, d Results of the detection of skin region

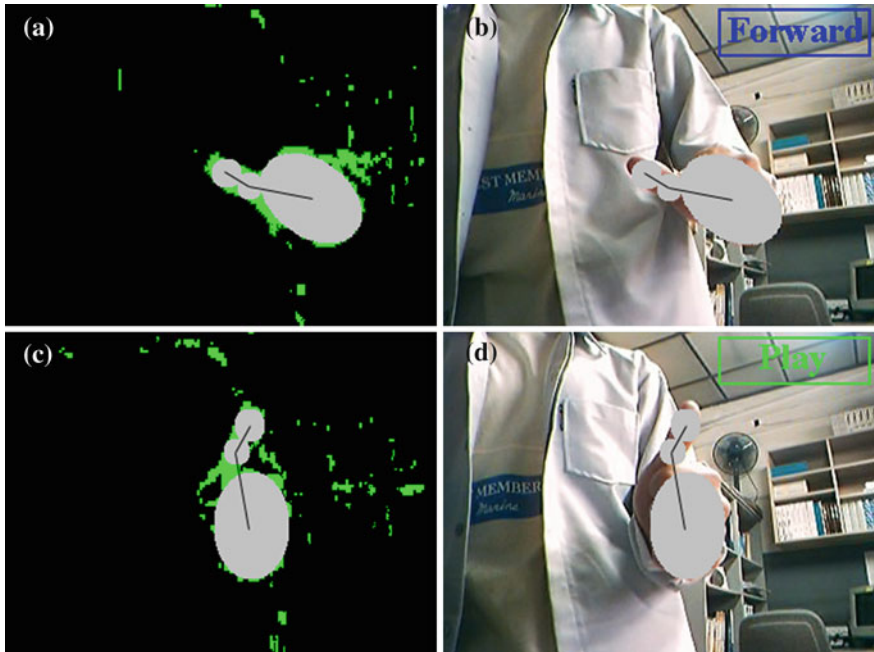


Fig. 4 a, b Results of shape disintegration, c, d Results of the recognition of hand signals

### 4 Results of the Experiment

Figure 3 shows images about the detection of a man’s skin color after an acquired image(320x240 RGB) is converted into a YcbCr model ( $77 \leq C_b \leq 127$ ,  $133 \leq C_r \leq 173$ ).

In Fig. 4a, c show the results of the display of the higher 3 elements by using shape disintegration; and b, d show the results of the recognition of the higher 3 elements by use of feature vectors.

### 5 Conclusion

The method, newly suggested in this research, is to recognize hand signals using such feature vectors that connect the central dots of primitive shape elements. These elements’ are as are the largest with the central dots of the other primitive shape elements after selecting the central dots of primitive shape elements, which are extracted using morphological shape disintegration. It’s expected that the said method will be widely applied to searches for video data and to interface designs related to the running of other electronic systems.



**Acknowledgments** Funding of this paper was provided by Namseoul University.

## References

1. Pavlovic VI, Sharma R, Huang TS (1997) Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Trans. PAMI* 19(7):677–695
2. Ahmad T, Taylor CJ, Lanitis A, Cootes TF (1997) Tracking and recognising hand gestures, using statistical shape models. *Image Vis Comput* 15:345–352, Elsevier
3. Wilson AD, Bobick AF (1999) Parametric hidden markov models for gesture recognition. *IEEE Trans PAMI* 21(9):884–900
4. Serra J (1986) Introduction to mathematical morphology. *Comput Vis Graph Image Process* 35(3):283–305
5. Serra J (1982) *Image analysis and mathematical morphology*. Academic Press, New York
6. Maragos P (1989) A representation theory for morphological image and signal processing. *IEEE Trans Pattern Anal Mach Intell* 11(6):586–599
7. Pitas I, Venetsanopoulos AN (1990) Morphological shape decomposition. *IEEE Trans Pattern Anal Mach Intell* 12(1):38–45
8. Pitas I, Venetsanopoulos AN (1992) Morphological shape representation, *Pattern Recog* 25(6):555–565

# Omnidirectional Object Recognition Based Mobile Robot Localization

Sungho Kim and In So Kweon

**Abstract** This paper presents a novel paradigm of a global localization method motivated. The proposed localization paradigm consists of three parts: panoramic image acquisition, multiple object recognition, and grid-based localization. Multiple object recognition information from panoramic images is utilized in the localization part. High level object information is useful not only for global localization but also robot-object interaction. The metric global localization (position, viewing direction) is conducted based on the bearing information of recognized objects from just one panoramic image. The experimental results validate the feasibility of the novel localization paradigm.

**Keywords** Object recognition · Localization · Omnidirectional camera

## 1 Introduction

A robot should have the ability to determine its global location in order to successfully handle self-initialization and kidnapping problem. Several approaches were proposed to handle such problem. Park et al. proposed a hybrid map of object and spatial layouts using a stereo camera to localize globally [1]. Angeli et al. proposed a topological visual SLAM (simultaneous localization and mapping) for determining global localization [2]. Visual words were used to handle global

---

S. Kim (✉)

LED-IT Fusion Technology Research Center and Department of Electronic Engineering,  
Yeungnam University, 214-1 Dae-dong Gyeongsan-si, Gyeongsangbuk-do, Korea  
e-mail: sunghokim@ynu.ac.kr

I. S. Kweon

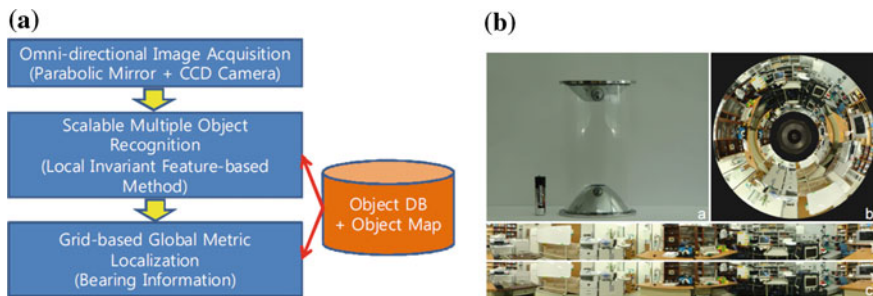
Department of Electrical Engineering and Computer Science, Korea Advanced Institute of  
Science and Technology, 373-1 Guseong-dong Yuseong-gu, Daejeon, Korea  
e-mail: iskweon@kaist.ac.kr

location, and odometry information was combined to give metric information. Ramisa et al. also proposed a topological localization method using affine invariant features [3]. Although these approaches can provide global location information, they used additional information, such as stereo and odometry, for global metric localization. An additional requirement is the fast global localization capability using just one image frame. Most approaches can achieve topological localization by recognizing objects or scenes from an image [4]. Metric localization is possible if there is a depth cue (stereo camera) or motion cue (structure from motion) [1, 4]. The last requirement is the capability of robot-object interaction for visual servoing. Robots should have object label and position information.

There are several paradigms for mobile robot localization. Initially, artificial landmark-based approaches were proposed [5, 6]. After then, the SLAM paradigm became a popular approach since it can build a map and localize itself simultaneously by using the extended Kalman filter and an invariant feature such as SIFT [7, 8]. Particle filter-based statistical estimation was also useful in SLAM approach. These paradigms were partially successful since they could estimate relatively accurate location information by matching low level features, such as corner points or invariant features, in multi-frames. However, the location estimation error can be large if they use only one frame. In addition, those approaches can not provide high level information for robot-object interaction.

Then, how can human visual systems (HVSs) localize themselves? HVSs can localize themselves and interact with environment robustly. Do HVSs recognize their locations by point matching as SLAM? Most people will say “No”. We surveyed the localization mechanisms of the HVS to get the answer or clue. Although accurate mechanisms are not disclosed, it is evident that object recognition and localization are strongly related according to experimental studies such as lesion of visual cortex (ventral stream and dorsal stream) [9, 10]. This observation means that object recognition and localization are strongly correlated and that they facilitate each other.

Motivated by such biological research results, we propose a novel localization paradigm using only high level object recognition information from one image. The paradigm consists of three parts: omnidirectional panoramic image acquisition, multiple object recognition and grid-based localization. Multiple object recognition is performed from a panoramic image, and mobile robot localization is then conducted using bearing information of objects. This paradigm can estimate both spatial position and viewing direction using only one image. [Section 2](#) overviews the proposed localization system and explains the multiple object recognition method. In addition [Sect. 2](#) represents the mobile robot localization algorithm using object information. [Section 3](#) experimentally validates the feasibility of the proposed paradigm, and [Sect. 4](#) concludes the paper.



**Fig. 1** **a** The proposed novel paradigm of localization using high level object information, **b** Omnidirectional stereo camera system

## 2 Object Recognition Based Localization

As shown in Fig. 1a, the proposed localization system consists of image acquisition, object recognition and global metric localization. The proposed localization system consists of an off-line database construction module and an on-line localization module. The object database and object-based map are constructed off-line. The object DB module contains learned local feature-based object models representing a 3D object as a set of views. Since it is based on a robust invariant feature, the learned models can handle geometrically, photometrically distorted objects in a general environment. The object-based map is built manually by accurately measuring object locations. On-line localization is conducted through the panoramic image acquisition module via an omnidirectional camera, multiple object recognition module and a bearing angle-based localization module. The large field of view is required for the object-based localization from one image. Although there can be several methods for getting an omnidirectional image, we adopt the parabolic mirror-based panoramic camera. After image acquisition occurs, we extract multiple object information (object label and position in image) by applying the local invariant feature-based method. We use object databases (DBs) that are learned to handle large numbers of objects. After such object recognition occurs, we can have the bearing (angle) information of each object. The final robot localization (spatial position and viewing direction) is estimated by intersecting the bearing information.

The proposed localization method utilizes the omnidirectional camera developed by Jang [11]. Figure 1b shows the omnidirectional camera system. It is composed of 2 parabolic mirrors and an IEEE 1394 camera ( $1600 \times 1200$  image resolution). We can acquire omnidirectional stereo images via the camera system. Currently, we use the upper rectified images for the purpose of recognizing objects in these images, which can give higher image resolution than lower images.

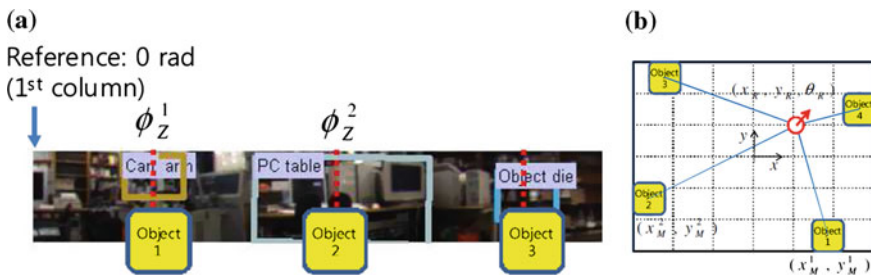
In the multiple object recognition module, we can recognize learned objects stored in the object database. Each recognized object can provide an object label and a bearing angle measurement. Since the resolution of a rectified image is

1800 × 161, the bearing measurement resolution of the top-line is 0.2 deg/pixel. Now, we introduce a powerful and efficient 3D object representation, learning and recognition method. Any 3D objects can be represented by a set of multiple views. Each view consists of local features. We apply the sharing concept to the features and views of scalable object representation [12].

How can we fully utilize the shared feature-based view clustering method in object recognition? Basically, we use the well-known hypothesis and verification framework. However, we modify it to recognize multiple objects via the proposed object representation scheme. We can get all possible matching pairs by NN (nearest neighbor) search in feature library. From these, hypotheses are generated by generalized Hough transform in CFCM (common frame constellation model) ID, scale (11 bins), orientation (8 bins) space [8], and grouped by object ID. Then, we determine whether to accept or reject the hypothesized object based on the bin size with an optimal threshold [13]. Finally, we select the optimal hypotheses that can be best matched to the object features in a scene.

In the localization module, the recognized object labels are used to achieve data association of objects in a map, and the intersection of bearing measurements is used to accomplish robot localization. Through object recognition, we can estimate the position of recognized objects in an image. Especially, the column position provides the bearing measurement ( $\phi_Z^i$ ) of  $i$ th object in panoramic images as shown in Fig. 2a. In this work, we regard the 1st column of an image as 0 radian. An object center is estimated by the similarity transform of a corresponding CFCM. Given a set of object labels and bearing measurements, the robot localization is defined as coordinate transformation from reference coordinates to robot coordinates in 2D space.

Let  $\{A\}_Z$  be a set of bearing measurements by a mobile robot through multiple object recognition,  $\{A\}_R$  be a set of model bearing measurements after coordinate transformation. Then the robot localization problem is to estimate  $T = (x, y, \phi)$  which is the coordinate transformation function from reference coordinates to robot coordinates as shown in Fig. 2b. Shimshoni proposed a direct estimation method based on linear constraints [14]. We applied this method but the estimation



**Fig. 2** Bearing measurement information of recognized objects in panoramic images: **a** Visual interpretation, **b** Localization problem is regarded as coordinate transformation from reference coordinates to robot coordinates

results are very unstable due to bearing measurement noise and a few number of measurements (usually 3–6). Fox et al. proposed a Monte Carlo localization method that approximates a posterior by a set of samples [15]. We also applied the latter method, but it takes time to converge. Instead, we use the grid-based localization method. If we divide the coordinate transformation space into moderate resolution (in current implementation  $\delta_x = \delta_y = 10$  cm,  $\delta_\phi = \pi/180$  rad), then robot location is estimated by Eq. (1).  $N$  denotes the number of recognized objects. If we specify the symbols, then the localization problem is the minimization problem of three dimensions as Eq. (2).  $\phi_R^i$  denotes the angle of model object  $i$  after transformation with  $T = (x_R, y_R, \theta_R)$  as shown in Eq. (3). We can get the optimal robot location with orientation information by minimizing Eq. (2).

$$\hat{T} = \min_T \left[ \sum_{i=1}^N \left( A_Z^{(i)} - A_R^{(i)}(T) \right) \right] \quad (1)$$

$$\left( \hat{x}_R, \hat{y}_R, \hat{\theta}_R \right) = \min_{(x_R, y_R, \theta_R)} \left[ \sum_{i=1}^N \left( \phi_Z^i - \phi_R^i(x_R, y_R, \theta_R) \right) \right] \quad (2)$$

$$\phi_R^i(x_R, y_R, \theta_R) = \tan^{-1} \left( \frac{y_M^i - y_R}{x_M^i - x_R} \right) + \theta_R \quad (3)$$

### 3 Experimental Results

We apply the object recognition-based localization method to a complex laboratory environment. There are bookshelf, PC table, air cleaner, wash stand, printer and so on. Note that the image quality of an individual object is very low. We use every two views for object modeling. The total number of objects is 9 with multiple views. According to the results of object learning, part clustering reduces the size by 44.2 %, while view clustering reduces the size by 39.8 %.

Figure 3a shows localization examples of a mobile robot, KASIRI IV, which can move accurately according to the planned path. In each result, the top image shows the recognized objects with object centers that are equal to the bearing measurements. In the bottom image, the red arrow represents the location (position with direction) of the mobile robot, and data association is linked by the dotted blue line. Note that multiple objects are recognized and used for robot localization. Figure 3b summarizes the overall localization performance. The red dotted line represents the true path of the mobile robot and the blue square represents the estimated robot location using our algorithm. The average location error is  $(x, y = 14.5$  cm,  $18.5$  cm), which is relatively large compared to those of the range sensor-based approaches or interesting point-based approaches (usually within 5 cm) in a  $10 \times 10$  m environment. However, our proposed system can



**Fig. 3** The overall localization performance of the test sequence: **a** Examples of robot localization using the proposed method, **b** Final localization results

provide high level information of an object that is useful for robot-environment interaction. Note that human visual systems (HVSs) can recognize relative locations with very low metric accuracy but can well interact in an environment with object information.

## 4 Conclusions and Discussions

In this paper, we proposed a new robot localization method using the object recognition method. Instead of fragile low level features, we regard objects as natural landmarks for localization. For this system, we introduce the multiple object recognition method based on a learned object model and grid-based localization using bearing measurements. The experimental results validate the feasibility of the proposed system. There are several research directions. Currently, we do not use the tracking of objects. If we utilize the temporal continuity, then we can get smoother localization. In addition, the map is generated manually. We have to investigate automatic object-map generation. If we combine it with topological localization, then the working space can be increased.

**Acknowledgments** This research was supported by Basic Science Research Program through the NRF funded by the Ministry of Education, Science and Technology (No. 2012-0003252) and by National Strategic R&D Program for Industrial Technology, Korea. It was also supported by the 2012 Yeungnam University Research Grants.

## References

1. Park S, Kim S, Park M, Park SK (2009) Vision-based global localization for mobile robots with hybrid maps of objects and spatial layouts. *Inf Sci* 179(24):4174–4198
2. Angeli A, Doncieux S, Meyer JA, Filliat D et al (2009) Visual topological slam and global localization. In: *ICRA '09: Proceedings of the 2009 IEEE international conference on robotics and automation*, pp 2029–2034
3. Ramisa A, Tapus A, de Mántaras RL et al (2008) Mobile robot localization using panoramic vision and combinations of feature region detectors. In: *Proceedings of the ICRA*, pp 538–543

4. Murillo AC, Guerrero JJ, Sagues C (2007) Topological and metric robot localization through computer vision techniques. In: Proceedings of the ICRA Workshop-from features to actions: unifying perspectives in computational robot vision, pp 79–85
5. Scharstein D, Briggs AJ (2001) Real-time recognition of self-similar landmarks. *Image Vis Comput.* 19(11):763–772
6. Jang G, Kim S, Lee W, Kweon I et al (2002) Color landmark based selflocalization for indoor mobile. In: Proceedings of the ICRA, pp 1032–1042
7. Durrant-Whyte H, Bailey T (2006) Simultaneous localisation and mapping (slam): part i the essential algorithms. *IEEE Robot Autom Magazine* 13:99–110
8. Lowe D (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vision* 60(2):91–110
9. Himmelbach M, Karnath H (2005) Dorsal and ventral stream interaction: Contributions from optic ataxia. *J. Cogn Neurosci* 17(4):632–640
10. Blangero A et al (2008) Dorsal and ventral stream interaction: Evidence from optic ataxia. *Brain Cogn* 67:2
11. Jang G, Kim S, Kweon I (2006) Single-camera panoramic stereo system with single-viewpoint optics. *Opt Lett* 31(1):41–43
12. Kim S, Kweon IS (2008) Scalable representation for 3D object recognition using feature sharing and view clustering. *Pattern Recogn* 41(2):754–773
13. Murphy-Chutorian E, Triesch J (2005) Shared features for scalable appearance-based object recognition. In: Proceedings of the seventh IEEE workshops on application of computer vision, vol 1 pp 16–21
14. Shimshoni I (2002) On mobile robot localization from landmark bearings. *IEEE Trans Rob* 18(6):971–976
15. Fox D, Burgard W, Dellaert F et al (1999) Monte carlo localization: Efficient position estimation for mobile robots. In: Proceedings of the national conference on artificial intelligence



# Gender Classification Using Faces and Gaits

Hong Quan Dang, Intaek Kim and YoungSung Soh

**Abstract** Gender classification is one of the challenging problems in computer vision. Many interactive applications need to exactly recognize human genders. In this paper, we are carrying out some experiments to classify the human gender in conditions of low captured video resolution. We use Local Binary Pattern, Gray Level Co-occurrence Matrix to extract the features from faces and Gait Energy Motion, Gait Energy Image for gaits. We propose to combine face and gait features with the combination classifier to enhance gender classification performance.

**Keywords** Gender Classification · Faces · Gaits · Local binary pattern · Gait Energy Motion · Gait Energy Image

## 1 Introduction

Nowadays, gender classification plays an important role in many practical applications in medical, social and security fields. And many mentioned applications critically depend on the correct gender classification. Based on gender, surveillance system can improve their performance of tracking and store managers can estimate the difference of male or female interest to increase profits. Automatic gender classification can be based on the human characteristics such as voice, faces and gaits.

Early studies about gender classification from faces began in 1990s with works of Golomb et al. [1] using multiple layer neural networks. Since then, many feature extraction algorithms have been developed to enhance the performance. Discrete cosine transform algorithm [2] is used to reduce the dimensionality of the data set.

---

H. Q. Dang (✉) · I. Kim · Y. Soh  
Department of Information and Communication Engineering,  
Myongji University, Yongin, South Korea  
e-mail: kit@mju.ac.kr

Some methods reduce the redundant information of high density such as local binary pattern (LBP) and principal component analysis [3].

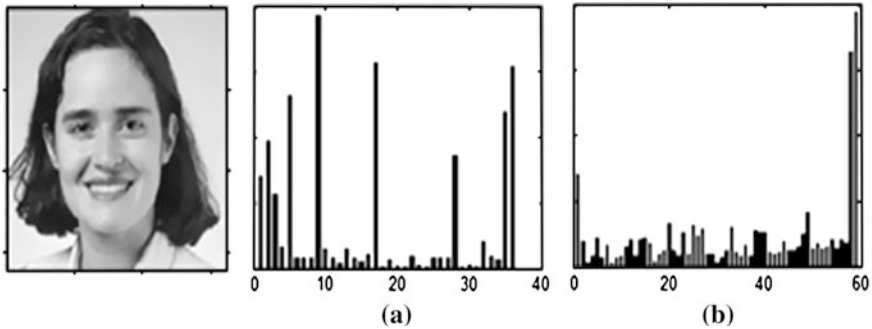
Meanwhile, gender classification by gait has received much attention in computer vision community due to its advantage of classification without attracting attention and distance limitation. Gait data can be captured at far distance without demanding physical human information. Therefore, there is a great potential for recognition applications based on cameras with low-resolution videos which are affected by the camera devices or indoor and outdoor environment. In gait, we have two kinds of features to extract. One is model-based features and the other is model-free features. Model-based features explore static and dynamic body parameters while model-free features just use binary silhouettes, without model of moving person requirement [4].

In this paper, we present the classification of gender with features extracted from face and human gait. We use LBP method and Grey Level Concurrence Matrix (GLCM) to extract face features. Gait Energy Motion (GEM) and Gait Energy Image (GEI) are used to extract gait features. Then we further propose to classify gender by combining them. Support Vector Machine (SVM) algorithm is used for classification with linear kernel and majority voting for a combination classifier. The frame size in the dataset is 320-by-240 pixel and the frame rate is 25 fps. Section 2 explains the feature extraction from face and gait, respectively. In Sect. 3, we experimented with several methods and the combinations of the methods depicted in previous sections. The conclusion is made in the last section.

## 2 Feature Extraction

### 2.1 Feature Extraction from Faces

Texture is an important characteristic to identify the regions of interest in an image. And the LBP operator is a gray-scale texture measure which summarizes the local spatial structure and gray scale contrast of an image. The original LBP operator, a nonparametric  $3 \times 3$  kernel, labels the pixels of an image by thresholding the  $3 \times 3$  neighborhood of each pixel with the center value and considers the results as a binary number. The advantage properties of the LBP are highly discriminative and invariance to the monotonic gray level changes under the rotation effect and the influence of illumination. In the case of rotated image, the sampling neighborhoods will be moved correspondingly along the perimeter of the circle around the center pixel with same direction of rotation of the image. The value of LBP label gets changed because the obtained binary number is shifted left or right. Keeping the LBP label constant at all rotation angles, the LBP value corresponds to the smallest shifted binary number. The rotation invariance approach gets 36 LBP labels in all 256 possibilities and it is presented as a histogram with 36 bins [5]. Another extension to the original operator is defined as uniform patterns, which can be used



**Fig. 1** a Rotation invariant LBP histogram. b Uniform LBP histogram

to reduce the length of the feature vector and implement a simple rotation invariant descriptor. A local binary pattern is called uniform if the binary pattern contains at most 2 bitwise transitions from 0 to 1 or vice versa. There is a separate label for each uniform pattern and all the non-uniform patterns are labeled with a single label. With  $3 \times 3$  neighborhoods, there are total of 256 patterns, 58 of them are uniform, which yields in 59 different labels [5]. Rotation invariant LBP histogram and uniform LBP histogram are shown in Fig. 1.

Grey Level Co-occurrence Matrix (GLCM) characterizes the texture by considering the spatial relationship of pixels over a sub-region of an image [6]. GLCM calculates how often different combinations of the reference pixel and neighbor pixel occur in the image. The combination defines the relationship between pixels in four directions such as horizontal, vertical, left and right diagonal. The number of occurrences a pixel with intensity  $i$  is adjacent to a pixel with intensity  $j$ , would be counted and stored in matrix with dimensions corresponding to the number of intensity values of an image. Because co-occurrence matrices are typically large and sparse, some statistic features can be extracted such as contrast, correlation and energy homogeneity to get more usefulness of feature. Contrast measures the local variations in the GLCM. Correlation measures the joint probability occurrence of the specified pixel pairs. Energy provides the sum of the squared elements in the GLCM. Homogeneity measures the closeness of the distribution of diagonal elements in the GLCM.

## 2.2 Feature Extraction from Gait

Dynamic features of human gait are used mostly in many approaches of individual recognition as well as gender classification. Two dynamic features of Gait Energy Image (GEI) and Gait Energy Motion (GEM) have proved its effectiveness for representing the characteristics of human gait. These features are from regular human walking, which is the repetitive motion of body with a differently stable

frequency for each person. In complicated environments, human movement is firstly extracted from the image by using background subtraction techniques using Gaussian mixture model or kernel density estimation. Some further image processing is applied to get silhouette images with normalization and horizontal alignment. Given the processed binary images at time  $t$  in a sequence, the gait energy image is defined in Eq. 1:

$$F(i, j) = \frac{1}{N} \sum_{t=1}^N I_t(i, j) \tag{1}$$

where  $I(i, j)$  is the intensity of a pixel and  $N$  is the number of frames in the sequence. In [7], a modification of GEI is explored by taking the motion between each frame of the sequence as spatial data for human gait, which is called Gait Energy Motion (GEM). It is defined as:

$$F(i, j) = \frac{1}{N} \sum_{t=2}^N |I_t(i, j) - I_{(t-1)}(i, j)| \tag{2}$$

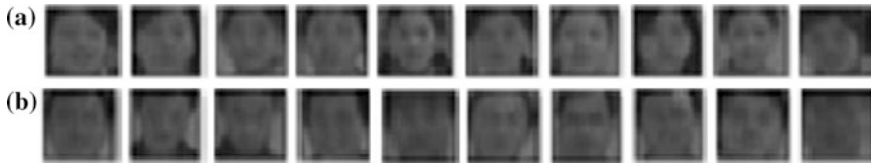
where  $N$  is the number of frames in a sequence of walking person,  $I(i, j)$  is the intensity of an image at time  $t$  and time  $(t - 1)$ . The difference between GEI and GEM is illustrated in Fig. 2.

### 3 Experiments

We carried out the experiments with the CASIA Gait Database of Dataset B collected by Institute of Automation, Chinese Academy of Science [8]. In our experiments, we used 29 female and 29 male gait dataset for training and testing. From 11 views of cameras, we just chose frontal view to get face information and silhouette sequences provided in Dataset B to get GEM. The images of volunteers

**Fig. 2** a Silhouette images. b GEI c GEM





**Fig. 3** Face dataset: **a** Female. **b** Male

were captured in  $320 \times 240$  resolutions with 25 fps of frame rate. We used the last frame of each video to detect the face regions of each subject. Therefore, detected face are different in appearance of expression, head pose variations, hair and glasses wearing. Faces are normalized as  $28 \times 28$  pixels images as shown in Fig. 3.

Human gait energy motions are calculated from the silhouette sequences in different of status walking: normal walking, carrying a bag and walking with a coat. 49 frames for each person are used to calculate GEM and GEI. With 29 female images and 29 male images, 80 % are used for training and 20 % for testing. As there are training images and test images, we selected images randomly, so that no overlapping could exist between them. We calculated average Correct Classification Rate (CCR). We used Support Vector Machine (SVM) as a classification method with linear kernel and used 10 cross fold validation as a training method. The results of testing with each feature of face and gait are shown in the Table 1a. Then we fused the label outputs using majority voting combining classifier [9]. The principal component analysis approach is used to reduce the feature dimensions for the combination. In our case, the average CCR of gender classification with face features gets lower performance because the noise has impacts to the extracted features. Better performance of 93.04 % is archived with the proposed combination of GEM and LBP as shown in the Table 1b.

**Table 1 a** Gender classification with each feature of face and gait.  
**b** Gender classification with the combination classifier

Method	Correct classification rate
GLCM	80.43 %
LBP	86.62 %
GEI	88.31 %
GEM	89.15 %
Method	Correct classification rate
GLCM + LBP	87.23 %
GEM + GEI	90.49 %
GEI + LBP	92.36 %
GEM + LBP	93.04 %

## 4 Conclusion

In this paper, we proposed the combination of face and gait for gender classification. Combining classifier appears as a natural step forward when a large amount of knowledge of single classifier models has been explored. Combining methods obtain better predictive performance than could be obtained from any of the constituent models. By combining face and gait using combination classifier, we increased the correct rate of gender classification for the low video resolution.

**Acknowledgments** This work (Grants No. C0005448) was supported by Business for Cooperative R&D between industry, Academy, and Research Institute funded by Korea Small and Medium Business Administration in 2012.

## References

1. Golomb B, Lawrence D, Sejnowski T (1990) Sexnet: a neural network identifies sex from human faces. In: Advance in neural information processing systems, California, vol 3, pp 572–577
2. Nazir M, Ishtaiq M, Batool A, Jaffar A, Mirza AM (2010) Feature selection for efficient gender classification. In: Proceedings of the WSEAS international conference, Wisconsin, pp 70–75
3. Fang Y, Wang Z (2010) Improving LDP features for gender classification. In: IEEE international conference on wavelet analysis and pattern recognition, pp 1203–1208
4. Wang J, She M, Nahavandi S, Kouzani A (2010) A review of vision-based gait recognition methods for human identification. In: international conference on digital image computing: techniques and applications, pp 320–327
5. Ojala T, Maenpaa T (2002) Multiresolution gray\_scale and rotation invariant texture classification with local binary pattern. *IEEE Trans Pattern Anal Mach Intell* 24:971–887 (Washington)
6. Clausi DA (2002) An analysis of co-occurrence texture statistics as a function of grey level quantization. *J Remote Sens* 28:45–62 (Canada)
7. Arai K, Asmara RA (2012) Human gait gender classification in spatial and temporal reasoning. *Int J Adv Res Artif Intell* 1:1–6
8. CASIA Gait Database. <http://www.cbrs.ia.ac.cn/English/index.asp>
9. Lam L, Suen CY (1997) Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Trans Syst Man Cybern* 27(5):553–568

# Implementation of Improved Census Transform Stereo Matching on a Multicore Processor

Jae Chang Kwak, Tae Ryong Park, Yong Seo Koo  
and Kwang Yeob Lee

**Abstract** Traditionally, sub-pixel interpolation in stereo-vision systems has been used for the block-matching algorithm. In this paper, Census transform algorithm which has been on area-based matching algorithm is improved and it's compared with existing census transform algorithm. Two algorithms are compared using Tsukuba stereo images provided by Middlebury web site. As a result, disparity map error rate is decreased from 16.3 to 11.8 %.

**Keywords** Census transform · Stereo matching · Area-based matching · Multicore processing

## 1 Introduction

The stereo vision is a method to extract image depth information using two different images that have been captured by right and left view points. Among overall procedures of stereo vision, a step to find matching points is called stereo

---

J. C. Kwak (✉)

Department of Computer Science, Seo Kyeong University, Seoul, Korea  
e-mail: jckwak@skuniv.ac.kr

T. R. Park · K. Y. Lee

Department of Computer Engineering, Seo Kyeong University, Seoul, Korea  
e-mail: trpark@skuniv.ac.kr

K. Y. Lee

e-mail: kylee@skuniv.ac.kr

Y. S. Koo

Department of Electronic Engineering, Dan Kook University, Yongin, Korea  
e-mail: yskoo@dankook.ac.kr

matching. The stereo matching is a core of stereo vision system. The Census Transform (CT) stereo matching algorithm is a method finding matching points from two images having different viewpoints using structural information of pixels in regions. CT algorithm has less computational complexity [1].

In this paper, the CT algorithm is improved and the improved algorithm has been compared with original CT algorithm for its accuracy of stereo matching. Also, the efficiency in parallel processing system using ARM 11 MP-Core conditions is proved through parallel processing of multi-core. In embedded environment, the effort to reduce computations is important as well as the accuracy of algorithm.

## 2 The Census Transform Algorithm

The CT algorithm [2] transforms images comparing intensity of pixels with their neighbors. The CT algorithm compares center pixels with their neighbor pixels and presents the result with bit string. For the comparison, the transformation to bit string can be expressed as Eq. (1).

$$P_{xy} = \begin{cases} 0 & \text{if } P_{center} \geq P_{xy} \\ 1 & \text{if } P_{center} < P_{xy} \end{cases} \quad (1)$$

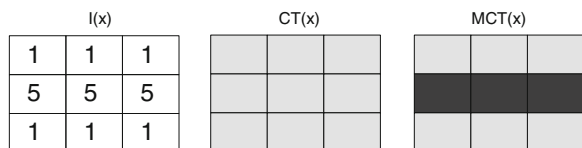
In Eq. (1), the  $P_{xy}$  means pixel intensity within sub-windows, and  $P_{center}$  means an intensity index of the center pixel. Each pixel in the sub-window is expressed with 0 and 1, comparing their intensity with the center pixel. Using the extracted  $P_{xy}$ , pixels are transformed to a bit string. Based on the pixel that has minimum Hamming Distance value, the disparity is extracted and the stereo matching is performed.

The MCT algorithm [3] is an improved version of CT algorithm. The MCT algorithm is an algorithm to reduce errors that can be caused by changes of intensity in source images. The MCT algorithm performs transformation based on the average values of pixels in sub window, while CT algorithm performs based on center pixels. The Fig. 1 shows difference between CT and MCT algorithms.

$$P_{xy} = \begin{cases} 0 & \text{if } P_{avg} \geq P_{xy} \\ 1 & \text{if } P_{avg} < P_{xy} \end{cases} \quad (2)$$

In Fig. 2, each values are determines to be 0 based on 5 in CT algorithm, and they are to be 0 or 1 based on average value, 2.3 in MCT algorithm. Thus, it has benefits to reduce errors that can happen according to changes of the intensity.

**Fig. 1** Difference of CT and MCT algorithms





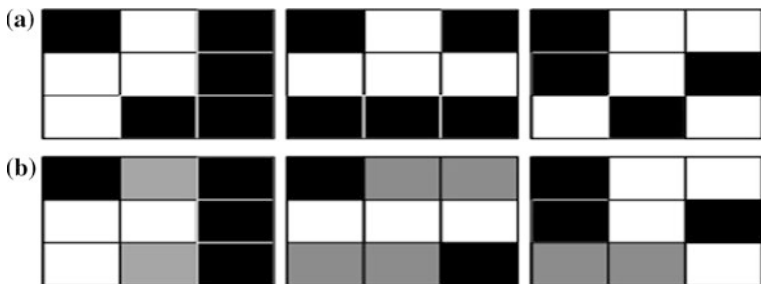


Fig. 2 Kernel index of CT a and proposed CT b algorithms

### 3 The Proposed Census Transform Algorithm

The proposed algorithm uses increased number of kernel index cases for more accurate stereo matching. As shown in Eqs. (1) and (2), previous algorithms express neighbors of center pixels as 0 or 1, so that there are 256 kernel indexes from 0 to 255. On the other hands, the proposed algorithm expresses more number of cases for kernel index, as Eqs. (3) and (4) show.

$$P_{xy} = \begin{cases} 2 & \text{if } P_{center} \geq P_{xy} - c \\ 1 & \text{if } P_{center} < P_{xy} + c \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$P_{xy} = \begin{cases} 2 & \text{if } P_{avg} \geq P_{xy} - c \\ 1 & \text{if } P_{avg} < P_{xy} + c \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$c$  in Eqs. (3) and (4) is a value for subdivision of ranges, so it sets the range of index pixel. Comparing to previous CT algorithms, the number of kernel index cases have been increased from  $28 = 256$  to  $38 = 6561$  in the proposed algorithm. For comparing pixel intensity, previous CT algorithms classify neighbors to small and big based on the intensity of center pixel as shown in Fig. 2a, however the proposed algorithm classifies neighbors to similar, small, and big for the higher precision as shown in Fig. 2b.

### 4 Multi-Core Processing

ARM 11 MP-Core system has four 320 MHz ARM11 processors and 32 Kb L1 command cache, 32 Kb L1 data cache, 1 Mb L2 share cache, and interrupt distributor. The consistency of L1 cache is managed by SCU (Snoop Control Unit). The ARM 11 MP-Core supports OpenMP [4] so that users can convert sequential code to parallel code only with directive insertion. The proposed algorithm is

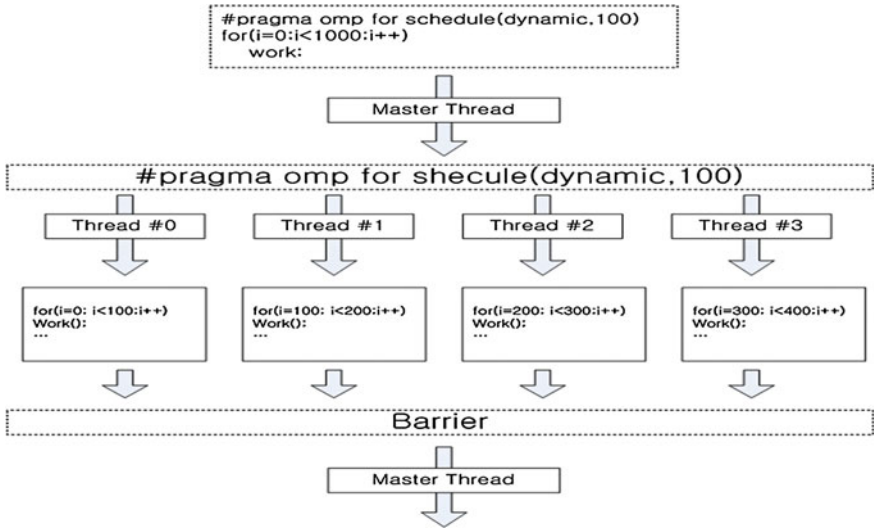


Fig. 3 Dynamic scheduling using OpenMP

paralleled using multi-core condition of OpenMP and the performance speed is compared.

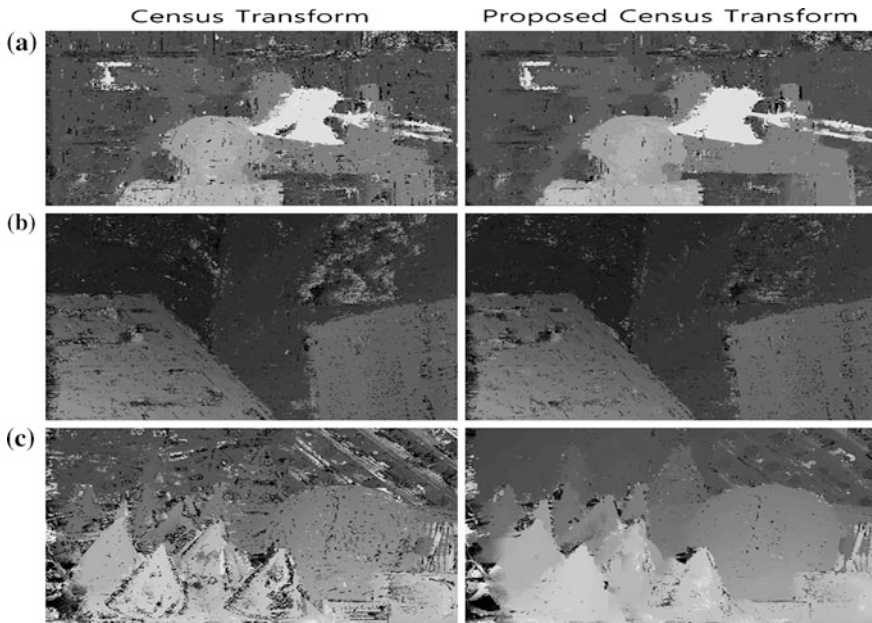
The parallel process divides an input image with the unit of x axes and distributes to multi core as shown in Fig. 3. Once a particular core finishes its allocated iteration, it returns to get another one from the iterations that are left. In this way, the core waiting time has been minimized using dynamic scheduling.

## 5 Experiment

For experiments, images and stereo pairs, provided by Stereo vision Research Page of Middlebury, are used. Using provided source images and ground truth of each images, errors in disparity map, block section, and discontinuous points are extracted. The disparity maps created by CT algorithm and the proposed algorithm have been compared. Sizes and search ranges for each image in experiments are shown on the Table 1. Figure 4 shows disparity maps. The left images are disparity maps extracted by previous Census Transforms Algorithm, and the right images are disparity maps extracted by the proposed Census Transform algorithm. In

Table 1 Parameters of stereo images

Image	Image size	Search range	Scale
Tsukuba	384 × 288	16	16
Venus	434 × 383	20	8
Cones	450 × 375	60	4



**Fig. 4** The result of Stereo matching. **a** Tsukuba. **b** Venus. **c** Cones

disparity map, error rates have been calculated using the method provided by Middlebury College excluding block section. The equation for error rate calculation is expressed as Eq. (5).

$$B = \frac{1}{N} \sum_{(x,y)} (|d_c(x, y) - d_T(x, y)| > \delta_d) \tag{5}$$

$d_c(x, y)$  is a displacement value at  $x, y$  of stereo matching and  $d_T(x, y)$  is a provided actual displacement value. The threshold  $\delta_d$  is set to be 1. If the absolute value of a gap between two displacement values at  $x, y$  in two disparity map is bigger than threshold, the pixel will be regarded as a bad one, and total number of bad pixels is divided by number of whole pixels,  $N$  to calculate the error rate in percentage. When Tsukuba is used as a source image, the error rate in the disparity maps of CT and PCT (Proposed CT) algorithms is presented on the Table 1. Also, error rates for MCT and PMCT (Proposed MCT) algorithms are compared on the Tables 2 and 3.

In Tsukuba image, the Census Transform algorithm has 39.5 % error rate in case of using  $5 \times 5$  sub windows, while the proposed algorithm showed 26.9 % error rate that has been improved 9 %. Moreover, the proposed algorithm shows better performance than the precious algorithms with step bigger sub-windows. The proposed algorithm has been experimented using ARM 11 MP-Core, and the

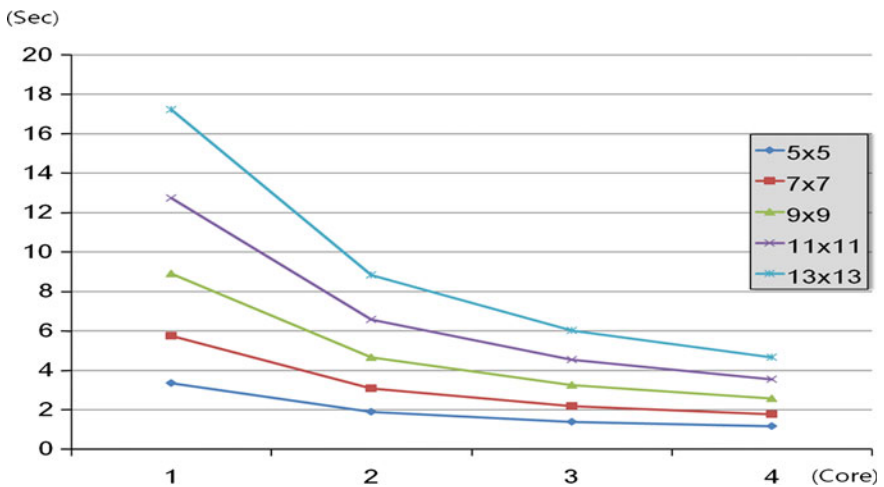
**Table 2** Comparison of error rate for CT and PCT

Size	CT			PCT		
	Non occ'	All	Disc'	Non occ'	All	Disc'
5 × 5	34.7	35.9	35.1	25.5	26.9	27.7
7 × 7	26.7	28	30.5	18.9	20.3	24.3
9 × 9	21.2	22.5	28.3	14.8	16.2	23
11 × 11	17.8	19	27.4	12.4	13.7	22.3
13 × 13	15.1	16.3	26.5	10.6	11.8	22.4

**Table 3** Comparison of error rate for MCT and PMCT

Size	MCT			PMCT		
	Non occ'	All	Disc'	Non occ'	All	Disc'
5 × 5	36.7	38.1	39.2	29.9	31.3	33
7 × 7	26.8	28.4	34.6	22.6	24.3	30.6
9 × 9	21.2	22.5	28.3	18.7	20.4	32.1
11 × 11	17.8	19	27.4	16.3	18.1	33.8
13 × 13	17.1	18.8	37.9	14.9	16.7	35.4

execution speed has been improved using parallel processing. When multi-core is applied for the proposed algorithm, the performance is shown in Fig. 5 for the parallel processing result using PCT with Tsukuba. The proposed algorithm with quad-core processor shows 3.69 times faster than the case of using single core processor.



**Fig. 5** Elapsed time according to the number of cores about PCT

## 6 Conclusion

In general, the development of stereo matching algorithm is focused on the improvement of accuracy rather than the execution speed. As the market of mobile devices is becoming prosperous, however, proper algorithm for embedded system is also important. In this paper, the improved version of CT algorithm is proposed for the stereo matching. The proposed algorithm is applied to multi-core processor and is compared with previous CT algorithm. Since the previous CT algorithm uses less kernel index cases, it has higher error rate. In case of the proposed CT algorithm, kernel index cases have been more diversified, so that the error rate has been decreased. In addition, the execution speed has been also improved through the parallel processing. There are two goals for further progress. First, constant values of the proposed algorithm should be extracted in variable ways according to the size of sub-windows. Second, hardware systems should be designed using the proposed CT algorithm, so that it performs faster stereo matching at embedded conditions.

**Acknowledgments** This work was sponsored by Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy (10039188, SoC platform development for smart vehicle info-tainment system) and Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy (10039145, the development of system semiconductor technology for IT fusion revolution).

## References

1. Humenberger M, Zinner C, Kubinger W (2009) Performance evaluation of a census-based stereo matching algorithm on embedded and multi-core hardware. In: Proceedings of the international symposium on image and signal processing and analysis, vol 6
2. Zabih R, Woodfill J (1994) Non-parametric local transforms for computing visual correspondence. In: Proceedings of the European conference on computer vision, pp 151–158
3. Fröba B, Ernst A (2004) Face detection with the modified census transform. In: Proceeding of IEEE Conference on automatic face and gesture recognition, pp 91–96
4. ARM11MPCore <http://www.arm.com/products/CPUs/ARM11MPCoreMultiprocessor.html>
5. Schastain D, Szeliski R (2002) Ataxonomy and evaluation of dense two-frame stereo correspondence algorithm. *Int J Comput Vis* 47(1):7–42

# A Filter Selection Method in Hard Thresholding Recovery for Compressed Image Sensing

Phuong Minh Pham, Khanh Quoc Dinh and Byeungwoo Jeon

**Abstract** Compressed sensing has been widely researched since the beginning of 2000s. Although there are several well-known signal recovery algorithms, its reconstruction noise cannot be avoided completely, thus requiring good filters to remove the noise in the reconstructing process. Since each different filter has its own advantages and disadvantages depending on specific reconstruction algorithm, the reconstruction performance can be varied according to the choice of filter. This paper proposes an inner filter selection method according to the sampling rate and the property of image to be sensed.

**Keywords** Compressed sensing · Wiener filter · Median filter

## 1 Introduction

In Nyquist-Shannon theorem, two times of signal bandwidth (Nyquist rate) is the slowest rate at which sampling of any band-limited signal should be done to guarantee perfect reconstruction [1]. However, in many low-cost practical applications of image or video, such a high Nyquist rate might be an expensive choice. David Donoho, Emmanuel Candes and Terence Tao introduced the compressed sensing [2, 3] which allows a signal to be sampled at a sub-Nyquist rate while still

---

P. M. Pham (✉) · K. Q. Dinh · B. Jeon  
School of Electrical and Computer Engineering, Sungkyunkwan University, Seoul, Korea  
e-mail: phamphuong@skku.edu

K. Q. Dinh  
e-mail: diqkhanh@gmail.com

B. Jeon  
e-mail: bjeon@skku.edu

attaining near-optimal reconstruction [4]. It relies on the two basic assumptions: signals are sparse and samples are linear functional.

Several factors affect the overall quality of reconstructed images—sparsity of signal, sub-rate, incoherence between sparsifying transform—measurement matrix, and smoothing algorithm (at the decoder’s side), etc. [2].

Some researchers [5] have designed adaptive smoothing algorithms, but they mainly focused on an iterative method which holds the estimation error of signal below a specified threshold. Both theoretical and practical studies proved that the selection of a sub-rate would directly impact reconstructed result. Moreover, some authors investigated the effects of different inner filters in the smoothing process [6].

In this paper, we investigate a filter selection method in the hard thresholding reconstruction algorithm for compressively sensed image. At each iteration, based on characteristic of image and sub-rate of transmitted signal, a better filter is chosen to obtain a higher PSNR value.

Our paper is organized as follows. Section 2 introduces some basic knowledge about the compressed sensing and the Block Based Compressed Sensing and Smoothed Projected Landweber reconstruction algorithm [5]. Section 3 represents the proposed method. Section 4 shows our experimental results. Finally, in Sect. 5 we conclude our work.

## 2 Background

### 2.1 Compressed Sensing Overview

Compressed sensing (CS) is a mathematical algorithm which reconstructs a signal with length  $N$  from  $M$  measurements where  $M \ll N$ . Suppose that  $x$  is a real-valued signal represented by an  $N \times 1$  vector,  $y$  is a sampled vector of length  $M$ :

$$y = \Phi x \quad (1)$$

where  $\Phi$  is called a measurement matrix of size  $M \times N$ , and the ratio  $M/N$  is called the sub-rate. Even if  $M$  is much smaller than  $N$ , if  $x$  is sparse, it is known to be able to reconstruct exactly or approximately  $x$  from  $y$  [2].  $x$  is called  $K$ -sparse if it has at most  $K$  non-zeros coefficients. Since natural signals may not be sparse, we may need to represent  $x$  in a transform domain as:

$$x = \Psi s \quad (2)$$

where  $\Psi$  is a transform matrix with  $N$  columns  $[\psi_1 | \psi_2 | \dots | \psi_N]$ , each column  $\{\psi_i\}$  ( $i = 1 \sim N$ ) is a basis vector of length  $N$ . If  $s$  satisfies Eq. (2) and has at most  $K$  non-zero coefficients ( $K \leq M \ll N$ ), we call  $x$  as  $K$ -sparse in a transform domain represented by  $\Psi$ .  $\Psi$  is named as sparsifying transform or sparsity basis.

The main issue in CS is how to reconstruct  $x$  from  $y$ . A large number of reconstruction algorithms were introduced for CS [5, 7–10]. Taking both the complexity and stability of reconstruction process into account, BCS-SPL (Block Based Compressed Sensing with Smoothed Projected Landweber) [5], which is formed by successively projecting and iterative hard thresholding, provides reduced computational complexity and possibly offers additional optimization criteria.

## 2.2 Block Based Compressed Sensing with Smoothed Projected Landweber

In essence, BCS-SPL combines Block Based Compressed Sensing (BCS) at the encoder’s side with Smoothed Projected Lanweber (SPL) reconstruction algorithm at the decoder’s side. At the encoder side, a natural image ( $x$ ) is divided into blocks of size  $B \times B$ . The  $j$ th block is denoted by  $x_j$ . Then the measurement is done for each block using a measurement matrix  $\Phi_B$  of size  $M_B \times B^2$  where  $M_B = B^2 \cdot \text{sub-rate}$ . The measurement, denoted by  $y_j$ , is computed as:  $y_j = \Phi_B x_j$ .

At the decoder side, the measurement of each vector is processed by using the SPL algorithm as below: For each block  $j$ ,

**Step 1: Compute initial reconstructed vector:**

Compute  $x_j^{(0)} = \Phi_B^T y_j$  and set iteration number  $i = 1$

**Step 2: Compute the reconstructed vector at iteration  $i$ :**

*Step 2.1: Set:  $x' = x^{(i-1)}$*

*Step 2.2: Filter  $x'$  to impose smoothness and remove noise in spatial domain:  
 $x' \leftarrow \text{wiener2}$*

*Step 2.3: Apply hard thresholding:  $x^{(i)} \leftarrow \text{hard thresholding}(x')$*

*Step 2.4: Check stopping condition:*

$$\tau = |D^{(i)} - D^{(i-1)}|$$

where  $D^{(i)}$  is a Mean Squared Error (MSE) and calculated by Eq. (3):

$$D^{(i)} = 1/\sqrt{N} \|x^{(i)} - x^{(i-1)}\|_2 \tag{3}$$

*If  $\tau < 10^{-4}$  go to Step 3, else go to Step 2.1 with  $i = i + 1$ .*

**Step 3: End**  $x^{(i)}$  is a reconstructed image for the block  $j$ .

Note that in this procedure, the compressed-sensed signal is supposed to be exactly available at the decoder side.



### 3 Proposed Method

As mentioned in Sect. 2.2, a Wiener filter is incorporated into the SPL algorithm to make a reconstructed image smoother but it sometime over-smoothes the image [11], and noise still exists in reconstructed image as in Fig. 1.

In spatial domain, a median filter is a very popular non-linear filter which can preserve edges, remove impulse noise (also known as salt and pepper noise), and avoid excessive smoothing [12]. Hence, in case image is prone to be over-smoothed, a median filter can be better to use than the Wiener filter. On the contrary, for an image with much texture, a reconstructed image using Wiener filter has higher quality than that using the median. In this paper, we design and implement a method which can flexibly choose the median filter or the Wiener filter in each Projected Landweber framework’s iteration. The proposed method is illustrated as in Fig. 2.

The measurement vector is processed via successive functional blocks:

- Linear Initialization: As Step 1 in SPL algorithm, an initial vector is computed by multiplying measurement vector with the transposed matrix  $\Phi_B: x_j^{(0)} = \Phi_B^T y_j$ .
- Projected Landweber: iterative hard thresholding operator for each block as Step 2.3 in SPL algorithm:  $x^{(i)} = \mathbf{hard\ thresholding}(x')$ .
- Filter Selection.

Correlation measure can give the similarity between two signals. After the Projected Landweber, we have  $x^{(i)}$  by hard thresholding  $x'$  (see step 2.3 in SPL algorithm): some coefficients of  $x'$  that are below a threshold will be replaced by “0”. Let  $A_j$  be the correlation of the  $j$ th blocks in  $x'$  and  $x^{(i)}$ . Obviously, the higher  $A_j$  is the more similar they are. Denote that  $R = \text{average}(A_j) / \text{maximum}(A_j)$  ( $j = 1 \sim \text{number of blocks}$ ). Due to this definition of  $R$ , we can observe the followings: If  $R$  is high, the values  $A_j$  are approximate, the number of coefficients



Fig. 1 A part of reconstructed image with sub-rate = 0.1. a Barbara. b Lena. c Cameraman

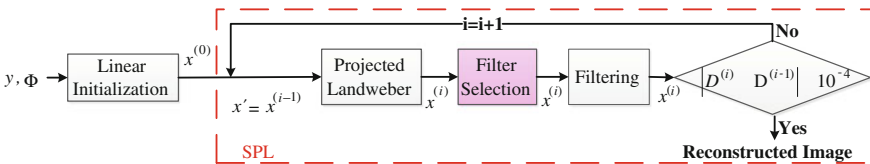
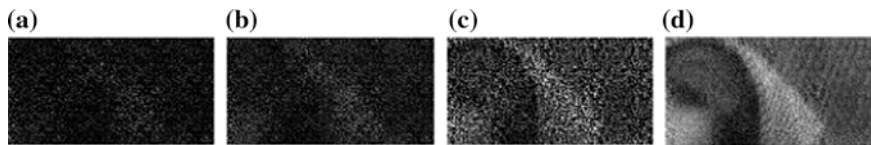


Fig. 2 Filter selection in an hard thresholding recovery of compressively-sensed image



**Fig. 3** A part of Barbara image. **a** and **c** are initial images with sub-rate 0.1 and 0.5, respectively, **b** and **d** are images after the first iteration of hard thresholding

which are less than the threshold in every block are close. In this case,  $x^{(i)}$  is seen to be smooth. In contrast, if there is a remarkable difference in the number of zeroed points from all partitioned blocks, there is significant difference between  $A_j$  values, and  $R$  becomes a small value. In this case,  $x^{(i)}$  is seen to be textured.

An initial vector is calculated by the equation:  $x_j^{(0)} = \Phi_B^T y_j = \Phi_B^T \Phi_B x_j$ . At low sub-rate ( $M \ll N$ ), noise is supposed to be at high level ( $\Phi_B^T \Phi_B$  is small), number of coefficients which are less than the threshold in blocks is large. After each iteration of hard thresholding, all blocks in image are changed equally. Therefore, image is smoother. When  $M$  progresses towards  $N$  (high sub-rate), noise becomes harder to perceive, and image converges fast to the original. Some blocks with many coefficients nearly “0” will have remarkable change after hard thresholding, some blocks with few coefficients less than a hard threshold will not change much after hard thresholding. The difference between changes of blocks is impressive. In this case, image is more textured (Fig. 3).

Based on the experiments and observations related to comparing  $R$  with sub-rate, we decide using a median filter if  $R >$  sub-rate, and using the Wiener filter otherwise.

- Filtering: Use a *filter* which is chosen in Filter Selection for  $x^{(i)}$ :

$$x^{(i)} \leftarrow \mathit{filter} (x^{(i)})$$

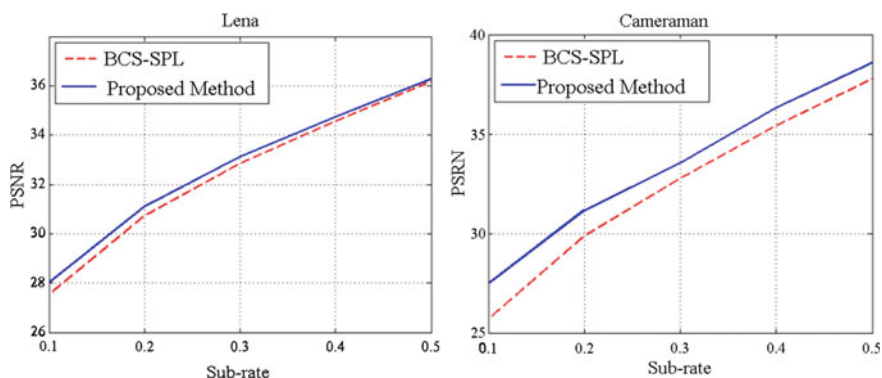
## 4 Experimental Results

In this paper, we evaluate the performance of our proposed method in comparison with original BCS-SPL at sub-rate from 0.1 to 0.5 with images of size  $512 \times 512$ . We use sparsifying transform as Discrete Wavelet Transform and block size  $32 \times 32$  to implement. Tested images are Barbara, Lena, Cameraman, and Girl.

Table 1 collects PSNR values of various reconstructed images when using the proposed method and when using BCS-SPL. In most cases, the proposed method introduces much higher PSNR than BCS-SPL. When the original image has texture as Barbara or Lena, the proposed method shows higher PSNR than BCS-SPL (0.5 dB at sub-rate = 0.1 and 0.2 dB at sub-rate = 0.5). Bold-faced in Table 2 shows significant improvements of the proposed method. In case image is smooth

**Table 1** PSNR of image with block size  $32 \times 32$  at sub-rate from 0.1 to 0.5

Images	Sub-rate	0.1	0.2	0.3	0.4	0.5
Barbara	Proposed method	22.9	24.0	25.2	26.7	28.2
	BCS-SPL Wiener filter	22.4	23.7	25.1	26.5	28.0
Lena	Proposed method	28.0	31.1	33.1	34.8	36.3
	BCS-SPL Wiener filter	27.5	30.7	32.9	34.6	36.1
Clown	Proposed method	<b>27.0</b>	<b>29.7</b>	<b>31.7</b>	<b>33.6</b>	<b>35.2</b>
	BCS-SPL Wiener filter	25.2	29.3	31.6	33.4	35.1
Cameraman	Proposed method	<b>27.5</b>	<b>31.2</b>	<b>33.5</b>	<b>36.3</b>	<b>38.6</b>
	BCS-SPL Wiener filter	25.7	29.9	32.8	35.4	37.8
Girl	Proposed method	29.8	32.6	34.4	36.2	37.9
	BCS-SPL Wiener filter	29.2	32.0	34.2	36.0	37.7

**Fig. 4** PSNR of lena image (*left*) and cameraman image (*right*)

(Girl), our idea produces objective quality higher than BCS-SPL (0.6 dB at sub-rate = 0.1 and 0.2 dB at sub-rate = 0.5).

Figure 4 shows results of coding Lena and Cameraman image. Our proposed method (solid lines) offers a higher PSNR than BCS-SPL (dotted lines).

## 5 Conclusion

This paper proposed a filter selection method in the hard thresholding recovery for compressed sensing image coding. It is shown that the proposed adaptive method inherits the advantage of both median and Wiener filter. In each iteration of hard thresholding, based on property of image and sub-rate, a better filter will be chosen. Consequently, it is more effective on reconstructed image's quality than using only Wiener filter in BCS-SPL algorithm.

**Acknowledgments** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-001-7578).

## References

1. Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley, New York
2. Donoho DL (2006) Compressed sensing. *IEEE Trans Inform Theory* 52(4): 1289–1306
3. Candes EJ, Romberg J, Tao T (2006) Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans Inform Theory* 52(2): 489–509
4. Baraniuk RG (2007) Compressive sensing [lecture notes]. *IEEE Signal Process Mag* 24(4):118–121
5. Fowler JE, Mun S, Tramel EW (2012) Block-Based compressed sensing of image and video. *Found Trends Signal Process* 4(4):297–416
6. Kumar S, Kumar P, Gupta M, Nagawat AK (2010) Performance comparison of median and Wiener filter in image de-noising. *Int J Comput Appl* (0975–8887) 12(4): 27–31
7. Candes E, Romberg J (2005)  $\ell_1$ -magic: recovery of sparse signals via convex programming. Technical report, California Institute of Technology
8. Figueiredo MAT, Nowak RD, Wright SJ (2007) Gradient Projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J Sel Topics Signal Process* 1(4):586–597
9. Chen SS, Donoho DL, Saunders MA (1998) Atomic decomposition by Basis Pursuit. *SIAM J Sci Comput* 20(1):33–61
10. Ji S, Xue Y, Carin L (2008) Bayesian compressive sensing. *IEEE Trans Signal Process* 56(6):2346–2356
11. Khireddine AK, Benmahammed, Puech W (2007) Digital image restoration by Wiener filter in 2D case. *Adv Eng Softw* 38(7):513–516
12. Church JC, Yixin C, Rice SV (2008) A spatial median filter for noise removal in digital images. In: *IEEE Southeastcon*, Alabama, pp 618–623

# Facial Expression Recognition Using Extended Local Binary Patterns of 3D Curvature

Soon-Yong Chun, Chan-Su Lee and Sang-Heon Lee

**Abstract** This paper presents extended local binary patterns (LBP) for facial expression analysis from 3D depth map images. Recognition of facial expressions is important to understand human emotion and develop affective human computer interaction. LBP and its extensions are frequently used for texture classification and face identification and detection. In the 3D surface analysis, curvature is very important characteristics. This paper presents an extension of LBP for modeling curvature from 3D depth map images. The extended curvature LBP (CLBP) is used for facial expression recognition. Experimental results using Bosphorus facial expression database show better performance by 3D curvature and the combination of 3D curvature and 2D images than by conventional 2D or 2D + 3D approaches.

## 1 Introduction

Facial expressions are one of the key components for understanding human emotional states. The recognition of human emotion is very important for affective computing, human robot interaction, and smart devices such as smart TV, smart phone, and smart lighting system. Conventionally, 2D image-based face detection, and feature extraction are used for facial expression recognition. Recently consumer 3D depth cameras like Kinect<sup>®</sup> are available to be used in personal computer. 3D depth information can be captured robust to illumination change and

---

S.-Y. Chun (✉) · C.-S. Lee

Department of Electronic Engineering, Yeungnam University, 214-1 Dae-dong, Gyeongsan-si, Gyeongsangbook-do 712-749, Korea

S.-H. Lee

Daegu Gyeongbuk Institute of Science and Technology, 50-1 Sang-ri, Hyeonpung-myeon, Dalseong-gue, Daegu 711-873, Korea

view change. Not only 3D depth information as well as 2D images is available from Kinect<sup>®</sup>. This paper presents a new feature for 3D depth map image and applies the feature for facial expression recognition from the combination of 2D and 3D depth map features.

Local binary patterns are introduced by Ojala et al. [1] for local shape analysis robust to illumination change. It is originally used for local texture analysis and applied for many other applications such as face identification, face detection, facial expression recognition [2]. A global description of texture from local descriptor can be achieved by dividing whole textures into local regions using regular grid and extracting LBP histograms from each sub regions independently [3, 4]. Patterns of oriented edge magnitudes are also used for face recognition from 2D texture images [5].

For the facial expression recognition from 2D, local binary patterns are also used [6]. Division of sub-region and weighting for dissimilarity measurement shows improvement of the facial expression recognition performance. Support vector machine, linear discriminant analysis, and linear programming are used for the recognition of facial expression from the LBP features and PCA subspaces.

In 3D depth map, depth value in z axis can be represented by gray value in x, y coordinate. Feature extraction in 2D gray image can be applied similarly to the gray value from 3D depth map. However, depth value changes smoothly and does not show clearly motion characteristics directly. Geometrically localized features and surface curvature features are used for better discrimination of facial expression from 3D [7]. 3D patch shape distance of curve shape [8] and histogram of curvature type [9] are also used for 3D facial expression recognition. Many of these features are related to the curvature of the 3D facial surfaces.

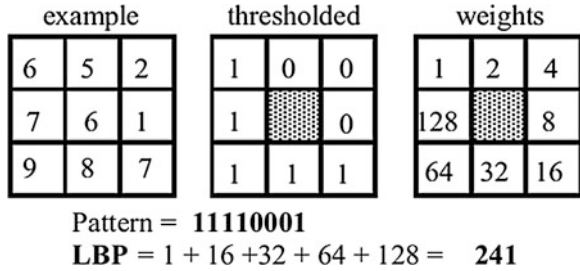
To improve the performance of the depth data using LBP, the difference of the depth data can also be coded in addition to the sign compared with the center [10], which shows improved performance in facial expression recognition from depth map image. However, the coding is complicated and does not use the characteristics of the curvature. This paper presents a new LBP extension which is directly connected to the curvature of the 3D facial surface and useful for facial expression recognition.

## 2 An Extension of LBP for 3D Surface Curvature

### 2.1 Introduction to Local Binary Patterns

Local binary patterns (LBP) are introduced for 2D image texture analysis. The basic principle of LBP operator was based on the assumptions that a texture has a pattern and its intensity, that is, its strength. As the texture pattern is more important and needs to be encoded invariant to the intensity variations, relative strength of neighbor points are described using binary patterns compared with

**Fig. 1** A simple example of LBP



central points. Figure 1 shows a simple example of LBP patterns for 3 × 3 rectangles. Threshold values are computed by comparing with the middle value 6. Weights are used to convert a threshold binary number to a decimal number.

The basic LBP operator was extended into multiscale using variations of radius of the sampling points and rotation invariance using circularly rotated code mapping into its minimum value [1]. Many other LBP variants are proposed in preprocessing, neighborhood topology, threshold and encoding, multiscale analysis, handling rotation, handling color, and so on [2].

## 2.2 Computation of Curvature and Curvature Local Binary Patterns

Geometrically, the curvature of straight line is defined to be zero. The curvature of a circle of radius R is defined to be the reciprocal of the radius R:

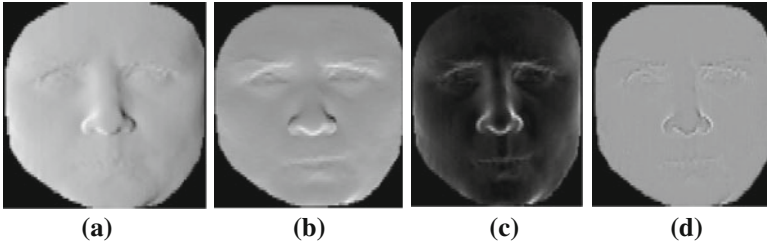
$$\kappa = \frac{1}{R} \tag{1}$$

For the plane curve given explicitly as  $y = f(x)$ , the curvature can be computed by Eq. (2). The equation can further be simplified when the slope is small compared with unity by Eq. (3).

$$\kappa = \frac{|y''|}{(1 + y'^2)^{3/2}} \tag{2}$$

$$\kappa \approx \left| \frac{d^2y}{dx^2} \right| \tag{3}$$

Primitive geometric features such as ridge, peak, saddle, convex hill can be used for facial expression recognition [9]. In order to estimate the geometry features of facial surface, fitting a smooth polynomial patch onto the local surface patch is required. When local patch height  $z(x, y)$  is approximated by polynomial surface as follows:



**Fig. 2** An example of depth map derivative and its magnitude. **a** dx, **b** dy, **c** drv and **d** 2nd drv

$$z(\bar{x}, \bar{y}) = \left( \frac{1}{2}A\bar{x}^2 + B\bar{x}\bar{y} + \frac{1}{2}C\bar{y}^2 + D\bar{x}^3 + E\bar{x}^2\bar{y} + F\bar{x}\bar{y}^2 + G\bar{y}^3 \right), \quad (4)$$

where  $\bar{x}$ ,  $\bar{y}$ , are local coordinate value. Weingarten matrix for the surface fitting becomes as follows:

$$W = \begin{bmatrix} A & B \\ B & C \end{bmatrix} \quad (5)$$

After the eigenvalue decomposition, the principal directions can be estimated.

In this paper, we approximate the curvature of 3D depth image from Eq. (3). When the magnitude of the derivative of x and y axis is given by Eq. (4), the LBP for the magnitude can approximate the second derivative because the LBP also compare the value of center and its neighborhood and have a role as a derivative. We call this approximation of 3D depth curvature as curvature local binary pattern (CLBP).

Computation of CLBP is easy and fast. First, the depth image derivative is computed for x axis and y axis as in Fig. 2a and b. Then, its magnitude is computed by Eq. (6) as in Fig. 2c. Conventional LBP or its variations are applied to the magnitude of the derivatives. Figure 2d shows the second derivative of the depth map image, which shows noisy characteristics of depth image data.

$$dz = \sqrt{dx^2 + dy^2} \quad (6)$$

### 3 Facial Expression Recognition Using Curvature LBP

This section explains how to recognize facial expressions using Curvature LBP from 3D depth image database. Bosphorus database [10] is used to evaluate the performance.



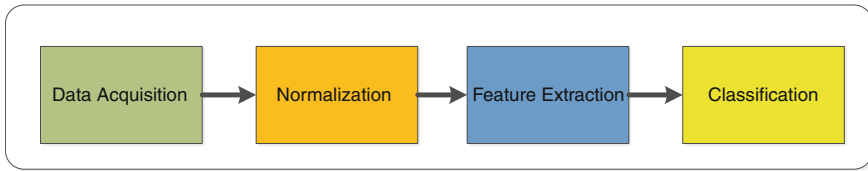


Fig. 3 Overview of facial expression recognition system from 2D and 3D

### 3.1 System Configuration

There are four steps to recognize facial expressions: data acquisition, normalization, feature extraction, and classification. 2D and 3D facial expression data and their landmark point data was collected from the Bosphorus database [10]. A mean landmark shape is estimated using Procrustes algorithm. Each 2D image and 3D depth data are normalized by geometric transformation to fit landmark points to the mean landmark shape. 2D appearance image, 3D depth gray image, 3D curvature image and their combinations are used. For the classification of facial expressions, Chi square distance-based NN classification, and support vector machine (SVM) are used. Figure 3 shows the whole steps for facial expression recognition.

### 3.2 Database for Facial Expression Recognition

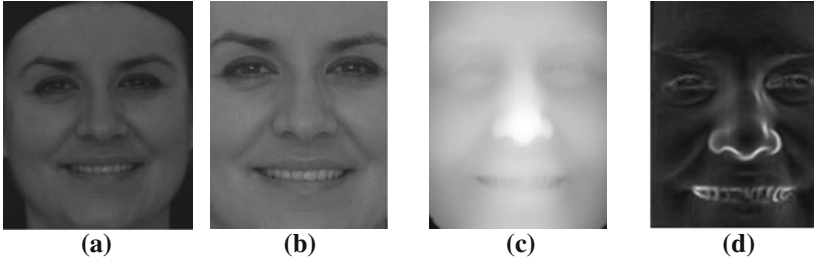
In this paper, Bosphorus facial expression database [10] is used. This database provides high resolution 2D images and low resolution 3D map, and their landmark points, 3D pose variations, facial action unit (AU), and facial expression types. 41 subjects whose data contains all six different facial expressions are used (Fig. 4).

### 3.3 Feature Extraction for Facial Expression Recognition

The 2D facial images and 3D depth images are normalized based on landmark points and cropped to remove background images and boundaries, which are not



Fig. 4 Sample examples of facial expression in 2D and 3D from Bosphorous DB



**Fig. 5** Cropped 2D, 3D, and 3D curvature images used for feature extraction. **a** Original image, **b** Cropped 2D image, **c** Cropped 3D depth image and **d** Cropped 3D curvature image

relevant to facial expression recognitions. A subject's original image, cropped 2D and 3D depth images are shown in Fig. 5a, b, and c. The cropped 3D curvature image in Fig. 5d represents the magnitude of the image derivatives in the x axis and y axis.

Not only conventional uniformity 2 LBP (LBP U2), but also uniformity 4 LBP (LBP U4), Local derivative pattern (LDP) [11], center-symmetric local binary patterns (CS-LBP) [12] are used to extract local pattern histogram from 2D, 3D, and 3D curvature images to compare performance in different type of LBPs.

### 3.4 Facial Expression Recognition

Extracted binary pattern features are histogram distributions according to applied local pattern descriptors and input source types. Chi square statistics of distance measurement in Eq. (7) and nearest neighborhood classifications are applied for facial expression recognition from different histogram distributions.

$$x^2(S, M) = \sum_{i,j} \frac{(S_{i,j} - M_{i,j})^2}{S_{i,j} + M_{i,j}} \quad (7)$$

In addition, SVM classifiers are learned after converting histogram distribution into vector representation. As shown in the experimental results, SVM performs better than nearest neighborhood search based on Chi square distance.

## 4 Experimental Results

We tested the performance of the proposed Curvature LBP and existing methods (LBP, CS LBP, LDP) for facial expression recognition using Bosphorus facial expression database. Leave one-subject out test method, which use 40 subject for

**Table 1** Facial expression recognition average results with 2D, 3D, 3D Curvature using Chi square method

	2D	3D	3D Curvature	2D + 3D	2D + 3D Curvature
LBP (u2)	60.98	54.88	65.85	62.20	65.85
LBP (u4)	57.72	59.73	63.01	61.38	<b>66.26</b>
LDP	58.54	58.13	55.69	58.94	63.41
CS LBP	60.98	52.85	60.98	61.38	60.98

training and one subject for testing, is used for the evaluation of the subject-independent facial expression recognition performance in different features and classifiers.

### 4.1 Facial Expression Recognition Performance in Different Feature

For the facial expression recognition, we used Chi square distance with LBP (u2), LBP (u4), CS LBP, LDP, CLBP. We tested the performance of each feature extraction methods using leave-one subject-out test method. Table 1 shows the estimated performance in the combination of 2D, 3D, and 3D curvature faces with four different feature extraction methods. In the experiment, images were resized into 110 × 150 pixels with 6 × 7 sub-regions. According to experimental result, 2D + 3D Curvature LBP (u4) shows the best performance in Chi square distance evaluation.

### 4.2 Facial Expression Recognition Performance by SVM

For better classification of facial expression, we used SVM with C-SVC type with RBF (Radial Basis Function) kernels. We used 120 images for training, others for evaluation among 246 images. We tested the performance of each feature extraction method. Table 2 shows the estimated facial expression recognition performance by SVM. According to experimental result, 2D + 3D CLBP (u2) shows the best performance in overall experiments.

**Table 2** Facial expression recognition results with 2D, 3D, 3D Curvature using SVM

	2D	3D	3D curvature	2D + 3D	2D + 3D curvature
LBP (u2)	76.19	65.08	64.29	69.05	76.98
LBP (u4)	67.46	65.87	63.49	69.84	75.40
LDP	70.63	59.52	58.73	24.29	74.60
CS LBP	74.60	65.08	62.70	71.43	76.19

## 5 Conclusions

This paper presented a new curvature binary pattern analysis to extract curvature information from 3D depth map images. The curvature binary from 3D depth image with the combination of 2D LBP shows the best performance in the facial expression recognitions from the Bosphorus facial expression database.

**Acknowledgments** This work was supported by the DGIST R&D Program of the Ministry of Education, Science and Technology of Korea (13-IT-03).

## References

1. Ojala T, Pietikainen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distribution. *Pattern Recognit* 29
2. Pietikainen M, Hadid A, Zhao G, Ahonen T (2011) *In computer vision using local binary patterns*. Springer (2011)
3. Ahonen T, Hadid A, Pietikainen M (2004) Face recognition with local binary patterns. In: *Proceedings of ECCV*, pp 469–481
4. Ahonen T, Hadid A, Pietikainen M (2006) Face description with local binary patterns: application to face recognition. *IEEE Trans PAMI* 28
5. Vu NS, Caplier A (2011) Enhanced patterns of oriented edge magnitudes for face recognition and image matching. *IEEE Trans Image Process* 21:135–1365
6. Shan G, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis Comput* 27(6):803–816
7. Sha T, Song M, Bu J, Chen C, Tao D (2011) Feature level analysis for 3D facial expression recognition. *Neurocomputing* 74:2135–2141
8. Maalej A, Amor BB, Daoudi M, Srivastava A, Berretti S (2011) Shape analysis of local facial patches for 3D facial expression recognition. *Pattern Recognit* 44:1581–1589
9. Wang J, Yin L, Wei X, Sun Y (2006) 3D facial expression recognition based on primitive surface feature distribution. In: *IEEE Conference on computer vision and pattern recognition*, pp 1399–1406
10. Savran A, Alyuz N, Dibeklioglu H, Gokberk O, Sankur B, Akarun I (2008) Bosphorus database for 3D face analysis. In: *Proceedings of the Workshop on BIOID*
11. Zhang B, Gao Y (2010) Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. *IEEE Trans Image Process* 19:533–544
12. Kalra P, Peleg S (2006) Description of interest regions with center-symmetric local binary patterns. In: *Proceedings of the ICVGIP, LNCS 4338*, pp 58–69

# Overview of Three and Four-Dimensional GIS Data Models

Tuan Anh Nguyen Gia, Phuoc Vinh Tran and Duy Huynh Khac

**Abstract** The paper focuses on brief presenting again the 3D, 4D GIS data models that have been proposed in the past. The paper uses the tables to compare models based on various criteria in the applications of 3D, 4D GIS. In 3D data models, there were some special models because they can represent the spatial objects by multiple levels of detail. The paper also compares the differences of these models in tables. These tables are the foundation for the developers to build the GIS applications.

**Keywords** 3D · 4D data model · GIS

## 1 Introduction

3D GIS are a system that be able represent, manage, and manipulate, analysis information links with 3D phenomena. 3D GIS data model is the key to 3D GIS [1] and is a big topic in the five major topics of 3D GIS: WebGIS, data presentation, spatial analysis, data model and data collection [2]. The selecting a data model to represent 3D objects 3D GIS for a specific application will determine methods to store, access. There are many models of the authors have suggested [1, 3–8] for spatial and [9–17] for spatial-temporal. The purpose of the paper provides an

---

T. A. N. Gia (✉)  
University of Science, Ho Chi Minh City, Vietnam  
e-mail: anhngt2003@yahoo.com

P. V. Tran · D. H. Khac  
University of IT, Ho Chi Minh City, Vietnam  
e-mail: Phuoc.gis@gmail.com

D. H. Khac  
e-mail: huynhkhacduy@gmail.com

overview of 3D, 4D data models have been proposed, compared the models on important criteria by tables. These tables will support to recognize the development trend of the 3D, 4D GIS data models in the future. The attempts to classify the models will be the foundation for the researches related to 3D GIS and 4D GIS. Previously several authors have made this issue for spatial model and spatial-temporal. However, the works still missed some of models appearing recently and some criteria.

The paper proposes four tables to compare and classify the spatial models: 3D objects are represented by its boundaries; voxel elements; a combination of the 3D basic block and by a combination of the above methods. The paper also proposed two tables to compare the spatial-temporal 3D, 4D models. The 3D, 4D models include either 2D spatial + 1D temporal or 3D spatial + 1D temporal. This paper structure includes three sections. [Section 1](#) introduces about the 3D, 4D GIS data models. [Section 2](#) has five comparison tables of models on many different criteria. [Section 3](#) describes and compares the 3D, 3.5D, 4D spatial-temporal data modes.

## 2 Compare Models

There are many data models, which were proposed in many past years. These modes are classified by four the approaches. They include B-REP, CGS, Voxel and hybrid method. B-REP approach has the models: 3D-FDS (Format Data Structure), TEN (Tetrahedral Network), OO Model (Object Oriented Model), SSM (Simplified Spatial Model), SOMAS (Solid Object Management System), UDM (Urban data Model), OO 3D (Object Oriented 3D), CITYGML model, LUDM model. Voxel has two modes: 3D Array and Octree and hybrid approach has also two models: V-3D, B\_REP + CSG.

### ***2.1 Comparing the Models on the Following Criteria: Surface Representation Method, Objects Inside Representation***

See [Table 1](#).

### ***2.2 Comparing the Models on the Following Criteria: Primitive Elements, Geometry Elements and Application of the Model***

See [Table 2](#).

**Table 1** Comparing models for 2.1

Style of model	Authors	Model	Surface representation	Objects inside representation
BREP	Molenaar 1990	3DFDS	Non triangular	No
	Pilouk 1996	TEN	Triangular	Yes
	Zlatanova 2000	SSM	Non triangular	No
	Delalosa 1999	OO	Triangular	Yes
	Pfund 2001	SOMAS	Non triangular	No
	Coors 2003	UDM	Triangular	No
	Shi et al. 2003	00 3D	Triangular	Yes
	Groger et al. 2007	City GML	Triangular	No
Voxel	Anh N.G Tuan 2011	LUDM	Non triangular	No
	NULL	3D Array	Non triangular	Yes
CSG	Meagher 1984	Octree	Non triangular	Yes
	Samet 1990	CSG	Non triangular	Yes
Combi nation	Xinhua et al. 2000	V-3D	Non triangular	Yes
	Chokri et al. 2009	B_REP CSG	Non triangular	No

**2.3 Comparing the Models on the Following Criteria: Spatial Structure, Direction, Measurement and Topology**

See Table 3.

**2.4 Comparing Models on Query Criteria: Attribute, Location and Topology**

See Table 4.

**2.5 Comparing Models for Representation: Lod, Curve Surface, Semantic, History**

See Table 5.

**3 The Spatial-Temporal 3D, 3.5D and 4D Models**

There were several spatial-temporal models proposed in the past. The Snapshot model was proposed in 1984 by Dangermon, Space Time Composite by Chrisman 1988, SpatioTime Cube by Szego 1987. Base-State was proposed by Armenakis

Table 2 Comparing the models for 2.2

Style of model	Model, primitive elements	Geometry elements	Main ideal	Data size	App
BREP	3DFDS. Point, Line, surface, body	Node, arc, face	2D GIS model	Large	3D UM
	TEN. Point, line, surface, body	Node, Arc triangle, tetra	Tetrahedron	Large	GA
	OO. 0-simplex, 1-2-3 simplex	Node, arc, face	Oriented-Object, n-simplex	Large	3D UM
	SSM. Point, line, surface, body	Node, face	Oriented-Object, topology	Small	3D WEB Urban
	SOMAS. Point, line, polygon, solid	Vertex, edge face, solid	Oriented-Object	Large	3D UM
	UDM. Point, line, surface, body	Node, face	Triangular	Small	3D UM
	OO3D. Point, line, surface, volume	Triangle, segment, node	Oriented-object, triangular	Large	3D UM
	CityGML. Point, curve, surface, solid	Polygon, linestring	Define standards of object	Large	3D UM
	LUDM. Point, line, surface, solid, LOD	Node, face	Represent LOD for solid	Large	3D UM
	Voxel	3D Array. Elements/ Octree, Voxel	None None	Partition object by array Partition object by voxel	Very large Very large
CSG	CSG. Basic 3D block	NULL	Combine basic 3D block	Small	CAD CAM
Combination	V3D. Point, line, surface, body, raster	Node, edge, Face	Combine BREP and raster	Large	3D UM
	B-REP + CSG. Point, Line, Surface, Body, Basic 3D block.	Linestring, Face	Combine BREP and CSG	Large	3D UM

3D UM 3D Urban management; GA Geographical application



**Table 3** Comparing the models for 2.3

Model	Spatial structure	Direction	Measurement	Topology
3DFDS	V	Yes	No	Yes
TEN	V	No	Yes	No
SSM	V	No	No	Yes
OO	V	Yes	No	No
SOMAS	V	Yes	No	No
UDM	V	Yes	Yes	No
OO 3D	V	No	No	No
CityGML	V	No	No	No
LUDM	V	No	Yes	No
3DArray	R	No	Yes	No
Octree	R	No	Yes	No
CSG	V	No	Yes	No
B_REP + CSG	V	No	Yes	No
V-3D	VR	No	No	No

**Table 4** Comparing the models for 2.4

Model	Attribute query	Position query	Topology relationship query
3DFDS	No	Yes	Yes
TEN	No	Yes	No
SSM	No	Yes	Yes
OO	No	Yes	No
SOMAS	No	Yes	No
UDM	No	Yes	No
OO 3D	No	Yes	No
CityGML	Yes	Yes	No
LUDM	No	Yes	No
3D Array	No	No	No
Octree	No	No	No
CSG	No	Yes	No
B_REP + CSG	Yes	Yes	No
V-3D	No	Yes	No

1992, Object- Oriented by Worboys. Three-Domain was proposed by Yuan 1994, Event-Oriented by Peugot 1995, History graph by Renolen 1996, STER by Tryfona 1997. MADS was proposed by Spaccapietra 1999, STUML by Tryfona 2000. Moving objects was proposed by R.H.Guting 2000, by Balovnev 2002. OO was proposed by Shouheil Khaddaj, Event Based by Shuo Wang 2005, Geotoolkit + Geodeform was in 2002 and TUDM in 2012. Tables 6 and 7 compare the models for the criteria: the number of dimension, change history of objects, the time types and their application.

**Table 5** Comparison the models for 2.5

Model	LOD	Curve surface	Semantic	History
3DFDS	1	No	No	No
TEN	1	Yes	No	No
SSM	1	No	No	No
OO	1	Yes	No	No
SOMAS	1	No	No	No
UDM	1	Yes	No	No
OO 3D	1	Yes	No	No
CityGML	5	Yes	Yes	No
LUDM	n	Yes	No	No
3D Array	1	Yes	No	No
Octree	1	Yes	No	No
CSG	1	Yes	No	No
B_REP + CSG	1	Yes	No	No
V-3D	1	No	No	No

**Table 6** Comparison between the spatial- temporal models

Model	The number of dimension	Object
Snapshot	3	Face
Space time composite	3	Face
Spatio time cube	3	Face
Base-state	3	Face
Object-oriented	3	Face
Three-domain	3	Face
Event-oriented	3	Face
History graph	3	Face
STER	3	Face
MADS	3	Face
STUML	3	Face
Moving objects	3	Point, face
OO	3	Face
Event based	3	Face
Geotoolkit + Geodeform	4	Surface
TUDM	4	Body, surface, line, Point

**Table 7** Comparison between TUDM and Geotoolkit + Geodeform

Model	Time	Application
Geotoolkit + Geodeform	Discrete, continuous	GM
TUDM	Discrete	UM

*UM* Urban management; *GM* Geology management

## 4 Conclusion

The main purpose of the paper is to present the development history, the main features of the 3D, and 4D data model. The models divided two groups: spatial model and spatial-temporal model. Each model has different advantages and limitations. The advantages of this model may be difficult for the other models. Choosing model depends on to develop a specific 3D, 4D GIS application. This paper classifies models for the nature of data structure on each model. The paper has created seven comparison tables to models on universal criteria in the areas of GIS, which was based on their characteristics. The tables again help the researchers have an overview of 3D, 4D GIS data model in past and recent years. It is the basis theory to envision in the next work of research and the important foundation for researchers to build models of 4D GIS later.

## References

1. Alias AR (2008) Spatial data modeling for 3D GIS. Springer, Berlin
2. Stoter J, Zlatanova S (2003) 3D GIS, where are we standing. In: Spatial, temporal and multi-dimensional data modeling and analysis
3. Tuan Anh NG, Vinh PT, Vu TP, Sy AT, Dang VP (2011) Representing multiple levels for objects in three-dimensional GIS model, iiWAS2011. ACM Press, Ho Chi Minh City, pp. 591–595 ISBN 978-1-4503-0784-0
4. Billen R, Zlatanova S (2003) 3D spatial relationships model: a useful concept for 3D cadastre. *Comput Environ Urban Syst* 27(4):411–425
5. Chokri K, Mathieu, K (2009) A simplified geometric and topological modeling of 3D building enriched by semantic data: combination of SURFACE-based and SOLID-based representations. In: ASPRS 2009 annual conference, Baltimore, Maryland
6. Coors V (2003) 3D-GIS in networking environments. *Comput Environ Urban Syst* 27:345–357
7. OGC (2007) City geography markup language (Citygml) encoding standard. Open Geospatial Consortium inc
8. Wang X, Gruen A (2000) A hybrid GIS for 3D city models. In: IAPRS, vol 23, Amsterdam
9. Anh N, VinhPT, Duy HK (2012) A study on four-dimensional GIS spatio-temporal data model. In: KSE 2012 the fourth international conference on knowledge and systems engineering, published in IEEE Xplore, Da nang Vietnam
10. Breunig M, Balovnev O, Cremers AB, Shumilov S (2002) Spatial and temporal database support for geologists—an example from the lower Rhine basin. *Neth J Geosci* 81(2):251–256
11. Pelekis N, Theodoulidis B, Kopanakis I, Theodoridis Y (2005) Literature review of spatio-temporal database models. *Knowl Eng Rev* 19:235–274
12. Raza A (2001) Object-oriented temporal GIS for urban applications. PhD thesis, University of Twente. ITC Dissertation 79. ISBN 90-3651-540-8
13. Thapa RB, Murayama Y (2009) Examining spatiotemporal urbanization patterns in Kathmandu Valley, Nepal: remote sensing and spatial metrics approaches. *Remote Sens J*. ISSN 2072-4292
14. Güting RH (2000) A foundation for representing and querying moving objects. *Geoinformatica ACM Trans Databases Syst* 25:1–42

15. Khaddaj S, Adamu A, Morad M (2005) Construction of an integrated object oriented system for temporal GIS. *Am J Appl Sci* 2:1584–1594
16. Wang S, Nakayama K, Kobayashi Y, Maekawa M (2005) An event-based spatiotemporal approach. *ECTI Trans Comput Inf Theory* 1:15–23
17. Zhang N (2006) Spatio-temporal cadastral data model: geo-information management perspective in China. Master thesis, International Institute for Geo-information science and earth observation enschede, The Netherlands

# Modeling and Simulation of an Intelligent Traffic Light System Using Multiagent Technology

Tuyen T. T. Truong and Cuong H. Phan

**Abstract** In this paper, we describe an approach of modeling and simulation based on multi-agent theory to build the model of traffic system. Our model presents a traffic system including traffic light system at a ‘+’ junction. This paper controls duration of red-light/green-light to release traffic congestion on roads. In addition, we establish six Vietnamese traffic rules in our model. The simulation model represents real traffic system in GAMA platform with GAML language. All systems are visualization in GUI to show the participant of vehicles and traffic light on roads.

**Keywords** Modeling · Multiagent-based simulation · Traffic light system · GAMA platform · GAML

## 1 Introduction

Nowadays, in Vietnam, many large cities have to face traffic jam such as Hanoi, HoChiMinh as well as CanTho city. Maybe the main reason of this problem is that the number of vehicles increases too fast. Besides, a traffic light system which is unchangeable duration of red/green-light—called normal traffic light system- are not useful if the density of vehicle on two roads is different. As a result, it is necessary to have an intelligent traffic light system which can change red/green duration of traffic light based on real time conditions. There were many researches that related to this topic. They used extension neural network, fuzzy logic and

---

T. T. T. Truong (✉) · C. H. Phan  
College of Information Technology, CanTho University, CanTho, Vietnam  
e-mail: tttuyen@cit.ctu.edu.vn

C. H. Phan  
e-mail: phcuong@cit.ctu.edu.vn

multi-agent to control traffic light system [1–4]. These articles had just mentioned how to control the time of traffic light system at ‘+’ junctions. They did not concentrate on the behaviours of vehicles. Besides, in Vietnam, almost traffic light systems are set up with unchangeable-duration of red-light/green-light. So that it is easy to occur traffic congestion. This problem is not only waste of time, money and fuel, but also directly affect the ecological environment and people as well. The intelligent traffic light systems can automatically adjust duration of the traffic light based on density of vehicle that existed on road. In fact, to establish an intelligent traffic light system will cost much higher in comparison with a normal traffic light system [3]. And, the efficiency of this system depends on the characteristics of particular crossroad. Besides, to install the initial parameters for the system is important too. Thus, before deploying an intelligent traffic light system at crossroad, we need to scrutinize. The previous researches about intelligent traffic light system did not toward a simulation software based on multi-agent simulation system. It is necessary to have a software to visualize and adjust duration of traffic light system to release traffic congestion. Moreover, in Vietnam, some traffic rules are quite difference in comparison with in other countries. So that, this paper describes the model of the traffic system in “+” junction which includes many types of vehicles and traffic rules in Vietnam only. And then, we use multi-agent technology to simulate the behaviour of vehicles on roads and control duration of traffic light based on the number of vehicles on roads with GAML language in GAMA platform.

This paper includes five sections. The [Sect. 1](#) is an introduction which briefly shows the motivation of this paper and some related works. Models will be presented in the [Sect. 2](#). The [Sect. 3](#) is simulation section. In this section, we describe how to simulate the models. Another [Sect. 4](#) is experimentation section, which shows some scenarios and evaluates the simulation results. The [Sect. 5](#) is conclusion.

## 2 The Model of Intelligent Traffic Light System

There are many types of vehicles such as car, truck, moto, bike and etc. In this paper, we divide them into three types: car, truck and moto. The attributes and behaviour of these types are nearly the same as shown in [Fig. 1](#). In some areas in Vietnam, motobike is allowed to turn right while the red-light is turning on. So that, we also set up parameter for this rule (true (by default) if allow to turn right while red-light is turning on otherwise false). The attributes and behaviours of vehicles, traffic light, and road are pointed out in [Fig. 1](#).

The moving direction of vehicle named  $h$  and its value is in  $[0..359]$ . In general, while moving, vehicle has to avoid collision with other neighbour vehicles. [Figure 2](#) shows the safety areas in left, right and front of vehicle. The safety areas are divided by two lines:  $f1$  and  $f2$  whereas  $(x, y)$  is current location of vehicle. The radius of this circle is depended on current speed of vehicle.

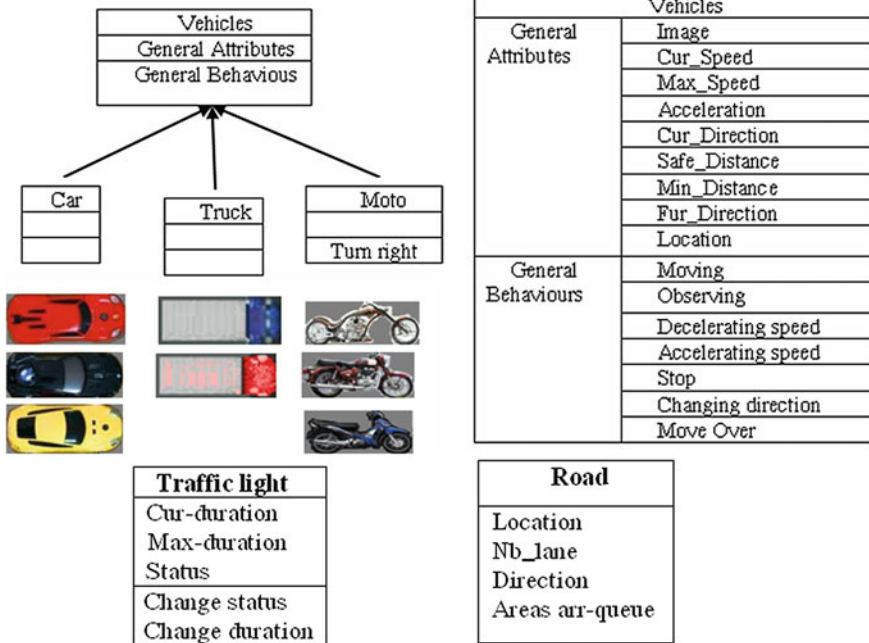
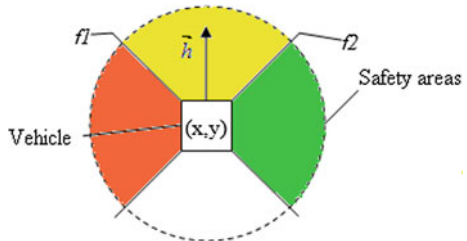


Fig. 1 The attributes and behaviours of vehicles, traffic light and road

Fig. 2 The safety areas of vehicles



Traffic light is also built as an agent as mentioned in Fig. 1. Its attributes are: status (value in [green, red, yellow]), the cur-duration (how many seconds are kept before changing another light’s status), the max-duration (the maximum duration of green-light status). The cur-duration is computed based on the density of vehicles in A, B, C, D arrival queue of previous step (in Fig. 3) by following rules:

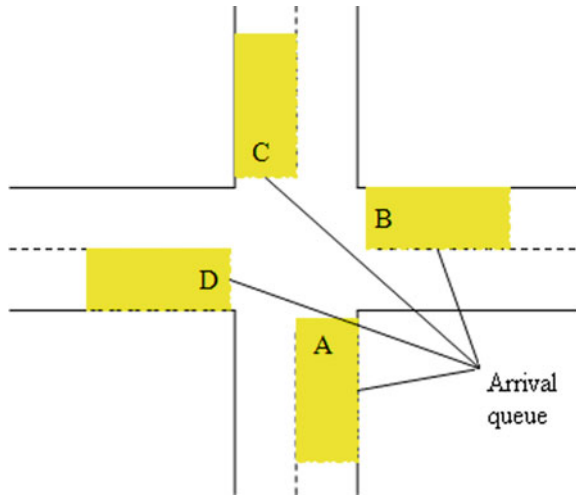


Fig. 3 The arrival queue at crossroad

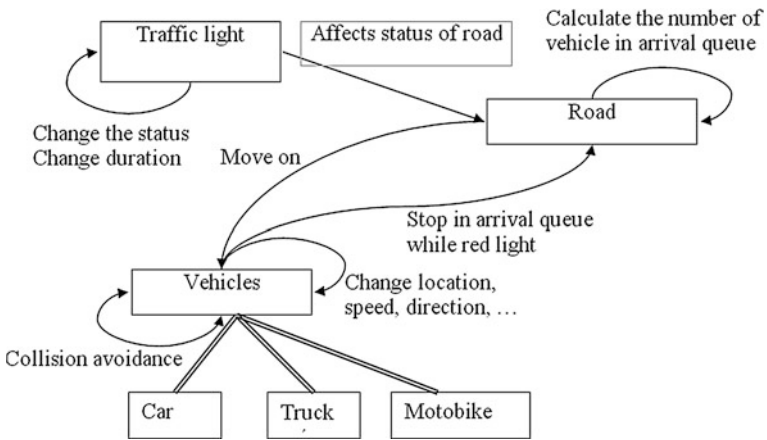


Fig. 4 The intelligent traffic light system model

$$\text{Cur-duration} = \begin{cases} \text{Minimum-time if the rate of density of vehicle in these arrival queue is nearly 0.} \\ (\text{Density of vehicle on B, D in previous light time cycle})(\text{Density of vehicle on A, C in previous light time cycle}) * \text{Avg\_duration.} \\ \text{Maximum-duration if the rate of density in these arrival queue is too large (if cur-duration > Maximum-duration).} \end{cases}$$



The relationship between agents is presented in Fig. 4.

Our traffic light system model is built to adjust the traffic rules in Vietnam. We establish six rules:

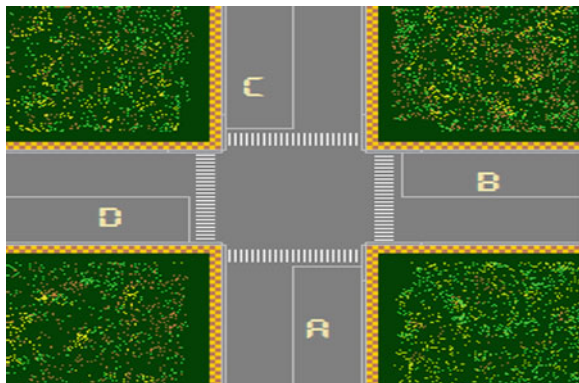
- Traffic light system must match the red-light and green-light duration to ensure that at crossroad, in specific time only vehicles on one road are allowed to go and the other must stop. For example: if vehicles on B, D roads are allowed to go, vehicles on A, C roads must stop.
- All vehicles have to move on right lanes.
- Vehicles must comply traffic light signals at crossroad. Vehicles stop in front of pedestrian bar at the crossroad while red-light is turning on. Especially, moto-bike can turn right while the red-light is turning on (as optional rule based on initial parameter of model).
- The speed of vehicles must lower than max-speed. Vehicles can change the speed to avoid a collision with the other vehicles or reach the ‘+’ junction.
- All vehicles must comply safe distance with neighbour vehicles.
- The red-light duration is equal to sum of the green-light duration and yellow-light duration at the crossroad.

### 3 Intelligent Traffic Light System Simulation

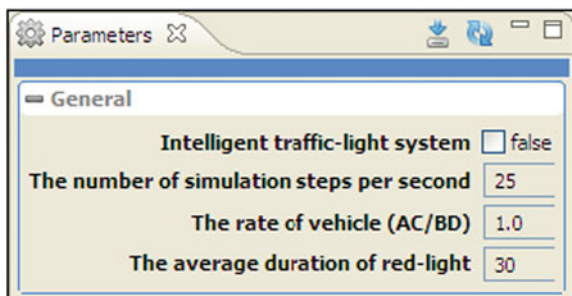
With the characteristics and rules mentioned above, we set up the simulation environment in a grid (GRID) as shown in Fig. 5. It represents ‘+’ junction of two roads and each road has three lanes. The A, B, C, D letters are the name of the arrival queue of roads at crossroad. Vehicles will appear from the rear of the roads and will move with six established rules as mentioned in part II. When vehicles move to the end of road, they will disappear.

Simulation model is established with the input parameters as shown in Fig. 6. The parameters of the model are:

Fig. 5 The grid environment



**Fig. 6** The parameters of simulation model



- The intelligent traffic light system (true/false): allow enable/disable intelligent traffic light system.
- The number of simulation steps per second.
- The rate of vehicle (AC/BD): The rate of the number of vehicle on the AC road per on BD road.
- Avg-duration: The average duration of red-light.

## 4 Experimentation

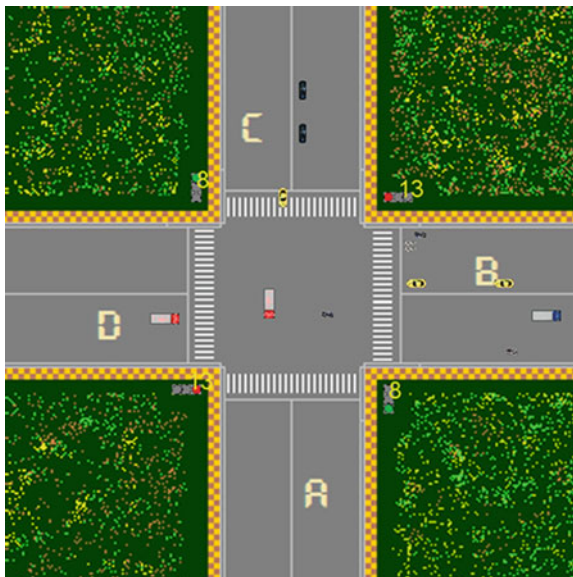
### 4.1 Scenario 1: Normal Traffic Light System

With density of vehicles on two roads are equal as initial parameters in Fig. 6, the simulation results after 7500 steps ( $\sim 5$  min) is not crowded (Fig. 7). If vehicle's density of two roads is equal, traffic jam has not occurred yet as shown in Fig. 7. After that, we change density of road BD up to 3 times (the rate of vehicle (AC/BD) = 0.33) and after step 9000th, there are many vehicles on BD road (Fig. 8). And then, at step 13500th ( $\sim 9$  min), traffic jam occurs at the crossroad as well as in BD road (Fig. 9).

### 4.2 Scenario 2: Intelligent Traffic Light System

When we change the input parameters as shown in Fig. 10 and after 21367 simulation steps ( $\sim 14$  min), traffic congestion has been still not occurred. The duration of red/green-light also changes depending on density of vehicles that exist on roads. Figure 10 shows the intelligent traffic system model which maintained for a long time (more than 14 min). Green-light duration can reach 44 s to release the traffic jams on the road.

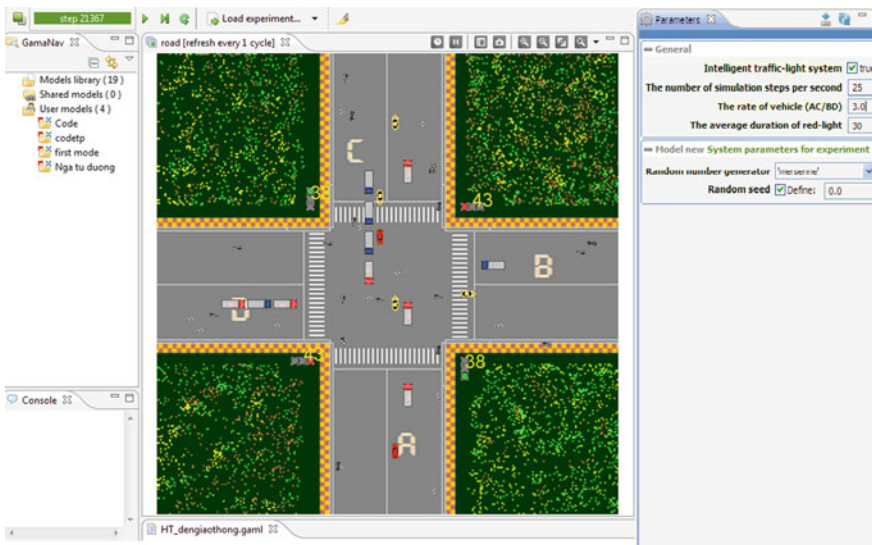
**Fig. 7** The simulation results after 7500 steps with the vehicle on AC and BD are equal



**Fig. 8** The simulation results after 9000 steps with the vehicles on AC and BD are different



**Fig. 9** The simulation results after 13500 steps with the vehicles on AC and BD are different



**Fig. 10** The results of intelligent traffic light system after 21367 steps

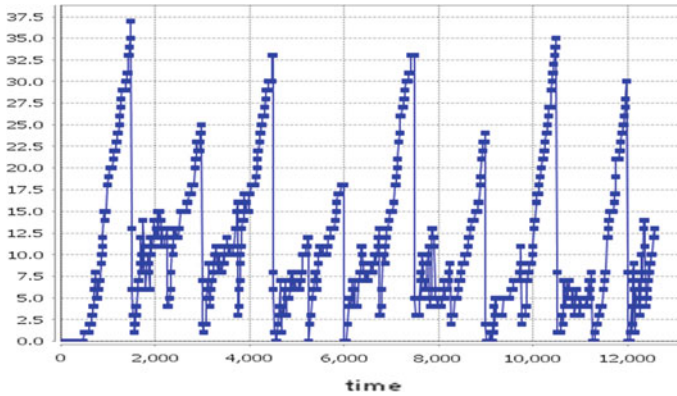


Fig. 11 The number of existed vehicles on normal traffic light system

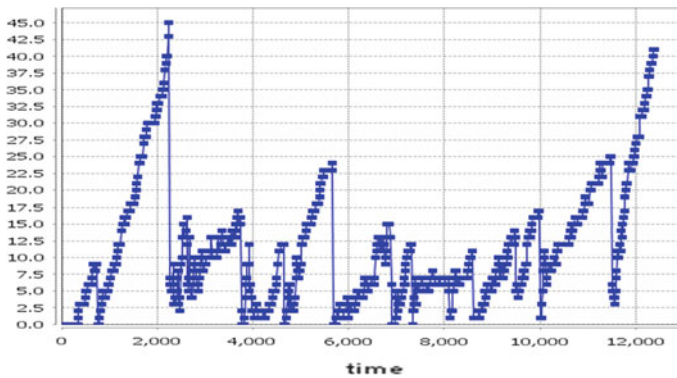


Fig. 12 The number of existed vehicles on intelligent traffic light system

### 4.3 The Comparison Simulation Results of the Two Traffic Lights System

Figures 11 and 12 illustrate the number of vehicles that are being existed on traffic system for the normal traffic light system and the intelligent traffic light system respectively. The results show that the intelligent traffic light system, the number of vehicles decreases faster and keeps close to the horizontal axis in comparison with the normal traffic light system. Through two simulation scenario results, we can see the effectiveness of intelligent traffic light system is better than normal traffic light system. That means the number of waiting vehicles at crossroad is reduced as a result, drivers can save money, time as well as keep good-atmosphere for residents who live near crossroad.

## 5 Conclusion

This paper presents a multiagent approach to simulate an intelligent traffic light system as well as a normal traffic light system. This article also builds simulation model in GAMA platform. All processes of participation of vehicles, traffic light system and traffic rules in Vietnam are visualized in our simulation model. Furthermore simulation results can help to prove the need for an intelligent traffic light system at '+' junction to save time, money and help protect the green environment. Our system can simulate to be able to choose a suitable duration of red-light/green-light in specific crossroad before establishing real traffic light system.

## References

1. Krajzewicz D, Brockfeld E, Jürgen M, Julia R, Rössel C, Tuchscheerer W, Wagner P, Wösler R (2005) Simulation of modern traffic lights control systems using the open source traffic simulation SUMO. In: Proceedings of the 3rd industrial simulation conference 2005, pp 299–302
2. Chao K-H, Lee R-H, Wang M-H (2008) An intelligent traffic light control based on extension neural network. In: Proceedings of the KES 2008, Part I, LNAI 5177, Springer, pp 17–24
3. Chi NN (2011) A method for smart traffic light control system design. In: Proceedings of the national conference on control and automation VCCA-2011, pp 639–644
4. Chinyere OU, Francisca OO, Amano OE (2011) Design and simulation of an intelligent traffic control system. *Int J Adv Eng Tech* 1(5):47–57
5. Drogoul A (2003) Multi-agent based simulation: where are the agents? Springer, 2581/2003, pp 43–49
6. Drogoul A (2008) Lecture notes simulation. Cantho University
7. Wooldridge M (2002) An introduction to multiagent systems. John Wiley & Sons Ltd, Chichester
8. <http://code.google.com/p/gama-platform/> (2013)

# A Numerical Approach to Solve Point Kinetic Equations Using Taylor-Lie Series and the Adomian Decomposition Method

Hag-Tae Kim, Ganduulga, Dong Pyo Hong and Kil To Chong

**Abstract** The point kinetic equations in nuclear dynamics, various analytical methods have been used. In this paper, a numerical approach of point kinetic equations using an inherently large sampling interval and multiple inputs is developed and analyzed. To implement this method, Taylor-Lie Series under the Zero Order Hold (ZOH) is used to approximate the neutron density and precursor concentrations at each corresponding time step. Afterwards, an additional technique, the Adomian Decomposition Method, is used based on its merit of algorithmic and computational advantages in carrying out the discretization.

**Keywords** Point kinetic equations • Numerical solution • Taylor-Lie series • Zero order hold (ZOH) approximation • Adomian decomposition method

## 1 Introduction

The point kinetics model is flawless such as problems mentioned in obtaining essential parameters that explain the reactor within the method. As point kinetic equations are systems which feature stiff nonlinear ordinary differential equations,

---

H.-T. Kim  
Korea Atomic Energy Research Institute, Daejeon, Republic of Korea

Ganduulga · K. T. Chong  
Department of Electronics Engineering, Jeonbuk National University,  
Jeonju, Republic of Korea

D. P. Hong  
Department of Mechanical Engineering, Jeonbuk National University,  
Jeonju, Republic of Korea

K. T. Chong (✉)  
Advanced Research Center for Electronics and Information, Jeonbuk National University,  
Jeonju, Republic of Korea  
e-mail: kitchong@chonbuk.ac.kr



by solving the equations, estimations of various system variables of the reactor core and the transient behavior of the reactor's power can be obtained [1]. In addition to solving these equations, the neutron density and delayed neutron precursor concentrations are solved in a tightly coupled reactor as a function of time. Many of the cases, however, use a model reactor with a minimum of six delayed precursor groups to solve the point kinetic equations which further results in a system consisting of seven or more coupled differential equations.

In the field of reactor kinetics, much research has focused primarily on finding the numerical solutions of point kinetic equations. Sanchez [2] proposed an A-Stable Runge–Kutta Method to combat the problem of stiffness. Kinard and Allen [3] devised a piecewise constant approximation (PCA) method to overcome the previously mentioned. Specifically, this method deals with point kinetic equations using the Zero Order Hold (ZOH) assumption for linearity within time-discretization intervals, allowing an exact solution by using exponential matrices to be obtained. The PWS method introduced by Sathiyasheela [4] was dependent on the truncated Taylor series or those of the exponential matrices under the Zero Order Hold (ZOH) assumption. Since then, several other meaningful attempts have been made such as the Taylor series method proposed by Nahla to solve point kinetic equations [1] with a small sampling interval needed for accurate results.

In this paper, a numerical solution based on the Zero Order Hold (ZOH) assumption and Adomian Decomposition Method is proposed in anticipation of providing a remedy to the problems mentioned above. More specifically, the proposed method makes use of the Taylor-Lie series of the neutron density and delayed precursor functions at each time step, then approximates the exact values using the Adomian Decomposition Method.

## 2 Discretization of the Point Kinetics Equation

The point kinetic equations with  $m$  delayed groups are [10, 11]:

$$\frac{dn(t)}{dt} = \frac{\rho(t) - \beta}{\Lambda} n(t) + \sum_{i=1}^m \lambda_i C_i(t) + F(t) \quad (1)$$

$$\frac{dC_i(t)}{dt} = \frac{\beta_i}{\Lambda} n(t) - \lambda_i C_i(t) \quad i = 1, 2, \dots, m \quad (2)$$

where  $n(t)$  is the time-dependent neutron density,  $\rho(t)$  the time-dependent reactivity function,  $\beta_i$  the  $i$ th delayed fraction,  $\beta = \sum_{i=1}^m \beta_i$  the total delayed fraction,  $F(t)$  the time-dependent neutron source function,  $C_i(t)$  the  $i$ th precursor density,  $\Lambda$  the neutron generation time, and  $\lambda_i$  the  $i$ th group decay constant.

The point kinetics equation can then be expressed with the state-space representation of the form:



$$\begin{aligned}
 x_1 &= n(t), & \dot{x}_1 &= \frac{dn(t)}{dt} \\
 x_2 &= C_1(t), & \dot{x}_2 &= \frac{dC_1(t)}{dt} \\
 x_i &= C_{i-1}(t), & \dot{x}_i &= \frac{dC_{i-1}(t)}{dt}
 \end{aligned} \tag{3}$$

$$u_1(t) = \rho(t), u_2(t) = F(t), u_3(t) \dots u_i(t) = 0 \tag{4}$$

Therefore, the system can then be rewritten as:

$$\begin{aligned}
 \dot{x}_1 &= -\frac{\beta}{\Lambda}x_1 + \lambda_1x_2 + \frac{1}{\Lambda}x_1 * u_1(t) + u_2(t) \\
 \dot{x}_2 &= -\frac{\beta_1}{\Lambda}x_1 + \lambda_1x_2 \\
 \dot{x}_i &= -\frac{\beta_{i-1}}{\Lambda}x_1 + \lambda_{i-1}x_i
 \end{aligned} \tag{5}$$

It is also assumed that systems (1) and (2) are driven by an input piecewise constant for the sampling interval, therefore, the zero order hold (ZOH) assumption holds true.

The solutions of Eqs. (1) and (2) are then expanded into a uniformly convergent Taylor series resulting in coefficients that hold true under the ZOH assumption while within the sampling interval. The solutions of Eqs. (1) and (2) are then easily identified using successive partial derivatives of the right hand side of each equation.

$$x_i(k + 1) = x_i(k) + \sum_{l=1}^{\infty} \frac{T^l}{l!} \left. \frac{d^l x_i}{dt^l} \right|_{t_k} = x_i(k) + \sum_{l=1}^{\infty} A_i^{[l]}(x(k), u(k)) \frac{T^l}{l!} \tag{6}$$

where  $x_1(k), x_2(k), \dots, x_i(k)$  are the values of the state vector  $x_1, x_2, \dots, x_i$  at time  $t = t_k = kT$  and  $A_1^{[l]}(x, u), A_2^{[l]}(x, u), \dots, A_i^{[l]}(x, u)$  produced by the following recursive procedure [12]:

$$\begin{aligned}
 A_1^1(x_1, u) &= f_1(x) + u_1g_1(x) + \dots + u_mg_m(x) \\
 A_2^1(x_2, u) &= f_2(x) + u_1g_1(x) + \dots + u_mg_m(x) \\
 A_i^1(x_i, u) &= f_i(x) + u_1g_1(x) + \dots + u_mg_m(x)
 \end{aligned}$$

$$A_i^{l+1}(x_i, u) = \frac{\partial A_i^l(x_i, u)}{\partial x_1} A_1^1 + \frac{\partial A_i^l(x_i, u)}{\partial x_i} A_i^1 + \frac{\partial A_1^{[l]}(x, u)}{\partial u_1} \frac{du_1}{dt} + \dots + \frac{\partial A_1^{[l]}}{\partial u_m} \frac{du_m}{dt} \tag{7}$$

### 3 Adomian Decomposition Approximations

Developed by G. Adomian, the Adomian Decomposition Method (ADM) possesses unique algorithmic and computational advantages, particularly when used to calculate approximations to numerical solutions of nonlinear differential or partial differential equations [5–9].

To approximate, the Adomian Decomposition Method must first be used to discrete the results utilizing the Taylor-Lie series. Within the discrete results of the point kinetic equations obtained, the  $i$ th-dimensional system should be considered given that the decomposition method requires that:

$$x_i(t) = \sum_{n=0}^{\infty} x_i^n(t) \tag{8}$$

When given the nonlinearities  $f_1, f_2, \dots, f_i$  under the zero order hold (ZOH) assumption, the corresponding Adomian polynomials are calculated such that:

$$f_i(x_1, x_2, u(k)) = \sum_{n=0}^{\infty} A_{fi}^n(x_1, x_2, u(k)) \tag{9}$$

where  $A_{f1}^n, A_{f2}^n, \dots, A_{fi}^n$  ( $n = 0, 1, 2, \dots$ ) are the corresponding Adomian polynomials of the nonlinearities  $f_1, f_2, \dots, f_i$ , respectively. Since we have [5–7],

$$\begin{aligned} N_1 &= \int_{t_0}^t f_1(x_1(s), x_2(s), \bar{u}) ds, \\ N_2 &= \int_{t_0}^t f_2(x_1(s), x_2(s), \bar{u}) ds, \\ &\quad \dots, \\ N_i &= \int_{t_0}^t f_i(x_1(s), x_2(s), \bar{u}) ds, \end{aligned}$$

the following can be acquired:

$$N_i = \int_{t_0}^t \sum_{n=0}^{\infty} A_{fi}^n(x_1, x_2, u(k)) ds = \sum_{n=0}^{\infty} \int_{t_0}^t A_{fi}^n(x_1, x_2, u(k)) ds \tag{10}$$

Also, the following can be attained [5–7]:

$$N_i = \sum_{n=0}^{\infty} \int_{t_0}^t A_{fi}^n(x_1, x_2, u(k)) ds = \sum_{n=0}^{\infty} A_i^n \tag{11}$$

Therefore, Eq. (11) can be used to calculate  $A_1^n, A_2^n, \dots, A_i^n$ :

$$A_i^n = \int_{t_0}^t A_{fi}^n(x_1, x_2, u(k)) ds \tag{12}$$

Using  $x^0 = g$ ,  $g_1 = x_1(t_0) + \int_{t_0}^t b_1 \bar{u} ds = x_1(t_0) + b_1(t - t_0)\bar{u}$ ,  $g_2 = x_2(t_0) + \int_{t_0}^t b_2 \bar{u} ds = x_2(t_0) + b_2(t - t_0)\bar{u}$ ,  $\dots$ ,  $g_i = x_i(t_0) + \int_{t_0}^t b_i \bar{u} ds = x_i(t_0) + b_i(t - t_0)\bar{u}$ , the zero-order terms can be calculated as [5–7]:

$$x_i^0(t) = x_i(k) + b_i(t - t_k)u(k) \tag{13}$$

Finally, using

$$\begin{aligned}
 x^1 &= L(x^0) + A^0(x^0), L_{1,1} = \int_{t_0}^t a_{11}x_1(s)ds, L_{1,2} = \int_{t_0}^t a_{12}x_2(s)ds, \\
 L_{2,1} &= \int_{t_0}^t a_{21}x_1(s)ds, L_{2,2} = \int_{t_0}^t a_{22}x_2(s)ds, A_1^n = \int_{t_0}^t A_{f1}^n(x_1, x_2, u(k))ds, \\
 A_2^n &= \int_{t_0}^t A_{f2}^n(x_1, x_2, u(k))ds, \dots, A_i^n = \int_{t_0}^t A_{fi}^n(x_1, x_2, u(k))ds,
 \end{aligned}$$

the following can be calculated:

$$\begin{aligned}
 x_1^1(t) &= L_{1,1} + L_{1,2} + A_1^0 \\
 &= a_{11} \int_{t_k}^t (x_1(k) + b_1(s - t_k)u(k))ds + a_{12} \int_{t_k}^t (x_2(k) + b_2(s - t_k)u(k))ds \\
 &\quad + \int_{t_k}^t A_{f1}^0(x_1(k) + b_1(s - t_k)u(k), x_2(k) + b_2(s - t_k)u(k))ds \tag{14}
 \end{aligned}$$

$$\begin{aligned}
 x_2^1(t) &= L_{2,1} + L_{2,2} + A_2^0 \\
 &= a_{21} \int_{t_k}^t (x_1(k) + b_1(s - t_k)u(k))ds + a_{22} \int_{t_k}^t (x_2(k) + b_2(s - t_k)u(k))ds \\
 &\quad + \int_{t_k}^t A_{f2}^0(x_1(k) + b_1(s - t_k)u(k), x_2(k) + b_2(s - t_k)u(k))ds
 \end{aligned}$$

or:

$$\begin{aligned}
 x_1^1(t) &= a_{11} \left( x_1(k)(t - t_k) + \frac{1}{2!}b_1(t - t_k)^2u(k) \right) \\
 &\quad + a_{12} \left( x_2(k)(t - t_k) + \frac{1}{2!}b_2(t - t_k)^2u(k) \right) \\
 &\quad + \int_{t_k}^t A_{f1}^0(x_1(k) + b_1(s - t_k)u(k), x_2(k) + b_2(s - t_k)u(k))ds \tag{15}
 \end{aligned}$$

$$\begin{aligned}
 x_2^1(t) &= a_{21} \left( x_1(k)(t - t_k) + \frac{1}{2!}b_1(t - t_k)^2u(k) \right) \\
 &\quad + a_{22} \left( x_2(k)(t - t_k) + \frac{1}{2!}b_2(t - t_k)^2u(k) \right) \\
 &\quad + \int_{t_k}^t A_{f2}^0(x_1(k) + b_1(s - t_k)u(k), x_2(k) + b_2(s - t_k)u(k))ds
 \end{aligned}$$

The above process presents higher-order terms in a recursive method using Adomian series [5–7]. This means that in all these expressions, time  $t$  enters as  $t - t_k$  owing to the autonomous nature of the differential equations under the zero order hold (ZOH) assumption. Therefore

$$\begin{aligned}
 x_1^n(t) &= x_1^n(x_1(k), x_2(k), u(k), t - t_k), x_2^n(t) = x_1^n(x_1(k), x_2(k), u(k), t - t_k), \dots, \\
 x_i^n(t) &= x_i^n(x_1(k), x_2(k), u(k), t - t_k), \text{ and}
 \end{aligned}$$

$$x_i(t) = \sum_{n=0}^{\infty} x_i^n(x_1(k), x_2(k), u(k), t - t_k) \tag{16}$$

Finally, by letting  $t = t_{k+1}$ , the above associative properties lead to the following exact sample-data representation:

$$x_i(k + 1) = \Phi_i(x_1(k), x_2(k), u(k)) = \sum_{n=0}^{\infty} x_i^n(x_1(k), x_2(k), u(k), T) \tag{17}$$

Approximate sample-data representations of order  $N$  are attributed to the finite truncation of  $N$  orders:

$$x_i(k + 1) = \Phi_i^N(x_1(k), x_2(k), u(k)) = \sum_{n=0}^N x_i^n(x_1(k), x_2(k), u(k), T) \tag{18}$$

In the previous analysis, the Adomian Decomposition Method was applied as an effective solution to nonlinear systems to the discrete results using Taylor-Lie series to approximate exact values.

### 4 Computational results

In order to verify the effectiveness, the proposed methods' results are compared with exact values. Secondly, the procedure is implemented using various initial conditions and input data such as single step, single ramp, multiple inputs of reactivity, and source function cases. Finally, the results of the simulation are presented and analyzed below:

The single step and single ramp reactivity are considered with all simulations executed using MAPLE software. The initial conditions used within the simulations are identical for simplicity and assumes a source-free equilibrium:

$$\dot{x}_1(0) = 1, \dot{x}_2(0) = \frac{\beta_i}{\lambda_i \Lambda} \quad i = 1 \dots 6$$

For single step reactivity, the parameters are defined as:

$$\Lambda = 0.00002$$

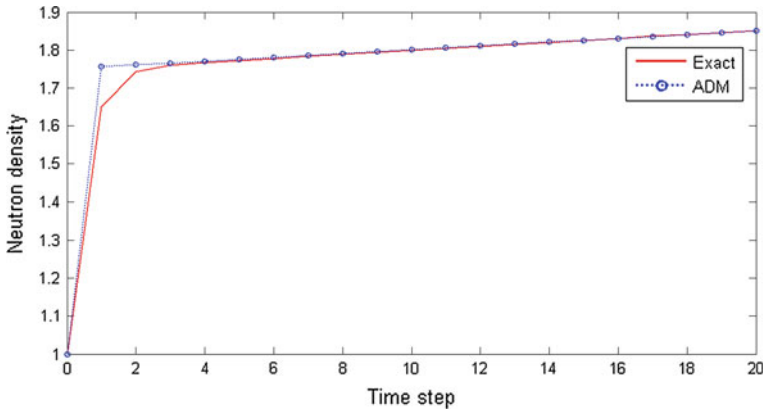
$$\beta = 0.007$$

$$\lambda_i = (0.0127, 0.0317, 0.115, 0.311, 1.4, 3.87)$$

$$\beta_i = (0.000266, 0.001491, 0.001316, 0.002849, 0.000896, 0.000182)$$

In addition, the one prompt subcritical  $\rho(t) = u(t) = 0.003$  is considered.

First, Fig. 1 shows the results of the proposed method (ADM) compared with exact values. Using a sampling time of 0.01 as well as a single step for input data,



**Fig. 1** The comparison of the exact values and the ADM results with single step input

**Table 1** The results using sampling time 0.01 with single step input

Time step	Exact	ADM	Error
1	1.649936078	1.75599774	0.106061662
4	1.766607476	1.771017817	0.004410341
7	1.783618313	1.786037893	0.00241958
10	1.799177796	1.801057968	0.001880172
13	1.814993972	1.816078039	0.001084067
16	1.830780794	1.831098109	0.000317315
19	1.84605792	1.846118175	6.0255E-05

it is evident that apart from time step 3 and before, the method shows no noticeable differences as time passes.

Table 1 outlines the errors and specified results of the ADM in comparison with exact values. Using the data, it is palpable that ADM is in line with exact values. Additionally, the RMSE of the ADM is an estimated 0.108177528.

## 5 Conclusion

In the previous paper, a new numerical solution for point kinetic equations, based on the zero order hold (ZOH) assumption and the Adomian Decomposition Method (ADM) in nuclear reactor dynamics was proposed. Moreover, the proposed method uses a Taylor-Lie series of the neutron density and delayed precursor functions at each corresponding time interval and then approximates the exact values using the ADM. Regarding the performance of the proposed numerical solution, detailed simulations using various initial conditions and input

data such as single step, single ramp, and multiple inputs were chosen. These simulations only verified the increased performance and accuracy. Furthermore, compared to analytical methods, the proposed method is far less complex and does not require small time intervals. For these reasons, solving point kinetic equations is simplified. In the future, this method will be developed further using the first order hold (FOH) assumption, a more complicated approach, to potentially increase accuracy and performance.

**Acknowledgments** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-038978) and (No. 2012-0002434).

## References

1. Nahla AA (2011) Taylor's series method for solving the nonlinear point kinetics equations. *Nucl Eng Des*
2. Sanchez J (1989) On the numerical solution of the point kinetics equations by generalized Runge–Kutta methods. *Nucl Sci Eng* 103:94–99
3. Kinard M, Allen EJ (2003) Efficient numerical solution of the point kinetics equations in nuclear reactor dynamics. *Ann Nucl Energy* 31:1039–1051
4. Sathiyasheela T (2008) Power series solution method for solving point kinetics equations with lumped model temperature and feedback. *Ann Nucl Energy* 36:246–250
5. Adomian G, Rach R, Meyers R (1991) Numerical algorithms and decomposition. *Comput Math Appl* 22:57–61
6. Adomian G (1991) A review of the decomposition method and some recent results on nonlinear equations. *Comput Math Appl* 21:101–127
7. Adomian G (1993) Solving frontier problems of physics: The decomposition method. Springer, London
8. Cherruault Y, Adomian G (1993) Decomposition methods. In: A new proof of convergence. *Math Comput Model* 18:103–106
9. Deeba E, Yoon JM (2002) A decomposition method for solving nonlinear systems of compartment models. *J Math Anal Appl* 266:227–236
10. Hetrick DL (1971) Dynamics of nuclear reactors. The University of Chicago Press, Chicago
11. Petersen CZ, Dulla S, Vilhena MTMB, Ravetto P (2011) An analytical solution of the point kinetics equations with time-variable reactivity by the decomposition method. *Prog Nucl Energy* 53:1091–1094
12. Kazantzis N, Chong KT, Park JH, Parlos AG (2005) Control-relevant discretization of nonlinear systems with time-delay using Taylor-Lie series. *J Dyn Syst Meas Contr* 127:153–159

# Regional CRL Distribution Based on the LBS for Vehicular Networks

HyunGon Kim, MinSoo Kim, SeokWon Jung and JaeHyun Seo

**Abstract** To protect the members of the vehicular networks from malicious users and malfunctioning equipments, certificate revocation list (CRL) should be distributed as quickly and efficiently as possible without over-burdening the network. The common theme among existing methods in the literature to reduce distribution time is to reduce the size of the CRL, since smaller files can be distributed more quickly. Our proposal has been concerned with the problem of how to reduce the size of CRL effectively. We propose a regional CRL distribution method that introduces partitioned CRLs corresponding to certificate authority (CA) administrative regions. A regional CRL includes only neighbouring vehicle's revoked certificates and distributed to vehicles within one CA region. Consequently, since there is no need to process full CRLs by all vehicles, the method can reduce computational overhead, long authentication delay, message signature and verification delay, and processing complexity imposed by full CRL distribution methods.

**Keywords** Regional CRL · VANET · PKI · CRL

---

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2011-0027-006).

---

H. Kim (✉) · M. Kim · S. Jung · J. Seo  
Department of Information Security, Mokpo National University,  
560 Muanno CheongGye-Myeon, Muan-Gun, Jeonnam 534-729, Korea  
e-mail: hyungon@mokpo.ac.kr

M. Kim  
e-mail: phoenix@mokpo.ac.kr

S. Jung  
e-mail: jsw@mokpo.ac.kr

J. Seo  
e-mail: jhseo@mokpo.ac.kr

## 1 Introduction

Vehicular network has been one of the emerging research areas and promising way to facilitate road safety, traffic management, and infotainment dissemination of drivers and passengers. However, without the integration of strong and practical security and privacy enhancing mechanisms, vehicular communication system can be disrupted or disabled, even by relatively unsophisticated attackers [1].

Security and privacy are essential components for the successful deployment of vehicle networks. Those components need to be carefully assessed and addressed in the design of the vehicular communication system, especially because of the life-critical nature of the vehicular network operation. The IEEE 1609.2 standard [2] defines security services for vehicular networks. It defines secure message formats and techniques for processing these secure messages using the public key infrastructure (PKI). In traditional PKI architecture, the most commonly adopted certification revocation scheme is to use CRL that is a list of revoked certificates stored in repositories prepared by CAs. In vehicular networks, the CA adds the identification of the revoked certificate(s) to a CRL. The CA then publishes the updated CRL to all vehicular network participants, and instructing them not to trust the revoked certificate. Timely access to revocation information is important for the robustness of its operation: message faulty, compromised, or otherwise illegitimate, and overall potentially dangerous, vehicles can be ignored.

CRL is straightforward and widespread used. Since the validity period of certificates is long and the number of users is immense, CRLs can grow extremely larger. A great amount of data needs to be transmitted especially, over the air. Also, according to the Dedicated Short Range Communication (DSRC) [3], each on-board unit (OBU) has to broadcast a message every 300 ms indicating its current position, speed, and the road conditions. In such scenario, each OBU may receive a large number of messages every 300 ms, and it has to check the current CRL for all the received certificates, which may incur long authentication delay depending on the size of CRL.

The common theme among existing methods in the literature to reduce distribution time is to reduce the size of the CRL, since smaller files can be distributed more quickly. To address these challenges, our proposal has been concerned with the problem of how to reduce the size of CRL effectively. We propose a regional CRL distribution method to reduce the size of CRL. The basic idea is that vehicles within a CA administrative region operate only one regional CRL, which include neighboring vehicle's revoked certificates.

The reminder of the paper is organized as follows: [Section 2](#) presents problem statements of CRL for vehicular networks and related works; [Sect. 3](#) introduces a proposed regional CRL distribution method, an algorithm used to make regional CRLs at location application server, and network architecture for obtaining precise vehicles location; and finally [Sect. 4](#) summarizes results and end with some conclusions.



## 2 Problem Statements and Related Works

To know global revocation information for all CAs and all vehicles in vehicular networks, full CRLs have to be distributed but, it would be a costly operation. Full CRLs could be large size, which is directly proportional to the revocation rate, the number of nodes in the system, and, the number of certificate used by each vehicle. This requires computational overhead, long authentication delay, message signature and verification delay, and processing complexity imposed by full CRL distribution methods. In addition, in a situation where the rate of revocation is very high, full CRLs will be large and will change often.

The size of CRL in vehicular networks is expected to be large for the following reasons: firstly, to preserve the privacy of the drivers, i.e., to abstain the leakage of the real identities and locations information of the drivers from any external eavesdropper [4, 5], each OBU should be preloaded with a set of anonymous digital certificates, where it has to periodically change its anonymous certificate to mislead attackers [6]. Consequently, an OBU revocation results in revoking all the certificates carried by that OBU leading to a large increase in the CRL size; secondly, the vehicular network scale is very large. CRLs for all the received certificates are stored in OBU of vehicle. Upon receiving signed or encrypted messages from neighboring road-side unit (RSU) or vehicles, each OBU has to check the current CRL for all the received certificates, which may incur long processing delay depending on the size of CRL. The ability to check a CRL for a large number of certificates in a timely manner forms an inevitable challenge to vehicle networks.

The problem of certificate revocation in vehicular networks has hardly attracted any attention. The [7] aims at achieving scalable and efficient mechanism for the distribution of large CRLs across wide regions by utilizing a very low bandwidth at each RSU. CRLs are encoded into numerous self-verifiable pieces, so vehicles only get from the RSUs those pieces of the CRLs. The [8] made many experiments on the size of CRL and how to distribute the CRL in the vehicular networks. The result said when the size of CRL is high the delay time for receiving it will be high. Another idea proposed in [9] said that CRL will store entries for less than a year old. This idea used to decrease the size of CRL, but still suffer from huge size.

## 3 Design of Regional CRL Distribution Method

The method involves segmenting a full CRL into regional CRLs according to a number of CA administrative regions; the regional CRLs will be significantly smaller than the full CRL. Each regional CRL includes only neighboring vehicle's revoked certificates and it would be distributed to vehicles within one CA administrative region. Consequently, all vehicles in vehicular networks may receive and process only relevant CRLs associated with their neighboring vehicles (Fig. 1).

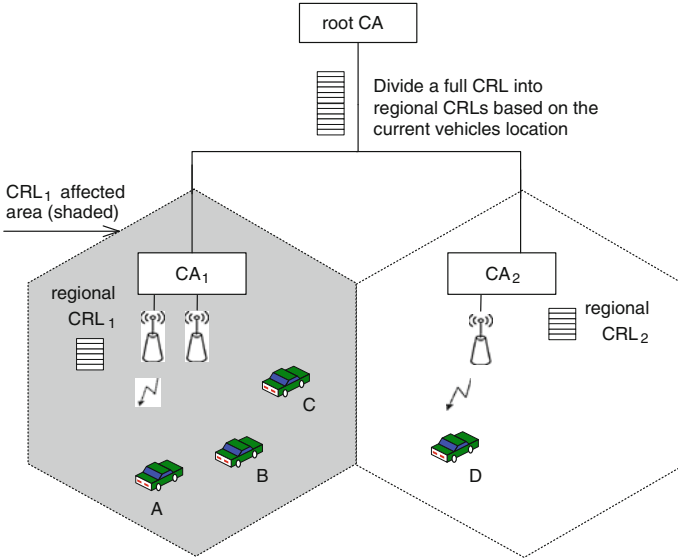


Fig. 1 Regional CRL distribution for reducing the size of CRL

### 3.1 Regional CRL Distribution Procedure

The proposed regional CRL distribution method is shown in Fig. 2. For segmentation, the root CA divides a full CRL into regional CRLs corresponding to the CA administrative regions. The regional CRL<sub>1</sub> is only distributed to vehicles within the CA<sub>1</sub> administrative region. Thus, the regional CRL<sub>1</sub> affects only CA<sub>1</sub> administrative region. Those vehicles would receive the regional CRL<sub>1</sub> and only need to process it.

However, since a regional CRL includes neighboring vehicle’s revoked certificates within a CA administrative region, the root CA has to know all vehicles location when it bundles all the revoked certificates corresponding to the given CA administrative region into single regional CRL. For this, vehicle’s mobility should be carefully considered since vehicles can move from one CA administrative region to another dynamically. Thus, how to efficiently know vehicles location corresponding to CA administrative regions represents a major challenge for the proposed method. Our approach envisioned to achieve this is via the location-based services (LBS) defined by open mobile alliance (OMA) [10]. The LBS can be a vehicular application that provides information and functionality to location application servers based on vehicles geographical location. LBS for the purpose of regulatory compliance and/or commercial services are already commonly supported in today’s deployed 2G and 3G wireless networks.

Figure 2 presents network architecture and regional CRL distribution procedure. We assume that each vehicle is capable of a LBS client called, location services client (LCS) [10]. Serving mobile location center (SMLC) and gateway

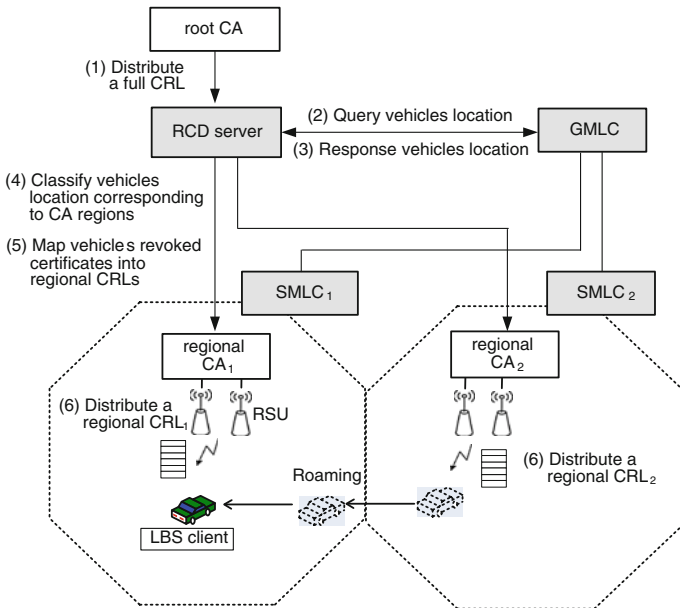


Fig. 2 Regional CRL distribution method

mobile location center (GMLC) are entities of vehicular networks and they provide functionality required to support LBS. The SMLC is responsible to calculate the location of vehicles and the GMLC is an entity obtaining geographic location information of all vehicles by interrogation with SMLCs. The regional CRL distribution server called RCD server, acts as a location application server [10]. To obtain all vehicles location from the GMLC, the mobile location protocol (MLP) [11] can be used that is an application-level protocol for getting the position of vehicles independent of underlying network technology.

The regional CRL distribution procedure involves the following steps: (1) the root CA sends a full CRL to RCD server, (2) the RCD server queries vehicles location to GMLC, (3) it obtains vehicles location from GMLC, (4) it classifies a full CRL into regional CRLs corresponding to CA administrative regions, (5) it maps vehicle’s revoked certificates based on the current location into the regional CRLs, (6) each regional CA distributes the regional CRL to vehicles within its region.

### 3.2 Algorithm for Making Regional CRLs

As a location application server, the RCD server performs main functionalities of the proposed method. Figure 3 presents an algorithm used to make regional CRLs at the RCD server involving the following steps: (1) receive vehicles location obtained in

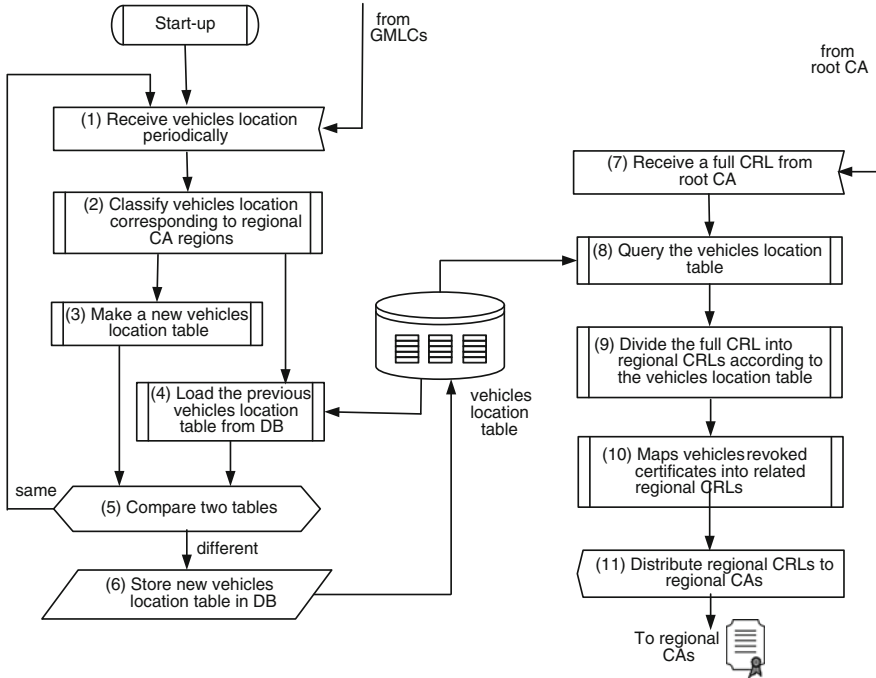


Fig. 3 Algorithm used to make regional CRLs at RCD server

the form of  $x, y$  coordinates from GMLCs periodically, (2) classify vehicles location corresponding to CA administrative regions, (3) make a new vehicles location table, for example  $\{CA_{1area} = \{Vehicle_1(x_1, y_1), Vehicle_2(x_2, y_2), \dots\},$  (4) load the previous vehicles location table from database, (5) compare the previous table and the new table. If two tables are the same then, repeat the first step. This means that vehicles do not move between CA administrative regions for given time period, (6) If different then, store the new table in database, (7) receive a full CRL from CA periodically, (8) when receive a full CRL, query the stored vehicles location table, (9) divide the full CRL into regional CRLs according to the vehicles location table, (10) maps vehicle’s revoked certificates into the related regional CRLs, (11) distribute them to the designated regional CAs and also the regional CAs distributes them to all vehicles within their regions.

### 3.3 Precise Vehicles Location

To obtain precise vehicles location, we can utilize combined access networks capable of LBS as shown in Fig. 4. The RCD server gathers vehicles location from different access networks such as WAVE network, 2G/3G network, 4G LTE

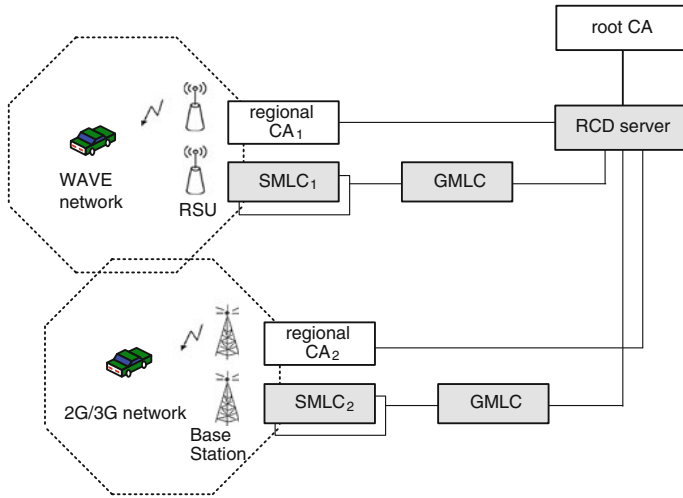


Fig. 4 Network architecture for obtaining precise vehicles location

network, WiBro, and WiFi etc. [2]. Combined LBS can allow the RDS server to know precise vehicles location and to make regional CRLs effectively.

### 4 Conclusions

We have presented a regional CRL distribution method that introduces regional CRLs partitioned by certificate authority administrative regions. The proposed method has been concerned with the problem of how to reduce the size of CRL effectively. The method involves segmenting a full CRL into regional CRLs according to a number of CA administrative regions; the regional CRLs will be significantly smaller than the full CRL. Each regional CRL includes only neighboring vehicle's revoked certificates and it would be distributed to vehicles within one CA administrative region. The advantage of the method is that it can reduce overhead, long authentication delay, message signature and verification delay, and processing complexity imposed by full CRL distribution methods since all vehicles in vehicular networks may receive and process only relevant small CRLs associated with their neighboring vehicles. The disadvantage is that location application server has to manage vehicles location corresponding to CA administrative regions in a real time manner. As a part of further work, we have been planning to evaluate performance analysis comparing full CRL distribution and regional CRL distribution method.

## References

1. Lin X, Lu R, Zhang C, Zhu H, Ho PH, Shen X (2008) Security in vehicular ad hoc networks. *IEEE Commun Mag* 46(4):88–95
2. IEEE Std 1609.2: Trial-Use standard for wireless access in vehicular environments—Security services for applications and management message. *IEEE Standard 1609.2* (2006)
3. Wasef A, Shen X (2009) MAAC: Message authentication acceleration protocol for vehicular ad hoc networks. In: *Proceedings of the IEEE Globecom 2009*, pp 1–6
4. Lin X, Sun X, Ho P-H, Shen X (2007) GSIS: A secure and privacy preserving protocol for vehicular communications. *IEEE Trans Veh Technol* 56: 3442–3456
5. Papadimitratos P, Kung A, Hubaux JP, Kargl F (2006) Privacy and identity management for vehicular communication systems: a position paper. In: *Proceedings of workshop on standards for privacy in user-centric identity management*
6. Laberteaux KP, Haas JJ, Hu YC (2008) Security certificate revocation list distribution for VANET. In: *15th ACM international workshop on VehiculAr InterNETworing (VANET)*, pp 88–89
7. Papadimitratos P, Mezzour G, Hubaux JP (2008) Certificate revocation list distribution in vehicular communication systems. In: *15th ACM international workshop on Vehicular InterNETworing (VANET)*, pp 1–10
8. Kamat P, Baliga A, Trappe W (2006) An identity-based security framework for VANETs. In: *Proceedings of the third ACM workshop on vehicular networks*
9. Raya M, Jungels D, Papadimitratos P, Aad I, Hubaux JP (2006) Certificate revocation in vehicular networks. In: *Laboratory for computer communications and applications (LCA) School of Computer and Communication Sciences*
10. OMA standard (2011) Secure user plan location. V3.0
11. OMA standard (2011) Mobile location protocol. V3.1

# Study of Reinforcement Learning Based Dynamic Traffic Control Mechanism

Zheng Zhang, Seung Jun Baek, Duck Jin Lee and Kil To Chong

**Abstract** A traffic signal control mechanism is proposed to improve the dynamic response performance of a traffic flow control system in an urban area. The necessary sensor networks are installed in the roads and on the roadside upon which reinforcement learning is adopted as the core algorithm for this mechanism. A traffic policy can be planned online according to the updated situations on the roads based on all the information from the vehicles and the roads. The optimum intersection signals can be learned automatically online. An intersection control system is studied as an example of the mechanism using Q-learning based algorithm and simulation results showed that the proposed mechanism can improve traffic efficiently more than a traditional signaling system.

**Keywords** Intelligent transportation system • Cooperative vehicle-highway systems • Reinforcement learning • Traffic control mechanism • Intersection signal control

---

Z. Zhang

Department of Mechanical Engineering, Xian Jiaotong University, Xian,  
Peoples Republic of China

S. J. Baek · K. T. Chong

Department of Electronics Engineering, Jeonbuk National University, Jeonju,  
Republic of Korea

D. J. Lee

Department of Mechanical Engineering, Jeonbuk National University, Jeonju,  
Republic of Korea

K. T. Chong (✉)

Advanced Research Center for Electronics and Information, Jeonbuk National University,  
Jeonju, Republic of Korea

e-mail: kitchong@chonbuk.ac.kr

## 1 Introduction

Intelligent Transportation Systems (ITS) utilizes synergistic technologies and systems engineering concepts to develop and improve transportation systems of all kinds [1]. Machine intelligence on the road has been a popular research area with the advent of modern technologies especially artificial intelligence, wireless communication and advanced novel sensors.

Current traffic signal control system design is based on historic traffic flow data which cannot adapt itself to the rapidly varying situations at a crossroad. In some extreme situations, there are no vehicles during a green light and lots of vehicles waiting at a red one.

Many researchers have proposed schemes to solve the afore-mentioned problems like Choy et al. [2] who introduced hybrid agent architecture for real-time signal control. He suggested in his paper a dynamic database for storing all recommendations of the controller agents for each evaluation period. Liu et al. [3] proposed a calculating method of intersection delay under signal control while Bao et al. [4] studied an adaptive traffic signal timing scheme for an isolated intersection. However all these papers solve the problem according to the history flow data but not the current information [5, 6].

This paper makes the following contributions in particular:

- (a) A novel traffic flow control mechanism is proposed based on the cooperation of the vehicle, road and traffic management systems. A roadside wireless communication network supports a dynamic traffic flow control method.
- (b) Reinforcement learning is introduced as the core algorithm to dynamically plan traffic flow in order to improve efficiency. A Q-learning based intersection traffic signal control system is studied as an example of the proposed mechanism.

## 2 Study of Intersection Signal Control

In this section, a Q learning algorithm will be used to create a real time cooperation policy for an isolated intersection control under the proposed Traffic Control Mechanism. The algorithm and the simulation are both described in detail. The result shows the advantage of the proposed method.

### 2.1 Q-Learning Algorithm

Q learning, a type of reinforcement learning, can develop optimal control strategies from delayed rewards, even when an agent has no prior knowledge of the effects of its actions on the environment [7].



The agent's learning task can be described as follows. We require that the agent learn a policy  $\pi$  that maximizes  $V^\pi(s)$  for all states  $s$ . We will call such a policy an optimal policy and denote it by  $\pi^*$

$$\pi^* \equiv \arg \max_{\pi} V^\pi(s), (\forall s) \quad (1)$$

To simplify notation, we will refer to the value function  $V^{\pi^*}(s)$  of such an optimal policy as  $V^*(s)$ .  $V^*(s)$  gives the maximum discounted cumulative reward that the agent can obtain starting from state  $s$ ; that is, the discounted cumulative reward obtained by following the optimal policy beginning at state  $s$ .

However, it is difficult to learn the function  $\pi^* : S \rightarrow A$  directly, because the available training data does not provide training examples of the form  $\langle s, a \rangle$ . Instead, the only training information available to the learner is the sequence of immediate rewards  $r(s_i, a_i)$  for  $i = 0, 1, 2, \dots$ . As we shall see, given this kind of training information it is easier to learn a numerical evaluation function defined over states and actions, then implement the optimal policy in terms of this evaluation function.

What evaluation function should the agent attempt to learn? One obvious choice is  $V^*$ . The agent should prefer state  $s_1$  over state  $s_2$  whenever  $V^*(s_1) > V^*(s_2)$ , because the cumulative future reward will be greater from  $s_1$ . The agent's policy must choose among actions, not among states. However, it can use  $V^*$  in certain settings to choose among actions as well. The optimal action in state  $s$  is the action  $a$  that maximizes the sum of the immediate reward  $r(s, a)$  plus the value  $V^*$  of the immediate successor state, discounted by  $\gamma$ .

$$\pi^*(s) = \arg \max_a [r(s, a) + \gamma V^*(\delta(s, a))] \quad (2)$$

where  $\delta(s, a)$  denotes the state resulting from applying action  $a$  to state  $s$ .

Thus, the agent can acquire the optimal policy by learning  $V^*$ , provided it has perfect knowledge of the immediate reward function  $r$  and the state transition function  $\delta$ . When the agent knows the functions  $r$  and  $\delta$  used by the environment to respond to its actions, it can then use Eq. (2) to calculate the optimal action for any state  $s$ .

Unfortunately, learning  $V^*$  is a useful way to learn the optimal policy only when the agent has perfect knowledge of  $\delta$  and  $r$ .

Let us define the evaluation function  $Q(s, a)$  so that its value is the maximum discounted cumulative reward that can be achieved starting from state  $s$  and applying action  $a$  as the first action. In other words, the value of  $Q$  is the reward received immediately upon executing action  $a$  from state  $s$ , plus the value (discounted by  $\gamma$ ) of following the optimal policy thereafter.

$$Q(s, a) \equiv r(s, a) + \gamma V^*(\delta(s, a)) \quad (3)$$

Note that  $Q(s, a)$  is exactly the quantity that is maximized in Eq. (3) in order to choose the optimal action  $a$  in state  $s$ . Therefore, we can rewrite Eq. (3) in terms of  $Q(s, a)$  as

$$\pi^*(s) = \arg \max_a Q(s, a) \quad (4)$$

Why is this rewrite important? Because it shows that if the agent learns the  $Q$  function instead of the  $V^*$  function, it will be able to select optimal actions even when it has no knowledge of the functions  $r$  and  $\delta$ . As Eq. (4) makes clear, it need only consider each available action  $a$  in its current state  $s$  and choose the action that maximizes  $Q(s, a)$ . This is exactly the most important advantages of  $Q$  learning, and also is the reason why we choose  $Q$  learning in this paper.

How should the  $Q$  learning algorithm be implemented? The key problem is finding a reliable way to estimate training values for  $Q$ , given only a sequence of immediate rewards  $r$  spread out over time. This can be accomplished through iterative approximation. To see how, notice the close relationship between  $Q$  and  $V^*$ ,  $V^*(s) = \max_{a'} Q(s, a')$ , which allows rewriting Eq. (3) as follows:

$$Q(s, a) = r(s, a) + \gamma \max_{a'} Q(\delta(s, a), a') \quad (5)$$

Equation (5) provides the basis for algorithms that iteratively approximate  $Q$ . In the algorithm,  $\bar{Q}$  will be the learner's estimate, or hypothesis of the actual  $Q$  function.  $\bar{Q}$  will be represented by a large table with a separate entry for each state-action pair. The table can be initially filled with random values (though it is easier to understand the algorithm if one assumes initial values of zero). The agent repeatedly observes its current state  $s$ , choose some action  $a$ , executes this action, then observes the resulting reward  $r = r(s, a)$  and the new state  $s' = \delta(s, a)$ . It then updates the table entry for  $\bar{Q}(s, a)$  following each such transition, according to the rule:

$$\bar{Q}(s, a) \leftarrow r(s, a) + \gamma \max_{a'} \bar{Q}(s', a') \quad (6)$$

Note that the above training rule uses the agent's current  $\bar{Q}$  values for the new state  $s'$  to refine its estimate of  $\bar{Q}(s, a)$  for the previous state  $s$ .

The iterative training rule (6) will be replaced by

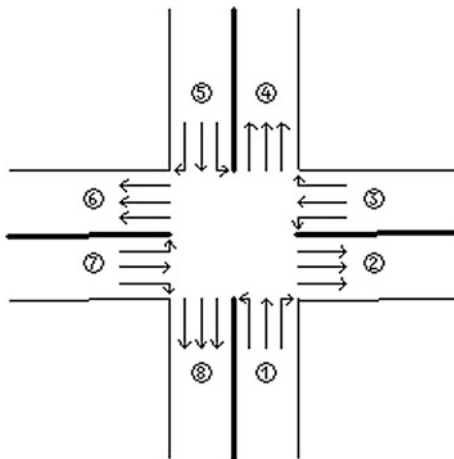
$$\bar{Q}(s, a) \leftarrow g(s, a) + \gamma \min_{a'} \bar{Q}(s', a'). \quad (7)$$

It means that the learning target is to minimize the  $Q$  function by minimizing the total cost when acting based on the optimum action sequences. This is exactly the algorithm used in this paper.

## 2.2 Model of the Intersection Signal System

A traffic system consists of various components, among which the traffic intersection is one of the most important [8]. Our method is applied to a traffic intersection that consists of two intersecting roads, each with several lanes and a set of synchronized traffic lights that manage the flow of vehicles, as shown in Fig. 1.

Fig. 1 Isolated intersection



In this intersection, the rule of traffic management is right-hand based, which is used in China and South Korea. The vehicles in lanes ①, ③, ⑤ and ⑦, are approaching the intersection. Vehicles in ②, ④, ⑥ and ⑧, are leaving the intersection. For each of the approaching lanes, there are three directions for vehicles to choose: turn left, turn right and go straight, as shown in Fig. 1.

We will not consider the turn right direction because it does not impact other directions. In order to make this problem easy to model, we will not consider the pedestrian crossing the road. It will be very easy to add an additional rule for a pedestrian under our proposed mechanism.

Therefore, this problem can be modeled as 8 queues for different paths, as shown in Table 1.

We assume that there are a random number of vehicles spreading on different queues at the beginning of a signal period. This is the initial state of the environment. The final state must be that all the vehicles in the initial state have crossed the intersection. The intersection signal control system is modeled as a leader agent to manage the actions of all vehicle agents around the intersection. Since the action libraries of vehicle agents include actions from A1 to A8, the leader agent can choose any one action or their reasonable combination to reach the final state.

If two of the actions from A1 to A8 are nonintervention, they are possible action combinations. We call these different combinations a signal phase. All possible combinations are shown in Table 2.

Therefore, the problem can be described as how to find the optimum sequence of action combinations to reach the final state. This is the main function of the intersection signal control agent.

For each of the discrete states from the initial state to the final state, the optimum policy will be independent of the previous state. The successor state will be deterministic after one action combination is done. Therefore, this problem can be modeled as a deterministic Markov decision process.

**Table 1** Basic action definition of different queues

Queue	Basic action symbol	Path
Que1	A1	① → ④
Que2	A2	① → ⑥
Que3	B1	⑤ → ⑧
Que4	B2	⑤ → ②
Que5	C1	③ → ⑥
Que6	C2	③ → ⑧
Que7	D1	⑦ → ②
Que8	D2	⑦ → ④

**Table 2** Action combination symbol

Phase	Action combination symbol	Component
Ph1	Ac1	A1 + A2
Ph2	Ac2	A1 + B1
Ph3	Ac3	B1 + B2
Ph4	Ac4	C1 + C2
Ph5	Ac5	C1 + D1
Ph6	Ac6	D1 + D2

### 2.3 Parameters of Learning Process

(1) *Cost function*

We suppose that the vehicle number is  $n$  at state  $s$ . After the selected action  $a$  completed, the current vehicle number will be  $n_1$ . The cost of this action depends on the waiting time  $t$ , and the remainder of vehicles  $n_1$ .

$$g(s, a) = n_1 \times (t + t_{transition}). \tag{8}$$

where  $t_{transition}$  equals one of the three numbers  $\{0, 1.5, 3\}$  shown in Table 3. The average time for each vehicle passing the crossroad is supposed to be 3 s.

**Table 3**  $t_{transition}$  of different phase transition

Phase transition type	Comment	$t_{transition}(s)$
No transition	Current phase is the same as the previous one	0
Half transition	$Ac1 \Leftrightarrow Ac2; Ac2 \Leftrightarrow Ac3;$ $Ac4 \Leftrightarrow Ac5; Ac5 \Leftrightarrow Ac6;$	1.5
Full transition	Phase transfer except half transition	3

(2) *Discount factor*

In the simulation we set the discount factor,  $\gamma = 0.8$ .

## 2.4 Simulation and Results

We wrote some MATLAB code to complete the simulation with the following configuration.

CPU: Intel Pentium 4 Processor 2.40 GHz,

Memory: 1047792 KB,

Operation System: Microsoft Windows XP Professional (SP3).

In order to show the advantage of our proposed mechanism, the traditional signal mechanism was introduced to create a comparative study. In the traditional mechanism, the signal phase transition is in a fixed sequence as shown by Ph1, Ph2, Ph3, Ph4, Ph5 and Ph6. However, our proposed method can determine the optimum phase sequence automatically based on the updated situation.

In the following, we will show the comparative result for three different periods  $T$  and different phase time interval  $t_{phase}$ .

In the above-mentioned tables,  $P_s$  is the simulation period series,  $NIV$  is the total number of vehicles at the initial state, Random Queues the number of vehicle queues that are randomly created,  $T_{IQ}$  is the time interval from the initial state to the final state for a Q learning method,  $T_{WQ}$  is the total waiting time for the Q learning method,  $T_{IT}$  is the time interval from the initial state to the final state for the traditional method,  $T_{IT} = 6 \times t_{phase}$ ,

$T_{WT}$  is the total waiting time for the traditional method,

$$P_{EI} = \frac{T_{IT} - T_{IQ}}{T_{IT}} \times 100 \% \quad (9)$$

Equation (9) determines the percent improvement in the traffic efficiency,

$$P_{WD} = \frac{T_{WT} - T_{WQ}}{T_{WT}} \times 100 \% \quad (10)$$

Equation (10) shows the percent decrease in total waiting time.

$OA$  is the optimum phase sequence from Q learning,  $TL$  is the running time of the Q learning program on the above mentioned computer.

### 2.5 Analysis of the Results

From Table 4, we find that all the running times of the Q learning program TL in every period are less than one second. This is short enough for the application of the intersection signal control system.

**Table 4** Simulation result when  $t_{phase} = 60$  s

$P_a$	$N_{IV}$	Random queues	$T_{IQ}$ (s)	$T_{WQ}$ (s)	$T_{WT}$ (s)	$P_{EI}$ (%)	$P_{WD}$ (%)	$O_A$	$T_L$ (s)
1	180	{20 20 40 20 20 20 20 20}	312	18900	24000	13.33	21.25	{4 5 6 1 2 3}	0.8438
2	175	{20 20 19 19 38 19 20 20}	303	17916	23280	15.83	23.04	{1 2 3 6 5 4}	0.9375
3	176	{20 20 18 18 40 20 20 20}	306	18048	23640	15.00	23.65	{1 2 3 6 5 4}	0.4375
4	202	{17 17 36 18 36 18 40 20}	345	28479	30600	4.17	6.93	{1 2 3 6 5 4}	0.8438
5	183	{38 19 16 16 17 17 40 20}	345	21939	26880	4.17	18.38	{3 2 1 4 5 6}	0.8906
6	189	{19 19 36 18 20 20 38 19}	351	23418	28380	2.50	17.48	{1 2 3 4 5 6}	0.9063
7	157	{14 14 16 16 38 19 20 20}	276	14331	22740	23.33	36.98	{3 2 1 6 5 4}	0.8750
8	174	{38 19 26 13 20 20 19 19}	282	18852	22020	21.67	14.39	{4 5 6 1 2 3}	0.9063
9	123	{30 15 14 14 13 13 12 12}	219	8916	14280	39.17	37.35	{4 5 6 3 2 1}	0.8906
10	133	{15 15 17 17 30 15 12 12}	234	10413	16440	35.00	36.66	{3 2 1 6 5 4}	0.9219
11	130	{11 11 34 17 22 11 12 12}	249	11007	16140	30.83	31.80	{6 5 4 1 2 3}	0.8594
12	152	{22 11 15 15 16 16 38 19}	285	14904	24480	20.83	39.12	{3 2 1 4 5 6}	0.8750
13	171	{36 18 32 16 24 12 22 11}	273	19292	19500	24.17	1.07	{4 5 6 1 2 3}	0.9531
14	183	{26 13 36 18 26 13 34 17}	300	22265	25560	16.67	12.89	{6 5 4 3 2 1}	0.9375
15	120	{32 16 20 10 20 10 6 6}	216	10221	11760	40.00	13.09	{6 5 4 1 2 3}	0.8750
16	128	{19 19 20 10 15 15 15 15}	219	9768	15900	39.17	38.57	{4 5 6 1 2 3}	0.8906
17	112	{14 14 9 9 28 14 16 8}	189	7905	14220	47.50	44.41	{1 2 3 4 5 6}	0.8906
18	100	{16 8 8 8 34 17 6 3}	195	6480	11760	45.83	44.90	{3 2 1 6 5 4}	0.9375
19	128	{16 16 32 16 16 8 16 8}	228	9960	15120	36.67	34.13	{4 5 6 1 2 3}	0.8906

(continued)

**Table 4** (continued)

$P_a$	$N_{IV}$	Random queues	$T_{IQ}$ (s)	$T_{WQ}$ (s)	$T_{WT}$ (s)	$P_{EI}$ (%)	$P_{WD}$ (%)	$O_A$	$T_L$ (s)
20	128	{15 15 30 15 147 16 16}	237	10431	16620	34.17	37.24	{6 5 4 1 2 3}	0.8906
21	108	{16 8 36 18 14 14 1 1}	189	6906	10860	47.50	36.41	{4 5 6 3 2 1}	0.8750
22	81	{5 5 24 12 18 9 4 4}	165	4365	10140	54.17	56.95	{6 5 4 1 2 3}	0.4844
23	109	{14 14 18 9 30 15 6 3}	207	7740	11400	42.50	32.11	{1 2 3 6 5 4}	0.4375
24	144	{9 9 16 16 40 20 17 17}	258	11772	21660	28.33	45.65	{3 2 1 6 5 4}	0.5156
25	77	{5 5 30 15 8 4 5 5}	156	3501	9060	56.67	61.36	{6 5 4 1 2 3}	0.4219

At the same time, the percent traffic efficiency improvement PEI, is located in [4.17 % 47.5 %], the percent total waiting time decrease PWD is located in [1.07 % 56.95 %]. The average percents of PEI are 32.2 % and the average percents of PWD are 37.5 %.

### 3 Conclusion

A new traffic control based mechanism based on a combination of machine learning and multiagent modeling methods is proposed for future intelligent transportation systems. The control systems, the vehicles, and some necessary roadside sensors are all modeled as intelligent agents in the proposed systems, therefore the ITS system will be a multiagent system. It is possible to improve the traffic control efficiency by some artificial intelligence algorithm.

The control method for an isolated intersection was studied specifically. The intersection signal was first modeled according to the proposed mechanism then a new algorithm based on reinforcement learning, especially Q-learning, was proposed and studied in detail. A simulation for such an intersection system was finally carried out and a comparative study with the traditional intersectional signal method was done.

Simulation results showed that the proposed intersection control mechanism can improve traffic efficiency by more than 30 % over the traditional method and simultaneously bring the drivers some benefit by decreasing the waiting time by more than 30 %. This proves that the proposed traffic control mechanism is applicable in the near future.

**Acknowledgments** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-038978) and (No. 2012-0002434).

## References

1. <http://www.ewh.ieee.org/tc/its/>
2. Choy MC, Srinivasan D, Cheu RL (2003) Cooperative, hybrid agent architecture for real-time traffic signal control. *IEEE Trans Syst Man Cybern Part A Syst Hum* 33(5):597–607
3. Liu G, Zhai R, Pei Y (2007) A calculating method of intersection delay under signal control. In: Proceedings of the 2007 IEEE intelligent transportation systems conference, Seattle, pp 1114–1119
4. Bao W, Chen Q, Xu X (2006) An adaptive traffic signal timing scheme for bus priority at isolated intersection. In: Proceedings of the 6th world congress on intelligent control and automation, Dalian, pp 8712–8716
5. Srinivasan D, Choy MC (2006) Cooperative multi-agent system for coordinated traffic signal control. *IEE Proc Intell Transp Syst* 153(1):41–50
6. Lee JH, Lee-Kwang H (1999) Distributed and cooperative fuzzy controllers for traffic intersections group. *IEEE Trans Syst Man Cybern C Appl Rev* 29:263–271
7. Mitchell TM (1997) *Machine learning*. McGraw-Hill, New York. ISBN: 0070428077
8. D'Ambrogio A et al (2008) Simulation model building of traffic intersections. *Simul Model Pract Theory*



# Understanding and Extending AUTOSAR BSW for Custom Functionality Implementation

Taeho Kim, Ji Chan Maeng, Hyunmin Yoon and Minsoo Ryu

**Abstract** AUTOSAR (Automotive Open System Architecture) is a de facto standard for automotive software development. It addresses crucial topics such as software architecture, application interfaces and development methodology, thereby providing a basic infrastructure for software development. However, the current AUTOSAR standard is too complex to learn and has significant dependence upon tool chains. As a result, it is very difficult to implement custom functionality in BSW (Basic Software) without special support from tool vendors. In this paper, we present how custom functionality can be implemented within AUTOSAR BSW obviating the need for tool vendor's support. We first examine the internal structure and function of AUTOSAR software stack with an emphasis on the interfaces and execution of BSW modules. We then describe how a new BSW functionality can be incorporated into AUTOSAR BSW. Our approach is illustrated through a simple BSW module implementation with EB tresos AutoCore and Infineon TriCore TC1797.

**Keywords** AUTOSAR · Automotive · Software architecture · BSW (basic software) · Custom functionality

---

T. Kim · J. C. Maeng · H. Yoon  
Department of EECS, Hanyang University, Ansan, Korea  
e-mail: thkim@rtcc.hanyang.ac.kr

J. C. Maeng  
e-mail: jcmaeng@rtcc.hanyang.ac.kr

H. Yoon  
e-mail: hmyoon@rtcc.hanyang.ac.kr

M. Ryu (✉)  
Department of CSE, Hanyang University, Ansan, Korea  
e-mail: msryu@hanyang.ac.kr

## 1 Introduction

AUTOSAR (Automotive Open System Architecture) is receiving wide attention as the best means of automotive software development. Since electronics and software are becoming increasingly important in vehicle technology, the vehicle industry faces new challenges for software development such as standardization of system-level functionality, component-based software development and integration, model-driven development, software reuse and maintainability. To overcome these challenges, AUTOSAR has been designed to address crucial topics such as software architecture, application interfaces and development methodology, thereby providing a basic infrastructure for automotive software development. Currently, many OEMs and suppliers are accepting AUTOSAR as a de facto standard for automotive software development.

AUTOSAR has many features from a software engineering point of view. One of them is to provide complete abstraction between applications and system software so that developers can focus on the design and implementation of application components. Such abstraction is made possible by AUTOSAR tools such as Vector's DaVinci and Elektrobit's EB tresos. These tools can be used to configure system-level functionality and automatically generate source code for Run-Time Environment (RTE) and Basic Software (BSW).

Although the abstraction supported by AUTOSAR is useful for application development, it also severely limits the accessibility of system software to developers. During automotive system development, engineers often need direct access to system software for many reasons such as system-level tracing and debugging, system-wide performance profiling and optimization. For instance, consider that a developer needs to analyze and optimize system-level performance. To do so, some custom functionality should be implemented within AUTOSAR BSW. However, implementing custom functionality in BSW is not easy without special support from the tool vendor since the current AUTOSAR standard is too complex to learn and has significant dependence upon tool chains.

In this paper, we describe how custom functionality can be implemented within AUTOSAR BSW. We first introduce the internal structure and behavior of AUTOSAR software stack with an emphasis on the interfaces and execution of BSW modules. We then present how a new BSW functionality can be incorporated into AUTOSAR BSW. Our approach is illustrated through a simple BSW module implementation with EB tresos AutoCore and Infineon TriCore TC1797. We implemented a BSW module that is able to blink LED while handling CAN (controller area network) interface interrupts. We could successfully incorporate this BSW module into EB tresos AutoCore without any technical support from the tool vendor.

## 2 Implementing Custom Functionality Within AUTOSAR BSW

The AUTOSAR software stack consists three layers, application, Run-Time Environment (RTE) and Basic Software (BSW). The application layer can be modeled as a set of interconnected software components (SW-C) that encapsulate application-specific functionality. Software components have well-defined interfaces through which components can communicate in various ways such as sender-receiver or client-server fashions. Software components also have runnables that can be invoked by an OS task or other software component. The notion of runnable is very similar to the Runnable interface in Java, which is intended to be executed by a separate thread of control.

The RTE layer plays a role of middleware between application and BSW layers. Its main responsibility is to provide communication services to application software, communications between software components and communications between software components and basic software modules.

The BSW layer is responsible for managing hardware resources and providing common services for application software. BSW is further divided into four parts, Microcontroller Abstraction Layer (MCAL), ECU Abstraction Layer (EAL), Services Layer and Complex Device Drivers (CDD). The Microcontroller Abstraction Layer is the lowest software layer in BSW and provides device drivers for microcontrollers and internal peripherals. The ECU Abstraction Layer provides device drivers for external devices. The Services Layer offers many functions including OS functionality, network management, memory management and diagnostic services. The Complex Drivers Layer provides special purpose functionality that is not specified within the AUTOSAR standard (Fig. 1).

In order to understand and extend BSW functionality, it is important to understand AUTOSAR OS and BSW scheduler in the Services Layer. The main

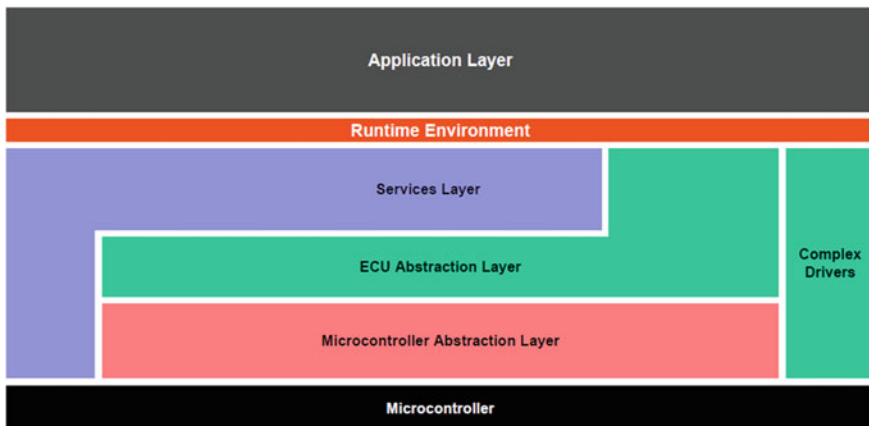
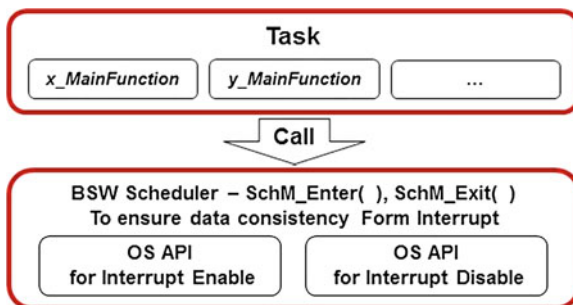


Fig. 1 Software layers in AUTOSAR

**Fig. 2** Invocation of BSW scheduler APIs to protect the critical section



role of AUTOSAR OS is the management and execution of tasks and ISRs. AUTOSAR OS is responsible for generating, executing and scheduling tasks at runtime. It is also responsible for invoking interrupt handlers.

BSW Scheduler, another important component in the Services Layer, provides two services. First, it executes main processing functions of BSWs in the form of tasks. Second, it provides mutual exclusion APIs needed for protecting critical sections in BSW modules. Specifically, BSW Scheduler provides “SchM\_Enter()” and “SchM\_Exit()” APIs that are used to disable and enable interrupts for protecting critical sections accessible by other interrupts (Fig. 2).

Since BSW modules are mostly used for hardware-related functionality, we need to know how their main processing functions and interrupts handlers are specified and executed when we create a custom BSW module. First, when we create a new BSW module, we must follow the AUTOSAR naming convention. Every main processing function in BSW module must be named as shown in the following code.

```
<Module name>_MainFunction_<Extension name>( );
```

“<Extension name>” is used to distinguish between main processing functions in a BSW module. Interrupt handlers should also be named as shown in the following code. Generally, “<Service name>” is used to express a service of an interrupt handler.

```
<Module name>_<Service name>( );
```

Second, interrupt handlers should be specified using AUTOSAR ISR macros. Each ISR can be associated with a specific interrupt type and appropriate interrupt handling code.

```
ISR ( /* interrupt type */ ) {
    /* Driver Module User Code to handle ISR */
}
```

Third, we may use `SchM_Enter(Instance, Exclusive Area)` and `SchM_Exit(Instance, Exclusive Area)` to avoid any possible data races in main processing functions.

```
SchM_Enter_<Module name>(Instance, ExclusiveArea);  
SchM_Exit_<Module name>(Instance, ExclusiveArea);
```

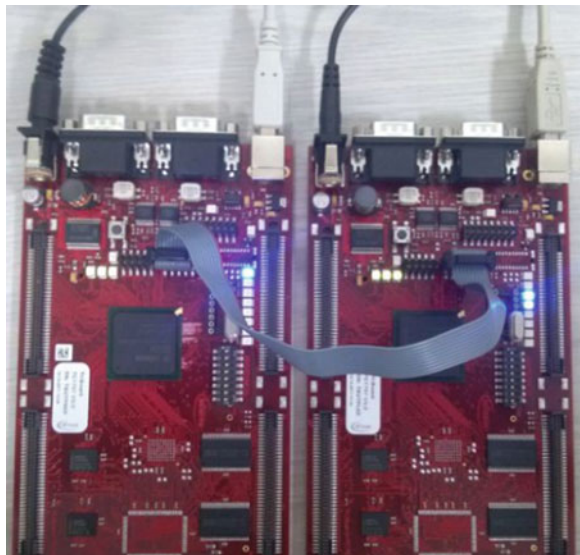
In above code, “Instance” is a main processing function in a BSW module “<Module name>”. In BSW modules, there are several types of critical sections such as initialization, writing and reading shared data. The second parameter, “Exclusive Area”, is used to specify the type of critical sections and invoke an appropriate API.

### 3 Case Study

We implemented a simple custom BSW module within EB tresos AutoCore on Infineon TriCore TC1797 hardware. It is able to blink LEDs while handling controller area network (CAN) interface interrupts. We could successfully incorporate this BSW module into EB tresos AutoCore without any technical support from AUTOSAR tool vendors (Fig. 3).

Specifically, we first implemented eight main processing functions for LED control as shown in the following code.

**Fig. 3** LED controller and CAN controller on Infineon TriCore TC1797



```

LedControl_MainFunction_1();
LedControl_MainFunction_2();
...
LedControl_MainFunction_7();
LedControl_MainFunction_8();

```

We then mapped the main processing functions onto a single task whose activation period is one second. Thus, all LED sequentially blinks every one second.

```

TASK (Ex_LedControl_Sequence_1second) {
    LedControl_MainFunction_1();
    LedControl_MainFunction_2();
    ...
    LedControl_MainFunction_7();
    LedControl_MainFunction_8();
}

```

We also implemented CAN message reception functionality in our BSW module by using an interrupt handler. We associated an ISR with the CAN\_RECEIVE interrupt. Whenever a CAN\_RECEIVE interrupt occurs, the ISR is invoked to blink LED.

```

ISR (CAN_RECEIVE) {
    Can_IsrReceiveHandler();
}

```

## 4 Conclusion

In this paper, we described how custom functionality can be implemented within AUTOSAR BSW. We first introduced the internal structure and behavior of AUTOSAR software stack with an emphasis on the interfaces and execution of BSW modules. We then presented how a new BSW functionality can be incorporated into AUTOSAR BSW. Our approach was illustrated through a simple BSW module implementation with EB tresos AutoCore and Infineon TriCore TC1797 hardware.

**Acknowledgment** This work was supported partly by Mid-career Researcher Program through National Research Foundation (NRF) grant NRF-2011-0015997 funded by the Ministry of Education, Science and Technology (MEST), partly by the IT R&D Program of MKE/KEIT [10035708, “The Development of Cyber-Physical Systems (CPS) Core Technologies for High Confidential Autonomic Control Software”], partly by Seoul Creative Human Development

Program (HM120006), and partly by the The Ministry of Knowledge Economy (MKE), Korea, under the Convergence Information Technology Research Center (CITRC) support program (NIPA-2013-H0401-13-1009) supervised by the National IT Industry Promotion Agency (NIPA).

## References

1. Daehyun K, Gwang-Min P, Seonghun L, Wooyoung J (2008) AUTOSAR migration from existing automotive software, In: International conference on control, automation and systems, pp 558–562
2. Wang D, Zheng J, Zhao G, Bo H, Liu S (2010) Survey of the AUTOSAR complex drivers in the field of automotive electronics. In: Intelligent computation technology and automation, pp 662–664
3. Diekhoff D (2010) AUTOSAR basic software for complex control units. In: Design automation and test in europe conference and exhibition, pp 263–266

# A Hybrid Intelligent Control Method in Application of Battery Management System

T. T. Ngoc Nguyen and Franklin Bien

**Abstract** This paper presents a hybrid adaptive neuro-fuzzy algorithm in application of battery management system. The proposed system employed the Cuk converter as equalizing circuit, and utilized a hybrid adaptive neuro-fuzzy as control method for the equalizing current. The proposed system has ability for tracking dynamic reactions on battery packs, due to taking advantages of adaptability and learning ability of adaptive neuro-fuzzy algorithm. The current output generated from learning process drives Pulse-Width-Modulation (PWM) signals. This current output is observed and collected for next coming learning process. The feedback line is provided for current output observation. The results demonstrate the proposed scheme has the ability to learn previous stages. Therefore, the proposed system has adaptability to deal with changing of working conditions.

**Keywords** Fuzzy logic · Adaptive neuro-fuzzy system · dc-dc converter · Battery equalization

## 1 Introduction

Nowadays, lithium-ion batteries are more and more replacing the traditional acid batteries. However, the lithium-ion battery cell provides very low voltage around 4–4.5 V that is not enough to supply in electronic vehicles. Therefore, the requirement of connecting lithium-ion battery cells is on demand. Normally, in

---

T. T. Ngoc Nguyen (✉) · F. Bien  
School of Electrical and Computer Engineering, Ulsan National Institute  
of Science and Technology, Ulsan, South Korea  
e-mail: ngocntt@unist.ac.kr

F. Bien  
e-mail: bien@unist.ac.kr



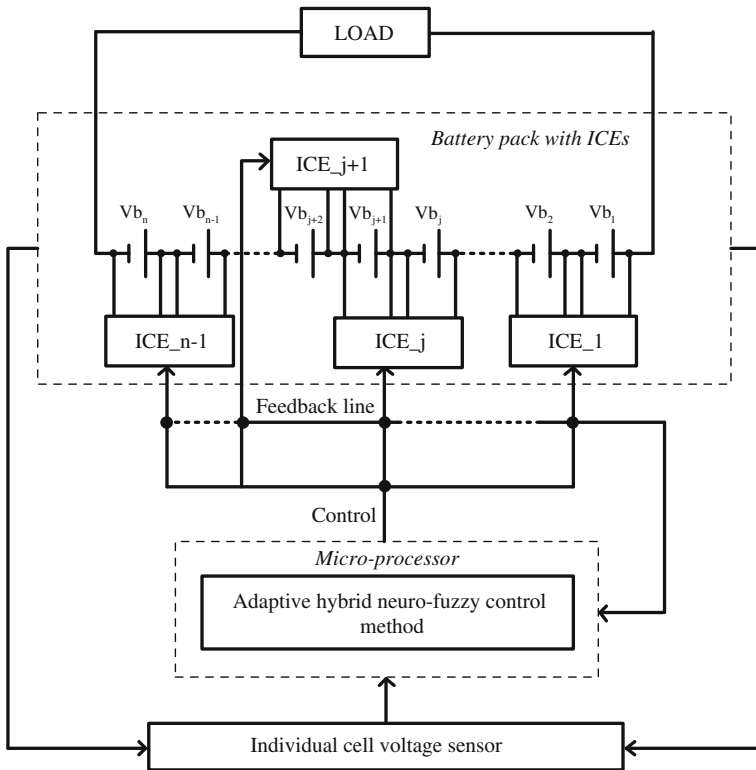
very high voltage systems, the hundreds of battery cells are connected together that are called serially-connected battery systems. The intelligent controllers are required for these systems due to imbalance between cells when they are connected in series. The differences in working conditions between cells cause severe problems such as high risk of explosion, depth discharging or over-charging. Therefore, balancing voltages between the battery cells is essential for cell protection and prolongation of the battery life. A number of research activities focusing on control method to equalizing battery voltages have been published. In recent times, integrated individual cell equalizer (ICE) is under development regarding equalization scheme [1–3]. The equalization control restricts battery cell's operation range to prevent over-discharging and over-charging, which cause subsequence damage to the active materials in the battery such as electrodes. During charging state, the battery voltage should be limited to this value. Otherwise, the internal pressure and temperature would produce a high risk of explosion.

On the other hand, the copper in the electrolyte are easy to be dissolved during over-discharging that shortens battery life. Normally, the low voltage threshold of a lithium-ion battery cell is in the range of 2.6–2.8 V. Obtaining maximum usable capacity from battery cells is one of purposes of battery controller, in order to have high energy efficiency. In the battery system using individual cell equalizers (ICE), the controller control equalizing current between two neighboring cells in an ICE. The imbalance problems caused by changes in the internal impedance and cell capacity are exacerbated while the battery is working. Meanwhile, most of previous control methods based on imbalanced state-of-charge (SOC) between cells, and gradients of ambient temperature of the battery pack. These characteristics are non-linear and difficult to model exactly due to objective and subjective parameters, such as temperature, workload, etc. As a result, these conventional methods are short of adaptability to the dynamic system.

The proposed scheme of equalization scheme employed adaptive neuro-fuzzy algorithm. The cell voltages are controlled by the driving pulse-width modulation (PWM) signals, based on the equalization algorithm scheme in the battery unit controller. This adaptive neuro-fuzzy algorithm provides off-line training, system tracking and estimation of the proposed BMS scheme. The advantage of the proposed system is inheriting adaptability of back propagation neuro network for training and estimating that are suitable with dynamic system like serially connected battery cells.

## 2 The Proposed Hybrid Adaptive Neuro-Fuzzy Control Method

Figure 1 show the proposed system, the framework of the proposed battery equalization system is based on individual cell equalizer (ICE) that balances each pair of the neighboring cells. In the proposed system, a complete cell voltage

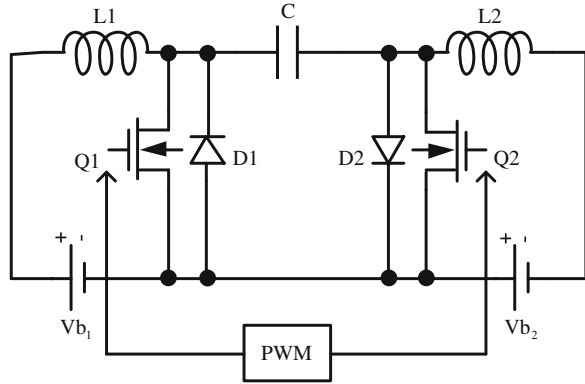


**Fig. 1** The proposed hybrid adaptive neuro-fuzzy control scheme

equalizer is performed using a bi-directional dc-dc converter modified from the Cuk converter. In conventional schemes, it is difficult to handle the battery cell model for describing the equalizing characteristic of the lithium-ion battery strings due to electrochemical reactions and ambient temperature. Fuzzy logic control has advantage in modeling the non-linear characteristic of equalizing current during charging and discharging. However, the limitation of fuzzy system is that it cannot automatically acquire the rules to make output decisions. In addition, it is challenging to model the membership functions in small change of voltage offsets with conventional fuzzy-based control methods that cause degradation accuracy of the system employing such control methods. In this work, combining neuron network in fuzzy logic control is used to tune membership functions of fuzzy system, which enables adaptability and learning ability. In the proposed system, a feedback line is provided to allow updating and tracking the changes of battery strings.

The equalizing circuit works on a cell-to-cell basis, therefore only one ICE<sub>1</sub> is analyzed for the proposed equalization as illustrated in Fig. 2. The voltage battery of cell 1 and cell 2 are represented by V<sub>b<sub>1</sub></sub> and V<sub>b<sub>2</sub></sub>, respectively. Control signal from PWM turns on or off MOSFET Q1 or Q2 with control frequency of  $f_s$  or

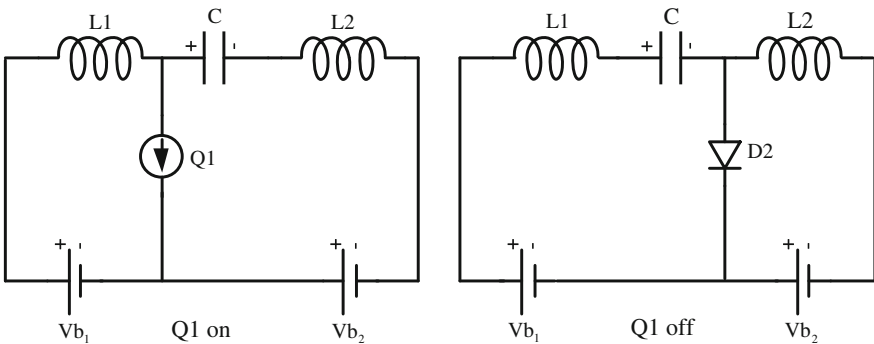
**Fig. 2** Individual cell equalization circuit



period of  $T_s = 1/f_s$ . Assuming  $V_{b1} > V_{b2}$ , battery cell 1 will be discharged and battery cell 2 will be charged. At the initial state, the voltage of the capacitor  $C$  equals to  $(V_{b1} + V_{b2})$ . During the first period ( $T_1$ ) while  $Q_1$  is turned on, the energy stored in the capacitor transfers to cell 2 and the inductor  $L_1$  stores energy transferred from  $V_{b1}$  of cell 1. Likewise, the inductor  $L_2$  also stores energy. In the remaining period ( $T_2$ ) of the PWM control signal,  $Q_1$  is turned off, and  $D_2$  is forced to turn on. Hence, the capacitor is charged by energy stored in  $L_1$ , and the energy stored in  $L_2$  continues to charge cell 2. The equivalent circuit for the equalization scheme is shown in Fig. 3. The energy transformation is continuous until the circuit reaches stable state. The quantitative analysis for the equivalent circuit for in Fig. 3 can be expressed with the following equations:

When  $Q_1$  is turned on: ( $0 \leq t < T_1$ )

$$\begin{cases} V_{b1} = L_1 \frac{di_{L1}}{dt} \\ V_{b2} = -L_2 \frac{di_{L2}}{dt} + \frac{1}{C} \int i_{L2} dt \end{cases} \quad (1)$$



**Fig. 3** Cell equalization circuit analysis

When Q1 is turned off: ( $T_1 < t \leq T_s$ )

$$\begin{cases} Vb_1 = L_1 \frac{di_{L1}}{dt} + \frac{1}{C} \int i_{L1} dt \\ Vb_2 = -L_2 \frac{di_{L2}}{dt} \end{cases} \quad (2)$$

Combining with the initial conditions and limitations, the average of inductor currents through L1 and L2 is computed as followed:

$$\begin{cases} I_{L1} = \frac{1}{2} T_s \left( \frac{Vb_1}{L_1} T_1^2 + \frac{V_C - Vb_1}{L_1} T_2^2 \right) \\ I_{L2} = \frac{1}{2} T_s \left( \frac{Vb_2}{L_2} T_2^2 + \frac{V_C - Vb_2}{L_2} T_1^2 \right) \end{cases} \quad (T_2 = T_s - T_1) \quad (3)$$

For simplicity, the value of L1 and L2 are chosen equally so that equal currents flow through these inductors. The output current  $I_{urm}$  from the adaptive neuron-fuzzy controller is used to switch the period of the PWM driving signal of the proposed scheme.

In the proposed control method, the battery cell voltage Vb and derivation of neighboring cells De are the inputs of the controller. The architecture of hybrid adaptive neuro-fuzzy algorithm consists of 5 layers.

- Layer 1 The first layer generates the membership functions of inputs {Vb, De}.

$$A_i(V_b) = \left[ 1 + \left( \frac{V_b - a_{i1}}{b_{i1}} \right)^2 \right]^{-1} \quad (4)$$

$$B_i(D_e) = \left[ 1 + \left( \frac{D_e - a_{i2}}{b_{i2}} \right)^2 \right]^{-1}$$

where  $\{a_i, b_i\}$  is the *parameter set* while A, B are linguistic labels.

- Layer 2: each node computes the firing strength of the associated rules. The outputs of these neurons are labeled by  $T$  because of choosing  $T$ -Norm for modeling, and are calculated as following:

$$\alpha_1 = A_1(V_b) \times B_1(D_e) = A_1(V_b) \wedge B_1(D_e) \quad (5)$$

$$\alpha_2 = A_2(V_b) \times B_2(D_e) = A_2(V_b) \wedge B_2(D_e)$$

The node in this layer are called rule nodes

- Layer 3: every node in this layer is labeled by  $N$  to indicate the normalization of firing levels.

$$\beta_1 = \frac{\alpha_1}{\alpha_1 + \alpha_2}, \beta_2 = \frac{\alpha_2}{\alpha_1 + \alpha_2} \tag{6}$$

- Layer 4: the outputs of these nodes at this layer are the product of the normalized firing level and the individual rule output of associated rule.

$$\beta_i s_i = \beta_i (p_i V_b + q_i D_e) \tag{7}$$

where  $\{p_i, q_i\}$  are the set of *consequent parameters* associated the *i*-rule.

- Layer 5: the single output of this layer computes the overall system output as the sum of all the incoming signals. The overall output is current output to drive PWM signals.

$$I_{turn} = \sum_i \beta_i s_i \tag{8}$$

The above layered structure is viewed as neural network. This structure can adapt its antecedent and consequent parameters to improve performance of the system by observing outputs. By employing these advantages of adaptive neuron-fuzzy architecture, the proposed battery equalization system has ability of learning and adaptation to deal with dynamic changes.

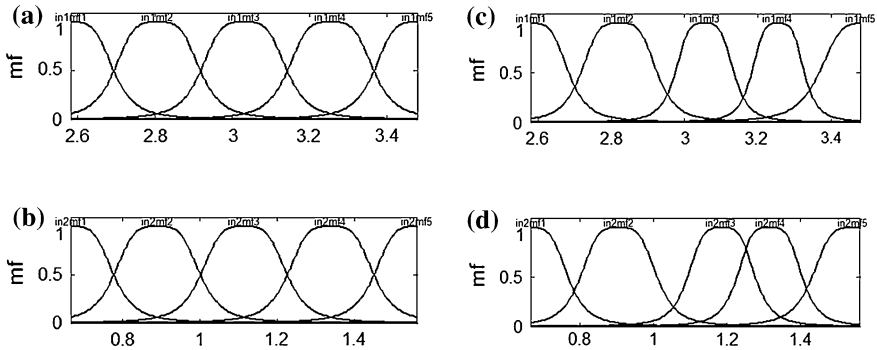
A hybrid learning rule based on least-square estimator (LSE) and gradient-descent method (GDM) is widely used [7]. A partition *p* of *i*-input is defined by giving an offset value  $a_i^p$  and distance  $b_i^p$ . Therefore, the total number of antecedent parameters to be trained with  $NP_i$  antecedents per input is  $2NP_i$ . Concerning the consequent parameters, the number of crisp consequents to be adapted is equal to the number of rules, where the maximum is  $m = \prod_{i=1}^n NP_i$  in which all possible rules are considered. The training data pair  $\{(V_b^k, D_e^k, I_{turn}^k), k \in (1, 2, \dots, K)\}$  is used in this learning process. After the initial antecedent parameters and consequence parameters are identified, GDM will optimize the bell-shaped membership functions by computing the error function:

$$E = \frac{1}{2K} \sum_{k=1}^K (I_{turn}^k - I_{turn}^{k'})^2 \tag{9}$$

where  $I_{turn}^k$  is the actual output of the network for the *k*-training data, and  $I_{turn}^{k'}$  is the desired output.

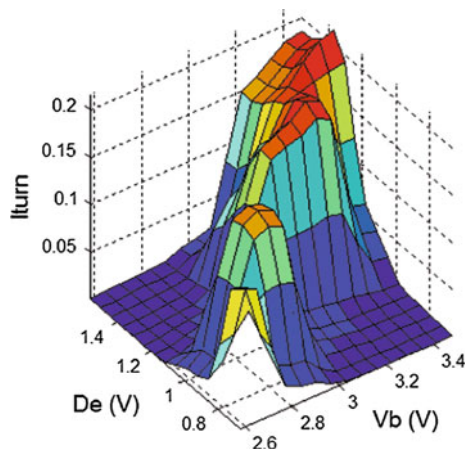
### 3 Experimental Results

The following part represents the experimental results of the proposed system. The initial training data set is composed of 1,000 pairs of input set  $\{V_b, D_e\}$  and current output  $I_{turn}$ . The membership functions for each input are determined after the training process as can be seen from Fig. 4. As illustrated in Fig. 4, after training



**Fig. 4** Membership functions of input voltages before and after learning process with initial training datasets. **a**  $V_b$  (V) in initialization. **b**  $De$  (V) in initialization. **c**  $V_b$  (V) after learning. **d**  $De$  (V) after learning

**Fig. 5** Control current output  $I_{turn}$  model from learning process according to input voltages  $De$  and  $V_b$



process, the membership functions of voltage inputs are changed by updating adaptive parameters. The output current  $I_{turn}$  is model based on adaptive membership functions of inputs in Fig. 5. In the experiments, the battery cell is the RHV-320064 model that has nominal voltage of 3.2 V. The circuit parameters are  $L1 = L2 = 200 \mu\text{H}$ ,  $C = 450 \mu\text{F}$ . The switching frequency for initialization is selected at 50 kHz, with duty cycle of 50 %.

### 4 Conclusion

The proposed system employing a hybrid adaptive neuro-fuzzy algorithm is demonstrated successfully. An intelligent learning and adaptable algorithm of hybrid adaptive neuro-fuzzy is employed for tuning driving current to control

PWM. Based on the proposed control method, the battery system has ability of learning previous stages and enabled adaptability to deal with dynamic reactions in serially connected battery cells. The experimental result demonstrated the advantages of the proposed system.

**Acknowledgment** This work was supported by the development program of local science park funded by the ULSAN Metropolitan City and the MEST (Ministry of Education, Science and Technology).

## References

1. Park HS, Kim CE, Kim CH, Moon GW, Lee JH (2009) A modularized charge equalizer for an HEV lithium-ion battery string. *IEEE Trans Ind Electron* 56(2):1464–1476
2. Lindemark B (1991) Individual cell voltage equalizers (ICE) for reliable battery performance. In: *Proceedings of the 13th Annual International Telecommunications Energy Conference, Kyotopp*, pp 196–201
3. Moore SW, Schneider PJ (2001) A review of cell equalization methods for lithium-ion and lithium polymer battery systems. In: *Proceedings of the SAE 2001 world congress, Detroit*
4. Lee YS, Cheng GT (2006) Quasi-resonant zero-current-switching bidirectional converter for battery equalization applications. *IEEE Trans Power Electron* 21(5):1213–1224
5. Lee YS, Jao CW (2003) Fuzzy controlled lithium-ion battery equalization with state-of-charge estimator. In: *IEEE International Conference on Systems Man Cybernetics, vol 5*, pp 4431–4438
6. Campo L, Echanobe J, Bosque G, Tarela JM (2008) Efficient Hardware/Software implementation of an adaptive Neuro-Fuzzy system. *IEEE Trans Fuzzy Syst* 16(3):761–778
7. Jang J (1993) ANFIS: adaptive-network-based Fuzzy inference system. *IEEE Trans Syst Man Cybern* 23(3):665–685

# Interpretation and Modeling of Change Patterns of Concentration Based on EEG Signals

JungEun Lim, Soon-Yong Chun and BoHyeok Seo

**Abstract** It is very important to understand the brain's biological cognition data processing mechanism for human cognitive ability and concentration enhancement. Based on biological data processing area and information flow, concentration indicators were defined to interpret the brain data processing mechanism in concentration by engineering, and cognitive concentration model based on this was proposed. The cognitive concentration model is the change of concentration patterns shown by EEG signal. The value of cognitive concentration model was verified with the EEG signals acquired from Subjects solving mathematical questions with different difficulties.

**Keywords** Brain information processing · Attention · Concentration · EEG · Cognitive concentration model

## 1 Introduction

Body organ engaging in 'thought', like cognition of things and thinking, is the brain. It is known that this brain constitutes the nervous system in which cells called neurons are entangled complicatedly [1]. Studies of brain science on human cognition and thinking are actively carried out. Among studies using this, various

---

J. Lim · B. Seo

School of Electrical Engineering and Computer Science, The Graduate School,  
Kyungpook National University, Daegu, Korea  
e-mail: euny1122@knu.ac.kr

B. Seo

e-mail: bhsuh@knu.ac.kr

S.-Y. Chun (✉)

Dongyang University, Yeongju, Korea  
e-mail: control@dyu.ac.kr



studies were attempted to increase human cognition processing ability related to memory, attention, and emotion, etc. [2–5]. Concentration is defined as examination of some stimulus or keeping attention to a stimulus selected. Attention is defined as concentration or focus of consciousness kept in mind vividly after selecting one among objects or thoughts [6]. In other words, attention can be said to be an ability to maintain attention to selected stimulus in a limited time. Thus, in a prior step to increase human cognition processing abilities, this paper defined general cognitive abilities, the maximum attention, and the duration of concentration. We set up a cognitive concentration model in order to interpret and model the change patterns of concentration. Also, it carried out a test to prove these using hard, normal and easy level questions in the field of mathematics.

## 2 Biological Cognition Mechanism

When you selectively focus on certain information, the frontal lobe, thalamus, and amygdala, etc. are involved [7–9]. The frontal lobe actions allow us to concentrate and process an assignment using attention, and among them, thanks to prefrontal cortex. If the actions of prefrontal cortex are not active, people cannot concentrate one thing for a long time paying attention and mind is easily dispersed. As a result of the brain image photography, the left brain prefrontal cortex of a person who is solving a hard question by concentrating attention was strongly activated, and this was activated when he or she processed information concentrating on anything other than mathematical question [8]. In a child with a problem in attention, the prefrontal cortex of the frontal lobe acts weakly while limbic system acts very actively, the limbic system decreases ability to pay attention and control emotions and increases instinctive and primary emotional roles. The amygdala, one of the limbic system can receive various data through thalamus before the cerebral cortex, and as emotional information flows directly into the amygdala, before the cerebral cortex analyzes and judges information, already one gets favor or displeasure for the stimulus [9].

In addition, in the process of filtering unnecessary stimuli other than important information, reticular formation and thalamus act, and the reticular formation is located at the center of the brain stem, controls awakening level, so the cerebral cortex can pay attention or process information by the actions of the reticular formation. Norepinephrine, a neurotransmitter for keeping functions of the reticular formation rapidly increases when an organism gets excited or nervous and when it is abnormally low, the awakening function of the reticular formation is deteriorated.

Thalamus located in inner center of the frontal brain takes sense information entering in each part of the body, selects important things only and concentrates and conveys command from the cerebral cortex to muscles. Approximately 100 million stimuli per second comes to the central nerves and among them, about 100 important stimuli are selected for the cerebral cortex to make a right decisions [8].

The reticular formation, thalamus and neurons in the multiple sensory systems of the brain serve as filters for numerous stimuli coming into the sensory receptors, and allow men to pay attention to appropriate stimuli. If the roles of the amygdala in the frontal lobe get stronger or the functions of the reticular formation decline by norepinephrine, with the problem in the thalamus’s ability of selecting information, there may be a problem in attention ability. Like this, attention consists of roles of various areas of the brain and neurotransmitters. Among existing studies on selective attention, Koch and Ullman in 1985 tried to show selective concentration using the concept of Saliency Map, and selecting Winner-Take-All through the most remarkable part and then make the part to concentrate. In the 1990s, Wolf and Cave insisted on Guided Search model of parallel process in pre-attention and sequential process in attention. This is gradually developed on the basis of visual attention model currently. However, it was concentrated on theories not developed to application system. Later in 1998, Itti made a four-step model combining the two methods [10]. Human cognition process can be viewed as a combination of bottom-up process of receiving external stimuli and top-down process of creating information inside the brain by active attention [11], and the systems of selective attention based on this biological cognition mechanism are proven as new cognition systems that combine human attention abilities with learning theory.

### 3 Propose of Cognitive Concentration Model

For an engineered model of selective attention, when there are data entered, each datum is processed after passing through a filter based on previously learned knowledge, which is estimated in a certain form, and the final conclusion is made according to the reliability of conclusions. This study aims to graft the concept of time in the engineer model of selected attention, and the inputting EEG power spectrum should be used for changing and providing model about the attention.

Figure 1 is the implementation of back-propagation algorithm for the single neuron as cognitive model. Once input, patterns are presented and attention is paid to one datum among them, the cost function  $E$  is defined Eq. (1). The distance between the desired value  $d_i$  of output neuron and the actual output (attended output) value  $y$  is minimized.

$$E = \frac{1}{2} \sum_i (d_i - y_i)^2 \tag{1}$$

The steepest descent gradient rule as cognitive weighting factor for each input datum  $x_i$  is expressed as the Eq. (2).

$$\frac{dw_{ij}}{dt} = -\mu \frac{\partial E_i}{\partial w_{ij}} \tag{2}$$

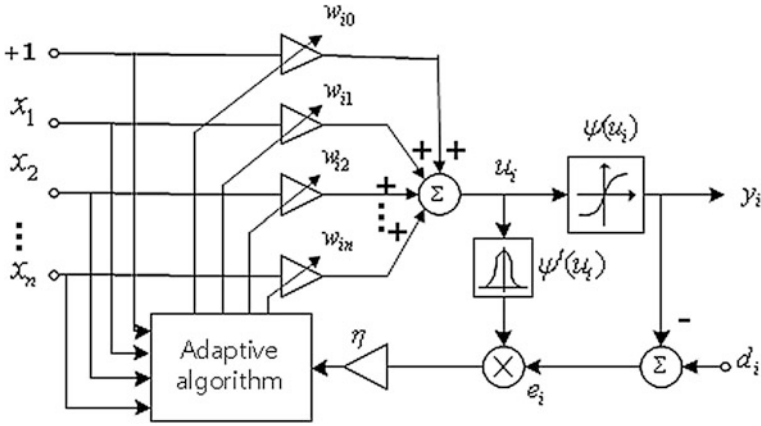


Fig. 1 Implementation of the single neuron as cognitive model

where  $\mu$  is a positive learning parameter determining the speed of convergence to targeted input datum  $x_i$ .

In this study, the cognitive ability in the daily life set as  $\mu$ , and it can be low or high according to age or experience, and can be different by people.

Pattern  $y$  with the greatest reliability to the kind of the entire pattern estimated as the outcome of cognition. The parameter  $\tau$  is the immersion constant and the cognitive concentration model is defined as the Eq. (3).  $y(t)$  means the concentration degree, and the bigger it is, the higher the concentration degree. But, it doesn't mean the relative value for ability of concentration. When a certain assignment is given, the maximum degree of concentration is defined as  $c_{max}$  the maximum immersion, and as of the point of time when an event takes place, the duration of concentration was defined as  $T$ . The input was used after measuring and processing EEG datum.

$$y(t) = C + (C - \mu)e^{-\tau t} \tag{3}$$

where  $c$  is immersion degree.

### 4 Experiment on the Change Patterns of Concentration

An experiment to verify the usefulness of the proposed cognitive concentration model was conducted as follows: the prefrontal lobes, Fp1 and Fp2 have important functions for cognition, thinking, and creativity, which play central roles of the brain functions related to learning behaviors [12]. Thus, Fp1 and Fp2 areas were measuring points of which the signals were measured using Biopac MP150 and EEG 100C. To interpret the change patterns of concentration, hard, normal and

easy mathematical questions were issued and the brain wave was measured while performing them step by step [13].

The collected EEG was classified to several band related to attention. The bands of the brain wave with high correlation to concentration are known as Beta wave, SMR wave, and high Beta wave. For the collected time domain data, using moving window analysis with Window size 2[s] and Overlap Ratio 40[%], data in each section were converted to frequency domain data and characteristics were examined. Power spectrum is found when time domain data was translated to frequency domain data. Discrete Fourier Transform (DFT) is as following Eq. (4):

$$H(f_n) = \sum_{k=0}^{N-1} h_k e^{-j2\pi n/N} = H_n \tag{4}$$

This equation is organized by Inverse Fast Fourier Transform (IFFT) as Eq. (5) and square the absolute value on both sides of this equation and then take  $\sum_{k=0}^{N-1}$  to sum to draw out Eq. (6).

$$h_k = \frac{1}{N} \sum_{n=0}^{N-1} H_n e^{-2\pi kn/N} \tag{5}$$

$$TotalPower \equiv \sum_{k=0}^{N-1} |h_k|^2 = \frac{1}{N} \sum_{n=0}^{N-1} |H_n|^2 \tag{6}$$

This means the total power value of the signal is the same in time domain or frequency domain, which is called Parseval theorem. One side power spectrum that satisfies this theorem is as Eq. (7).

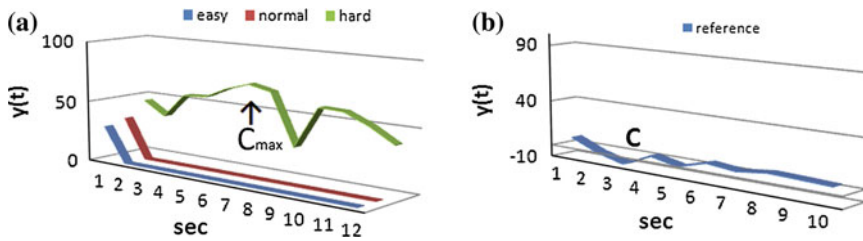
$$P(f_0) = P(0) = \frac{1}{N^2} |H_0|^2 \tag{7}$$

Using this, the brain wave by each was found and when the subjects perform hard, normal and easy assignments, changes of the brain wave by time were compared.

As elements with a big difference during concentration, beta wave and theta wave were selected to use in the verification of the cognitive concentration model.

## 5 EEG Interpretation and Cognitive Concentration Model Verification

To verify the defined model using the brain wave, the change patterns of concentration were interpreted with the difference between the selected theta and beta waves by comparing each power spectrum of the brain wave. Figure 2a shows change of  $y(t)$  by each difficulty. As the assignment got more difficult, the performing time and difficulty of the assignment got higher along with the value of  $y(t)$ . However, it decreased again as time passed. This can be interpreted that after a certain time, concentration fell. Figure 2b shows the value to find the basic brain



**Fig. 2** a Concentration by difficulty; b general cognitive ability

wave before performing assignments, which was defined as the mean after 10 time of execution. It had the value of  $|y(t) \leq 10|$ . Approximately it had a value in  $25 \leq y(t) \leq 35$  for easy questions,  $35 \leq y(t) \leq 45$  for normal ones and  $30 \leq y(t) \leq 70$  for hard ones. This is consistent with the result that they could solve assignments with a lower difficulty without great attention, and the higher the difficulty was, the higher the attention should get. Duration of immersion  $T$  appears around 10 s. This differs depending on the degree of individual learning.

## 6 Conclusions

This study defined general cognitive ability, maximum attention, and duration of concentration. As a prior step to increase human cognition processing ability, we defined cognitive concentration model to interpret and modeling change pattern of concentration, the ability to keep attention to stimulus selected for a limited time. In addition, using hard, normal and easy questions in mathematical area, the brain wave experiment was conducted to prove this. It measured the brain wave, converted it to frequency domain, examined characteristics by time and verified the usefulness of the cognitive concentration model.

As a result of this paper, using human cognitive concentration model to enhance attention will be a foothold to system development that can get a better result in emotional stability and enhancement of academic record through enhancing attentiveness.

## References

1. Crick F (1996) The astonishing hypothesis: the scientific search for the soul. *Contemp Psychol* 41(5):427–428
2. Robbins J, Wired for miracles (neurofeedback therapy). *Psychology Today*, May 1st
3. Anna W (1995) *High performance mind*. Tarcher Putnam, New York
4. Nak CL (1992) *Correlates of EEG hemispheric integration*. PhD Indiana University
5. Jung-Eun L, Bo-Hyeok S, Soon-Yong C (2012) Study on EEG feature extraction under LED color exposure to enhance the concentration. *Adv Eng Forum* 2–3:261–265

6. Serman MB (1977) Sensorimotor EEG operant conditioning and experimental and clinical effects. *Pavlov J Biol Sci* 12(2):65–92
7. You-Me K (2001) *The Journal of Elementary Education studies* 8(1):1–32
8. Man-sang F (2007) *Creating an intelligent brain*. Jisiksanupsa, Seoul
9. Joo-yun C (1998) Educational applications of cognitive sciences discoveries about learning/memory. *J Elem Educ* 12(2):5–27
10. Itti L, Koch C, Niebur E (1998) Model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Pattern Anal Mach Intell* 20(11):1254–1259
11. Posner M, Raichle M (1994) *Images of mind*. Scientific American Library, New York
12. Lawson AE (1997) The role of the prefrontal lobes in scientific reasoning. *J Korea Assoc Sci Educ* 17(4):525–540
13. Jung-Eun L, Un-ho J, Bo-hyeok S, Soon-yong C (2012) Measurements of color effects on the brain activity for eye-attention, ICROS

# Design of Autonomic Nerve Measuring System Using Pulse Signal

Un-Ho Ji and Soon-Yong Chun

**Abstract** As for studies on autonomic nerve reactions, many researchers have published study results for long time, but in the standpoint of reproducibility and usefulness, no satisfactory outcomes have been obtained so far. As a new way to measure the change in autonomic nerve system, the method of using pulse signal the oriental medicine diagnosis measure is proposed. Specifically the autonomic nerve reaction measuring device to verify the usefulness of the proposed method is designed directly, and new method to process the measured pulse signals is proposed.

**Keywords** Autonomic nerve · Radial pulse · Pressure pulse wave · Fast fourier transform

## 1 Introduction

Methods to measure the change of human body autonomic nerve include electrocardiogram, heartbeat variability, skin resistance, electromyogram, electroencephlogram and electro-oculogram etc., and such methods predict the change of autonomic nerve by analyzing the utilization of sympathetic nerve and parasympathetic nerves [1–7]. Such methods can compare the utilization degrees of sympathetic nerve and parasympathetic nerve by identifying the characteristics like increase in the number of heartbeats and decrease in variance of heartbeats that occur due to activation of sympathetic nerve. Existing measurement methods

---

U.-H. Ji · S.-Y. Chun (✉)  
Dongyang University, Yeongju-si, Gyeongbuk, Korea  
e-mail: control@dyu.ac.kr

U.-H. Ji  
e-mail: jiunho@hotmail.com

are useful to analyze the utilized degree of synthetic nerve and parasympathetic nerve comparatively, and express the utilized degrees of synthetic nerve and parasympathetic nerve by particular disease or internal/external stimulations. Also, as for the comparison of activity degrees of synthetic nerve and parasympathetic nerve which are determined through repetitive measurements, the activity degrees measured before and after measurement were found to be constant at two levels “increase” and “decrease”. However, it is judged that the methods require complementation for use as measurement methods to provide reliable data for doctors who use instruments clinically in quantitative aspect. That is because the bio-electrical characteristics of human body are liable to change, but are kept constant by the activities of autonomic nerve system that works to keep the constancy of human body [8]. Therefore, in order to diagnose the medical condition of disease by measuring the change in autonomic nerve system, there needs be a method to specifically judge the long term change in the condition of human body consisting the inside based on the change in the reaction of autonomic nerve system rather than just the level of activity like in existing method. It is judged that to that end, measuring and analyzing the change in pulse that occurs as change in human body autonomic nerve would be useful. Pulse is characteristic that it changes by the speed and flow rate of blood flow the characteristics of blood vessel linked to all elements of human body, and the change in the characteristics of blood vessel occurs as the blood circulates human body elements. The method that can be utilized the most effectively with the use of the correlation of the characteristics of pulse change and human body is pulse method a method effectively used in oriental medicine. Pulse signals which are considered as important measure for diagnosis of lesions in oriental medicine, signals that include the characteristics of blood vessel flow rate and pressure that changes according to elements conditions as the blood circulates human body, are judged as capable to identify the characteristics of various changes in human body occurring by the change in autonomic nerves.

In this study, method to judge the change in human body elements based on the change in human body autonomic nerve reactions by measuring pulse signals was proposed, and after designing of system for measurement of pulse signal, pulse signals measured through experiment were analyzed.

## **2 Measuring Autonomic Nerve Reaction**

### ***2.1 Actions to Measure Autonomic Nerve Reactions***

Methods used the most as device to measure autonomic nerve reaction include the method of measuring the variance of heartbeat [1–3]. In general, HRV can be calculated from the change in ECG signal R–R interval, and the power spectrum can be divided into three parts depending on frequency areas: [LF(0.01–0.08 Hz),



MF(0.08–0.15 Hz), and HF(0.15–0.5 Hz)]. While LF reflects the activity of sympathetic nerve system mostly and those of parasympathetic nerve system a little, HF reflects the activities of parasympathetic nerve system exclusively, so LF/HF has been used as indicator to measure the balance of sympathetic nerve system and parasympathetic nerve system. It is known that MF represents mixed activities of sympathetic nerve system and parasympathetic nerve system, but it represents the activities of parasympathetic nerve system more. Bio electrical testing method is a method that recognizes human body as a circuit with electrical characteristics and analyzes the physiological/pathological characteristics by detecting skin resistance value and conductivity when static current or static voltage is flown based on the fact that sin electrical activity is closely related to sweat gland and autonomic nerve activity [9–13].

Such measurement of change in autonomic nerve system have been subjects in many studies by being recognizes as important device to predict diseases that can occur in human body and diagnose health condition, but there has been found no device to measure, predict and analyze specific changes in health condition and judgment of particular disease symptom. To judge change in health condition and judge particular disease symptom after analyzing the result of measurement of change in autonomic nerve system, specific actions for measurement and analysis are judged required, and to that end, in this study, the method of measuring and analyzing pulse signal used as important diagnosis action in oriental medicine is proposed.

### 2.2 Designing of Autonomic Nerve Reaction Measurement System Using Pulse Signal

Figure 1a shows the overall composition of measurement system designed and produced as action to measure pulse signal in this study. Figure 1b shows the pulse signal measurement system for measurement of autonomic nerve reaction in the method proposed in this study.

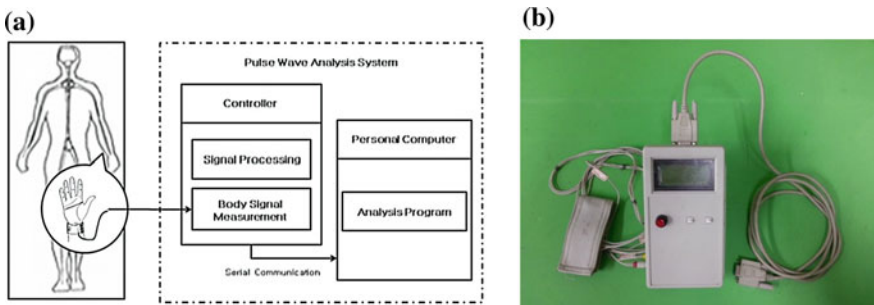


Fig. 1 a Block diagram of measurement system. b Pulse measurement system

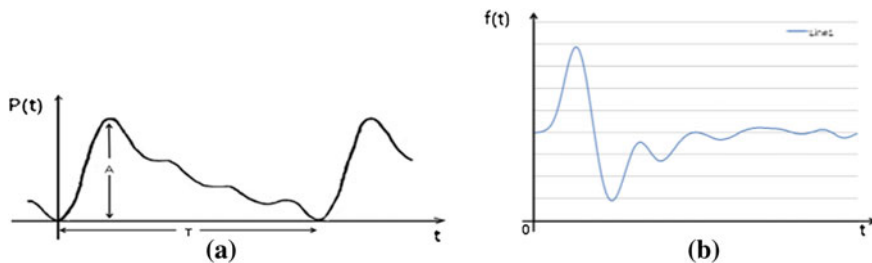


Fig. 2 a Human body pulse. b Differential pulse

Pulse signals, signals of waves measured in hipbone artery of wrist, measure the capacity pulse of hipbone artery utilizing piezoelectric element. Pulse is wave that occurs by the change in the blood vessel amount of blood and speed of blood flow by constriction and relaxation of heart, and one cycle of pulse occurs through a series of process like beginning of constriction of left ventricle, culmination of construction, expansion of aorta wall, reduction of blood vessel and elastic wave of heart valve and heart muscle etc. Also, pulse signals include the characteristics that they pass capillary vessels of body organs. Figure 2a shows the flow of pulse signal.

Here, ‘T’ indicates the cycle of pulse and ‘A’ the maximum point of constriction. In particular, the maximum point of constriction in pulse, which constitutes the largest interval and an important point in detecting change in heartbeat, can be used in analyzing the activity of autonomic nerve system. Figure 2b shows the differential pulse put out through the system designed in this study. Differential pulse can be used to identify the characteristics of change in blood vessel capacity from beginning of heart constriction to culmination of constriction, and effectively represents the characteristics of change of volume occurring due to the expansion of aorta wall, reduction of blood efflux and resilience of heart valve and heart muscle after reduction of ventricles constriction. Analysis of such elements is used as indicator for evaluation of blood vessel age [14]. The procedure of processing the pulse signals in the system designed in this study can be summed up as in Fig. 3.

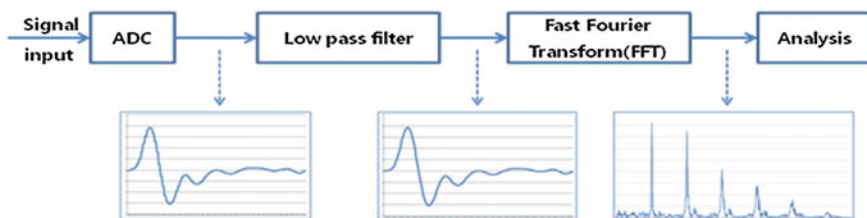


Fig. 3 The procedure of processing pulses

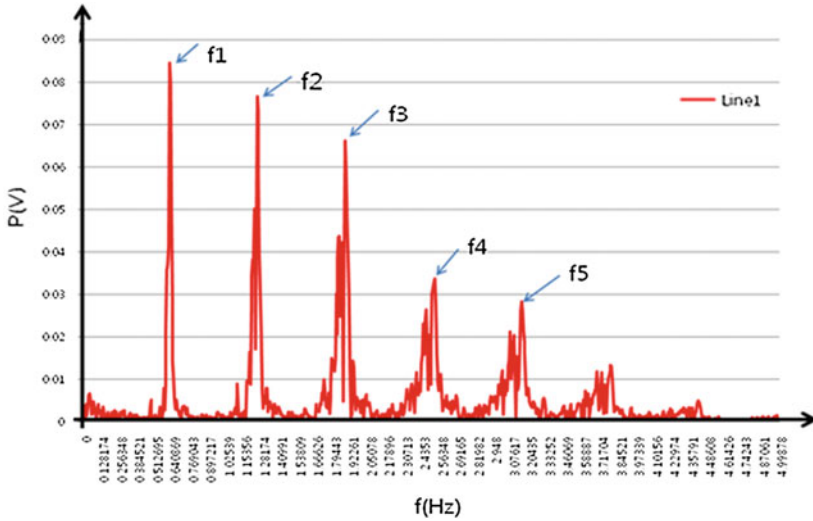


Fig. 4 The FFT of differential pulse

To remove the noise elements that can be included in the procedure of processing signals in measurement system, 2nd low bandwidth filter was used, and to create indicator for autonomic nerve reaction, FFT results of differential pulse were used. Figure 4 shows the data obtained through FFT transformation, and as shown in Fig. 4, the power of particular frequency bandwidth appear high in FFT

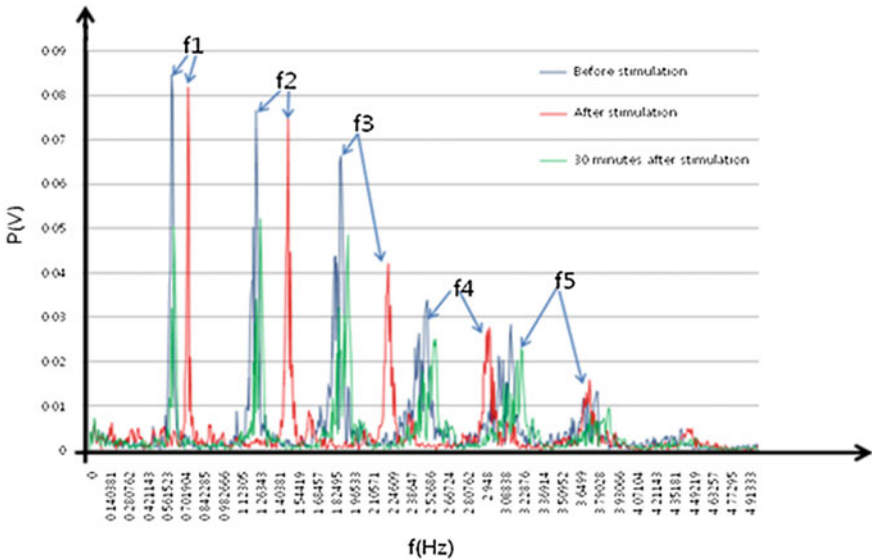
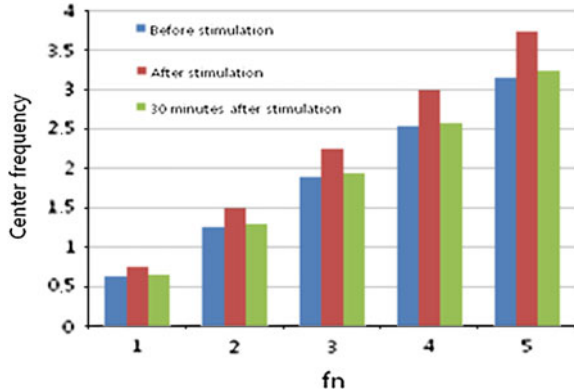


Fig. 5 The characteristics of pulse signals before and after outside stimulation

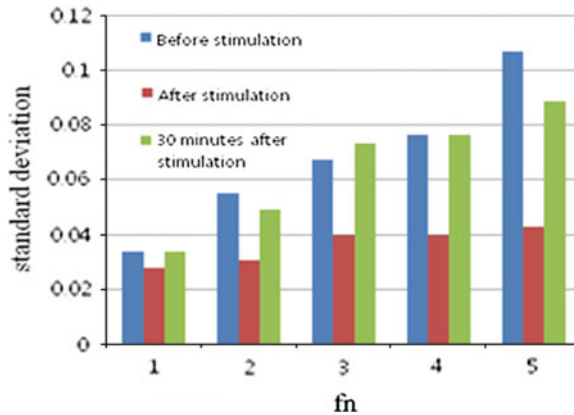
**Table 1** The characteristics indicator of change in pulse signal

	f1	$\pm\epsilon 1$	F2	$\pm\epsilon 2$	F3	$\pm\epsilon 3$	F4	$\pm\epsilon 4$	F5	$\pm\epsilon 5$
Before stimulation	0.6225	0.0305	1.2512	0.0732	1.8859	0.1037	2.5268	0.1037	3.1494	0.1647
After stimulation	0.7446	0.0305	1.4892	0.0366	2.2399	0.0427	2.9907	0.0610	3.7353	0.0549
30 min after stimulation	0.6408	0.0366	1.2878	0.0610	1.9348	0.1098	2.5756	0.1159	3.2348	0.1342

**Fig. 6** Central frequency locations of f1 through f5 before and after stimulation and after 30 min rest



**Fig. 7** Standard deviations before and after stimulation and after 30 min rest



signal. The central frequencies in this part are defined respectively as f1, f2, f3, f4 and f5, and the change width of the central frequencies occurring by repetitive measurement as standard deviation  $\pm \epsilon$ .

### 2.3 Measurement Experiment of Autonomic Nerve Reaction

To check the potential for measurement of autonomic nerve reactions through the pulse signal measurement system designed in this study, experiment was conducted. As for experiment method, the characteristics appearing usually and the characteristics of change in autonomic nerves for particular outside stimulation were analyzed comparatively.

**Measurement of pulse signal:** In the experiment, as for outside stimulation element to raise autonomic nerve change, the experiment participant was set to run about 400 m distance with full force, and with the characteristics of the usual pulse

signals of experiment participant measured 10 times repeatedly before experiment, the averages were taken and compared with experiment data. Figure 5 shows the comparison results.

**Results of pulse signal analysis:** Table 1 shows the results of analyzing the characteristics of pulse before and after outside stimulation from the results of Fig. 5.

Figure 6 shows the comparison of the central frequency location of  $f_n$  that occurs before and after stimulation and 30 min rest in graph. It is found that the central frequencies of  $f_1$  through  $f_5$  increase before and after stimulation, and after 30 min rest, the central frequency moves toward usual frequency range again.

Figure 7 shows that once stimulation is applied due to change in standard deviation occurring before and after stimulation and 30 min rest, standard deviation decreases.

### 3 Conclusion

In this study, a paper of experiment study on the device to measure autonomic nerve system reaction, experiment device to measure autonomic nerve reaction was designed and produced. Through the experiment of using the produced measurement system, the usefulness of autonomic nerve system reaction measurement was found through experiment.

Unlike in existing method that indicated the activity degree of sympathetic nerve and parasympathetic nerve generally occurring depending on the change in autonomic nerve system, in this study a device that measures and analyzes the change in autonomic nerve system was proposed with the use of pulse signals used as important diagnosis measure in oriental medicine.

As for analysis of pulse signal, the data measured from sensor are amplified and filtered through hardware signal processing part, and the processed results were made digital through ADC. The digitalized values were analyzed on frequency components through FFT, and  $f_1$ – $f_5$ ,  $\pm\varepsilon_1$ – $\pm\varepsilon_5$  the quantitative evaluation indicators were specified from the analysis results.

For experiment to measure autonomic nerve reaction, change in human body was induced through exercise as outside stimulation element, and through analysis of pulse signals measured here, the changes were measured and analyzed by being divided into before stimulation, after stimulation and after rest.

As a result of analysis, as shown in Figs. 6 and 7, evaluation indicator changed due to the change in autonomic nerve system.

Based on this study results, future studies must be done on the characteristics of autonomic nerve system that reacts according the inside/outside stimulation or condition change of human body for development of medical advances.

## References

1. Lee Y-H (2004) Designing of oriental medicine diagnosis system by bio-electrical reaction. Korea Ocean IT Soc J 420-429 (Book 8, 2nd edn)
2. Park C-W (2004) A study on the effects of far-infrared ray heat on human autonomic nerve function. Korea Med Eng Soc J 25(6): 623-628 (87th edn)
3. Lee J-H (2000) Designing of the time frequency analysis system of heartbeat change signal for evaluation of autonomic nerve system operation. Yonsei University, Seoul
4. Choi E-M (2006) A study on the skin resistance variance of ovarian insufficiency using bio-electrical autonomic reaction measurer. Korea Orient Med Gynecol Soc J 19(3): 247-256
5. Oh D-H (2003) Mental physiology of post-trauma stress disability. Ment Health Study 22:24-37
6. Han T-R (1994) The effects of heat sympathetic reaction on nerve conduction and autonomic nerve function. Korea Rehabil Med J 18(1): 28-34
7. McCraty R, Atkinson M, Tiller WA, Rein G, Watkins AD (1995) The Effects of Emotions on Short-Term Power Spectrum Analysis of Heartbeat Variability. Am J Cardiol 76: 1089-1093
8. Lee B-C (1995) Analysis of the characteristics of bio non-linear dynamic system using chaos theory. Yonsei University, Seoul
9. Park Y-J (2001) A study on skin resistance variance. Korea Orient Med Diagn J 5(2): 365-376
10. Boucsein W (1992) Electrodermal Activity. Plenum press, New York, pp 1-42
11. Nam D-H (2001) The effects of Korean adults males and females with healthy deep breath capability on skin electrical autonomic reaction. Korea Orient Med Diagn J 5(1): 139-152
12. Park C-W (1990) Autonomic nerve pharmacology. Seoul National University Press, Seoul, pp 63-80
13. Chun S-Y, Ji U-H (2004) Measurements of the current change on acupuncture spots at the meal time. IEEE, CBMS 2004, p 59
14. Takazawa K, Tanaka N, Fujita M, Matsuoka O, Saiki T, Aikawa M, Tamura S, Ibukiyama C (1998) Assessment of vasoactive agents and vascular aging by the second derivative of photoplethysmogram waveform. Am Heart Assoc Hypertens 32(2) 365-370

# Semiconductor Monitoring System for Etching Process

Sang-Chul Kim

**Abstract** In this paper, we developed the semiconductor monitoring system for the etching process. Process monitoring techniques has an important role to give an equivalent quality and productivity to produce semiconductor. The proposed monitoring system is mainly focused on the dry etching process using plasma and it provides the detailed observation, analysis and feedback to managers.

**Keywords** Etching · Plasma · Optical emission spectroscopy (OES) · End-point detecting · Real-time monitoring · Dynamic linked library (DLL) interface

## 1 Introduction

The semiconductor industry began to develop in earnest from the mid-1950s, interlinking with space development and weaponering. At the present time it has entered the microelectronic age, with the need for subminiature, ultra-light, high-reliability electronic components. Here at, many countries are keenly competing with each other for it [1]. In particular, there has been an increasing interest in endpoint sensing to prevent defects in the manufacturing process. A defect in the semiconductor etching process causes great economic loss. The real-time observation of the semiconductor etching process, the application of systematic process control and the maximization of automatic process control (APC) efficiency are required for the prevention of defects. In this regard, this study is on a monitoring system to implement the foregoing. [Section 2](#) is to deal with involved technologies and usable platforms. [Section 3](#) is to explain the system structure, and [Sect. 4](#) is to show the test results. Lastly, chapter [Sect. 5](#) is to make a conclusion and make mention of future plans.

---

S.-C. Kim (✉)

School of Computer Science, Kookmin University, Seoul, Korea  
e-mail: sckim7@kookmin.ac.kr



## 2 Related Works

A semiconductor etching process is to etch the wafer, made with processed silicon, with a circuit, which is divided into wet etching and dry etching. In the case of wet etching; a chemical solution is applied to the membrane of the wafer. It was widely used early in the semiconductor industry. However, it causes undercut as shown in Fig. 1a because chemical reactions occur vertically as well as horizontally.

The biggest problem is that it is difficult to form the vertical core layer because the pattern line width becomes narrower. That is why plasma-based dry etching has been widely used in recent times [1]. The dry etching is based on a chemical reaction between elements in plasma and the surface of the wafer, and the reaction is accelerated as active species in plasma crash against the surface. Since the process can be controlled in atomic unit, it is possible to etch the wafer in any shape as shown as in Fig. 1b [2].

The existing etching process, putting electrodes into plasma, mostly used the Langmuir probe to measure electron density, electron temperature and plasma potential, a interferometer to make material waves into single waves and thus to observe the interference fringes of neutral atoms and a mass spectrometer to gauge the mass of the ion made from specimen particles in a vacuum.

The optical emission spectroscopy (OES) sensor as shown in Fig. 2 makes a real-time analysis of the spectrum of gas caused by the plasma reaction in a semiconductor etching process, and the processed data to interface devices [2]. The OES sensor has the merit of being installed outside process equipment in which the plasma reaction occurs. Thus, it is not affected by internal variables such as voltage, frequency, electromagnetic field and gas pressure.

## 3 Proposed Monitoring System

The system, shown in Fig. 3, is largely divided into three parts. The OES sensor that transmits data after observing and analyzing the semiconductor-etching process, Dynamic-linked library (DLL) Interface that processes data, transmitted from the sensor, and transmits them to the monitoring system, and the monitoring system that reprocesses the transmitted data in accordance with the logic made by the administrator and checks whether the process proceeds as planned.

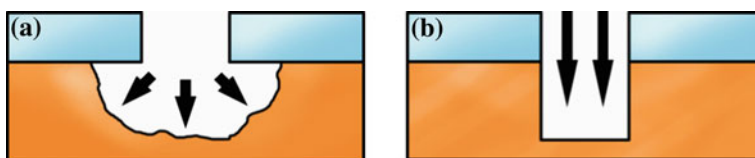


Fig. 1 a Model of wet-etching. b Model of dry-etching

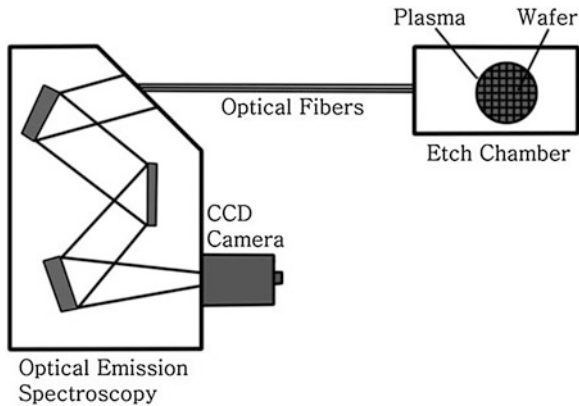


Fig. 2 Optical emission spectroscopy (OES) sensor

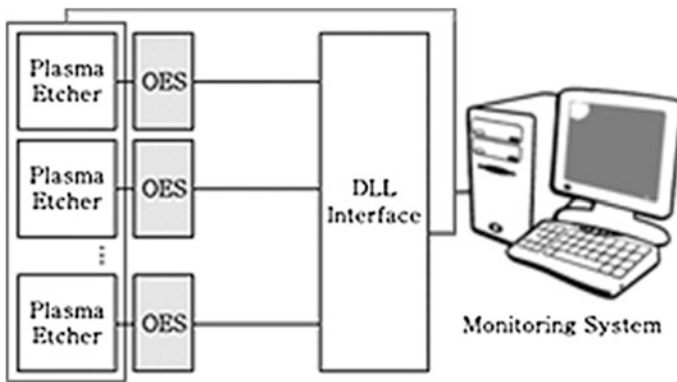


Fig. 3 System architecture

The OES sensor is installed outside plasma etch equipment. The inside can be observed through an optical fiber cable connected from transparent lens to an OES sensor. The observed image falls on the charge-coupled device (CCD) array, which is transmitted from USB interface to DLL Interface [2]. In this study, the monitoring system is equipped with the OES sensor that can detect 200–900 nm wavelength, which is equivalent to ultraviolet, visible and infrared lights. The OES sensor divides the 200–900 nm wavelength into 3648 sections and measures the strengths. Specifically, the CCD Array is formed with a length of 3648.

The OES sensor is controlled by DLL Interface. The OES sensor provides DLL Interface with information in the pre-work stage. The sensor information is composed of sensor ID, reserved name, wavelength and wavelength table. In case many sensors are interlocked with each other, respective sensors can be distinguished through the transmitted information. Figure 4 defines the structure that saves sensor information in DLL Interface.

```
typedef struct OesUnitInformation {  
    s32 id;  
    char model_name[256];  
    char serial[256];  
    s32 wave_length;  
    double wave_length_table[4096];  
} OesUnitInformation;
```

Fig. 4 Structure of OES sensor information

The main class of monitoring system is App class. It can approach to all the objects in the system, but at the same time, allows all of them access to itself, which can be implemented as the points of objects in the system are listed in App class. Figure 5 shows the structures of objects in the monitoring system. In the monitoring system, all the information related to OES sensors is organized in the object of the sensor module (more hereof later).

The user interface of monitoring system is largely divided into three parts, *configuration dialog*, *chart view* and *recipe view*. *Configuration dialog* is implemented. *Chart view* is composed of a full spectrum chart and a time trend chart as shown in Fig. 6. The full spectrum chart includes the data source list that tells of the sources of respective wavelengths. The time trend chart includes the equation list premade by the administrator. If an item is chosen in the data source list, the corresponding wavelength range is highlighted in the full spectrum chart. In case an item is selected in the equation list, it is shown in the time trend chart in chronological order after data was processed within the range predetermined by the administrator. *Recipe view* is to show the actual state of process; to be specific; it shows where the process is on the scenario made by the administrator.

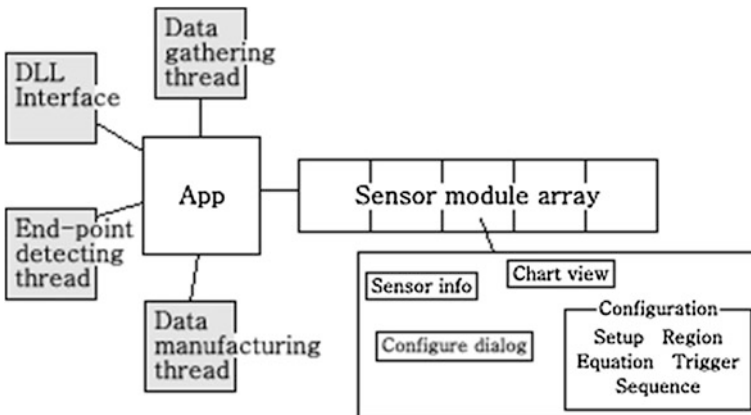


Fig. 5 Map of structured objects

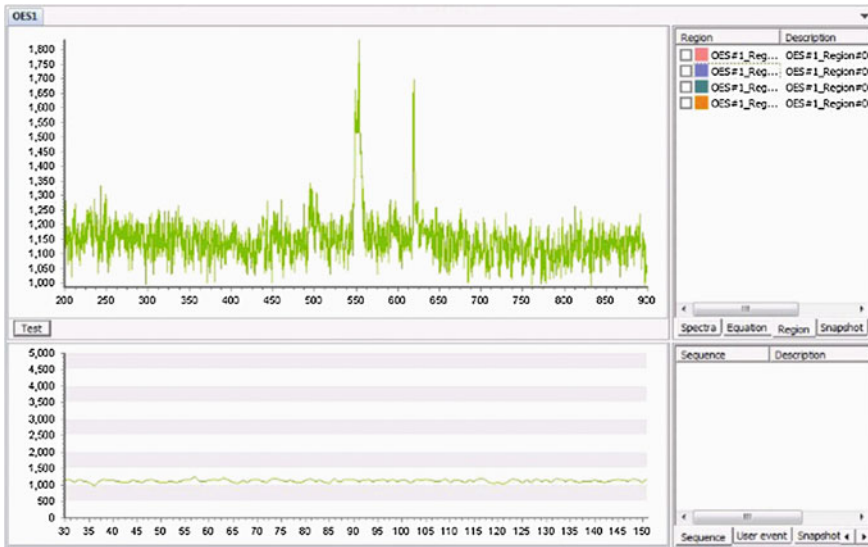


Fig. 6 Chart view

The ultimate purpose of the semiconductor etching process monitoring system is to detect the endpoint. The endpoint is detected as per the scenario made by the administrator. The administrator can get various forms of data that were useably processed, by using the setups of *configuration*. The scenario, based on the data, is a recipe for the corresponding process. A scenario is made by the administrator in *sequence* in *configuration*, and the logic that the scenario is applied to the monitoring system.

The monitoring system judges whether the manufactured data meet what was stipulated in the present sequence item or the described technique, whenever they are updated. The index is *chart data* that reflects the present process. With the completion of endpoint detection (EPD), the EPD signal is transmitted to process equipment to stop the process.

## 4 Experimental Results

The OES hardware, used in the present system, makes it possible to transmit data at intervals of at least 10 ms. But the interface, implemented in the monitoring system, was so designed that it could receive data at intervals of at least 100 ms, and thus memory can be shared without conflicts when data are processed in the system. Accordingly, the test should be conducted at a sampling per 10 counts, as mentioned above, and at intervals of 100 ms, the maximum performance of interface. Since it is realistically difficult to conduct the test in a real process, it is

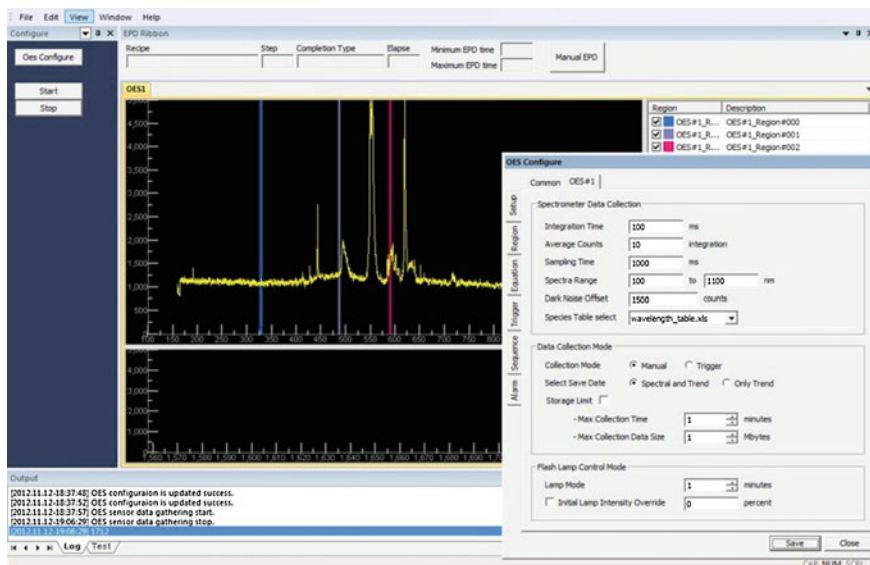


Fig. 7 The monitoring system after test

necessary to check whether data are normally gathered, whether the processing results, produced by configuration, are accurately reflected in *chart view* and how long the test is on the log. Figure 7 shows that the data have been successfully gathered and processed as per given time and environment.

The test is focused on checking how often data samplings are conducted and whether it meets the expected value. The source code is to put on record when data begin to be gathered, when data gathering is finished, and how many data samplings are conducted while data are gathering, of which records are saved in the log. In result, data samplings were conducted 1712 times for 28.5 min, which means that data samplings were conducted 60 times per min. It met the data sampling cycle determined in advance. The results showed that the data processing method could be controlled in real time as to the processing environment, and that data reliability could be heightened through normalization even in the shortest process cycle. As a result, it was proved to be highly effective to heighten control efficiency in different processing environments.

## 5 Conclusion

The monitoring system, developed in this study, has superiority over others. First, the optical sensor can accurately observe the inside of process equipment from outside it. Second, logic could be made once to  $N$  times by use of *equation* and thus data could be processed with almost infinite numbers. Lastly, the automated

check system prevents damage, taking primary measures in unforeseen circumstances. The semiconductor etching process monitoring system observes the visible state of process, analyzing sensor data, and at the same time conducts real-time logic that diagnoses the present process, and thus detects the endpoint. In this study, the monitoring system was focused on the semiconductor etching process, but in reality it is expected to be applied to more fields.

**Acknowledgments** This research was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2011 (Grants No. 0045590) and the Global Leading Technology Program of the Office of Strategic R&D Planning (OSP) funded by the Ministry of Knowledge Economy, Republic of Korea (grant number: 10042421).

## References

1. Kang Ho Y An Introduction of Etch Process, hynix semiconductor
2. Roawen C OES-based sensing for plasma processing in IC manufacturing. Doctor of Philosophy in Engineering, Electrical Engineering and Computer Sciences, University of California at Berkeley

# Enhancing the Robustness of Fault Isolation Estimator for Fault Diagnosis in Robotic Systems

Ngoc-Bach Hoang and Hee-Jun Kang

**Abstract** Fault diagnosis and fault tolerant control are increasingly importance in robotic systems. A number of researchers have proposed the generalized observer scheme for fault isolation when a fault happened. One of the key issues in this scheme is based on the sensitive of the residual with the corresponding adaptive threshold. In this paper, we present a new method to derive the adaptive threshold in order to enhance the robustness of fault isolation estimator and reduce the fault isolation time. Mathematical proof and computer simulation are performed to show the effectiveness of the proposed method.

**Keywords** Fault detection · Fault isolation · Nonlinear observer · Robotics

## 1 Introduction

Nowadays, robotic systems are widely used in complex engineering applications which demand very high performance, productivity and safe operation. Robotic fault can potentially result both in the loss of productivity and in unsafe operation of the system. Moreover, difficult and often dangerous environments limit the ability of humans to perform any supervisory and corrective tasks. Hence, automated fault detection, isolation and accommodation play a key role in the operation of modern robotic systems.

A number of researchers have worked on the problem of designing automated fault diagnosis schemes for robotic systems using analytical redundancy methods.

---

N.-B. Hoang

Graduate School of Electrical Engineering, University of Ulsan, Ulsan 680-749, Korea  
e-mail: hoangngocbach@gmail.com

H.-J. Kang (✉)

School of Electrical Engineering, University of Ulsan, Ulsan 680-749, Korea  
e-mail: hjkang@ulsan.ac.kr

A neural network based learning methodology is described in [1]. In [5, 6], a robust fault diagnosis scheme for abrupt and incipient faults in nonlinear uncertain dynamic systems is proposed. The extension of [5] to the case of robotic systems is presented in [7]. The architecture in [5, 6] consists of a bank of nonlinear adaptive estimators, one of which is used for the detection and the approximation of a fault, whereas the rest are used for online fault isolation. First, a nonlinear observer is designed based on the robotic model. The fault detection is carried out by comparing the observer states with their signatures. Then, multiple state observers are constructed based on possible fault function set. The fault isolation is implemented by checking each residual generated by observer state estimation and the corresponding threshold. One of the most importance criteria in fault diagnosis is the fault isolation time, which refers to the time taken by the fault isolation scheme to identify a fault that has occurred. In order to increase the capability of successful isolation and decrease the fault isolation time, we propose a more robust threshold in this paper. Both mathematical prove and simulation results are presented to show the effectiveness of our proposed threshold.

This paper is organized as follows: In Sect. 2, the robot dynamics in the presence of faults is presented. In Sect. 3, the design of fault detection and isolation scheme, including the derivation of new robust adaptive threshold, is described. In Sect. 4, the effectiveness of the proposed robust threshold is demonstrated via computer simulations. Section 5 has some concluding remarks.

## 2 Fault in Robot Manipulator

In the presence of a fault, the robot dynamics can be represented by [7]

$$\ddot{q} = M^{-1}(q)[\tau - V_m(q, \dot{q}) - F(\dot{q}) - G(q) - \tau_d] + \beta(t - T)\phi(q, \dot{q}, \tau) \quad (1)$$

The fault profile  $\beta(t - T)$  is a diagonal matrix of the form

$$\beta(t - T) = \text{diag}(\beta_1(t - T), \beta_2(t - T), \dots, \beta_n(t - T)) \quad (2)$$

Each time profile can be represented by

$$\beta_i(t - T) = \begin{cases} 0 & t < T \\ 1 - e^{-\alpha_i(t-T)} & t \geq T \end{cases} \quad (3)$$

where  $\alpha_i (i = 1, 2, \dots, n)$  are unknown constants that represent the development of fault.

The following assumptions will be used throughout this paper:

- (1) The system states remain bounded after the occurrence of the fault:  $q(t), \dot{q}(t) \in L_\infty$ .



- (2) The friction satisfies  $F(\dot{q}) = F_v\dot{q} + F_d$  with  $\|F_d\| \leq k_B\|\dot{q}\| + k_F$ , ( $k_B, k_F > 0$ ), where  $F_v$  and  $F_d$  are the coefficient matrix of viscous friction and the dynamic friction term, respectively.
- (3) The load disturbance is bounded by  $\|\tau_d\| \leq \tau_B$ , where  $\tau_B$  is a known constant.

### 3 Fault Diagnosis Scheme

#### 3.1 Fault Detection

Let  $x = \dot{q}^T(t)$ . The dynamic model of Eq. (1) can be rewritten as

$$\dot{x} = M^{-1}(q)[\tau - V_m(q, \dot{q}) - F_v\dot{q} - F_d - G(q) - \tau_d] + \beta(t - T)\phi(q, \dot{q}, \tau) \quad (4)$$

The following estimated model is considered

$$\dot{\hat{x}} = M^{-1}(q)[\tau - V_m(q, \dot{q}) - F_v\dot{q} - G(q)] + \Lambda(\hat{x} - x) \quad (5)$$

where  $\hat{x} \in R^n$ ,  $\Lambda = \text{diag}(-\lambda_1, -\lambda_2, \dots, -\lambda_n)$  are the estimation vector of  $x$  and a stable matrix, respectively. From Eqs. (4) and (5), we can derive the error dynamic equation

$$\dot{\varepsilon} = \Lambda(\hat{x} - x) - M^{-1}(F_d + \tau_d) + \beta(t - T)\psi(q, \dot{q}, \tau) \quad (6)$$

By taking the norm of Eq. (6), we obtain the upper bound of  $\|\varepsilon\|$

$$\|\varepsilon\| \leq \alpha e^{-\nu t} \|\varepsilon_0(0)\| + \int_0^t \alpha e^{-\nu(t-\xi)} (k_B\|\dot{q}\| + k_F + \tau_B) d\xi = {}^D\varepsilon \quad (7)$$

*Fault detection decision scheme:* The decision about the occurrence of the fault is made when the residuals  $\|\varepsilon\|$  exceeds its corresponding threshold  ${}^D\varepsilon$  [7].

#### 3.2 Fault Isolation with Robust Threshold Derivation

After a fault is detected, the following isolation estimators are activated:

$$\dot{\hat{x}}^s = \Lambda^s(\hat{x}^s - x) + M^{-1}(q)[\tau - V_m(q, \dot{q}) - F_v\dot{q} - G(q)] + \hat{\phi}^s(q, \dot{q}, \tau, \hat{\theta}^s) \quad (8)$$

$$\hat{\phi}^s(q, \dot{q}, \tau) = [(\hat{\theta}_1^s)^T g_1^s(q, \dot{q}, \tau), \dots, (\hat{\theta}_n^s)^T g_n^s(q, \dot{q}, \tau)]^T \quad (9)$$

where  $\hat{x}^s$  and  $\hat{\theta}_i^s$  are the estimated states and parameters  $x$  and  $\theta_i^s$ , respectively.  $\Lambda_s = \text{diag}(-\lambda_1, \dots, -\lambda_n)$  ( $\lambda_1, \dots, \lambda_n \geq 0$ ) is a stable matrix. The online updating law

for  $\hat{\theta}_i^s$  is  $\dot{\hat{\theta}}_i^s = P_{\Theta_i^s} \{ \Gamma_i^s g_i^s(q, \dot{q}, \tau) \varepsilon_i^s \}$  where  $\Gamma_i^s$ ,  $g_i^s(q, \dot{q}, \tau)$ ,  $\varepsilon_i^s = (x_i^s - \hat{x}_i^s)$  ( $i = 1, \dots, n$ ) are the positive definite learning rate, the corresponding smooth vector field, the error state, respectively.

From Eq. (8), the error dynamics is given by

$$\dot{\varepsilon} = \Lambda^s \varepsilon - M^{-1}(F_d + \tau_d) + \beta(t - T)\phi^s(q, \dot{q}, \tau) - \hat{\phi}^s(q, \dot{q}, \tau, \hat{\theta}^s) \tag{10}$$

Thus, each element of state estimation error is given by:

$$\dot{\varepsilon}_i^s = -\lambda_i \varepsilon_i - \sum_{j=1}^n m v_{ij} (f_{dj} + \tau_{dj}) + (1 - e^{-\alpha_i(t-T)}) (\theta_i^s)^T g_i^s(q, \dot{q}, \tau) - (\hat{\theta}_i^s)^T g_i^s(q, \dot{q}, \tau) \tag{11}$$

$$\begin{aligned} \varepsilon_i^s = & \varepsilon_i^s(T_d) \exp(-\lambda_i(t - T_d)) + \int_{T_d}^t \exp(-\lambda_i(\xi - T_d)) \sum_{j=1}^n m v_{ij} (f_{dj} + \tau_{dj}) d\xi \\ & + \int_{T_d}^t \exp(-\lambda_i(\xi - T_d)) ((1 - e^{-\alpha_i(\xi-T)}) (\theta_i^s)^T g_i^s(q, \dot{q}, \tau) - (\hat{\theta}_i^s)^T g_i^s(q, \dot{q}, \tau)) d\xi \end{aligned} \tag{12}$$

In this paper, a new threshold functions are derived as below:

The following function is defined to use later for deriving the threshold as:

$${}^1h_i = \begin{cases} g_i^s(q, \dot{q}, \tau) & (g_i^s(q, \dot{q}, \tau) \geq 0) \\ 0 & (g_i^s(q, \dot{q}, \tau) < 0) \end{cases}; {}^2h_i = \begin{cases} 0 & (g_i^s(q, \dot{q}, \tau) \geq 0) \\ g_i^s(q, \dot{q}, \tau) & (g_i^s(q, \dot{q}, \tau) < 0) \end{cases} \tag{13}$$

Because  $[\theta_i^s]$  is assumed to belong to a known compact set  $\Theta_i^s \subset R^{q_i^s}$ , there exist two values  ${}^m\theta_i, {}^M\theta_i$  the minimum and maximum of  $\Theta_i^s \subset R^{q_i^s}$ . So we get the inequality:

$${}^m\theta_i \leq \theta_i^s \leq {}^M\theta_i \leq 0 \quad \text{or} \quad 0 \leq {}^m\theta_i \leq \theta_i^s \leq {}^M\theta_i \tag{14}$$

Moreover, we assume that there exists a lower bound of fault evolution rate which satisfies  $\alpha_i \geq \bar{\alpha}_i$ , so that the following inequality can be established

$$(1 - e^{-\bar{\alpha}_i(t-T_d)}) \leq (1 - e^{-\alpha_i(t-T_d)}) \leq 1 \tag{15}$$

If  $0 \leq {}^m\theta_i \leq {}^M\theta_i$ , from Eq. (14) and (15), we have

$$m_i = (1 - e^{-\bar{\alpha}_i(t-T_d)}) {}^m\theta_i \leq (1 - e^{-\alpha_i(t-T_d)}) \theta_i^s \leq {}^M\theta_i = M_i \tag{16}$$

Or if  ${}^m\theta_i \leq {}^M\theta_i \leq 0$  then

$$m_i = {}^m\theta_i \leq (1 - e^{-\alpha_i(t-T_d)}) \theta_i^s \leq (1 - e^{-\bar{\alpha}_i(t-T_d)}) {}^M\theta_i = M_i \tag{17}$$

Now, let's combine Eqs. (13) (16) and (17) and put them in the matrix form

$$(m_i)^T \cdot^1 h_i + (M_i)^T \cdot^2 h_i \leq (1 - e^{-\alpha_i(t-T_d)})(\hat{\theta}_i^s)^T g_i^s \leq (M_i)^T \cdot^1 h_i + (m_i)^T \cdot^2 h_i \quad (18)$$

Thus, the following threshold (upper bound and lower bound) can be chosen for isolation decision

$$\begin{aligned} U_{\varepsilon_i} = & \exp(-\lambda_i(t - T_d))e_i(T_d) + \int_{T_d}^t \exp(-\lambda_i(\xi - T_d)) \sum_{j=1}^n |mv_{ij}|(f_{Bj}|\dot{q}_j| + f_{Fj} + \tau_{Bj})d\xi \\ & + \int_{T_d}^t \exp(-\lambda_i(\xi - T_d))((M_i)^T \cdot^1 h_i + (m_i)^T \cdot^2 h_i - (\hat{\theta}_i^s)^T g_i^s(q, \dot{q}, \tau))d\xi \end{aligned} \quad (19)$$

$$\begin{aligned} L_{\varepsilon_i} = & \exp(-\lambda_i(t - T_d))e_i(T_d) - \int_{T_d}^t \exp(-\lambda_i(\xi - T_d)) \sum_{j=1}^n |mv_{ij}|(f_{Bj}|\dot{q}_j| + f_{Fj} + \tau_{Bj})d\xi \\ & + \int_{T_d}^t \exp(-\lambda_i(\xi - T_d))((m_i)^T \cdot^1 h_i + (M_i)^T \cdot^2 h_i - (\hat{\theta}_i^s)^T g_i^s(q, \dot{q}, \tau))d\xi \end{aligned} \quad (20)$$

The threshold given by Eqs. (19) and (20) can be implemented with a linear filter with the transfer function  $1/(s + \lambda_i)$  and with the input

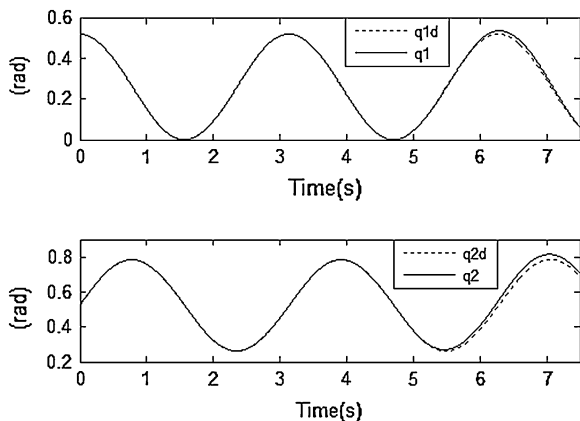
$$U_{I_i} = -(\hat{\theta}_i^s)^T g_i^s(q, \dot{q}, \tau) + ((M_i)^T \cdot^1 h_i + (m_i)^T \cdot^2 h_i) + \sum_{j=1}^n |mv_{ij}|(f_{Bj}|\dot{q}_j| + f_{Fj} + \tau_{Bj}) \quad (21)$$

$$L_{I_i} = -(\hat{\theta}_i^s)^T g_i^s(q, \dot{q}, \tau) + ((m_i)^T \cdot^1 h_i + (M_i)^T \cdot^2 h_i) - \sum_{j=1}^n |mv_{ij}|(f_{Bj}|\dot{q}_j| + f_{Fj} + \tau_{Bj}) \quad (22)$$

### 4 Simulation Results

A two-link planar robotic system now is used to illustrate for the performance of our proposed threshold. We consider 2 types of possible faults being given by

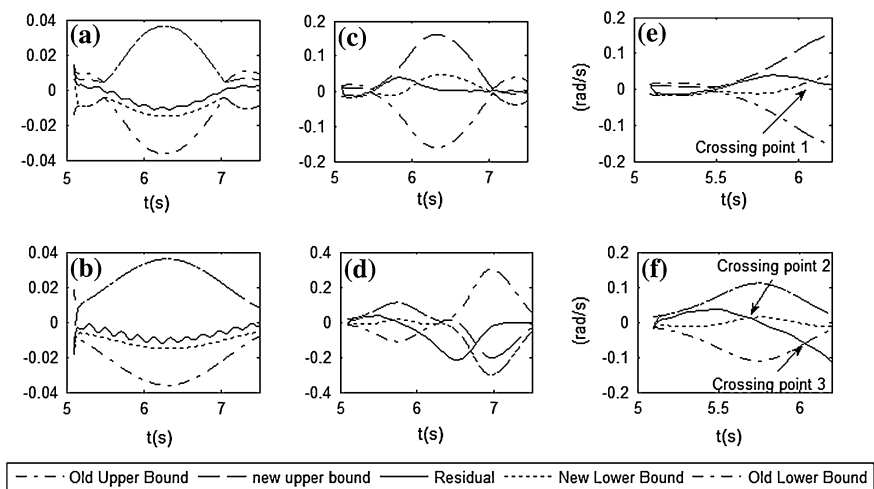
$$F = \left\{ \left[ \begin{array}{c} 10 \sin(3q_1) \cdot \dot{q}_2 \\ 15q_1 \end{array} \right] \cdot \left[ \begin{array}{c} 20 \sin(3q_1) \cdot \dot{q}_2 \\ 20 \sin(3q_1) \cdot \dot{q}_1 \end{array} \right] \right\}$$



**Fig. 1** Position normal control with fault occurrence

The fault #1 is assumed to be triggered at  $T = 5(s)$ . Figure 1 shows the position control when fault #1 occurs at  $T = 5(s)$  without a fault accommodation scheme. By using fault detection scheme, a fault is detected at  $T_d = 5.1(s)$ .

After the fault is detected, two observers are activated. The outputs of two observers are showed in Fig. 2. Figure 2a and b belong to observer 1, Fig. 2c and d belong to observer 2. In Fig. 2, old upper bound and old lower bound stand for the derived threshold in [5–7], new upper bound and new lower bound stand for our proposed threshold. The fault #1 is successful isolated at  $T_{isol} = 5.67(s)$ . In order to show the effective of our derived threshold, we enlarge Fig. 2c and d in Fig. 2e,



**Fig. 2** Fault isolation. **a, b** Output of fault 1 observer. **c, d** Output of fault 2 observer. **e, f** Enlarge of **c, d**

f. In Fig. 2e, f if our threshold is applied, the fault is successfully isolated at  $T_2 = 5.67(s)$  by checking crossing point 2, earlier than  $T_3 = 6.05(s)$  by checking crossing point 3. This results show that our derived thresholds are better than which are proposed in [5–7].

## 5 Conclusions

In this paper, we present a new method to derive the adaptive threshold for fault isolation in robotic systems. With our derived threshold, the capability of successful isolation is improved and the fault isolation time is reduced. Hence, the fault tolerant controller is activated sooner to reduce the effect of fault to the robotic systems. The detailed simulation for two-link manipulator has been given to show the effectiveness of the proposed method.

**Acknowledgments** This work was supported by the Ministry of Knowledge Economy under the Human Resources Development Program for Convergence Robot specialists and under the Robot Industry Core Technology Project.

## References

1. Vemuri AT, Polycarpou MM (1997) Neural-network-based robust fault diagnosis in robotic systems. *IEEE Trans Neural Netw* 8:1410–1420
2. Wang H, Daley S (1996) Actuator fault diagnosis: an adaptive observer-based technique. *IEEE Trans Autom Control* 41:1073–1078
3. Gang T, Xiaoli M, Joshi SM (2000) Adaptive state feedback control of systems with actuator failures. In: *Proceedings of the American control conference*, vol 4, pp 2669–2673
4. Visinsky ML, Cavallaro JR, Walker ID (1994) Expert system framework for fault detection and fault tolerance in robotics. *Comput Electr Eng* 20:421–435
5. Xiaodong Z, Parisini T, Polycarpou MM (2004) Adaptive fault-tolerant control of nonlinear uncertain systems: an information-based diagnostic approach. *IEEE Trans Autom Control* 49:1259–1274
6. Xiaodong Z, Polycarpou MM, Parisini T (2002) A robust detection and isolation scheme for abrupt and incipient faults in nonlinear systems. *IEEE Trans Autom Control* 47:576–593
7. Huang SN, Kok Kiang T (2008) Fault detection, isolation, and accommodation control in robotic systems. *IEEE Trans Autom Sci Eng* 5:480–489

# Software-Based Fault Detection and Recovery for Cyber-Physical Systems

Jooyi Lee, Ji Chan Maeng, Byeonghun Song, Hyunmin Yoon,  
Taeho Kim, Won-Tae Kim and Minsoo Ryu

**Abstract** Cyber-physical systems demand higher levels of reliability for several reasons. First, unlike traditional computer-based systems, cyber-physical systems are more vulnerable to various faults since they operate under harsh working conditions. For instance, sensors and actuator may not always obey their specification due to wear-out or radiation. Second, even a minor fault in cyber-physical systems may lead to serious consequences since they operate under minimal supervision of human operators. In this paper we propose a software framework of fault detection and recovery for cyber-physical systems, called Fault Detection and Recovery for CPS (FDR-CPS). FDR-CPS focuses on specific types of faults related to sensors and actuators, which seem to be the likely cause of critical system failures such as system hangs and crashes. We divide such critical failures into four classes and then present the design and implementation of FDR-CPS that

---

J. Lee (✉) · J. C. Maeng · B. Song · H. Yoon · T. Kim  
Department of EECS, Hanyang University, Seoul, Korea  
e-mail: jylee@rtcc.hanyang.ac.kr

J. C. Maeng  
e-mail: jcmaeng@rtcc.hanyang.ac.kr

B. Song  
e-mail: bhsong@rtcc.hanyang.ac.kr

H. Yoon  
e-mail: hmyoon@rtcc.hanyang.ac.kr

T. Kim  
e-mail: thkim@rtcc.hanyang.ac.kr

W.-T. Kim  
Embedded SW Research Division, ETRI, Daejeon, Korea  
e-mail: wtkim@etri.re.kr

M. Ryu  
Department of CSE, Hanyang University, Seoul, Korea  
e-mail: msryu@hanyang.ac.kr

can successfully handle the four classes of critical failures. We also describe a case study with quadrotor to demonstrate how FDR-CPS can be applied in a real world application.

**Keywords** Cyber-physical system • Reliability • Fault • Detection • Recovery

## 1 Introduction

A cyber-physical system is defined as a network of interconnected physical and computational elements. In typical cyber-physical systems, physical elements are monitored and controlled through sensors and actuators while computational elements process the information collected by sensors and send appropriate commands to actuators. This style of system organization is very similar to automatic control system structures, and thus enables various levels of autonomy in cyber-physical systems.

Cyber-physical systems demand higher levels of reliability for several reasons. First, unlike traditional computer-based systems, cyber-physical systems are more vulnerable to various faults since they operate under harsh working conditions. For instance, sensors and actuators may not always obey their specification due to wear-out or radiation. Second, even a minor fault in cyber-physical systems may lead to serious consequences since they operate under minimal supervision of human operators.

In this paper, we propose a software framework of fault detection and recovery for cyber-physical systems, called FDR-CPS (Fault Detection and Recovery for CPS). The primary goal of FDR-CPS is to provide resilience against critical faults that can cause system hangs or crashes. To that end, FDR-CPS focuses on specific types of faults that are related to sensors and actuators. Note that device-related faults are the main cause of critical failures like system crashes in many computer-based systems [1]. Since most software is written assuming the reliability of hardware devices, a single device fault may seriously impact the whole system. In this work we divide such critical failures into four classes and then present the design and implementation of FDR-CPS that can successfully handle the four classes of critical failures. We also describe a quadrotor case study to show how FDR-CPS can be applied in practice.

## 2 Critical Failures in Cyber\_Physical Systems

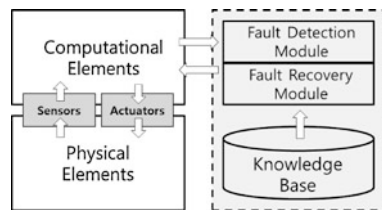
Sensors and actuators are highly error-prone since they are often exposed to harsh physical environments. There exist many sources of faults such as wear-out, EMI and radiation that may cause various types of failures in sensors and actuators. Among them, we identify four classes of critical failures as follows.

- **Indefinite waiting:** Some specific event being waited for does not occur. For instance, a sensor or actuator may not be able to generate interrupts because of some internal failure while the program is waiting for the interrupt. In this case, the system may wait indefinitely. We will show this type of failure can be easily handled by using a timeout mechanism.
- **Infinite loop:** The program’s workflow goes into infinite loop. For instance, a stuck-at fault may occur at some I/O channel that is referenced by a conditional loop. For instance, the program cannot escape from an infinite loop if a device’s busy flag has a stuck-at fault and does not satisfy the loop’s stopping condition. In this case, we can use a counter to detect and finish infinite loop iterations.
- **Erroneous sensor data:** Sensor data are not correct. This type of failure can be further divided into two subtypes. First, sensor data are inaccurate due to some reasons like noises or variations in ambient conditions. Second, sensor data are invalid because they exceed the normal range or telling inaccurate data constantly due to some internal failure.
- **Repeated operation failures:** A sensor or actuator repeatedly fails to complete program’s requests. For instance, a sensor or actuator may constantly return error code as a return value in response to program’s request. In case a transient fault causes this type of failure, we may recover the device through a reset leveraging the shadow driver technique [5].

### 3 Fault Detection and Recovery

FDR-CPS consists of three main components, fault detection module, fault recovery module and knowledge base. The fault detection module provides a set of skeleton functions that should be specialized by the programmer. A skeleton function is associated with each type of failure and provides a generic structure needed to diagnose a specific failure. The fault recovery module also provides a set of skeleton functions, each of which provides a generic structure to handle a specific type of failure. The fault detection and recovery modules may consult the knowledge base for failure handling policies and guidelines. The knowledge base contains several policies and guidelines about fault identification, logging, reporting and recovery (Fig. 1).

Fig. 1 FDR-CPS architecture





### 3.1 Direct Code Insertion

In order to apply FDR-CPS, we need to implement fault handling functionality in the target CPS program's code as well as specializing skeleton functions. FDR-CPS offers two options of implementation, direct code insertion (DCI) and function call insertion (FCI). The DCI technique is to insert some source code directly into the target program's source code. This technique is more efficient than FCI since it does not incur function call overhead.

Figure 2 illustrates the application of DCI technique to NS8390 ethernet device driver in Linux kernel 3.5.4. The original device driver had a potential infinite loop, but we inserted code to count the number of iterations at line 6, report and recover from line 8 through line 9 and escape the infinite loop at line 10.

Figure 3 illustrates another application of DCI technique. The example came from DE600 ethernet driver in Linux kernel 3.5.4. This example shows how invalid data can be detected and handled.

**Fig. 2** Illustration of DCI for infinite loop handling

```

1 static void el2_get_8390_hdr(...)
2 {
3   ...
4   while ((inb(E33G_STATUS) & ESTAT_DPRDY) == 0)
5     {
6       if(!boguscount--)
7       {
8         pr_notice("...", dev->name);
9         el2_reset_8390(dev);
10        goto blocked;
11      }
12      ...
13    }

```

**Fig. 3** Illustration of DCI for invalid data failure handling

```

1 static void de600_rx_intr(...)
2 {
3   size = de600_read_byte(RX_LEN, dev);
4   size += (de600_read_byte(RX_LEN, dev) << 8);
5   size -= 4;
6   ...
7   if ((size < 32) || (size > 1535)) {
8     printk(...);
9     if (size > 10000)
10      adapter_init(dev);
11    return;
12  }
13  ...

```

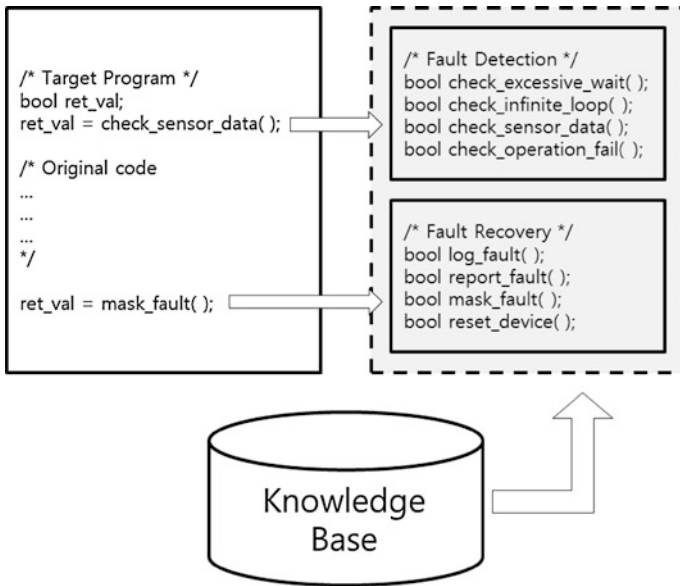


Fig. 4 Illustration of FCI for fault detection and recovery

### 3.2 Function Call Insertion

Function Call Insertion (FCI) may cause some performance overhead due to function call processing. However, it provides many benefits from a software engineering point of view such as modularity and standardized interfaces. Figure 4 shows the brief overview of FCI technique. With FCI, programmers just need to insert some function calls at appropriate places while minimizing source code modification.

## 4 Case Study

In order to evaluate our approach, we conducted a case study with a commercial quadrotor, called AR. Drone. The target has ARM9 processor, Linux 2.6.27, various types of sensors including accelerometer, gyrometer and ultrasound altimeter, and electric motors for driving high efficiency propellers. We implemented a prototype of FDR-CPS in the quadrotor’s control program, and also constructed a simple knowledge base with several criteria needed for diagnosing some specific faults. One of them is that “if the quadrotor keeps moving up for more than 30 s while the setpoint value does not change, there is a stuck-at fault at ultrasound altimeter.”

Initially, the quadrotor was configured to be hovering 50 cm high from the ground. We then injected a stuck-at fault into the ultrasound altimeter so that it tells the aircraft is 30 cm high from the ground. After this, we could observe the target kept moving up to reach the setpoint. FDR-CPS detected this fault by using a time-out mechanism and the knowledge base. FDR-CPS then started its recovery operation and the aircraft immediately returned to the original correct setpoint.

## 5 Conclusion

In this paper we proposed a software framework of fault detection and recovery for cyber-physical systems, called FDR-CPS (Fault Detection and Recovery for CPS). FDR-CPS focuses on specific faults related to sensors and actuators, which seem to be the likely cause of critical system failures such as system hangs and crashes. We divided such critical failures into four classes and then presented the design and implementation of FDR-CPS that can successfully handle the four classes of critical failures. We also described the quadrotor case study how FDR-CPS can be applied in real world applications.

**Acknowledgments** This work was supported partly by Mid-career Researcher Program through NRF (National Research Foundation) grant NRF-2011-0015997 funded by the MEST (Ministry of Education, Science and Technology), partly by the IT R&D Program of MKE/KEIT [10035708, “The Development of CPS (Cyber-Physical Systems) Core Technologies for High Confidential Autonomic Control Software”], partly by Seoul Creative Human Development Program (HM120006), and partly by the MKE (The Ministry of Knowledge Economy), Korea, under the CITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0401-13-1009) supervised by the NIPA (National IT Industry Promotion Agency).

## References

1. Kadav A, Renzelmann MJ, Swift MM (2009) Tolerating Hardware Device Failures in Software. In: Proceedings of the ACM SIGOPS 22nd symposium on operating systems principles, pp 59–72
2. Graham S (2002) Writing drivers for reliability, robustness and fault tolerant systems. <http://www.microsoft.com/whdc/archive/FTdrv.ms>
3. Ploski J, Rohr M, Schwenkenberg P, Hasselbring W (2007) Research issues in software fault categorization. ACM SIGSOFT Softw Eng Notes 32(6): 1–8 (article No. 6)
4. Ball T, Bounimova E, Cook B, Levin V, Lichtenberg J, McGarvey C, Ondrusek B, Rajamani SK, Ustuner A (2006) Thorough static analysis of device drivers. In: Proceedings of the 1st ACM SIGOPS/EuroSys European conference on computer systems, pp 73–85
5. Candea G, Fox A (2003) Crash-only software. In: Proceedings of HotOS IX: The 9th workshop on hot topics in operating systems

# Sample Adaptive Offset Parallelism in HEVC

Eun-kyung Ryu, Jung-hak Nam, Seon-oh Lee, Hyun-ho Jo and Dong-gyu Sim

**Abstract** We propose a parallelization method for SAO, in-loop filter of HEVC. SAO filtering proceeds along CTB lines and there exists data dependency between inside and outside of CTB boundaries. Data dependency makes data-level parallelization hard. In this paper, we equally divided an entire frame into sub regions. With a little amount of memory, proposed method shows 1.9 times of performance enhancement in terms of processing time.

**Keywords** Sample adaptive offset · SAO · SAO parallelism · HEVC parallelism · Multi-core parallelism · In-loop filter

## 1 Introduction

Needs for the realistic video service, such as higher resolution and higher quality, are increasing with the developments of multimedia-related hardware and software technologies. In addition, requirements for the advanced video coding standard have arisen on the multimedia industrial market. Dependent on the demands, Joint

---

E. Ryu · J. Nam · S. Lee · H. Jo · D. Sim (✉)  
Computer Engineering, Kwangwoon University, Seoul, Republic of Korea  
e-mail: dgsim@kw.ac.kr

E. Ryu  
e-mail: dms0314@kw.ac.kr

J. Nam  
e-mail: qejixfyza@kw.ac.kr

S. Lee  
e-mail: seon-oh@kw.ac.kr

H. Jo  
e-mail: idjhh@kw.ac.kr

Collaborative Team on Video Coding (JCT-VC) which is composed by ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG) has been standardizing HEVC, whose target coding efficiency is twice better than that of H.264/AVC [1].

To introduce the state-of-art video codec to various multimedia industries as soon as standardization process of HEVC is completed, researches on algorithms lowering computational complexity and methods of optimization for real time processing are indispensable. Even during standardizing process, parallel processing schemes for real-time processing have been contributed and accepted for HEVC. Thinking of trends on multi-core and various mobile devices and high complexity of HEVC, researches on parallelization for HEVC should be done.

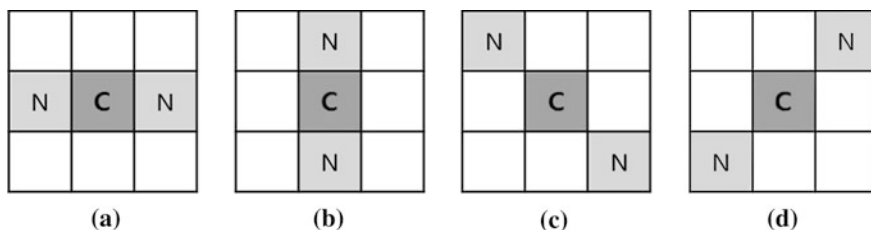
Lowering quantization error of reconstructed frames, SAO has dependency in pixel decoding process. To develop SAO with parallel scheme, we should think of a feature of SAO dependency.

In this paper, we propose a parallel core assigning method which resolves CTB-level dependency of SAO. We assign the same number of CTBs into each core.

This paper is organized as follows. In Sect. 2, SAO method is introduced. In Sect. 3, dependency between adjacent CTBs is analyzed, and the proposed method is shown. Experimental results and evaluation are given in Sect. 4. Finally, conclusions are stated in Sect. 5.

## 2 Sample Adaptive Offset for HEVC

New in-loop filter, SAO has been adopted for HEVC. Different to the other tools of video codec, quantization process causes data loss between original and reconstructed videos. Larger transforms and longer-tap interpolation of HEVC than those of the other video codecs can introduce ringing artifacts due to quantization error of transformed coefficients and loss of high frequency components. Ringing artifacts or quantization errors are compensated with parameters within compressed bitstream during SAO process such as Edge Offset (EO) and Band Offset (BO) (Fig. 1).



**Fig. 1** Four types of edge offset class: **a** 1-D 0-degree, **b** 1-D 90-degree, **c** 1-D 135-degree **d** 1-D 45-degree

**Table 1** Category conditions

Category	Conditions
1	$C < N1 \ \&\& \ C < N2$
2	$(C < N1 \ \&\& \ C == N2) \ \parallel \ (C == N1 \ \&\& \ C < N2)$
3	$(C > N1 \ \&\& \ C == N2) \ \parallel \ (C == N1 \ \&\& \ C > N2)$
4	$C > N1 \ \&\& \ C > N2$
0	None of the above

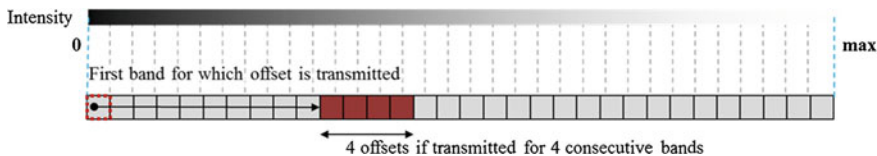
EO reduces ringing artifacts and quantization errors by making local valleys and concave corners become smoother. Four classes are used for EO and the classification is based on edge direction derived from comparison between current and neighboring pixel values on the encoder side. Side information for EO indicates which one of four classes is applied and instructs offset values to be added.

For a given EO class, each sample inside the CTB is classified into one of five categories. The current sample value, labeled as ‘C’, is compared with its two neighbors along the selected 1D pattern. The classification rules for each sample are summarized in Table 1. Categories 1 and 4 are associated with a local valley and a local peak along the selected 1D pattern, respectively. Categories 2 and 3 are associated with concave and convex corners along the selected 1D pattern, respectively [2]. If the current CTB does not belong to EO categories 1–4, then it are category 0, and SAO is not applied. Category determining process needs neighboring samples which are not processed by SAO and this referencing structure derives a data dependency problem for multi-core parallelization.

BO reduces quantization errors by adding offset values to all samples of the selected bands. The starting band position and offsets of four consecutive bands determined on the encoder side is signaled to decoder. One offset is added to all samples of the same band. In case of BO, neighboring samples are not required. So, we need not consider data dependency between adjacent CTBs for BO parallelization [2] (Fig. 2).

First of all, data dependency derived from coding mode using neighboring samples or context should be thought for the data-level parallelization of video codec. To parallelize SAO, it is important to know that the sample classification is based on comparison between current samples and neighboring samples whose values are not filtered by SAO.

Figure 3 shows data dependency between adjacent CTBs for EO of SAO. Neighboring samples for EO used for reference should be not processed by SAO. In case of raster order SAO process, boundary samples of CTBs should be stored



**Fig. 2** Bands

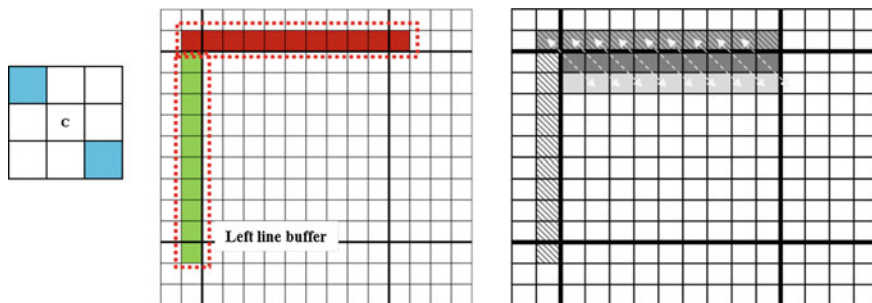


Fig. 3 Data dependency between adjacent CTBs for EO of SAO

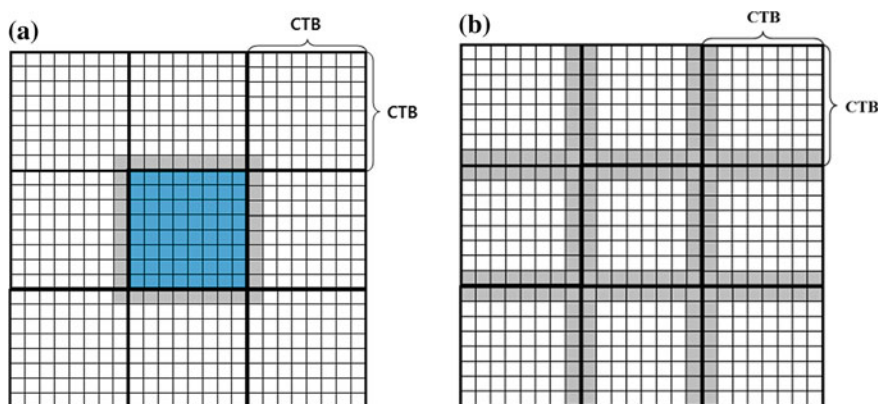


Fig. 4 Four directional data dependency for parallel processing. a Data dependency of a CTB. b Line buffer in a frame to remove the data dependency

until SAO filtering for all of the adjacent CTBs is completed. HM software using raster order for SAO utilizes additional memory buffer to store left and upper boundary samples.

### 3 Proposed Method

Different to a case of raster order SAO process which has left and upper directional data dependency, there exists four directional data dependency for parallel processing of SAO. As shown in Fig. 4a, decoder should store four directional neighboring samples for all of CTBs and it allows parallel processing.

Conventional data-level parallelizing methods for SAO use temporal memory buffer storing an entire frame [3]. Amount of memory for an entire frame causes overhead on the decoder side. For the reason, we used line buffer memory which

temporally stores two pixel lines composed of inside and outside of CTB boundaries in both vertical and horizontal directions, as shown in Fig. 4b. The line buffer needs 1/16 amount of memory than an entire frame buffer, which allows parallel SAO filtering on the decoder side. In this paper, we selected a method dividing a frame into same size of areas among the conventional data-level parallelizing methods.

## 4 Experiment Results

### 4.1 Experiment Conditions

To evaluate the performance of the proposed method, three test sequences from common test conditions of HEVC were used. Test sequences were encoded using all intra, low delay (LD) and random access (RA) modes. Values of quantization parameters used to encode are 22, 27, 32 and 37 Table 2.

**Table 2** Experiment conditions

Operating system	Microsoft Windows 7 (32 bit)
CPU	Intel Core™2 Quad CPU
RAM	4 GB
Profile	Main
Reference software	HM 9.0
OpenMP version	OpenMP 2.0
Compiler	MS Visual Studio 2008 Release mode
#Iteration	3
Measurement	Decoding time (ms)
QP	22, 27, 32, 37

**Table 3** Time saving in all intra mode

Sequence	QP	Reference	Proposed	Speed up
BasketballDrive	22	2544.9	1170.6	2.2 x
	27	2783.0	1228.4	2.3 x
	32	2182.5	1050.1	2.1 x
	37	1360.6	760.0	1.8 x
KristenAndSara	22	1501.8	639.8	2.3 x
	27	1266.5	549.1	2.3 x
	32	798.7	414.0	1.9 x
	37	409.3	262.3	1.6 x
NebutaFestiva	22	1632.9	877.0	1.9 x
	27	2685.2	1226.3	2.2 x
	32	3482.1	1448.5	2.4 x
	37	2660.1	1323.3	2.0 x
Average				2.1 x



**Table 4** Time saving in low delay mode

Sequence	QP	Reference	Proposed	Speed up
BasketballDrive	22	2282.1	1035.9	2.2 x
	27	1274.9	730.8	1.7 x
	32	459.8	317.4	1.4 x
	37	217.2	182	1.2 x
KristenAndSara	22	393.3	214.7	1.8 x
	27	178.6	114.4	1.6 x
	32	59.4	51.5	1.2 x
	37	42.8	41.5	1.0 x
NebutaFestiva	22	1431.8	732.8	2.0 x
	27	1042.4	477.4	2.2 x
	32	1183.2	526.5	2.2 x
	37	839.9	448.6	1.9 x
Average				1.7 x

**Table 5** Time saving in random access mode

Sequence	QP	Reference	Proposed	Speed up
BasketballDrive	22	2118.3	964	2.2 x
	27	835.2	493.3	1.7 x
	32	341	195	1.7 x
	37	200	132.9	1.5 x
KristenAndSara	22	340.1	165.8	2.1 x
	27	149.4	77.5	1.9 x
	32	88.8	52.8	1.7 x
	37	43	31	1.4 x
NebutaFestiva	22	1631.4	858.9	1.9 x
	27	2545	1007.9	2.5 x
	32	2974.6	1070.7	2.8 x
	37	1036.4	462.8	2.2 x
Average				1.9 x

## 4.2 Experimental Results

Tables 3, 4 and 5 show performance of HM 9.0 and proposed method in terms of processing time of only SAO module for all intra, low delay and random access coding mode. Generally, lower QP values cause more SAO enable CTBs while the higher QP values derive the less number of SAO enable CTBs. So, reduced processing time for higher QP cases shows higher performance enhancement. The proposed method shows 1.9 times of performance enhancement in terms of average processing time.

## 5 Conclusion

We propose a parallelization method for SAO, in-loop filter of HEVC. SAO filtering for CTBs needs neighboring sample information, which derives data dependencies. For the parallelization, though conventional methods use temporal memory buffer for an entire frame to resolve data dependency problem of SAO, which causes memory overhead problems. In this paper, we used only 1/16 amount of memory than a frame buffer for data-level parallelization. The proposed method of data-level parallel processing for multi-core processors shows 1.9 times than the conventional method of HM 9.0 in terms of processing time.

**Acknowledgments** This research was partly supported by the Samsung Electronics, and partly supported by the Ministry of Knowledge Economy (MKE), Korea, under the Information Technology Research Center (ITRC) support program (NIPA-2012-H0301-12-1011) supervised by the National IT Industry Promotion Agency (NIPA).

## References

1. Wiegand T, Ohm J-R, Sullivan GJ, Han W-J, Joshi R, Tan TK, Ugur K (2010) Special section on the joint call for proposals on high efficiency video coding (HEVC) standardization. *IEEE Trans Circuits Syst Video Technol* 20(12):1661–1666
2. Fu C-M, Alshina E, Alshin A, Huang Y-W, Chen C-Y, Tsai C-Y, Hsu C-W, Lei S-M, Park JH, Han W-J (2012) Sample adaptive offset in the HEVC standard. *IEEE Trans Circuits Syst Video Technol* 22(12):1755–1764
3. Chi CC, Alvarez-Mesa M, Juurlink B, Clare G, Henry F, Pateux S, Schierl T (2012) Parallel scalability and efficiency of HEVC parallelization approaches. *IEEE Trans Circuits Syst Video Technol* 22(12):1827–1838

# Comparison Between SVM and Back Propagation Neural Network in Building IDS

Nguyen Dai Hai and Nguyen Linh Giang

**Abstract** Recently, applying the novel data mining techniques for anomaly detection-an element in Intrusion Detection System has received much research alternation. Support Vector Machine (SVM) and Back Propagation Neural (BPN) network has been applied successfully in many areas with excellent generalization results, such as rule extraction, classification and evaluation. In this paper, we use an approach that is entropy based analysis method to characterize some common types of attack like scanning attack. A model based on SVM with Gaussian RBF kernel is also proposed here for building anomaly detection system. BPN network is considered one of the simplest and most general methods used for supervised training of multilayered neural network. The comparative results show that with attack scenarios that we create and through the differences between the performance measures, we found that SVM gives higher precision and lower error rate than BPN method.

**Keywords** Back propagation neural network • Denial of service • Entropy • RBF kernel • Support Vector Machine

---

N. D. Hai (✉)

School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam  
e-mail: haidnguyen0909@gmail.com

N. L. Giang

Department of Communication and Computer Networks, Hanoi University of Science and Technology, Hanoi, Vietnam  
e-mail: giangnl@soict.hut.edu.vn

## 1 Introduction

Intrusion Detection Systems (IDS) have become popular tools for identifying anomalous and malicious activities in computer systems and networks. Anomaly detection is a key element of intrusion detection and other detection system in which perturbations from normal behavior suggest the presence of attacks, defects etc. Anomaly detection is performed by building a model that contains metrics derived from system operation and flagging any observation as intrusive that has a significant deviation from the model.

In this paper, we consider two common types of attack including Bandwidth flood like ICMP flood, TCP flood, and Scanning attack like TCP scan, Null scan. There are two problems we need to resolve, one problem is that the amount of traffic data does not allow real-time analysis of details and other is that the specific characteristics events are not known in advance. So, there is a need for analysis methods that are real-time capable and can handle large amounts of traffic data. We will apply an entropy based approach proposed in [1], that determines and reports entropy contents of traffic parameters such as IP addresses, port. Changes in the entropy content indicate a massive network event.

In order to improve the performance of Anomaly-based IDS, many Artificial Intelligence methods such as neural network have been widely used. These methods are based on an empirical and have some disadvantages such as local optimal solution, low convergence rate, and epically poor generalization when the number of class samples is limited [2]. In 1990s, Support Vector Machine (SVM) was introduced as a new technique for solving a variety of learning, classification and prediction problems. SVM is a set of related supervised learning methods used for classification and regression. SVM originated as an implementation of Vapnik's [3] structural risk minimization principle which minimizes the generalization error. The paper aims at investigating the capabilities of SVM when compared with BPN network model for building anomaly-based Intrusion Detection System.

This paper is organized as follows: [Section 2](#) explains the fundamental idea related to entropy based analysis method and characteristics of two types of attack. [Section 3](#) deals with the foundation of SVM and BPN network, in this section, we have procedures for anomaly-based intrusion detection. A discussion of implementation details and experimental results is given in [Sect. 4](#). [Section 5](#) give concluding remarks and discuss related work.

## 2 Entropy Based Analysis Method

In this paper, we consider two common types of attack including Bandwidth flood and Scanning Attack. In bandwidth flood, a small number of hosts send large amounts of traffic to a single destination. So, characteristic of this type of attack is that number of packets in some particular protocol such as TCP, UDP, ICMP

increase significantly. We can use number of packets in TCP, UDP and ICMP protocols as attributes of input data of classifier.

Scan attacks can be characterized by using entropy based analysis method. Generally speaking entropy is a measure of how random a dataset is. The more random it is, the more entropy it contains. Entropy contents of a sequence of values can be measured by representing the sequence in binary form and then using data compression on that sequence. The size of the compressed object corresponds to the entropy contents of the sequence. If the compression algorithm is perfect in the mathematical sense, the measurement is exact. The change in IP addresses characteristics seen on flow level. When scanning hosts try to connect to a lot of other hosts, the source IP addresses of the scanning hosts will be seen in many flows and since they are relatively few hosts, the source IP address fields will contain less entropy per address seen then normal traffic. On the other hand the target IP addresses seen in flows will be much more random than in normal traffic. A similar thing happens on the port level. If an attacker scans for a specific vulnerability, these scans often have to go to a specific target port. If scanning traffic with this characteristic becomes a significant component of the overall network traffic, the entropy contents of the destination port fields in flows seen in the network will decrease significantly. So IP address and port entropy change can be used to detect scanning attack.

We compared three different lossless compression methods, the well-known bzip2 [4] and gzip [5] compressors as well as the LZO (Lempel–Ziv–Oberhumer) [6] real-time compressor. We did not consider lossy compressors. Bzip2 is slow and compresses very well, gzip is average in all regards and lzo is fast but does not compress well. Direct comparison of the three compressors on network data shows that while the compression ratios are different, the changes in compressibility are very similar. Because of its speed advantage LZO was selected as preferred algorithm for our work.

In our experiments we choose one minute as measurement interval length. We collect packets in each time interval, create string arrays containing IP addresses, ports and then by using LZO compression algorithm we get compression ratio. These ratios are represented as content entropy of IP addresses, ports (Table 1).

**Table 1** Compression rate represents entropy

---

```
# Compresses all attributes


---


my $csip    = compress $strsip;    #contain source IP addresses
my $csport  = compress $strsport;  #contain source ports
my $cdip    = compress $strdip;    #contain destination IP addresses
my $cdport  = compress $strdport;  #contain destination ports
# get ratio ($strsip,$strsport,$strdip,$strdport);

$this-> {_ratio_sip} = length($csip)/length($strsip);
$this-> {_ratio_sport} = length($csport)/length($strsport);
$this-> {_ratio_dip} = length($cdip)/length($strdip);
$this-> {_ratio_dport} = length($cdport)/length($strdport);
```

---

### 3 Classification Methods

We will describe briefly two classification methods that we use in this paper.

#### 3.1 Overview of SVM

Support Vector Machines (SVMs) are a useful technique for data classification. A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one “target value” (i.e. the class label) and several “attributes” (i.e. the feature or observed variables). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes. Given a training set of instance-label pairs  $(x_i, y_i)$ ,  $i = 1, \dots, l$  where  $x_i \in R^n$  and  $y \in \{1, -1\}^l$ , SVM require the solution of the following optimization problem:

$$\min_{w, b, \xi} \frac{1}{2} w^T w + C \sum_{i=1}^l \xi_i \text{ Subject to } y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

The decision function is formulated in terms of these kernels:

$$f(x) = \text{sign} \left( \sum_{i=1}^n \alpha_i^* y_i (\phi(x) \phi(x_i) - b^*) \right) = \text{sign} \left( \sum_{i=1}^n \alpha_i^* y_i (K(x, x_i) - b^*) \right).$$

In general, the Gaussian RBF kernel is a reasonable first choice. This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear.

#### 3.2 Overview of Back Propagation Neural Network

Back Propagation Neural (BPN) is a common method of training artificial neural networks so as to minimize the objective function. Back propagation works by approximating the non-linear relationship between the input and the output by adjusting the weight values internally. A BPN model consists of an input layer, one or more hidden layers, and output layer. There are two parameters including learning rate ( $0 < \alpha < 1$ ) and momentum ( $0 < \eta < 1$ ) required to define by user. The theoretical results showed that one hidden layer is sufficient for a BPN to approximate any continuous mapping from the input patterns to the output patterns to an arbitrary degree freedom [7]. The selection nodes of hidden layers primarily affect the classification performance.

## 4 Empirical Illustration

### 4.1 Data Set

In this study, the measured attributes are (11 attributes): (1–2) Entropy-compression rate of the source/destination IP address and (3–4) source/destination port, (5) number of packets, (6–7) total/average size of the packets, (8) standard deviation of packet size and (9–11) number of TCP/UDP/ICMP packets. So each instance will be represented by a vector including 11 attributes and the input of each classifier is differential vector of current vector and reference vector which refer to normal state.

### 4.2 Experiment

In this section, we will test the system's ability of detecting anomaly-based intrusion activities using two methods: SVM and BPN. We will proceed on the four attack scenarios including ICMP flood, TCP flood, UDP flood and port scan. Each attack will change significantly the number of ICMP, TCP, UDP packets and entropy.

#### 4.2.1 Testing Environment

The system was tested on virtual LAN 100 Mps environment using VMware tool, including two Window XP computers and a Ubuntu computer installed the Anomaly IDS. These computers are connected to each other through a virtual switch.

#### 4.2.2 Testing Scenarios

Two Window XP computers implement TCP flood, UDP flood, ICMP flood refer to bandwidth flood attacks using tools like hping3, udpflood.exe, ping respectively or scan port in range 1–300 on Ubuntu computer installed anomaly IDS. Our program will collect and analysis packets in order to detect anomalous in traffic. Fig. 1.

As we can see, the characteristic of the port scan attack is that entropy source port, destination change significantly compared with the normal state.

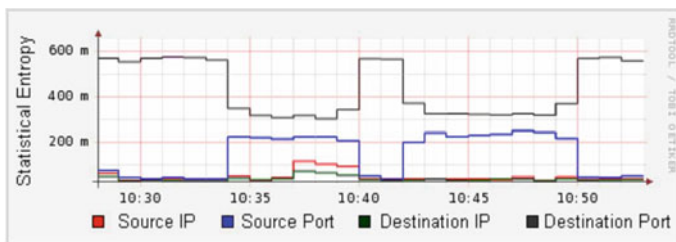


Fig. 1 Statistical entropy when port scanning

### 4.2.3 Experimental Result

With the attack scenarios as above, we tested IDS’s ability of detecting some common attack types using two methods SVM and BNP.

#### (1) SVM method

We use **Algorithm::SVM** library [8] with 5-fold cross-validation to find the best parameter  $C$  and  $\gamma$  [3]. The results confirmed that the classification precision of the SVM with RBF kernel was high as 98 % when  $\gamma$  and  $C$  are 0.000015 and 2, respectively. Then we use the best parameter  $C$  and  $\gamma$  to train the whole training set, we have 9 support vectors, the outputs are:

Accuracy = 99 % (49/50) (classification), Iteration = 123.

#### (2) BPN method

We use **Algorithm::NNFlex** library to implement BPN method. In the training model, there are 11 attributes as input nodes, four hidden nodes and one output node standing for normal (+1) or abnormal (−1).  $>0$  as a target value of the normal and  $<0$  as a target value of abnormal. We adopted the range of 0.6–0.9 and 0.1–0.4 to be the decisions of learning rate and momentum. Through several trail-and-error experiments, the structure of 11-4-1 model had the best performance. The outputs are:

Accuracy = 92 % (46/50), Iteration = 411, Time training: 45 s.

We conducted simulating attacks with different scenarios then collecting 200 data samples used for predicting process. The prediction result is listed in Table 2. SVM method shows the best overall prediction accuracy level at 99 % and BPN method, shows the best overall accuracy level at 96 %.

Here we define the attack detection rate and false-positive rate as [9].

Table 2 Detected attack rate and false-positive rate

Methods	Attack		Normal		Attack detection rate (%)	False positive rate (%)
	Detected	Real	Misclassified	Real		
SVM	50	50	2	150	100	1.3
BPN	48	50	6	150	96	4



## 5 Conclusion

This study constructed an Anomaly-based intrusion detection model based on Support Vector Machine and Back Propagation Neural Network. First, we introduce an approach based on entropy to characterize some kinds of scanning attack. Second, we simply define the SVM and BPN methods. Third, a proposed for building an Intrusion Detection System using SVM and BPN is discussed. Finally, experimental results proved the validity of our proposes and we found that SVM method has better accuracy with lower misclassification rate than MLP method based on these results.

## References

1. Nychis G, Sekar V, Andersen DG, Kim H, Zhang H (2008) An empirical evaluation of entropy-based traffic anomaly detection. In: Proceedings of the 8th ACM SIGCOMM conference on internet measurement
2. Yuan, SF, Chu FL (2006) Support vector machine based on fault diagnosis for turbo-pump rotor. *Mech Syst Signal Process* 20:939–952
3. Ben-Hur A, Weston J (2010) A user's guide to support vector machines. *Methods Mol Biol* 609:223–239
4. The bzip2 and libbzip2 official home page. <http://sources.redhat.com/bzip2/>
5. The gzip home page. <http://www.gzip.org/>
6. <http://www.oberhumer.com/opensource/lzo/>. LZO compression library
7. Randall SS, Dorsey RE (2000) Reliable classification using neural networks: a genetic algorithm and back propagation comparison. *Decis Support Syst* 30:11–22
8. <http://search.cpan.org/~lairdm/Algorithm-SVM-0.13/>
9. Liao Y, Vermuri VR (2002) Use of k-nearest neighbor classifier for intrusion detection. *Comput Secur* 21:439–448
10. Chang CC, Lin CJ (2009) LIBSVM: a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. 18 Nov 2009
11. Fausett L (1994) *Fundamentals of neural networks: architectures, algorithms and applications*. Prentice-Hall, New Jersey

# Anomaly Detection with Multinomial Logistic Regression and Naïve Bayesian

Nguyen Dai Hai and Nguyen Linh Giang

**Abstract** Intrusion Detection by automated means is gaining widespread interest due to the serious impact of Intrusions on computer system or network. Several techniques have been introduced in an effort to minimize up to some extent the risk associated with Intrusion attack. In this paper, we have used two novel Machine Learning techniques including Multinomial Logistic Regression and Naïve Bayesian in building Anomaly-based Intrusion Detection System (IDS). Also, we create our own dataset based on four attack scenarios including TCP flood, ICMP flood, UDP flood and Scan port. Then, we will test the system's ability of detecting anomaly-based intrusion activities using these two methods. Furthermore we will make the comparison of classification performance between the Multinomial Logistic Regression and Naïve Bayesian.

**Keywords** DoS · Logistic regression · Naïve Bayesian · Intrusion detection system

## 1 Introduction

Intrusion Detection is a process of gathering intrusion related knowledge that occurred in the computer networks or systems and analyzing them for detecting future intrusions. Intrusion Detection can be divided into two categories: Anomaly

---

N. D. Hai (✉)

School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam  
e-mail: haidnguyen0909@gmail.com

N. L. Giang

Department of Communication and Computer Networks, Hanoi University of Science and Technology, Hanoi, Vietnam  
e-mail: giangnl@soict.hut.edu.vn

detection [2] and Misuse detection. The former analyses the information gathered and compares it to a defined baseline of what is seen as “normal” service behaviors, so it has ability to learn how to detect network attacks that are currently unknown. Misuse detection is based on signatures for known attacks, so it is only as good as the database of attack signatures that it uses for comparison. Misuse detection has low false positive rate, but can not detect novel attacks. However, anomaly detection can detect unknown attacks, but has high false positive rate.

The Naïve Bayesian (NB) method is based on the work of Thomas Bayesian. In Bayesian classification, we have a hypothesis that the given data belongs to a particular class. We then calculate the probability for the hypothesis to be true. This is among the most practical approaches for certain types of problems. The approach requires only one scan of the whole data.

A Multinomial Logistic Regression (MLR) model is used for data in which the dependent variable is unordered or polytomous, and independent variables are continuous or categorical predictors. This type of model is therefore measured on a nominal scale and was introduced by McFadden (1974). Unlike a binary logistic model in which a dependent variable has only a binary choice (e.g., presence/absence of a characteristic), the dependent variable in a multinomial logistic model can have more than two choices that are coded categorically, and one of the categories is taken as the reference category.

In this paper, we propose two methods MLR and NB in building anomaly-based IDS and compare the performance of two linear classifier of Naïve Bayesian (NB) and multinomial Logistic Regression (MLR) based on attack scenarios which we created, and search for the characteristics of the data that determine the performance. The comparison between LR and MNB has been studied theoretically by Ng and Jordan (2002).

This paper is organized as follows: [Sect. 2](#) deals with the description of data set for our experiment. [Section 3](#) deals with foundation of methods including naïve Bayesian, multinomial logistic regression, In this section we will consider the problem of applying the two methods in building anomaly-based IDS. In [Sect. 4](#), we give an illustration and experimental results with four attack scenarios. It help in understanding of this procedure, a demonstrative case is given to show the key stages involving the use of the introduced concepts. [Section 5](#) is conclusion.

## 2 Dataset

Our data set is created by the following activities:

**Data collection activity:** collection attribute-value of the flow in terms of packet data (IP, port, TCP, UDP, ICMP). Based on these attributes, the program will build Profile (bin level) which contains the characteristic parameters for network traffic in a given time, including: (1–2) Entropy compression rate of the source/destination IP address, (3–4) Entropy compression rate of the source/destination port, (5) number of packets, (6) total size of the packets, (7) average size of packets, (8) standard

deviation of packet size, (9) number of TCP packets, (10) number of UDP packets and (11) number of ICMP packets.

**Statistical analysis activity:** This activity is based on the data have been analyzed from the data collected to build the corresponding bin arrays. The bin is divided into the following levels: hours, days, months correspond to the three classes of data is the current class, reference class and the differential classes:

**Cur\_bin:** represent for each instance “bin” (bin is the smallest time unit, in my program one minute). These instances is continuously created in the processes monitoring network traffic.

**Ref\_bin:** represents the reference model corresponding to one unit of time reference. Reference model is adaptably updated, based on values of Cur\_bin in the absence of intrusion detection.

**Dif\_bin:** represents the difference between the current value and the reference value and is the input of classifiers.

### 3 Methods

#### 3.1 Naïve Bayesian

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. Naïve Bayesian classifiers allow the representation of dependencies among subsets of attribute [9]. Through the use of Bayesian networks has proved to be effective in certain situations, the result obtained, are highly dependent on the assumption about the behavior of the target system, and so a deviation in these hypotheses leads to detection errors, attributable to the model considered [10]. The NB classifier work as follows: Let T be a training set of samples, each with their class labels. There are k classes  $C_1, C_2, \dots, C_k$ , each sample is represented by an n-dimensional vector  $X = \{X_1, X_2, \dots, X_n\}$ .

Given a sample X, The classifier will predict that X belongs to the class having the highest a posteriori probability, conditional on X. That is X is predicted to belong to the class C, if and only if  $P(C_i|X) > P(C_j|X)$  for  $1 \leq j \leq m, j \neq i$ .

By bayes' theorem, we have  $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$ . As P(X) is the same for all classes and only  $P(C_i)$  are not known, then it is commonly assumed that the classes are equally likely, that is,  $P(C_1) = P(C_2) = \dots = P(C_m)$  we would therefore maximize  $P(X|C_i)$ .

In order to reduce computation in evaluating  $P(X|C_i)$ . The naïve assumption of class conditional independence is made. Mathematically this means that  $P(X|C_i) \approx \prod_{k=1}^n P(X_k|C_i)$ . The probabilities  $P(X_k|C_i)$  can easily be estimated from the training set. If X is continuous-valued, then we typically assume that the values have a Gaussian distribution with a mean  $\mu$  and standard deviation  $\sigma$ . So that

$P(X_k|C_i) = g(X_k, \mu_{ci}, \sigma_{ci})$ . We need to compute  $\mu_{ci}, \sigma_{ci}$  in training stage. In order to predict the class label of  $X$ ,  $P(X|C_i)P(C_i)$  is evaluated for each class  $C_i$ . The classifier predicts that the class label of  $X$  is  $C_i$  if and only if it is the class that maximizes  $P(X|C_i)P(C_i)$ .

### 3.2 Multinomial Logistic Regression

A multinomial logistic regression model is used for data in which the dependent variable is unordered or polytomous, and independent variables are continuous or categorical predictors. This type of model is therefore measured on a nominal scale and was introduced by McFadden (1974). Unlike a binary logistic model in which a dependent variable has only a binary choice (e.g., presence/absence of a characteristic), the dependent variable in a multinomial logistic model can have more than two choices that are coded categorically, and one of the categories is taken as the reference category. This study used “0” (normal) as the reference category. Suppose  $y_i$  is the dependent variable with five categories for individual connection  $i$ -th, and the probability of being in category  $s$  ( $s = “1”$  [TCP flood], “2” [ICMP flood], “3” [UDP flood], “4” [Scan Port]) can be denoted  $\pi_i^{(s)} = \Pr(y_i = s)$  with the chosen reference category,  $\pi_i^{(0)}$ . Then, for a simple model with one independent variable  $x_i$ , a multinomial logistic regression model with logit link can be represented as:

$$\log\left(\frac{\pi_i^{(s)}}{\pi_i^{(0)}}\right) = \beta_0^{(s)} + \beta_1^{(s)}x_i, s = 1, 2, 3, 4.$$

An alternative way to interpret the effect of an independent variable,  $x$ , is to use predicted probabilities  $\pi_i^{(s)}$  for different of  $x$ :

$$\pi_i^{(s)} = \frac{\exp(\beta_0^{(s)} + \beta_1^{(s)}x_i)}{1 + \sum_{k=1}^4 \exp(\beta_0^{(k)} + \beta_1^{(k)}x_i)}.$$

Then, the probability of being in the reference category, “0” (normal), can be calculated by subtraction:

$$\pi_i^{(0)} = 1 - \sum_{k=1}^4 \pi_i^{(k)}$$

## 4 Experiment and Results

In this section, we summarize our experimental results to detect network intrusion detections using Naïve Bayes and Multinomial Logistic Regression over dataset we created based on four attack scenarios including: TCP flood, ICMP flood, UDP flood and Port Scan.

### 4.1 Purpose of Study

The objective of this study is to detect some common attack types in computer systems and networks. We furthermore make the comparison of classification performance between the NB and MLR model.

### 4.2 Dataset

In this study, the measured attributes are (in particular, 11 attributes): entropy compression rate of the source/destination IP address and source/destination port, number of packets, total/average size of the packets, standard deviation of packet size and number of TCP/UDP/ICMP packets, So each instance will be represented by a vector including 11 attributes and the input of each classifier is differential vector of current vector and reference vector which refer to normal state (Table 1).

### 4.3 Experiment

We will test the system's ability of detecting anomaly-based intrusion activities using two methods: Naïve Bayes and Multinomial Logistic Regression. We will proceed on the four attack scenarios including ICMP flood, TCP flood, UDP flood and port scan. Using with each attack will change significantly the number of ICMP, TCP, UDP packets and entropy source/target.

**Table 1** Number of examples in dataset we created

Attack types	Training samples
Normal	110
TCP flood	205
ICMP flood	200
UDP flood	150
Scan port	180

### 4.3.1 Testing Environment

The system was tested on virtual LAN 100 Mps environment using VMware tool, including two Window XP computers and a Ubuntu computer installed the Anomaly IDS. These computers are connected to each other through a virtual switch.

### 4.3.2 Testing Scenarios

Two Window XP computers implement TCP flood, UDP flood, ICMP flood refer to bandwidth flood attacks using tools like hping3, udpflood.exe, ping respectively or scan port in range 1–300 on Ubuntu computer installed anomaly IDS. Our program will collect and analysis packets in order to detect anomalous in traffic.

### 4.3.3 Experimental Results

A “confusion matrix” is sometime used to represent the result of, as shown in Table 2 (Naïve Bayes) and Table 3 (Multinomial Logistic Regression). The advantage of using this matrix is that is not only tells us how many got misclassified but also what misclassification occurred. We define the Accuracy, Detection rate and false-alarm:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad Detection - rate = \frac{TP}{TP + FP}$$

$$False - Alarm = \frac{FP}{FP + TN}$$

FN: False Negative, TN: True Negative, TP: True Positive and FP: False Positive (Table 4).

**Table 2** Confusion matrix for naïve bayes

Actual	Predicted normal	Predicted TCP flood	Predicted ICMP flood	Predicted UDP flood	Predicted scan port	Accuracy (%)
Normal	110	0	0	0	0	100
TCP flood	1	201	0	0	3	98
ICMP flood	2	0	197	1	0	98.5
UDP flood	0	0	1	147	2	98
Scan port	1	2	1	4	172	95.6

**Table 3** Confusion matrix for multinomial logistic regression

Actual	Predicted normal	Predicted TCP flood	Predicted ICMP flood	Predicted UDP flood	Predicted scan port	Accuracy (%)
Normal	110	0	0	0	0	100
TCP flood	0	204	0	0	1	99.5
ICMP flood	1	0	203	1	0	99
UDP flood	0	0	0	149	1	99.3
Scan port	1	0	0	1	177	98.33

**Table 4** Comparison between BN and MLR

	Naïve bayes		Multinomial logistic regression	
	Detection rate	False alarm	Detection rate	False alarm
Normal	100	1	100	1
TCP flood	99	0.33	100	0
ICMP flood	98.99	0.4	100	0
UDP flood	96.7	0.625	98.7	0.66
Port scan	97	0.45	98.8	0.5

## 5 Conclusion

This study constructed an Anomaly-based Intrusion Detection Model based on Naïve Bayes and Multinomial Logistic Regression algorithm. We also experiment IDS's ability of detection using both these methods in the data sets that we created based on four attack scenarios including ICMP flood, UDP flood, TCP flood and Scan Port. The experimental results show that both two methods give very high accuracy and could be applied in practice. However, this is still only the initial test, and more research is needed, in the future we will continue to improve and experiment in a real network environment.

## References

1. Lippmann R, Haines JW, Fried DJ, Korba J, Das K (2000) The 1999 DARPA off-line intrusion detection evaluation. *Comput Netw* 34:597–595
2. Stillerman M, Marceau C, Stillman M (1999) Intrusion detection for distributed systems. *Commun ACM* 42(7):62–69
3. Chang CC, Lin CJ (2009) LIBSVM: a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. 18th November 2009
4. Anderson J (1980) Computer security threat monitoring and surveillance. James P. Anderson Co, Washington
5. Yu Y, Hao H (2007) An ensemble approach to intrusion detection based on improved multi-objective genetic algorithm. *J Softw* 18(6):1369–1378



6. Luo J, Bridges SM (2000) Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection. *Int J Intell Syst* 15(8):687–703
7. Barbard D, Wu N, Jajodia S (2001) Detecting novel network intrusions using bayes estimator. In: *Proceeding of the 1st SIAM international conference on data mining*
8. Kuchimanchi G, Phoha V, Balagani K, Gaddam S (2004) Dimension reduction using feature extraction methods for real-time misuse detection systems. In: *Fifth annual IEEE proceedings of information assurance workshop*, pp 195–202
9. Han J, Kamber M, (2012) *Data mining: concepts and techniques*. Elsevier, San Francisco
10. Garcia-Teodoro P, Díaz-Verdejo JE, Maciá-Fernández G, Vázquez E (2009) Anomaly-based network intrusion detection: techniques, systems and challenges. *Comput Secur* 28(1–2):18–28
11. Phoha VV (2002) *The springer dictionary of internet security*. Springer, New York
12. Vapnik VN (1999) *Statistical learning theory*. Wiley-Interscience, New York

# Implementation of Miniaturized Automotive Media Platform with Vehicle Data Processing

Sang Yub Lee, Sang Hyun Park, Duck Keun Park, Jae Kyu Lee  
and Hyo Sub Choi

**Abstract** Among the variety of vehicle technology trend issues, the biggest one is focus on automotive network system. Especially, optical network system is mentioned. The outstanding point in optical network system for vehicular environment is that realization of automotive media platform. This paper is introduced the implementation of media platform which is consist of optical network as called Media Oriented System Transport (MOST) and engraftment of vehicle data processing module collected from On-Board Diagnosis (OBD) via Controller Area Network (CAN) transformed into Extensible Marked Language (XML) schema structure. Also, for the reliability of the development, it is executed the MOST compliance test of designed platform.

**Keywords** In vehicle network system · MOST · OBD · CAN · XML · Automotive media platform · Compliance Test

---

S. Y. Lee (✉) · S. H. Park · D. K. Park · J. K. Lee · H. S. Choi  
Jeonbuk Embedded System Research Centre, Korea Electronics Technology Institute,  
Dunsan-ri, Bongdong-eup, Wanju-gun, Jeolabuk-do 565-902, Republic of Korea  
e-mail: syubleee@keti.re.kr

S. H. Park  
e-mail: shpark@keti.re.kr

D. K. Park  
e-mail: parkdk@keti.re.kr

J. K. Lee  
e-mail: jae4850@keti.re.kr

H. S. Choi  
e-mail: hschoi@keti.re.kr

## 1 Introduction

Most of people want to be experienced in high quality audio streaming service while they drive. According to customer's demands, MOST system is developed to provide an efficient and cost effective fabric to transmit audio data between any devices attached to the harsh environment of automobile. For the simply network setup, it can be obtained the convenient installation, high fuel efficiency from the usage of plastic optical fiber and high reliability without electro-magnetic problems. In realization terms of car audio system, the streaming service gathering automobile information via CAN be enables to tune the volume automatically for their status in automobile. To develop the self-contained audio system, the complex information having vehicular status is needed to be accessed and interlocked.

Section 2 is MOST network system and MOST network service is explained. Vehicle data processing module which transforms into XML data format from vehicular status information to be collected by CAN network is described in Sect. 3. In Sect. 4, developed platform and demonstration of automotive media system linked with vehicle communication module are shown. Particularly, for MOST network system, compliance test is executed whether the designed platform is working properly as a network device or not. This paper is concluded in Sect. 5.

## 2 MOST Network System

MOST is the de-factor standard for efficient and cost effective networking of automotive multimedia and infotainment system [1, 2]. The current MOST standards released MOST150, 150 means that 150 Mbps network bandwidth with quality of services is available. To satisfy the demands from various automotive applications, MOST network system provides three different message channels: control, synchronous used in streaming service and asynchronous channel only for packet data transmission. Described in Fig. 1, proposed network system is consist of MOST devices with CAN bus line.

### 2.1 MOST Network Services for External Host Controller

Network service is based on the network interface controller and provides a programming interface for the application which basically consists of the function blocks. In this paper, it comprise designed platform for transferring streaming data in synchronous channel. It contains mechanisms and routines for operating and managing the network and it ensures dynamic behavior of the network. The network service is implemented on the external host controller as shown in Fig. 2.

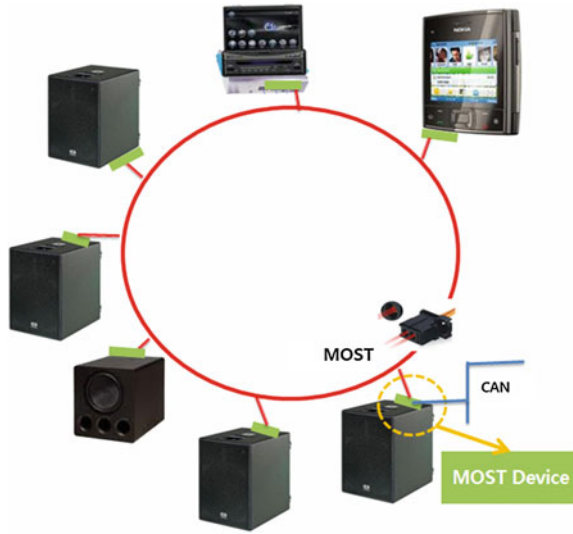


Fig. 1 Automotive media service model

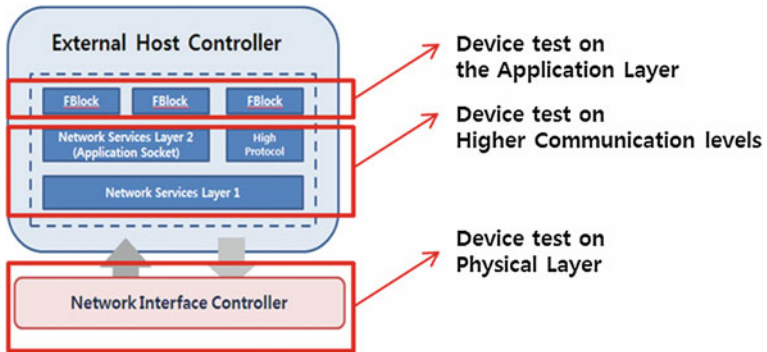


Fig. 2 The development structure applied to platform and compliance test

Being explained the network services; MOST devices have the unique interface which transfers the data between the physical layer network and processor as an external host controller [3]. Working in conjunction with the clock manager, the network port recovers the network clock for time synchronization. And then, it is executed for decoding the received data and delivered to the microprocessor. Transferred data is routed into appropriate memory destination on and off in platform.

## **2.2 MOST Compliance Test**

The integration of network system and modules is particularly challenging due to their complexity. An effective measure is to introduce a certification test process so that the devices and modules are tested before they are integrated into the system. As described in Fig. 2, MOST compliance test process is divided into three courses: device test on physical layer, higher communication levels and application layer.

With regard to the physical layer compliance, the measurement point is to check the signal characteristics for constancy. For the normal behavior, power, error, ring break diagnosis and network management, higher communication level compliance tests are performed. Distinctive features and the dynamic behaviors are tested in the application layer compliance test.

## **3 Vehicular Status Information**

### **3.1 Vehicle Data Processing Module with XML Schema**

As described in this paper, the specific module collecting and analyzing automobile data is called as the vehicle data processing module. In the processing module, vehicle data transmission and channel connection is served in data convergence engine. Existed analysis and process of vehicle information has a limit that it cannot be used in the mobile or telematics terminal, since these devices cannot support CAN communication sockets. Also, as the number of sensors connected with CAN bus are increasing, data management and transmission are faced with high bus load traffic. It causes the system breaking and errors. In order to recover these problems, XML data classification method is proposed. Data Convergence Engine (DCE) provides the process of XML type of collecting data via OBD. As depicted Fig. 3, at the first step, Data Processing Module set configuration parameter from the targeted platform and defines data format and schema to be fit XML format on. And then, from the VSF Library, the variety of the vehicle information such as automobile speed, gear shift and steering etc. are regenerated to XML regards to the type of schema. Through the schema fetcher and reader, data exchange and extraction are performed in Data Processing Module (Fig. 4).

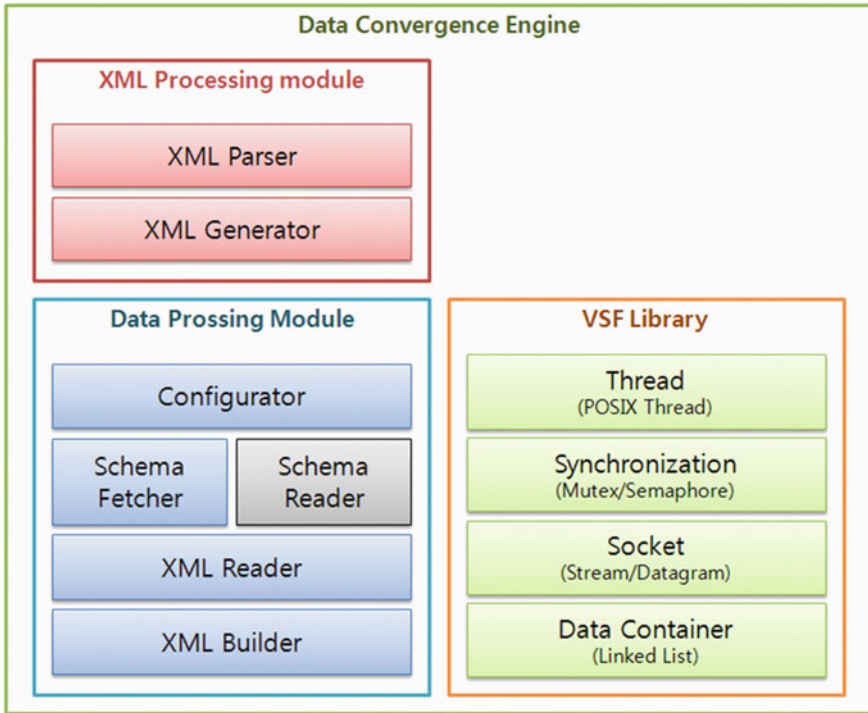


Fig. 3 Data convergence engine

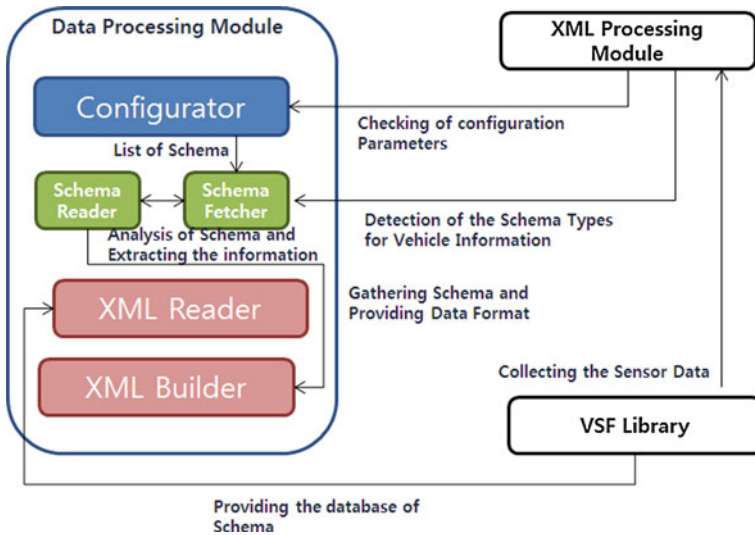


Fig. 4 The flow of data processing module

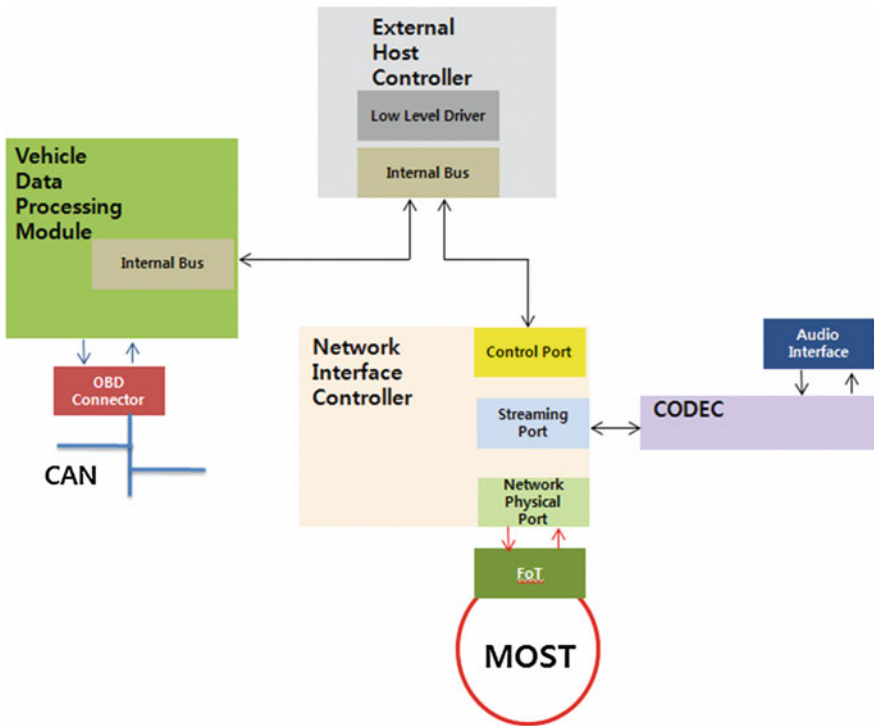


Fig. 5 System architecture of media gateway

## 4 Implementation of Automotive Media Platform

### 4.1 System Architecture

The external host controller communicates with network interface controller via I2C bus and connected with OBD via CAN bus [4, 5]. As shown below, the streaming port included in network interface controller can be used for stereo audio exchange between the network and physical audio port. The external host controller may support for managing the streaming audio exchange remotely.

As described in Fig. 5, in order to make the data connection path, Data transferred from the internal bus accesses into embedded memory space. External host controller is used for low graded microprocessor and internal local bus. It is facilitated commercial product.

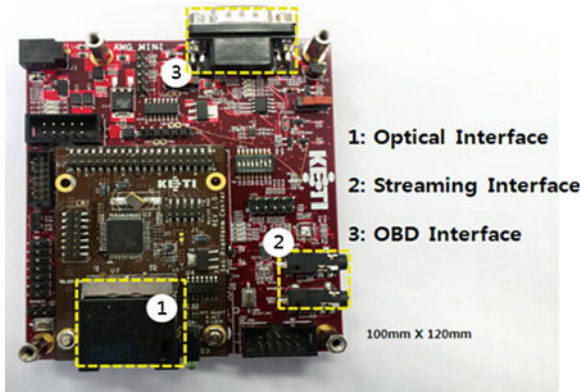


Fig. 6 Streaming service platform

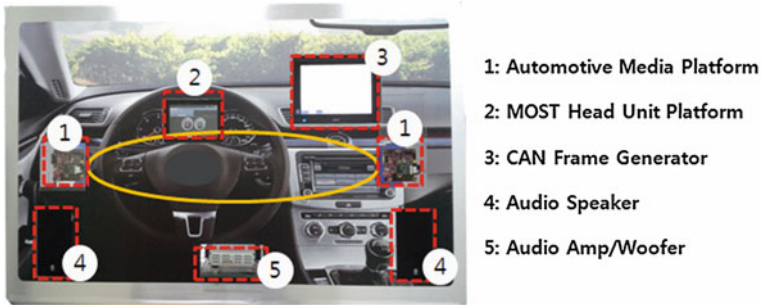


Fig. 7 System demonstration

### 4.2 Developed Media Platform

As mentioned above that external host controller having network service for network management and schema structure generation for vehicle information processing includes XML converting block. XML converting block conducts exchange of the command line used to produce assembler and C header files from the XML. Being in this method, the XML converting block has to be embedded in external host controller. With this convenient property, for other in-vehicle network devices connecting with heterogeneity communication, it is useful exchanging vehicle status information between data processing terminals which can show indication information in car. In Fig. 6, the interfaces of automotive media platform are shown. And system demonstration is built with designed platform connected with audio speaker and amplifier, vehicle parameter generator which makes the virtual automobile environment via OBD interface in Fig. 7.



**Measurement of SP2 output signal of the DUT(5\_Tmax\_Umin\_Pmax\_DTmin)**

T <sub>23°C</sub> /U <sub>typ</sub> /P <sub>opt3</sub> (max) / Duty Cycle <sub>min</sub>		Setup 1				
Parameter	Min	Max	Unit	Test Value	Result	
<b>Optical power</b>						
Optical output power Popt2	-8.5	-1.5	dBm		not tested	
<b>Data consistency</b>						
unlock ET Block		-	status		not tested	
bit error ET Block		0	error		not tested	
unlock PhLSTT		-	status		not tested	
bit error PhLSTT		0	error		not tested	
<b>Signal Integrity</b>						
Rise time 20% 80% (tr2)	-	0.5	UI	0.246	passed	
Fall time 80% 20% (tf2)	-	0.5	UI	0.199	passed	
Transferred Jitter (Jtr2) (RMS)		112	ps	10.113	passed	
Positive Overshoot		0	Error	0	passed	
Undershoot 2UI		0	Error	0	passed	
Undershoot 3UI		0	Error	0	passed	
Undershoot 4UI		0	Error	0	passed	
Undershoot 5UI		0	Error	0	passed	
Undershoot 6UI		0	Error	0	passed	
Alignment Jitter acc. To Eye MASK	-	0	Error	0	passed	
Bit Rate	147.4265	147.4854	Mbits/s	147.455	passed	

Fig. 8 The result of physical layer compliance test

```

                expected          received
                result            result      State
DUT responds correctly to all queries
DUT responds correctly to all queries
DUT ok

The DUT has passed the test
    
```

Fig. 9 The summary of higher communication level compliance test

### 4.3 Compliance Test Result

Figure 8 shows the test result for physical layer compliance for MOST150. Based on an analysis of the higher communication level, the compliance test cases were determined as network behavior, system configuration and notification method in Fig. 9.

## 5 Conclusion

This paper is introduced the development of the automotive media platform based on optical network system. Particularly, optical network system, MOST, to be optimized sound level depending on vehicle environment is satisfied with reducing the weight and ensuring the reliability for free of electro-magnetic problems. With vehicle status information, designed network platform realized and demonstrated

that self-contained MOST speaker and woofer can be controlled and tuned their volume level automatically to be served to passengers more conveniently. Especially, outstanding point for implementation is represented XML data interchange from vehicle status information. For being ensured reliability of designed platform, it is performed compliance test of physical and higher communication level.

**Acknowledgments** This work was supported by the IT R&D program of MKE/KEIT [1004091, the development of Automotive Synchronous Ethernet combined IVN/OVN and Safety Control System for the 1 Gbps class].

## References

1. Grzemba A (2011) MOST book from MOST25 to MOST150. MOST cooperation. Franzis, Deggendorf
2. Lee SY, Park SH, Choi HS, Lee CD (2012) MOST network system supporting full-duplexing communication. IEEE ICACT, Korea, pp 1272–1275
3. Lee SY, Kim BC, Choi HS, Lee CD (2012) Development of automotive media streaming device with MOST network system. SMA2012, China
4. Otto S, Rindha R, Jan L (2007) Communication in automotive system principles, limits and new trends for vehicles, airplanes and vessels. IEEE ICTON, pp 1–6
5. Godavarty S, Broyles S, Parten M (2000) Interfacing to the on-board diagnostic system. In: Proceedings of IEEE vehicular technology conference, 52nd-VTC, vol 4, pp 24–28

# Design of Software-Based Receiver and Analyzer System for DVB-T2 Broadcast System

M. G. Kang, Y. J. Woo, K. T. Lee, I. K. Kim, J. S. Lee  
and J. S. Lee

**Abstract** In this paper, a receiver and an analyzer system of Digital Video Broadcasting\_2nd Generation Terrestrial (DVB-T2) were designed using Digital Video Broadcasting\_2nd Generation Satellite (DVB-S2), and Digital Video Broadcasting\_2nd Generation Cable (DVB-C2). This software-based receiver and analyzer system were implemented by memory sharing techniques for minimization of system overloads.

**Keywords** DVB-S2 · DVB-T2 · DVB-C2 · Receiver · Analyzer

---

M. G. Kang (✉) · Y. J. Woo (✉)  
Hanshin University, #411 Hanshinda-gil, Osan-si 447-491, Korea  
e-mail: Kangmg@hs.ac.kr

Y. J. Woo  
e-mail: cosch0610@gmail.com

K. T. Lee  
Korea Electronics Technology Institute, 10FL, Electronics Center, #1599,  
Sangam-dong, Mapo-gu, Seoul, Korea  
e-mail: ktechlee@keti.re.kr

I. K. Kim  
Innodigital Co. Ltd., #907, KINS Tower, 25-1 Jeongja-dong, Bundang-Gu, Seongnam-Si  
463-782, Korea  
e-mail: ikkim@innodigital.net

J. S. Lee  
Haesung Optics Co. Ltd., 3B/3L 921 GosaeK-dong, Kwonsun-gu, Suwon-si 441-813, Korea  
e-mail: jsyi@hso.co.kr

J. S. Lee  
LG CNS Co. Ltd., LG Twin Towers, West 9F, #20, Yoido-dong,  
Youngdungpo-gu, Seoul 150-721, Korea  
e-mail: jsunglee@lgcns.com

## 1 Introduction

As Analogue Switch off (ASO) announced an end to the era of analogue TV, broadcasting system throughout the world is swiftly being changed into digital. In the case of European digital broadcast, DVB-S and DVB-C were established in 1994. And DVB-T was established in early 1997 and experimental broadcast took place for the first time in 1998 in UK.

Most European countries are fully covered with digital television and many has switched off analogue TV. According to the DVB organization, DVB 2nd Generation Broadcasting System performance gaining over DVB 1st Broadcasting System is around 30 % at the same transponder bandwidth and emitted signal power.

By utilizing broadcasting frequency that will be newly acquired by ASO, DVB 2nd Generation Broadcasting System is expected to provide multi-channelled High-Definition TeleVision (HDTV) service and data service in various forms.

This paper realizes integrated receiver and analyzer system that can receive DVB-S2/T2/C2 data using File or Ethernet packet and analyzed received frame. Also, this paper introduces technology about DVB 2nd Generation Broadcasting System.

## 2 S/W Design and Analysis of DVB-T2 Receiver

### 2.1 DVB 2nd System Overview

DVB-S2 standard is based on DVB-S, and DVB-S2 is envisaged for broadcast services including standard and HDTV, interactive services including in internet access, and data content distribution.

DVB-S2 have two key features that were compared to the DVB Standard are a powerful coding scheme based on a LDPC code.

Next is Variable Coding and Modulation (VCM) and Adaptive Coding and Modulation (ACM) modes, which allow optimizing bandwidth utilization by dynamically changing of transmission parameters [1].

DVB-T2 standard is an expanded form of DVB-T standard, its core being improving digital TV transmission rate, as well as robustness and flexibility of transmission network usage. It also focuses on receiving digital TV contents via mobile devices and receiving broadcasting service smoothly while on transportation such as subway, train, and bus [2].

DVB-C2 standard is broadcasting transmission of digital television over cable. To compare with DVB-C, by using LDPC coding and new modulation techniques, DVB-C2 is greater than DVB-T2, showing 30 % higher spectrum efficiency under the same conditions, and the gaining in downstream channel capacity will be greater than 60 % [3].

### 2.2 Baseband Frame and Input Modes

By principle, DVB standards are developed to be compatible with each other. Therefore, if an outstanding solution is developed, this solution can also be adopted by other standards. According to such principal, Baseband Frame for data packaging established in DVB-S2 is also used in DVB-T2, DVB-C2.

DVB 2nd Generation Broadcasting System input mode utilizes Generic Encapsulated Stream (GSE), Generic Continuous Stream (GCS), Generic Fixed-length Packetized Stream (GFPS) in addition to MPEG2-TS (Transport Stream).

These 4 types of input modes are transmitted separately or collectively through a transmission method titled Physical Layer Pipes (PLPs) [5].

### 2.3 Processing Design of DVB 2nd Receiver and Analyzer

DVB-T2 Receiver and analyzer system received input streams, depending on whether user chooses File or Ethernet stream. Input stream is transmitted in DVB Piping format. Layer signal information is parsed from the reconstructed X2MI frame which then is displayed on the application windows, and the system type is extracted on Layer signal information.

Baseband Frame is an extracted base on L1 information. From the extracted baseband frame header information, stream input method is acquired. Decoding process per each input method follows. Basic processing procedure of receiver and analyzer system is shown in Fig. 1.

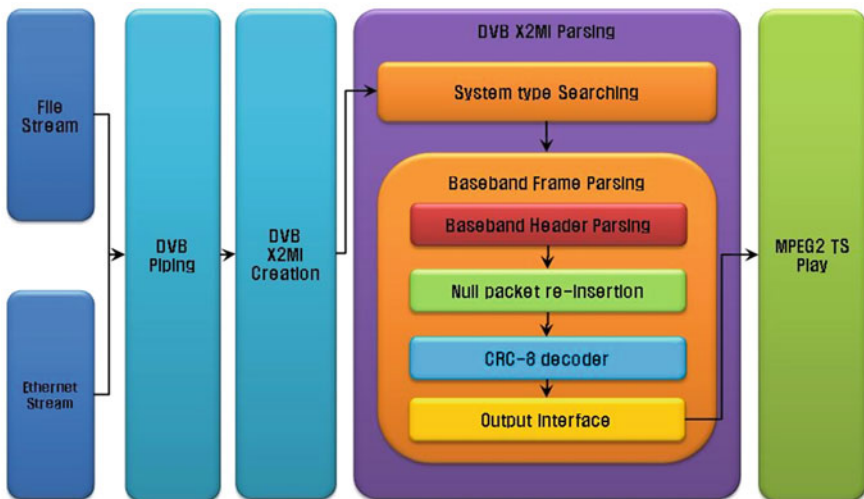


Fig. 1 Process diagram of DVB 2nd generation broadcasting receiver and analyzer system

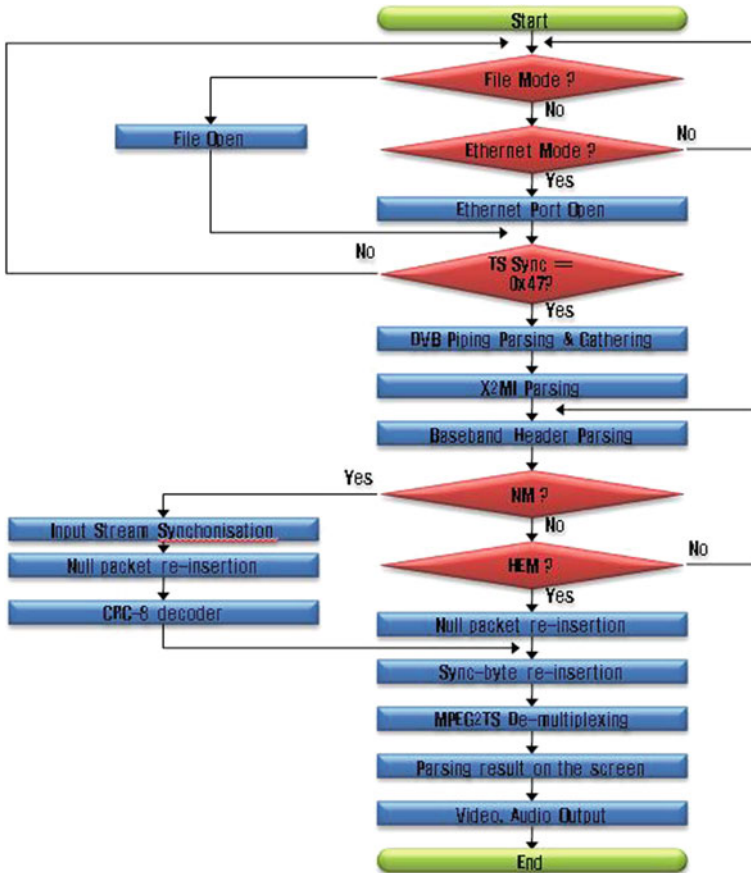


Fig. 2 DVB 2nd generation broadcasting receiver and analyzer system flow chart [6, 7]

To explain the processing procedure from paper in further detail, stream that needs to be processed according to their File or Ethernet packets mode is selected through realized application.

As inputted DVB 2nd stream is transmitted through DVB-Piping method, Input packet is parsed using TS module and X2MI frame restructured. Layer signal acquired from restructured X2MI frame is parsed and used to understand inputted broadcast system type and overall frame structure.

Baseband Frame is extracted and structured from this restructured X2MI frame. Composed Baseband Frame Header is parsed to determine system parameters (e.g., input types (TS or GSE), Mode types (Normal Mode or High Efficiency Mode), and whether additional modules (Input Stream Synchronization (ISSY)), Null Packet Deletion (NPD) is needed). According to acquired Header information, ISSY, NPD, CRC8 decoder and Sync-byte re-insertion module are applied.

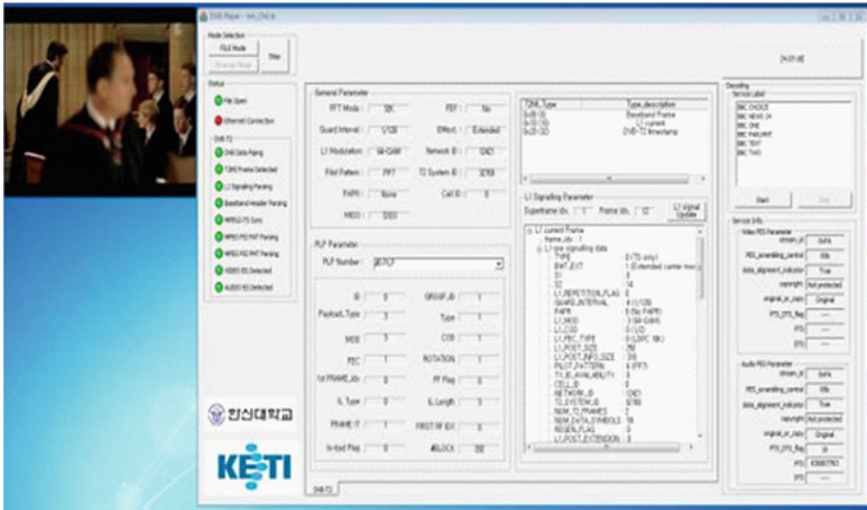


Fig. 3 Performance analysis of receiver and analyzer software result

This TS packet is restructured from Baseband Frame's Data Field. TS module is once again used on these restructure TS packet to compose Program Specific Information (PSI)/Service Information (SI) table. System flow procedure is shown in Fig. 2.

These supposed DVB-T2 receiver, and analyzer system have minimized the system overload by reconstructing TS system used as transmitting method of DVB-S2, DVB-T2, and DVB-C2 into a common module block. Method tiled X2MI, it's made from changed T2MI feature.

This realization result depends on the stream transmission through File or Ethernet packets, and on DVB 2nd Generation broadcast output analysis window on X2MI structure that transmits system type information, baseband frame and information of actual contents in Fig. 3.

This system analyzes transmission information of the system that transmits DVB-S2, DVB-T2 and DVB-C2 and confirms normal function by replaying the transmitted contents.

### 3 Conclusion

In this paper, the design, implementation of a receiver, and analyzer for DVB 2nd Generation Broadcasting System with window system based software were shown. These receiver and analyzer system were suggested for normal function by analyzing input system and Baseband Frame of received stream that varies depending on each system and playing the contents successfully.

**Acknowledgments** This work was supported by Hanshin University, and Korea government (MKE; Ministry of Knowledge Economy, #10039717/#10035547, NIPA; IT Industry Promotion Agency, and KOFST; The Korean Federation of Science and Technology Societies, 2012 Science and Technical Support specialists Supporters (H5701-12-1002), Small and Medium Business Administration (#S2064061), and KORIL-RDF: Korea-Israel Industrial R&D Foundation).

## References

1. ETSI EN 302 307 (2009) Digital Video Broadcasting (DVB), Second generation framing structure, channel coding and modulation systems for Broadcasting, Interactive Services, News Gathering and other broadband satellite applications (DVB-S2)
2. ETSI EN 302 755 (2009) Digital Video Broadcasting (DVB), Frame structure channel coding and modulation for a second digital terrestrial television broadcasting system (DVB-T2)
3. ETSI EN 302 769 (2009) Digital Video Broadcasting (DVB), Frame structure channel coding and modulation for a second generation digital transmission system for cable system (DVB-C2)
4. Seo JW, Kang MG, Jeon ES, Kim DK (2011) Performance evaluation of a DVB-T2 receiver with iterative de-mapping and decoding in MISO transmission mode. KSII 12(2)
5. ETSI TS 102 773 (2009) Digital Video Broadcasting (DVB), Modulator Interface (T2-MI) for a second generation digital terrestrial television broadcasting system (DVB-T2)
6. Kang M et al (2012) Method of processing DVB-T2/S2/C2 piping-format broadcasting signals with memory sharing, and computer-readable recording medium with broadcasting signal processing program. USA Patent Pending (PCT/KR#2012/004913)
7. Paik JH et al (2011) Design of window based DVB-T2 receiver with file, and ethernet modes. KSII International conference on internet (ICONI)



# Age-Group Classification for Family Members Using Multi-Layered Bayesian Classifier with Gaussian Mixture Model

Chuh0 Yi, Seungdo Jeong, Kyeong-Soo Han and Hankyu Lee

**Abstract** This paper proposes a TV viewer age-group classification method for family members based on TV watching history. User profiling based on watching history is very complex and difficult to achieve. To overcome these difficulties, we propose a probabilistic approach that models TV watching history with a Gaussian mixture model (GMM) and implements a feature-selection method that identifies useful features for classifying the appropriate age-group class. Then, to improve the accuracy of age-group classification, a multi-layered Bayesian classifier is applied for demographic analysis. Extensive experiments showed that our multi-layered classifier with GMM is valid. The accuracy of classification was improved when certain features were singled out and demographic properties were applied.

**Keywords** Age-group classification · Gaussian mixture model · Feature selection

---

C. Yi

Research Institute of Electrical and Computer Engineering, Hanyang University,  
Seoul, Korea  
e-mail: d1luck@hanyang.ac.kr

S. Jeong (✉)

Department of Information and Communication Engineering, Hanyang Cyber University,  
Seoul, Korea  
e-mail: sdjeong@hycu.ac.kr

K.-S. Han · H. Lee

Electronics and Telecommunications Research Institute, Smart TV Service  
Research Team, Daejeon, Korea  
e-mail: kshan@etri.re.kr

H. Lee

e-mail: hkl@etri.re.kr

## 1 Introduction

Due to the advent and popularization of interactive TV and internet TV, there is a need for research into user profiling based on viewer watching history to support targeted advertising [1]. In a previous study, we proposed a method for computing TV viewer preferences for nine types of genre based on viewing history data [2]. To avoid including meaningless behavior such as random/aimless channel changing, we modeled the preference function as a beta distribution. However, extensive experiments showed that simple preference profiling based on an averaged preference for a few genres was not enough to recognize age-groups. There are several reasons for this: TV watching data have a multi-modal characteristic, no feature by itself can be used to identify an age-group, and one person's watching data are typically corrupted by other members of his/her family (watching data are almost never limited to one person but rather include the behavior of multiple viewers).

Here, we propose four methods to overcome these difficulties. First, the types of preferences are more specifically divided, as they are divided into 70 types of genre by the hour and day of the week. This results in 645 features. Then, to reflect the multi-modal distribution of the data, the distribution of features according to viewer age-group is modeled using a Bayesian Gaussian mixture model (GMM). Second, we propose a multilayered classifier to improve the performance of the system through the analysis of demographic properties. Third, because whole features complicate the classification of viewer age-groups, we propose a feature-selection method based on weighting calculated with a training set. Finally, *a priori* knowledge based on the demographic composition of the training data is used in the classification process. In experiments, TV viewer age-groups were inferred using our trained, multi-layered classifier. We evaluate and discuss the results of the classification below.

## 2 The Proposed Method

### 2.1 Basic Bayesian Model

Preference values, which are used as observation data in the present study, are computed by dividing each viewer's watching time by the total viewing time. Preference features consist of 70 types of program genre, eight time periods in a day, and 7 days of the week. Hence, a total of 645 profiling selected features are used in the training and classification steps.

Every extracted feature is assumed to be independent and thus is individually processed in the training step. If observation  $F$  has  $N$  number of data, the posterior probability is expressed as in Eq. (1), where  $C$  is a class. In the present work, the class is the age-group of the TV viewer.

$$p(C|F_1, F_2, \dots, F_N) = \frac{p(C)p(F_1, F_2, \dots, F_N|C)}{p(F_1, F_2, \dots, F_N)} \quad (1)$$

Using the chain rule of conditional probability, Eq. (1) can be rewritten as Eq. (2):

$$p(C|F_1, F_2, \dots, F_N) \propto p(C) \prod_{i=1}^N p(F_i|C) \quad (2)$$

where  $p(C)$  is the *a priori* probability, and  $p(F_i|C)$  is a likelihood. In this paper, the demographic statistics of an age-group are used as  $p(C)$ . We assume that the likelihood has a multimodal distribution. Therefore, we model the distribution using GMM, as shown in Eq. (3):

$$p(F_i|C = c_a) = \sum_{j=1}^{K_a} w_j \cdot N(F_i|\mu_j, \Sigma_j) \quad (3)$$

where  $w_j$  is the prior probability (weight) of the  $j$ th Gaussian,  $\sum_{j=1}^{K_a} w_j = 1$  and  $0 \leq w_j \leq 1$ , and  $K_a$  is a peak number for class  $c_a$ . To make it possible to select  $K_a$  automatically, the Figueiredo–Jain algorithm is used in the training step [4].

The goal of this research is classifying an age-group for each person of family members and the outcome of single classifier based on basic Bayesian model is an existence of respective age-group; therefore, several classifiers are used in this paper to classify age-groups for all members of family.

## 2.2 Multi-Layered Structure

Our proposed multi-layered classifier assumes that preprocessing of data may be helpful for classifying family members into age-groups. To this end, the system first determines whether anyone in the family is younger than 10 years of age. This is the most important criterion because the composition of a family is highly affected by the presence of this age-group. Therefore, this classifier is at the highest level (C1). If no one is in this age-group, the system next identifies anyone over 60 years of age. This is the second-most important criterion and itself makes up the next highest level of classification (C2). All subsequent classifiers are at the C3 level; they include groups 10–19, 20–29, 30–39, 40–49, and 50–59 years of age (hereafter, the 10s, 20s, 30s, 40s, and 50s age-groups, respectively). Hence, a total of 17 classifiers are represented. Figure 1 shows the proposed system.

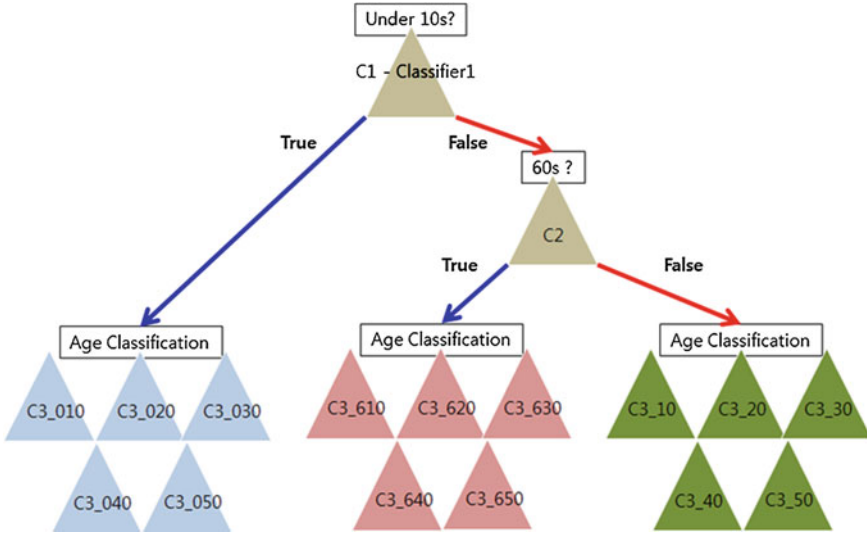


Fig. 1 The proposed multi-layered classifier

### 2.3 Feature Selection

The feature-selection step verifies the usefulness of each viewing feature (i.e., each combination of genre, day, and time of day) using the training dataset. A score is used to weight each feature based on its estimated usefulness for classification. Here, we assume that each feature is independent of all others.

Equation (4) is a utility function that determines whether a feature can be used to produce a true-positive age-classification result based on the training dataset, where  $C_1$  represents a class of positive data, and  $C_2$  is a class of negative data. Equation (5) does the same for true-negative results. The result of these two utility functions is 1 when a feature always correctly determines true-positive and true-negative results, respectively.

$$r_{i,a} = \begin{cases} 1, & p(C = c_1 | f_i \in F_{c_1}) > p(C = c_2 | f_i \in F_{c_1}) \\ 0, & p(C = c_1 | f_i \in F_{c_1}) < p(C = c_2 | f_i \in F_{c_1}) \end{cases} \quad (4)$$

$$s_{i,b} = \begin{cases} 1, & p(C = c_2 | f_i \in F_{c_2}) > p(C = c_1 | f_i \in F_{c_2}) \\ 0, & p(C = c_2 | f_i \in F_{c_2}) < p(C = c_1 | f_i \in F_{c_2}) \end{cases} \quad (5)$$

Second, the scoring functions shown in Eqs. (6) and (7) are used to compute the reliability of each feature for correctly classifying the entire training dataset.

$$score(F_i, C = c_1) = \frac{1}{|M|} \sum_{a=1}^M r_{i,a} \quad (6)$$

$$score(F_i, C = c_2) = \frac{1}{|M|} \sum_{b=1}^M s_{i,b} \quad (7)$$

where  $M$  is the number of data points in the training dataset. If a score is above a given threshold, then the feature is considered reliable for classification. We determine this threshold using a heuristic searching process. In other words, through an exhaustive search, the values of  $score(F_i, C = c_1)$  and  $score(F_i, C = c_2)$  are chosen as the threshold when the accuracy of the posterior probability  $p(C|F_{training})$  is highest based on the training dataset  $F_{training}$ .

## 2.4 A Prior Knowledge for Demographic Data

Next, the age-group classification is processed using the trained classifiers, where  $F_{s \in \{S\}}$  is the set of features selected during the previous steps described above. Maximum *a posteriori* (MAP) estimation is applied as shown in Eq. (8):

$$c^* = p(C) \prod_{i=1}^N p(F_{s,i}|C) \quad (8)$$

To prevent computational underflow, Eq. (8) is transformed using a logarithmic expression, as shown in Eq. (9):

$$c^* = \log\{p(C)\} + \sum_{i=1}^N \log\{p(F_{s,i}|C)\} \quad (9)$$

where  $p(C)$  is the *a priori* probability, and we use the ratio according to age-group for demographic statistics which is analyzed with the training data.

## 3 Experimental Result

We tested our system on a dataset of the TV watching history of 2060 family members from 689 families. The time period considered was April 1 to June 30, 2012 (13 weeks/91 days). The maximum family size was 6. For the training step, we selected 85 % of the dataset (585 families/1758 members) as the training set. The remaining data points (104 families/302 members) were used as the test set. The experiments were run on a Windows 7 operating system using MATLAB 2010. A modified version of GMMBayes Toolbox [5] was used for GMM training and classification.

Table 1 shows the results of the age-group classification for family members. Method 1, which used unprocessed, raw data, did not work for the 30s and 50s

**Table 1** Accuracy of age-classification experiments

Ages (s)	Method 1 (%)	Method 2 (%)	Method 3 (%)
<10	45.19	47.12	50.96
10	64.04	67.31	68.27
20	59.62	77.88	76.92
30	–	44.23	43.27
40	43.27	57.69	59.62
50	–	63.46	63.46
60	65.38	65.50	62.50
<b>Average</b>	<b>55.50</b>	<b>60.03</b>	<b>60.71</b>

age-groups and had a 55.50 % average accuracy. Method 2 used the feature-selection method proposed in this paper and had a 60.03 % average accuracy; additionally, it worked well for the 30s and 50s age-groups. Method 3, which also used *a priori* knowledge estimations, as described above, had an average accuracy of 60.71 %. These results indicate that our proposed feature-selection and *a priori* knowledge steps are meaningful for age-group classification.

The goal of this research is the classification of TV viewer's age-group for family members and even was not known how many people live in a house together. For the ideal watching history data, the viewer should be forced to log in the TV with a remote controller manually. However, in the real situation, almost viewers did rarely log in or change their ID every time. It caused corruption to the viewing history data by other members in the family. Thus, this point is a huge obstacle for an accurate classification of the age-group for family members. We also suffer from this obstacle for classification in this research; however we try to overcome these difficulties with the proposed four approaches. As a result, our proposed method could be reasonably classified the age-group for family members.

## 4 Conclusions

Our proposed probabilistic model for age-group classification based on complex viewing history data combines Bayesian classifiers with GMM methods. In experiments, our proposed system classified age-groups with reasonable accuracy and well enough to be incorporated into a targeted advertising system.

**Acknowledgments** This work was supported by the Electronics and Telecommunications Research Institute (ETRI) R&D Program of Korea Communications Commission (KCC), Korea [11921-03001, "Development of Beyond Smart TV Technology"].

## References

1. Spangler WE, Gal-Or M, May JH (2003) Using data mining to profile TV viewers. *Commun ACM Mob Comput Oppor Chall* 46(12):66–72
2. Lee S, Park S, Hong J, Yi C, Jeong S (2012) Inference for the preference of program genre using audience measurement information. In: *International conference on information and knowledge, engineering*, pp 224–225
3. Wonneberger A, Schoenbach K, Meurs LV (2009) Dynamics of individual television viewing behavior: models, empirical evidence, and a research program. *Commun Stud* 60(3):235–252
4. Figueiredo MAT, Jain AK (2002) Unsupervised learning of finite mixture models. *IEEE Trans Pattern Anal Mach Intell* 24(3):381–396
5. GMMBayes Toolbox, Gaussian mixture model learning and Bayesian classification. <http://www.it.lut.fi/project/gmmbayes/>

# Enhancing Utilization of Integer Functional Units for High-Throughput Floating Point Operations on Coarse-Grained Reconfigurable Architecture

Manhwee Jo, Kyuseung Han and Kiyoung Choi

**Abstract** Supporting floating point operations on coarse-grained reconfigurable architecture (CGRA) becomes essential as the increase of demands on various floating point inclusive applications such as multimedia processing, 3D graphics, augmented reality, or object recognition. However, efficient support for floating point operations on CGRA has not been sufficiently studied yet. This paper proposes a novel technique to enhance utilization of integer units for high-throughput floating point operations on CGRA with experimental results.

**Keywords** Reconfigurable · CGRA · Floating point · Resource utilization

## 1 Introduction

Not many years ago, phones are used just for calling, but today they are used for playing audios/videos, surfing web, image processing, and enjoying games. Not only phones but also tablets have been gaining popularity rapidly and are now a part of our daily lives. However, as the functionality of such mobile devices becomes more diverse and complex, supporting them with limited resources is a big challenge. Multicores are not enough to meet the requirement of compute-intensive programs even though they are suitable for running several control-intensive problems simultaneously. ASICs can hardly support various programs since we

---

M. Jo (✉) · K. Han · K. Choi  
Department of EECS, Seoul National University, Seoul, Korea  
e-mail: manhwee@dal.snu.ac.kr

K. Han  
e-mail: darprin@dal.snu.ac.kr

K. Choi  
e-mail: kchoi@dal.snu.ac.kr



cannot put tens or hundreds of them into a single chip. In addition to that, we have encountered another challenge. While conventional compute-intensive programs such as multimedia applications are based on integer calculation, new ones such as 3D graphics, augmented reality, object recognition, or face recognition require real number operations. Thus, efficient support for both fixed- and floating-point operations with limited resources is also important in future embedded systems.

For the first challenge mentioned above, coarse-grained reconfigurable architecture (CGRA) [1–5] has been proved to be one of viable solutions since it can provide performance and flexibility at the same time. It consists of abundant processing elements (PEs) so it can accelerate the execution of programs by parallel processing. In addition, it can run various applications by changing the functionality of hardware dynamically through the reconfiguration of the PEs and the interconnections between them. Thus we focus on CGRA in this paper.

Regarding the second challenge, especially for CGRA, only a few researches have been conducted on CGRA related to floating point operations. Considering that introducing floating-point units in addition to abundant integer units wastes area and power (either of integer units or floating point units are always unused), the approaches in [6–8] share FUs between two types of operations. However, they try only simple approaches resulting in an inefficient use of functional resources. In this paper, we propose a novel architecture that can utilize FUs for both types of operations. In particular, we propose a way to better utilize integer FUs when performing floating point operations on CGRA.

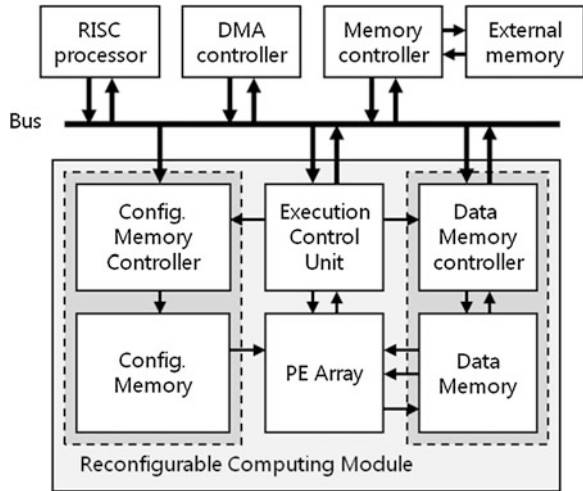
The organization of the paper is as follows. [Section 2](#) presents a base coarse-grained reconfigurable architecture, and [Sect. 3](#) describes a conventional approach to implement floating point operations on the CGRA. [Section 4](#) proposes an approach to enhance the utilization of FUs in PEs. [Section 5](#) compares the base architecture and the proposed architecture in terms of performance and cost. Then [Sect. 6](#) concludes the paper with future directions.

## 2 Base Architecture

The base coarse-grained reconfigurable architecture (Fig. 1) consists of a general purpose RISC processor, reconfigurable computing module (RCM), hardware for transferring data between local memory and external memory such as direct memory access (DMA) unit. RCM executes compute-intensive code blocks while the RISC processor executes control-intensive code blocks.

RCM has two-dimensional array of PEs, each of which contains functional units such as adder/subtractor, comparator, and shifter, and local register file. Execution control Unit of RCM fetches the configuration for the PE array so that each PE in the PE array executes different operations for its configuration. Each PE has FUs such as adder or shifter, and a local register file. A PE also has an output register connected to the network in the PE array. A PE can communicate with

**Fig. 1** A coarse-grained reconfigurable architecture



other PEs in the same row and column via the network. Input and output data for the PE array are stored in local data memory which hides the access time to the external global memory.

### 3 Conventional Approach

Since each PE has only integer FUs inside, a couple of PEs can cooperate in order to execute floating point operations [6]. An floating point operations are divided into a bunch of integer operations and they are executed on a pair of PEs, named an FPU-PE cluster. Specifically, floating point numbers to be calculated are divided into two parts: signed mantissa part and exponent part, then one PE (mantissa PE:  $PE_M$ ) in a FPU-PE cluster manipulates mantissa parts while the other PE (exponent PE:  $PE_E$ ) operates exponent parts. PEs in the cluster store the intermediate data in their local registers and communicate with each other via network interconnection between them while running floating point operations. Each PE has a finite state machine (FSM) for configuring the behaviour of the datapath during multi-cycle floating point operations rather than fetching configuration bits from the configuration memory every cycle. The floating point operations implemented in this work include floating point addition (FADD), subtraction (FSUB) and multiplication (FMUL). The latencies of FADD and FSUB are 8 cycles, while that of FMUL is 4 cycles. In FMUL operation,  $PE_M$  multiplies two floating point mantissa parts using integer multiplier module, which is shared by all the PEs in a row for reducing area cost [3].

## 4 Proposed Approach

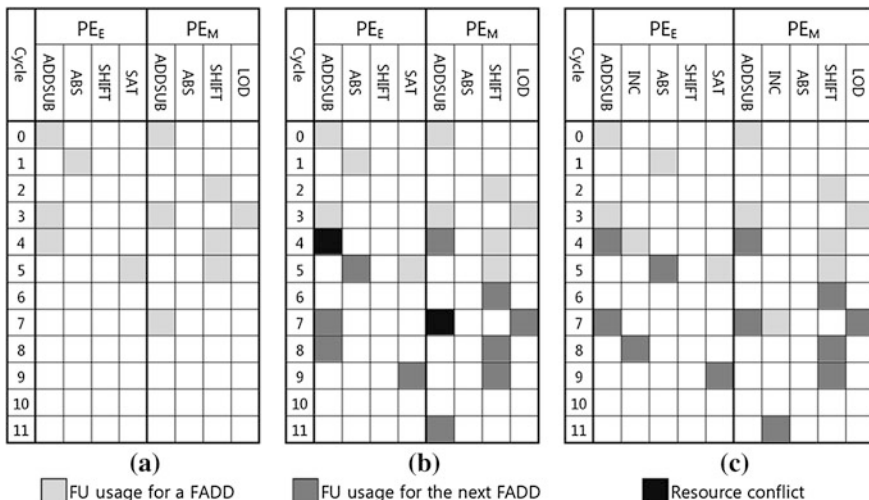
Although using integer FUs in order to perform floating point operations increases utility of the architecture, the long latencies of the floating point operations limit the performance of floating point applications (since the base architecture is optimized for running various single cycle integer operations, scheduling integer operations split from a floating point operation requires unnecessary idle cycles due to extra mux and functional unit delays). In order to mitigate it, we propose to share FUs that are temporarily freed during a floating point operation with another floating point operation. Simply, it is overlapping two floating point operations with different operands by sharing FUs between the two floating point operations. That can be done with minimal overhead of additional control logic and registers. Such overlapping of two different floating point operations improves performance by enhancing the utilization of FUs when multiple independent floating point operations are to be executed.

During the execution of floating point operations, each PE employs its FUs to manipulate the input values, a temporal register to keep the intermediate value, and output register to communicate with each other. Since all the FUs in each PE are not used every cycle during execution of a floating point operation, they can be used for another floating point operation. Unlike the FUs, which are frequently freed due to data dependency, two specific registers are utilized almost every cycle so that they cannot be shared with other floating point operations. This problem is solved by inserting an additional register with dedicated communication channel for communication in an FPU-PE cluster for another floating point operation to be overlapped. For intermediate values, there is no need of inserting additional register, but utilizing one of the registers in the local register file in the PE works. Thus just dedicated interconnects to store intermediate value to a register in the register file are added.

For the simplicity of the FSM in each PE, time difference (overlap) between two floating point operations is fixed as a half of the latency of the operation, i.e., 4 cycles for FADD, and 2 cycles for FMUL.

Since the proposed approach cannot be applied to two operations that have between them, we expect it to be mainly applied in unrolled iterations of kernels. In general, however, FADD and FMUL in a kernel program are likely to have data dependency, and thus overlapping FADD and FMUL is not considered in this paper.

Figure 2a shows the usage of FUs of PEs in an FPU-PE cluster for an FADD operation. When another FADD operation starts four cycles later, there are resource conflicts at two different positions shown as black boxes in Fig. 2b. There are two alternative ways of avoiding the conflict: duplication of the conflicted FU, or executing the following floating point operation one cycle later. Fortunately, both of the two conflicts are for incrementing the input value by 1 for rounding. Thus inserting an incrementer instead of another adder/subtractor is good enough for overlapping two floating point operations with 4 cycles of time difference (Fig. 2c).



**Fig. 2** a Usage of integer FUs in FPU-PE cluster for a FADD, b two overlapping FADDs with conflict, and c two overlapping FADDs with no conflict

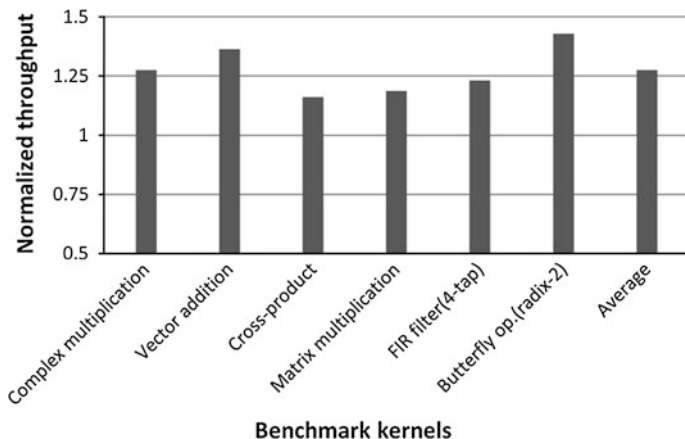
On the other hand, there is one conflict on adder/subtractor unit when overlapping two 4-cycle FMUL operations with 2 cycles of time difference. The conflict is also due to a rounding operation, and thus the incrementer inserted for the conflict on FADD is utilized.

The FSM is also changed to support overlapping two floating point operations. Since the cycle difference is fixed, there is no need of parallelizing FSMs but only merging states of the two overlapping floating point operations is needed.

## 5 Experimental Results

Figure 3 shows the performance of the simple functions running on the existing architecture and the architecture where the proposed approach is applied. Employed benchmark functions are basic arithmetic functions which are frequently used in signal-processing, multimedia applications, 3D graphics, and other mathematical algorithms. Normalized throughput shown in Fig. 3 is throughput ratio of the proposed architecture to the base architecture for each function. By applying the proposed approach of sharing FUs between two floating point operations, we obtain up to 42.9 % better performance and 27.5 % better performance on average compared to the base architecture.

Both of the proposed architecture and the existing architecture are synthesized by Synopsys DesignCompiler with TSMC 130 nm technology library. Each architecture has  $8 \times 4$  array of PEs, one column of shared integer multipliers (i.e., eight multipliers), and configuration and data memory blocks. Synthesis result



**Fig. 3** Normalized throughput of the proposed architecture compared to the base architecture

shows that the area overhead caused by the proposed approach is about 13.8 % at the same clock speed of 330 MHz. It is only about 2.9 % of the whole RCM with configuration and data memory blocks. The area overhead is from FSM and decoder on the control path as well as additional registers and incrementers.

To consider both performance and area cost at the same time, we simply introduce a term, gain, as throughput ratio divided by area ratio. The gain of the proposed approach to the base architecture is 1.12, which indicates about 12 % improvement over the base architecture.

## 6 Conclusion

This paper proposes a novel mechanism to utilize integer functional units during execution of floating point operations. By sharing functional units, we can support floating point operation with minimal overhead. Especially the proposed overlapping technique enhances the utilization of functional units so that we can accelerate floating point kernels effectively. Experimental results show that our approach has 12 % better throughput-to-area ratio compared to the previous sharing schemes. As a future work, sharing FUs among different types of floating point operations, or operations with data dependency could be researched in order to reach further improvements.

**Acknowledgments** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2012-0006272).

## References

1. Hartenstein R (2001) A decade of reconfigurable computing: a visionary retrospective. In: Proceedings of the 4th conference on design, automation and test in Europe. IEEE Press, New York, pp 642–649
2. Singh H, Lee MH, Lu G, Bagherzadeh N, Kurdahi FJ, Filho EMC (2000) Morphosys: an integrated reconfigurable system for data-parallel and computation-intensive applications. *IEEE Trans Comput* 49(5):465–481
3. Kim Y, Kiemb M, Park C, Jung J, Choi K (2005) Resource sharing and pipelining in coarse-grained reconfigurable architecture for domain-specific optimization. In: Proceedings of the 8th conference on design, automation and test in Europe. IEEE Press, New York, pp 12–17
4. Novo D, Moffat W, Derudder V, Bougard B (2005) Mapping a multiple antenna SDM-OFDM receiver on the ADRES coarse-grained reconfigurable processor. In: Proceedings of the 9th IEEE workshop on signal processing systems design and implementation. IEEE Press, New York, pp 473–478
5. Hasegawa Y, Abe S, Matsutani H, Amano H, Anjo K, Awashima T (2005) An adaptive cryptographic accelerator for IPsec on dynamically reconfigurable processor. In: Proceedings of the IEEE conference on field-programmable technology. IEEE Press, New York, pp 163–170
6. Jo M, Arava VKP, Yang H, Choi K (2007) Implementation of floating-point operations for 3D graphics on a coarse-grained reconfigurable architecture. In: Proceedings of the 20th IEEE international SOC conference. IEEE Press, New York, pp 127–130
7. Syed MA, Schueler E (2006) Reconfigurable parallel computing architecture for on-board data processing. In: Proceedings of the 1st NASA/ESA conference on adaptive hardware and systems. IEEE Press, New York, pp 229–236
8. Brunelli C, Garzia F, Rossi D, Nurmi J (2010) A coarse-grain reconfigurable architecture for multimedia applications supporting subword and floating-point calculations. *J Syst Architect* 56(1):38–47

# An Improved Double Delta Correlator for BOC Signal Tracking in GNSS Receivers

Pham-Viet Hung, Dao-Ngoc Chien and Nguyen-Van Khang

**Abstract** Multipath is one of main error sources in code tracking in global navigation satellite system. One of the first approaches to mitigate multipath is called Double Delta Correlator (DDC). This technique originally applied to Global Positioning System signals (C/A signals) which are Binary Phase Shift Keying ones and have a single-peak autocorrelation function. Latter, it also applied to new GNSS signals which are binary offset carrier (BOC) modulated ones. It shows a good code multipath performance for mid-delayed and long-delayed multipath in both cases. For short-delayed multipath, there is still remaining error. This paper presents an Improved DDC that could reduce the error for short-delayed multipath signals. This improved DDC method is proposed for BOC signals. The simulation results show a better performance of proposed method than conventional DDC.

**Keywords** BOC signal · Double delta correlator · Multipath mitigation

## 1 Introduction

Currently, binary offset carrier (BOC) modulation has been recommended to some new navigation signals such as Galileo L1OS (OS for Open Service), Global Positioning System (GPS) future L1C (New Civilian L1) [1]. The BOC ( $m, n$ ) signal is created by modulating a sine wave carrier with the product of a spreading code and a square wave subcarrier. The parameter  $m$  stands for the ratio between the subcarrier frequency and the reference frequency  $f_0 = 1.023$  MHz, and  $n$  stands for the ratio between the code rate and  $f_0$ .

---

P.-V. Hung (✉) · D.-N. Chien · N.-V. Khang  
School of Electronics and Telecommunications, Hanoi University of Science  
and Technology, Hanoi, Vietnam  
e-mail: hung.phamviet@hust.edu.vn

The most common multipath mitigation techniques are based on tracking loop structure which contains carrier phase locked loop (PLL) and code delay locked loop (DLL) [2]. Conventionally, two correlators spaced at one chip from each other are used in delay estimator. It tries to track the delay of the direct signal by correlating the down-converted received signal with replicas of local generating codes in the receiver. However, the classical estimator fails to estimate the multipath error envelope (MEE) accurately. Thus several evolutions have been introduced in the literature in order to mitigate the influence of MP signals, especially in short delayed multipath scenarios. Examples of these are Narrow Correlator [3], DDC [4], early/late slope technique, Early1/Early2 tracker [5]. Initially, these techniques were proposed to GPS signals which are Binary Phase Shift Keying (BPSK) ones and have triangular autocorrelation function.

Among these above mentioned MP mitigation techniques, DDC have the best code multipath performance for medium-to-long delayed MP [4]. However, DDC are still not good enough for short-delayed MP environment such as urban canyons, which is a key motivation for present researchers.

The purpose of this paper is to introduce an improved DDC method suitable for general BOC signals. It is based on conventional DDC. The proposed method will estimate the tracking error of DDC when tracking signal in closely spaced MP environment. Then, the estimate is applied to the DDC as modification in order to reduce tracking error.

The rest of this paper is organized as follows. In Sect. 2, an overview of DDC is presented. In Sect. 3 the improved DDC is proposed. The simulation results of conventional DDC and improved DDC in multipath delay estimation is discussed in Sect. 4, and finally Sect. 5 concludes this paper.

## 2 Double Delta Correlator Description

As mentioned above, the conventional delay estimator uses 03 correlators so-called: Early (E), Prompt (P) and Late (L). The space between early and late correlator is one (01) chip. For this classical approach, the multipath envelope error could not be estimated accurately. Narrowing chip space is one of the solutions for the improvement of MP error. Whereas the classical discriminator uses 1 chip space, the narrow correlator reduces it to 0.1 chip space [3]. However, band-limitation of the pre-correlation filter rounds the autocorrelation peak so that the space between correlators could not be too small [6].

Also narrowing the chip space, DDC uses 05 correlators in the tracking loop instead of 03 ones (two early, one in prompt and two late) [4]. The mathematical definition of correlators of DDC is shown as

$$\begin{aligned}
 P &= R(\tau); \\
 E_1 &= R(\tau - d/2); & E_2 &= R(\tau - d) \\
 L_1 &= R(\tau + d/2); & L_2 &= R(\tau + d)
 \end{aligned} \tag{1}$$



where  $P$  relates to the prompt correlator,  $E_1$  and  $L_1$  relate to the inner correlators with a chip spacing of  $d$ , whereas  $E_2$  and  $L_2$  relate to the outer correlators with a chip spacing of  $2d$ . And,  $R$  is defined as autocorrelation function (ACF) of BOC signal.

In case of carrier phase locked tracking, the output of the DDC discriminator is equal to [7]

$$D_{DDC} = (E_1 - L_1) - (E_2 - L_2)/2 \quad (2)$$

However, in multipath environment, the signal reaching the receivers consist one direct path and  $M - 1$  multipath. For theoretical analysis, we set up the dedicated scenario with one direct path and only one multipath and a constant signal attenuation factor. Analytically, the direct signal and multipath signal can be calculated separately and the total correlation function is given as [8]

$$R_{total}(\tau) = R(\tau) + \alpha R(\tau + \delta) \quad (3)$$

where  $\alpha$  is the multipath to direct signal amplitude ratio and  $\delta$  is the delay of the multipath signal relative to the direct signal.

Consequently, in the scenario of the presence of multipath, the ACF of BOC (1,1) is distorted and discriminator output of DDC is equal to zero at non zero delay ( $\tau \neq 0$ ).

Assuming the chip space is  $d = 0.1 \text{ chip}$ , the ACF of BOC (1,1) signal in multipath environment with  $\delta = 0.08$  and  $\alpha = 0.5$  is illustrated in Fig. 1. The discriminator output of DDC in that scenario is shown in Fig. 2.

The impact of multipath on code tracking is represented in Fig. 3. It illustrates the maximum error resulting from one multipath with certain delay, amplitude and phase, so-called MEE. In case the MP signal and the LOS signal are in-phase, it is corresponded to upper part of the Fig. 3 [2]. On the other hand, if the MP signal and LOS signal are out of phase, it is corresponded to the lower part of the figure.

As shown in Fig. 3, even though the DDC mitigates perfectly the error for medium-long delayed MP, but for short-delayed MP ( $\delta < 0.2 \text{ chip}$ ), it cannot remove it and the error is significant. The proposed method in this paper tries to reduce the short-delayed MP error.

### 3 The Improved Double Delta Correlator

In order to mitigate the MP in short-delayed MP, the proposed method evaluates the DDC tracking error, then, it is applied to DDC. The tracking error evaluation is based on the geometric analysis of ACF for several MP delays. When DDC code tracking is accuracy, the P correlator is located on the true correlation peak, it means that DDC remove perfectly the error. When DDC code tracking is wrong, the P correlator is located on right side (left side) of the true correlation peak, depending on in-phase (out-phase) MP signals.

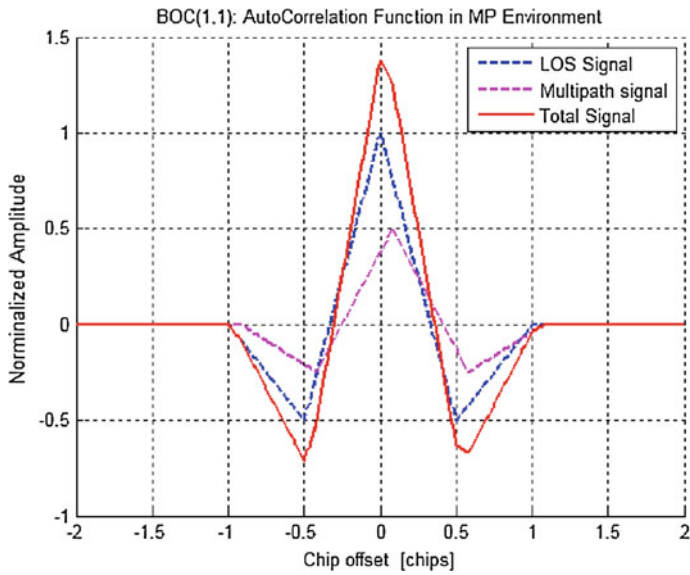


Fig. 1 ACF of BOC (1,1) signal in MP environment

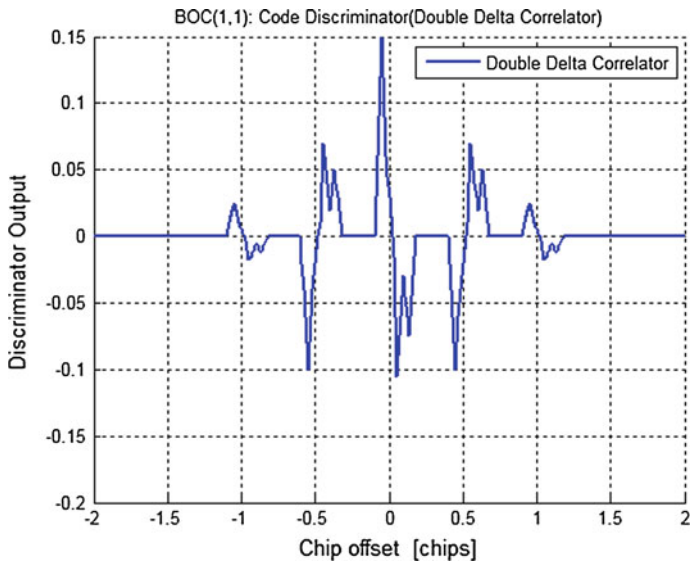


Fig. 2 Discriminator output with DDC for BOC (1,1) signal

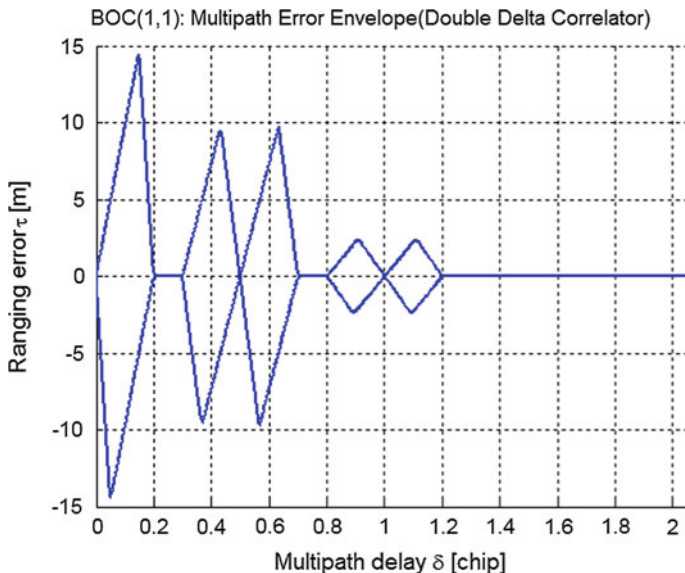


Fig. 3 MEE for BOC (1,1) signal with DDC ( $d = 0.1$  chip)

### 3.1 In-Phase MP signals

The ACFs of BOC (1,1) in MP environment where MP signal is in-phase with LOS signal is shown in Fig. 4. All parts of the all ACFs (LOS ACF, MP ACF and total signal ACF) are modeled as mathematical equations.

Setting  $\varepsilon$  is DDC tracking error,  $a$  is the maximum amplitude of LOS ACF,  $b$  is the maximum amplitude of MP ACF,  $\delta$  is the MP delay and  $d$  is the chip space.

The ACF of the received BOC (1, 1) signal is the sum of LOS ACF and MP ACF. As illustrated in Fig. 4, the shape of the distorted ACF can be separated into 09 lines in order to model each line by an equation as

$$\begin{aligned}
 d_1 &= y_1; & d_2 &= y_1 + y_5; & d_3 &= y_2 + y_5 \\
 d_4 &= y_2 + y_6; & d_5 &= y_3 + y_6; & d_6 &= y_3 + y_7 \\
 d_7 &= y_4 + y_7; & d_8 &= y_4 + y_8; & d_9 &= y_8
 \end{aligned}
 \tag{4}$$

In the scenario of the short-delayed MP ( $d = 0.1$  chip,  $0.05 < \delta < 0.2$ ) DDC fails tracking the true correlation peak, so the  $P$  correlator is located on the right of the true peak. Consequently,  $P$  correlator output is on the line of  $d_5$ ,  $E_1$  and  $E_2$  are on the line of  $d_4$ ,  $L_1$  and  $L_2$  are on the  $d_6$ .

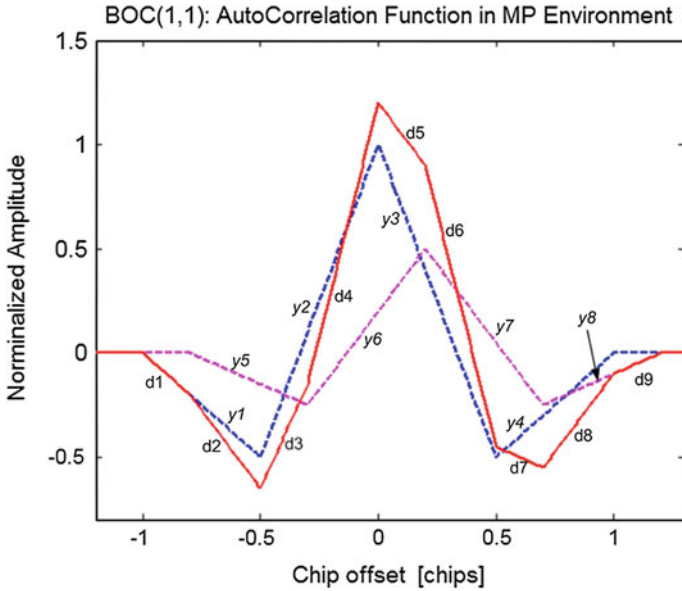


Fig. 4 Geometry of ACF for BOC (1,1) signal with a in-phase MP

Based on Eq. (4), the amplitudes of all correlator outputs are calculated as

$$\begin{aligned}
 P &= (-a + b)3\varepsilon + a + (1 - \delta)b \\
 E_1 &= 3(a + b)(\varepsilon - d) + a + (1 - \delta)b \\
 E_2 &= 3(a + b)(\varepsilon - 2d) + a + (1 - \delta)b \\
 L_1 &= -3(a + b)(\varepsilon + d) + a + (1 + \delta)b \\
 L_2 &= -3(a + b)(\varepsilon + 2d) + a + (1 + \delta)b
 \end{aligned}
 \tag{5}$$

Solving Eq. (5), the DDC tracking error can be expressed as

$$(E_1 - E_2) - (P - E_1) = 6a\varepsilon
 \tag{6}$$

As seen in Eq. (6), the left side of it could be used as an indicator of DDC tracking error with a scale of  $6a$ . The left side is the difference between the distance of  $E_1, E_2$  and the distance of  $E_1, P$ . If the difference is zero, it means that  $E_1$  is located at the midpoint of  $E_2$  and  $P$ , resulting in  $E_1, E_2$  and  $P$  are on the same line. Consequently, there is no error in tracking loop ( $\varepsilon = 0$ ) and  $P$  correlator output is located on the true correlation peak. When there is tracking error ( $\varepsilon > 0$ ),  $P$  correlator output has to move to the right side of true peak due to  $E_1 - E_2$  is larger than  $E_1 - P$ .

However, the tracking error should be calculated without depending on the maximum amplitude of LOS ACF as well as MP ACF. Also derived from Eq. (5), the sum of maximum amplitude of LOS ACF and the one of MP ACF is calculated as

$$a + b = (L_1 - L_2)/3d \tag{7}$$

In almost natural cases, the amplitude of direct component is larger than the one of MP components and thus  $a \geq b$ . Therefore, Eq. (7) could be changed as

$$2a \geq (L_1 - L_2)/3d \tag{8}$$

Taking Eq. (8) into Eq. (6), we yield

$$\varepsilon \leq d \frac{(E_1 - E_2) - (P - E_1)}{L_1 - L_2} \tag{9}$$

Therefore, the maximum of DDC tracking error could be chosen as the modification of the output of the DDC discriminator. The output of proposed DDC discriminator is written as

$$\begin{aligned} D_{\text{Improved}} &= D_{\text{DDC}} + \text{Modification} \\ &= (E_1 - L_1) - \frac{E_2 - L_2}{2} + d \frac{(E_1 - E_2) - (P - E_1)}{L_1 - L_2} \end{aligned} \tag{10}$$

### 3.2 Out-Phase MP Signals

The geometry of ACF for BOC (1,1) in case of an out-phase MP signal with a half of the LOS amplitude is shown in Fig. 5.

There is no difference between LOS signal in two cases (in-phase MP and out-phase MP).

For MP ACF, there is the difference between two cases. In  $[-1 + \delta; 1 + \delta]$  chip offset, ACF is separated into 04 lines, each line is modeled by an equation, respectively. From left side to right side, 04 lines are represented as

$$\begin{aligned} y_5 &= b(x + 1 - \delta); & y_6 &= -b(3x + 1 - \delta) \\ y_7 &= b(3x - 1 - \delta); & y_8 &= -b(x - 1 - \delta) \end{aligned} \tag{11}$$

As the same as the scenario of in-phase MP signal, the ACF of the received BOC (1,1) signal is the sum of LOS signal and MP signal. The geometry of ACF is also separated into 09 parts of line and they are mathematical modeled as Eq. (4).

When the short-delayed out-phase MP signal ( $d = 0.1 \text{ chip}$ ,  $0.05 < \delta < 0.2$ ) is taken into account, the  $P$  correlator output is moved to the left of the true correlation peak. Resulting in the  $P$  is on the  $d_4$  line,  $E_1$  and  $E_2$  are also on the  $d_4$  line, whereas  $L_1$  is on the  $d_5$  line and  $L_2$  are on the  $d_6$  line. Therefore, based on Eqs. (4) and (11), the amplitudes of all correlator outputs are calculated as

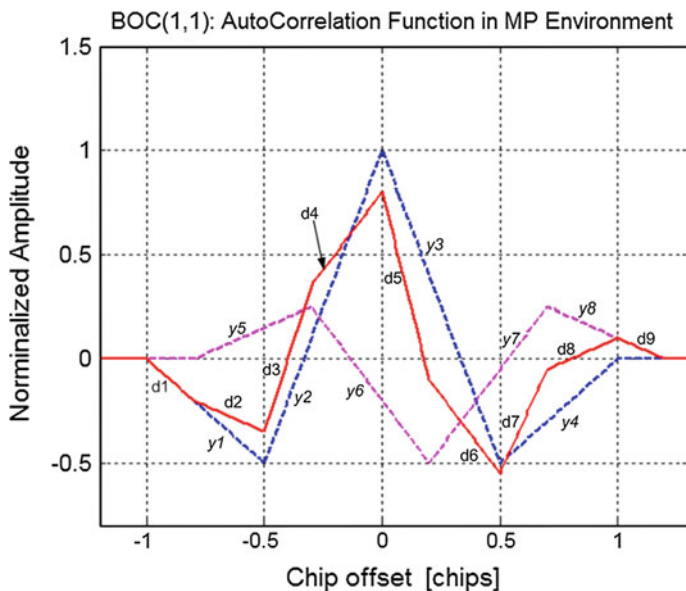


Fig. 5 Geometry of ACF for BOC (1,1) signal with a out-phase MP

$$\begin{aligned}
 P &= (a - b)3(-\varepsilon) + a - (1 - \delta)b \\
 E_1 &= 3(a - b)(-\varepsilon - d) + a - (1 - \delta)b \\
 E_2 &= 3(a - b)(-\varepsilon - 2d) + a - (1 - \delta)b \\
 L_1 &= -3(a + b)(-\varepsilon + d) + a - (1 - \delta)b \\
 L_2 &= -3(a - b)(-\varepsilon + 2d) + a - (1 + \delta)b
 \end{aligned}
 \tag{12}$$

As shown in [4], the difference between the distance of  $L_1$ ,  $L_2$  and the distance of  $L_1$ ,  $P$  may be used as an indicator of DDC tracking error. However, in this case,  $L_1$  and  $L_2$  are on two different lines and causes  $(L_1 - L_2) - (P - L_1)$  could not be used. Another Late correlator should be created for this purpose with the chip spacing  $d/2$  from  $P$  correlator and so called  $L_{12}$ . Therefore,  $L_{12}$  are on  $d_5$  and the amplitude of  $L_{12}$  output could be calculated as

$$L_{12} = -3(a + b)(-\varepsilon + d/2) + a - (1 - \delta)b
 \tag{13}$$

Hence, the DDC tracking error could be written by solving Eq. (12) and Eq. (13) as

$$(P - L_{12}) - (L_{12} - L_1) = -6a\varepsilon
 \tag{14}$$

The sum of maximum amplitude of LOS ACF and the one of MP ACF in this case is also derived from above equation as

$$(P - L_{12}) - (L_{12} - L_1) = -6a\epsilon \quad (15)$$

Finally, with the same calculation in in-phase case, the exactly DDC tracking error could be expressed as

$$\epsilon \leq d \frac{(L_{12} - L_1) - (P - L_{12})}{2(L_{12} - L_1)} \quad (16)$$

Therefore, the output of improved DDC discriminator is constructed as

$$\begin{aligned} D_{\text{Improved}} &= D_{\text{DDC}} + \text{Modification} \\ &= (E_1 - L_1) - \frac{E_2 - L_2}{2} + d \frac{(L_{12} - L_1) - (P - L_{12})}{2(L_{12} - L_1)} \end{aligned} \quad (17)$$

Equations (10) and (17) is the output of improved DDC discriminator. With these modifications, the DDC tracking error could be reduced for MP delays  $0.05 < \delta < 0.2$ .

## 4 Simulation Results

The simulations have been carried out for BOC (1,1) modulated signals for an infinite front-end bandwidth. In MEE analysis, several simplifying assumptions have been made in order to ensure the error source is only multipath signals. Such assumptions are: zero Additive White Gaussian Noise (AWGN), ideal infinite-length PRN codes and zero residual Doppler [9]. Moreover, the scenario is only one direct signal plus one multipath signal so that the analysis expression for MEE is not complicated. The simulations parameters are: number of paths is 02 paths; Chip spacing is  $d = 0.125$  chip; the ratio between the amplitudes of MP signal and direct signal is  $\alpha = 0.5$ .

For above configuration, as shown in Fig. 6, the ranging error due to multipath signal is about 15 m in the short-delayed MP scenario (chip-delayed  $0.05 < \delta < 0.2$ ). The improved DDC has reduced ranging error from 15 to 10 m for both cases in-phase (upper part) and out-phase (lower part) MP signals. For medium and long-delayed MP ( $0.5 < \delta < 1.0$ ), the performance of improved DDC is similar to the conventional DDC. It means that, the improved DDC focuses on improving the performance of DDC in the scenario of short-delayed MP signals. It's very important because the short-delayed MP signal occurs frequently.

If the chip spacing is reduced, the DDC tracking error of the conventional DDC as well as the improved DDC are reduced. Therefore, the ranging error is also reduced. However, the ranging error of the improved DDC is always smaller than the one of the conventional DDC.

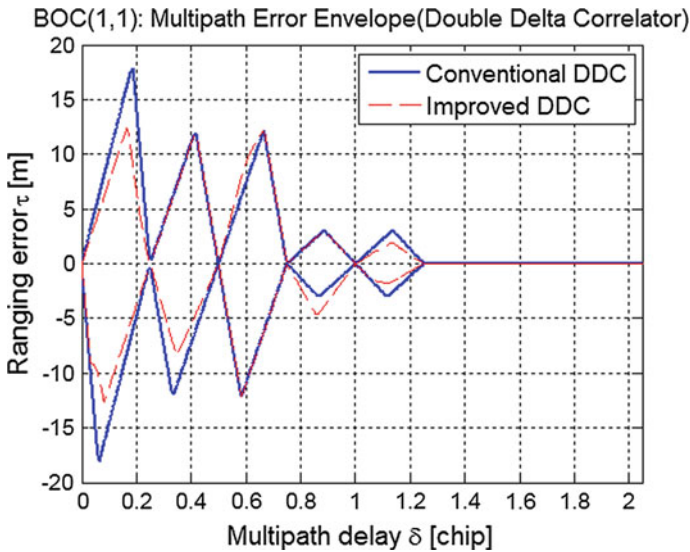


Fig. 6 MEE of improved DDC for BOC (1,1) signal

## 5 Conclusions

Multipath is one of major sources of ranging error in precise application in GNSS receivers. Many multipath mitigation methods have been researched and proposed in theoretical case. In non-parametric approach, Double Delta Correlator provides the best performance of multipath mitigation. It reduces almost multipath errors in the medium and long delayed multipath environment. For short-delayed multipath, DDC only achieves the multipath mitigation effect same as Narrow Correlator [4]. In this study, the improved DDC was proposed. It has improved the multipath mitigation performance in short-delayed multipath more than 30 % in comparison with the conventional DDC.

## References

1. Betz JW (2001) Binary offset carrier modulations for radio navigation. *J Inst Navig* 48(4):227–246
2. Borre K, Akos DM, Bertelsen N, Rinder P, Jensen SH (2007) A software-defined GPS and Galileo receiver—a single-frequency approach. Birkhäuser, Boston
3. Dierendonck AJV, Fenton P, Ford T (1992) Theory and performance of narrow correlator spacing in a GNSS receiver. *J Inst Navig* 39(3):265–283
4. Irsigler M, Eissfeller B (2003) Comparison of multipath mitigation techniques with consideration of future signal structures. In: The 16th international technical meeting of the satellite division of the institute of navigation (ION GNSS '03), pp 2584–2592



5. Dierendonck AJV, Braasch MS (1997) Evaluation of GNSS receiver correlation processing techniques for multipath and noise mitigation. The 1997 national technical meeting of the institute of navigation, pp 207–215
6. Zhu X, Chen X, Chen X (2011) Comparison between strobe correlator and narrow correlator on MBOC DLL tracking loop. In: Instrumentation and measurement technology conference (I2MTC), pp 1–4
7. Tawk Y, Botteron C, Jovanovic A, Farine PA (2010) Performance comparison of different correlation Techniques for the AltBOC modulation in multipath environments. In: 2010 IEEE international conference on communications (ICC), pp 1–6
8. Jovanovic A, Tawk Y, Botteon C, Farine PA (2010) Multipath mitigation techniques for CBOC, TMBOC and AltBOC signals using advanced correlators architectures. Position location and navigation symposium (PLANS), pp 1127–1136
9. Bhuiyan MZH, Lohan ES (2010) Advanced multipath mitigation techniques for satellite—based positioning applications. Int J Navig Obs 1–15 Hindawi Publishing Corporation

# Implementation of Automatic Failure Diagnosis for Wind Turbine Monitoring System Based on Neural Network

Ming-Shou An, Sang-June Park, Jin-Sup Shin, Hye-Youn Lim  
and Dae-Seong Kang

**Abstract** The global action began to resolve the problem of global warming. Thus, the wind power has been emerged as an alternative energy of existing fossil fuel energy. The existing wind power has limitation of location requirements and noise problems. In case of Korea, the existing wind power has difficulties on limitation of location requirements and the noise problems. The wind power turbine requires bigger capacity to ensure affordability in the market. Therefore, expansion into sea is necessary. But due to the constrained access environment by locating sea, the additional costs are occurred by secondary damage. In this paper, we suggest automatic fault diagnosis system based on CMS (Condition Monitoring System) using neural network and wavelet transform to ensure reliability. In this experiment, the stator current of induction motor was used as the input signal. Because there was constraint about signal analysis of large wind turbine. And failure of the wind turbine is determined through signal analysis based wavelet transform. Also, we propose improved automatic monitoring system through neural network of classified normal and error signal.

**Keywords** Wavelet transform · Neural network · Automatic failure diagnosis · CMS · Wind turbine monitoring system

## 1 Introduction

Due to the maturity of wind turbine technology, the unit cost of wind power was decreased. Therefore, the unit cost of wind power is similar to the cost of existing fossil fuels compared to other renewable energy. The global action also began to

---

M.-S. An · S.-J. Park · J.-S. Shin · H.-Y. Lim · D.-S. Kang (✉)  
Department of Electronics Engineering, Dong-A University, 840 Hadan 2-Dong,  
Saha-Gu, Busan, Korea  
e-mail: dskang@dau.ac.kr

resolve the problem of global warming. Thus, the wind power has been emerged as an alternative energy of existing fossil fuel energy. Because of technological advances over the past 20 years, the wind turbines have become bigger. However, the larger turbines inevitably increase tower height and blade length. Also, the components of wind turbine increase mechanical and electrical permit capacity. Consequently, failure rate of turbine would increase [1]. The existing wind power has limitation of location requirements and noise problems. In case of Korea, the existing wind power has difficulties on limitation of location requirements and the noise problems. The wind power turbine requires bigger capacity to ensure affordability in the market. Therefore, expansion into sea is necessary, but there are some problems for the operation of large wind turbines at sea. The first, mechanical and electrical failures of larger wind turbine are increased. Second, due to the constrained access environment by locating sea, the additional costs are occurred by secondary damage. Therefore, monitoring technology of wind turbine is essential for utilization and reliability [2]. In this paper, failure of the wind turbine is determined through signal analysis based wavelet transform. Also real-time signal analysis was made through wavelet transform. Feature information of classified signal pattern through signal analysis had been learned by using neural network algorithm to implement automatic fault diagnosis system.

## **2 Condition Monitoring Techniques of Wind Turbine**

The surveillance system of wind power is classified into supervisory control and data acquisition (SCADA) system and condition monitoring system (CMS). The SCADA system remotely performs control function of wind turbine in conjunction with turbine controller. It is an essential component to consist of wind power system. However, the current wind turbines are difficult to identify the operating condition because SCADA systems are made differently per the turbine manufacturer.

The other hand, the CMS is prevention system that diagnoses malfunction in advance by closely monitoring, analyzing and predicting component of wind turbine. In the past, the CMS was regarded as an optional component. In case of large and wind turbine located sea, the CMS is recognized as an essential component because of the matter of credibility. In this paper, the stator current of induction motor was used as a input signal. Because the signal analysis of large offshore wind turbines is constrained. Overall, we must ensure the reliability by suggesting automatic fault diagnosis system based on CMS through neural network on the feature information of signal patterns and analysis using the wavelet transform.

### 2.1 Signal Analysis Using Wavelet Transform

The wavelet analysis appeared by integrating special techniques individually developed to meet special purpose belonging to signal processing system. Basic techniques of computer vision using multi-resolution analysis method, sound and video compression using sub-band coding technique and applied mathematics using wavelet series are developed recently into wavelet theory’s special applications. The wavelet transform can understand that input signals are separated into the set of the basis function. The set of basis function used to wavelet transform can be obtained through expansion, reduction and parallel transference of time axis about basis function of wavelet. Basis function of wavelet indicates band-pass filter of special form. And the relative bandwidth invariability of wavelet transform is satisfied by expansion and reduction of temporal axis about wavelet basis function. Thus, the scale is called instead of frequency band in wavelet transform. Unlike Fourier transform, wavelet transform includes high resolvent ability about the scale of signal. Therefore, wavelet transform is time–scale transform.

Wavelet equation about time-domain and frequency-domain is

$$\phi_{ab}(x) = 1/2\phi((x - b)/a), \tag{1}$$

where a means expansion and b which means movement indicates the temporal position. The more a is increased, the more resolution of frequency is increased. The scaling is an expanding or reducing signal. Large-scaling signal is expanded. And small scaling signal responds to the compression [3]. Integration for entire signal of Fourier transform causes high-frequency component by diversion in the finite space, it’s difficult to handle flexibly non-stationary function because the analysis about variation of frequency has limit. To overcome these demerits, various analytical methods were devised and wavelet transform among the rest is efficiently used [4]. After studying orthogonal basis functions, discrete wavelet transform (DWT) is developed. DWT which indicates two space of separated frequencies for the signal analysis is defined as follows:

$$x(t) = \sum_{n=-\infty}^{\infty} a_{0,n}(t), g_{0,n}(t) + \sum_{j=0}^{\infty} \sum_{n=-\infty}^{\infty} d_{j,n}(t), h_{j,n}(t), \tag{2}$$

where  $a_{0,n}(t)$  is component of the signal,  $g_{0,n}(t)$  represents low-pass filter (LPF),  $d_{j,n}(t)$  is detail component of the signal and  $h_{j,n}(t)$  is high-pass filter (HPF).

In Fig. 1, it is possible to regenerate two-band about the filtering signal. In other words, if separation of 2-level is finished, the original signal is separated into frequency of four bands [5].

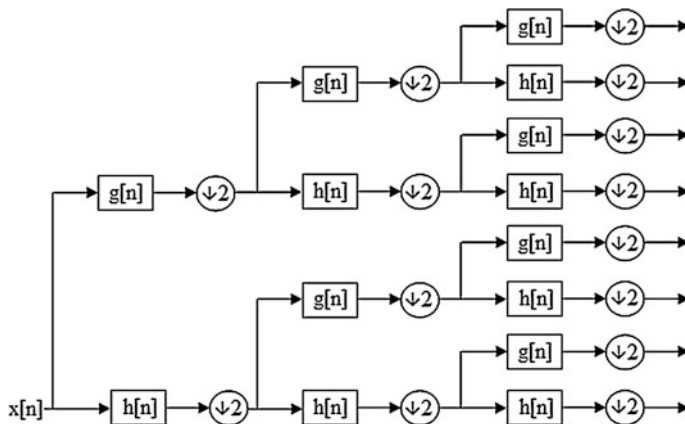


Fig. 1 2-level separation of image signal

### 2.2 Feature Extraction of Signal Pattern Through Signal Analysis

The wavelet transform is used for noise rejection in signal analysis. In this paper, because of constrained environments, the signal analysis of large wind turbines was substituted for input signal using the stator current of an induction motor. Figure 2 shows the removed noise of input signal using the wavelet transform.

There is very wide variety of wavelet filter according to its purpose. Figure 3 also shows that the coefficients of the filter increase in accordance with N value of Db(N). In this paper, Db(4) is used to analyze input signal of stator current removed noise. Very short basis functions need to suggest discontinuity of input signal, and very long low-frequency basis functions need to analyze frequency

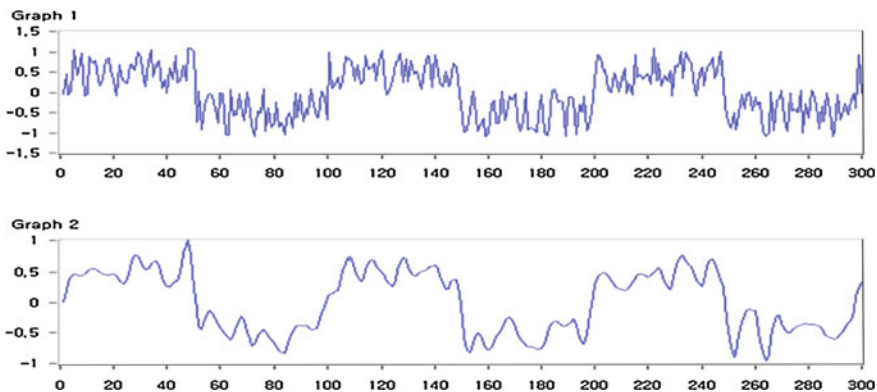


Fig. 2 Noise removal using the wavelet transform

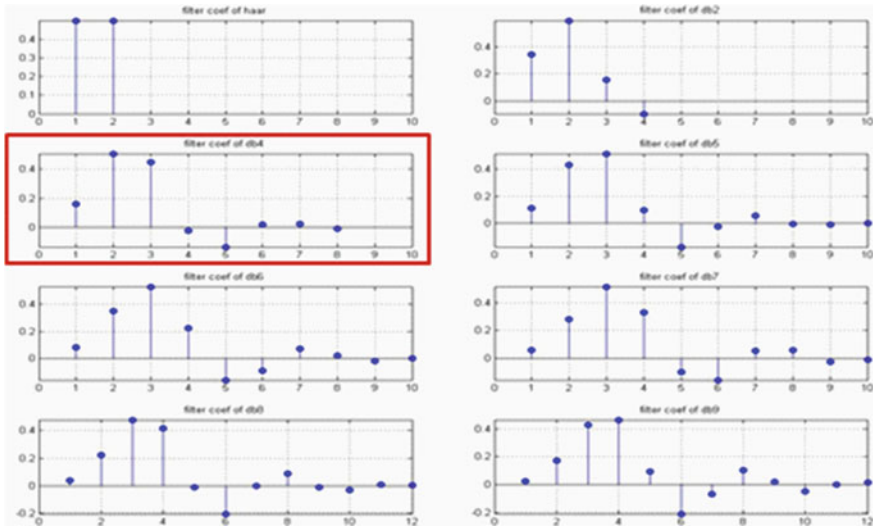


Fig. 3 Db(N) DWT filters

accurately. We simultaneously use two basis functions based on Daubechies wavelet. So, we easily can get the ambiguous feature information on the signal analysis of time–frequency domain.

### 2.3 Neural Network Modeling

The primary function of neuron calculates input and weighted summing NET of connection strengths in artificial neural network. And the output comes out by the activation function. Therefore, output of neuron is different depending on the activation function.

Step and sigmoid functions are simple activation functions. When the input is over the threshold value, the output is a function that is activated as 1. The sigmoid function for the change of the value has a form to infinitely approach 0 and 1. In other words, the sigmoid function linearly translates disorderly nonlinear values in neural network model.

### 2.4 Neural Network Learning Algorithm of Back Propagation

The back propagation (BP) algorithm which also called error back propagation algorithm is used to be applied multiple neural network. BP is an universal neural network algorithm used to various fields.

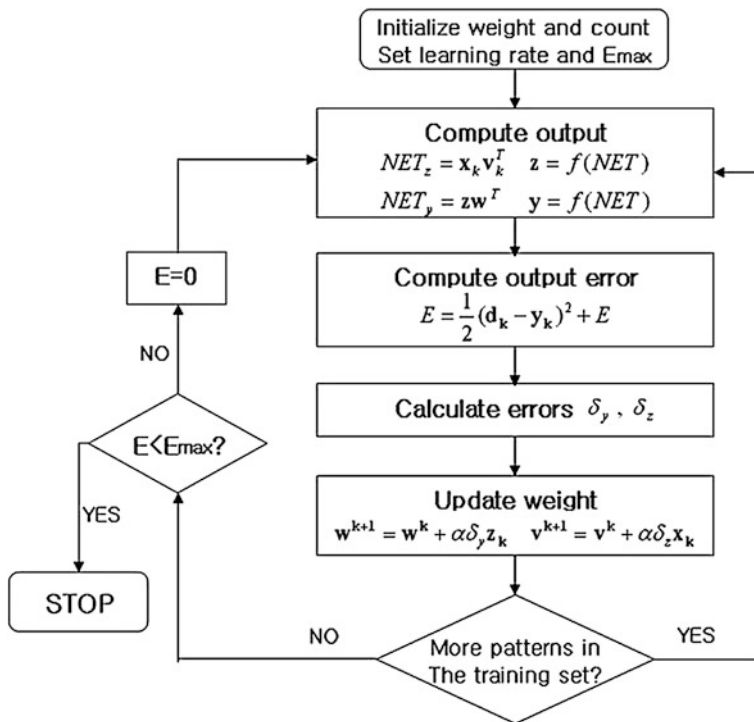


Fig. 4 The flow chart of BP algorithm

The renewal of the connection strength is the most important part in the learning algorithm. The BP algorithm consists of the forward and backward steps. As with other neural network learning algorithm, the learning is made by renewal of the connection strength. Figure 4 is a three-tier structure of BP neural network. The input layer enters classified signal pattern through signal analysis into neural network. And it is multiplied with connection weights connected to the hidden layer. The multiplied values are passed to hidden layer. Through repetition of this process, the end result of output layer is obtained. Error is calculated through the subtraction of the output and the target figure [6].

### 3 Experimental Results of Automatic Fault Diagnosis System

In this paper, to implement automatic fault diagnosis of offshore wind turbine, we conducted an experiment to improve the reliability by suggesting neural network algorithm and wavelet transform based on CMS for signal analysis.

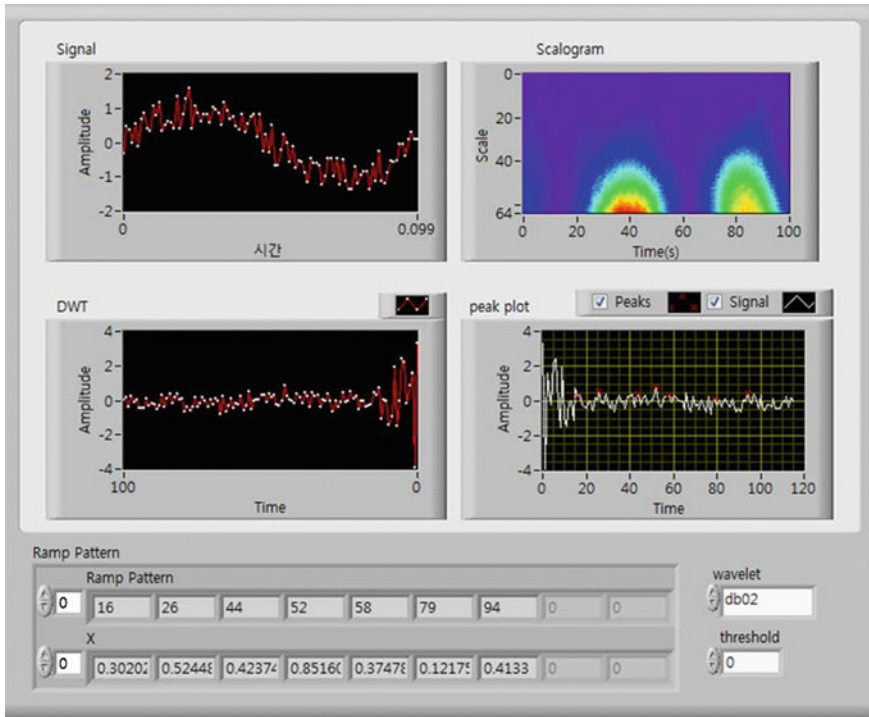


Fig. 5 The automatic failure diagnosis system based on LabView

Figure 5 shows the overall system of automatic failure diagnosis through learning using feature information of classified signal pattern. In this paper, we extracted peak value of classified signal using peak detection function based on LabView for extraction of feature information. And, extracted feature information is used as input to the learning of neural network algorithm to implement automatic failure diagnosis system of offshore wind turbine. In this study, we early diagnose a complex fault signal that occurs over a long time rather than a simple failure. And there are aims to reduce secondary damage. Also, we look forward to applying to a variety of monitoring environments.

**Acknowledgments** This work was supported by the Human Resources Development of the Korea Institute of Energy Technology Evaluation and planning (KETEP) grant funded by the Ministry of knowledge Economy, Republic of Korea (No. 20114010203060).



## References

1. Robi P The wavelet tutorial-fundamental concepts and an overview of the wavelet theory, 2nd edition
2. Park JY (2012) Development of wind power integrated condition monitoring system, Korea Electrical Contractors Association, pp 56–63, Feb 2012
3. Kim CH, Kim H, Ko YH, Byun SH, Aggarwal RK, Allan TJ (2002) A novel fault-detection technique of high-impedance arcing faults in transmission lines using the wavelet transform. *IEEE Trans Power Deliv* 17(4):921–929
4. Mallat S (1991) Zero crossings of a wavelet transform. *IEEE Trans Inf Theory* 37(4):1019–1033
5. Wenxian Y, Tavner PJ, Michael W (2008) Wind Turbine condition monitoring and fault diagnosis using both mechanical and electrical signatures. In: Proceedings of the 2008 IEEE/ASME international conference on advanced intelligent mechatronics, pp 1296–1301, July 2008
6. He Q, Du DM (2007) Fault diagnosis of induction motor using neural networks. In: Proceedings of the 6th international conference on machine learning and cybernetics. vol 2, pp 1090–1095, Aug 2007

# Development of Compact Microphone Array for Direction-of-Arrival Estimation

Trình Quốc Võ and Udo Klein

**Abstract** Direction-of-arrival estimates are required in many applications such as automatic video camera steering and multiparty teleconferencing for beam forming and steering to suppress noise and reverberation and improve speech intelligibility. Ambient noise and multiple reflections of the acoustic source signal significantly degrade the performance of time-difference-of-arrival (TDOA) methods to localize the sound source using only two microphones. In this work, we investigate the performance of a multichannel cross-correlation coefficient (MCCC) algorithm for the estimation of the direction-of-arrival (DOA) of an acoustic source in the presence of significant levels of both noise and reverberation. Simulations and initial experimental results confirm that the DOA estimation robustness and complexity is suitable for a practical micro-phone array using miniature MEMS microphones and an FPGA implementation of the MCCC algorithm.

**Keywords** Direction of arrival estimation · Microphone arrays · Signal processing algorithms · Time of arrival estimation

## 1 Introduction

The basic idea to solve the DOA problem is to determine the time difference of a sound source signal arriving at two microphone locations. If the sound source is located in the far-field and the distance  $b$  between the two microphones is known,

---

T. QuốcVõ · U. Klein (✉)  
School of Electrical Engineering, International University, Vietnam National University,  
HCMC, Ho Chi Minh City, Vietnam  
e-mail: u.klein@ieee.org

the TDOA  $\tau_{12} = (b \cos \theta)/v_a$  is directly related to the DOA angle  $\theta$ . Here,  $v_a$  is the sound velocity in air.

In order to improve the estimate of the DOA in noisy and reverberant environments many algorithms have been proposed using multiple microphones. The linear spatial prediction method [1], the multichannel cross-correlation coefficient algorithm [1], and the broadband multiple signal classification (MUSIC) method [2] all employ the correlation of the aligned microphone signals to estimate the TDOA. The minimum entropy (ME) method [3] uses higher order statistics that could be more suitable for non-Gaussian source signals such as speech. The most reliable TDOA estimation performance in reverberant environments is achieved by adaptive blind multichannel identification (ABMCI) [4], which relies on the blind identification of the real reverberant impulse response functions of the SIMO system, consisting of the single signal source and multiple microphones. Both ME and ABMCI are considered to give better results than cross-correlation based techniques, although at the cost of higher computational requirements, thereby increasing the hardware complexity and cost.

Because of its relatively low computational complexity the MCCC method has been selected for a low-cost hardware implementation of a microphone array for DOA estimation of an acoustic source in the presence of significant levels of both noise and reverberation. The MCCC method uses the redundancy of the microphone signals by applying an extension of the generalized cross-correlation proposed by Knapp and Carter [5]. In order to evaluate the performance of the MCCC method the algorithm has been implemented in MATLAB<sup>®</sup> and simulation results are presented in this paper. Following our simulations we have set up a demonstration system using an analog MEMS microphone array and an analog-to-digital (A/D) converter as the interface to the MCCC computer program. First experimental results indicate that the simulated performance can be achieved with compact low-cost hardware using a digital MEMS microphone array and an FPGA implementation of the MCCC algorithm.

## 2 The MCCC Algorithm

The microphone array consists of  $L$  microphones in a linear equidistantly spaced array, from the 1st to the  $L$ th microphone. The delay between the 1st and the  $l$ th microphones is then given by  $f_l = (l - 1)\tau$  where  $\tau$  is the time delay between two neighboring microphones. For the application of the MCCC algorithm, we consider the column vector of the aligned signals at the  $L$  microphones

$$\mathbf{x}_{1:L}[n - f_L(m)] = [x_1[n - f_L(m) + f_1(m)] \quad x_2[n - f_L(m) + f_2(m)] \cdots x_L[n]]^T$$

with  $m/f_s = \hat{\tau}$  as a guess for the delay, where  $f_s$  is the sampling frequency. The spatial correlation matrix of the microphone signals can be factored as

$$\mathbf{R}_{m,1:L} = \mathbb{E}\{\mathbf{x}_{1:L}[n - f_L(m)] \cdot \mathbf{x}_{1:L}^T[n - f_L(m)]\} = \mathbf{D}\tilde{\mathbf{R}}_{m,1:L}\mathbf{D}$$

with the diagonal matrix  $\mathbf{D}$  and the symmetric matrix  $\tilde{\mathbf{R}}_{m,1:L}$  defined as

$$\mathbf{D} = \begin{bmatrix} \sqrt{\mathbb{E}\{x_1^2[n]\}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\mathbb{E}\{x_L^2[n]\}} \end{bmatrix}, \quad \tilde{\mathbf{R}}_{m,1:L} = \begin{bmatrix} 1 & \cdots & \rho_{m,1L} \\ \vdots & \ddots & \vdots \\ \rho_{m,L1} & \cdots & 1 \end{bmatrix},$$

and the cross-correlation coefficients between  $x_k[n - f_i(m)]$  and  $x_l[n - f_k(m)]$

$$\rho_{m,kl} = \frac{\mathbb{E}\{x_k[n - f_i(m)]x_l[n - f_k(m)]\}}{\sqrt{\mathbb{E}\{x_k^2[n]\}\mathbb{E}\{x_l^2[n]\}}}, \text{ with } k \text{ and } l = 1, 2, \dots, L.$$

The multichannel cross-correlation coefficient is now defined as [1]

$$\rho_{m,1L}^2 = 1 - \det \tilde{\mathbf{R}}_{m,1:L}.$$

The delay estimation is then based on maximizing the cross-correlation coefficient  $\rho_{m,1L}^2$  or by minimizing the cost function, defined as the determinant of the matrix  $\tilde{\mathbf{R}}_{m,1:L}$ , with respect to the guessed delay  $m$ .

### 3 Simulation

**Simulation Parameters.** In order to simulate the reverberant acoustic environment the image-source method for room acoustics has been employed [6]. Walls, ceiling, and floor of the room are characterized by frequency-independent and incident-angle-independent reflection coefficients. The dimensions of the room are chosen to be 5 m by 5 m by 2.5 m. Reflection coefficients  $r_i$  ( $i = 1, 2, \dots, 6$ ) are varied between 0 and 0.8. The sound source is located at the position (4.0 m, 2.5 m, 1.5 m). For the simulations, up to eight microphones are placed in parallel with the  $x$ -axis and with a spacing of 11 cm. The first microphone is located at (2.88 m, 0.5 m, 1.3 m) and the last is at (2.11 m, 0.5 m, 1.3 m). Gaussian noise is added to each microphone signal with all noise signals uncorrelated both with the source signal and with the noise at the other microphones. The signal-to-noise ratio (SNR) has been varied between  $-5$  dB and 20 dB. A 20 s recorded speech signal was sampled at  $f_s = 16$  kHz with 16-bit resolution and has been used as the sound source for the simulations. All data processing has been performed on signal frames of 512 samples, corresponding to a sampling length of 32 ms.

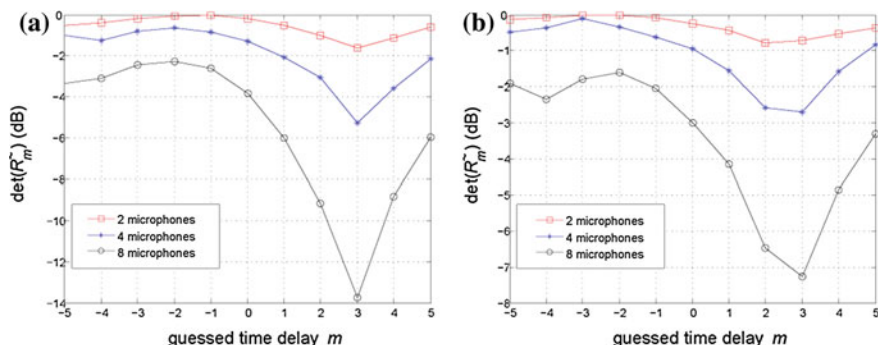
The desired resolution of the estimated DOA angle has been chosen to be better than 20 degrees. This means the DOA estimation system needs to be able to detect a maximum delay  $M$  between two neighboring microphones of 5 samples. The relation between the minimum spacing  $b_{\min}$  between two neighboring

microphones, the sampling frequency  $f_s$  and the maximum delay  $M$  is given by  $b_{\min} \geq Mv_a/f_s$ . The resulting minimum distance between two neighboring microphones is  $b_{\min} = 10.7$  cm, leading to the selected microphone separation of  $b = 11$  cm.

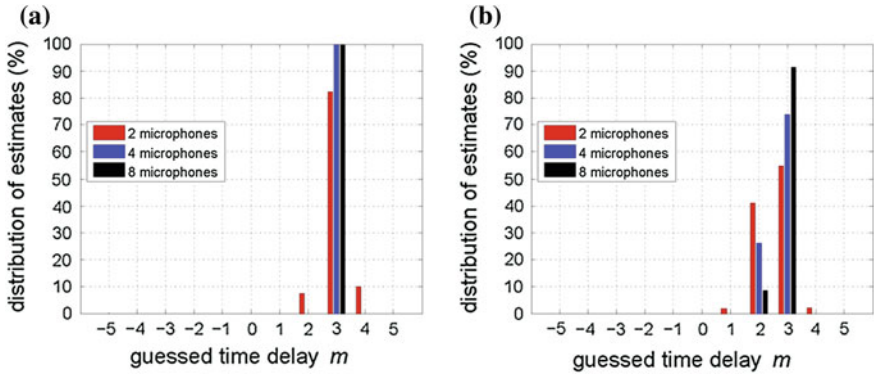
**Simulation Results.** The robustness of the MCCC algorithm for a given source signal depends on the SNR, the degree of reverberation, and the number of microphones in the array. Figure 1a shows the cost function as a function of the guessed delay  $m$  in a reverberant-free environment with 0 dB SNR and using two, four, and eight microphones, respectively. Distinctive minima exist at the correct delay estimate of  $m = 3$ . For a strongly reverberant environment comparable to a typical meeting room, the cost functions in Fig. 1b have much less pronounced minima and for  $L = 2$  microphones the minimum at  $m = 2$  results in an incorrect delay estimate.

Typical distributions of the time delay estimates for repeatedly applying the MCCC algorithm in a reverberant-free environment and in a strongly reverberant environment with 0 dB SNR are shown in Fig. 2. In a reverberant-free environment and with two microphones, only 80 % of the estimates correspond to the true delay of 3 samples while four- and eight-microphone arrays do not show any incorrect estimates. In a strongly reverberant environment, even an eight-microphone array only achieves 80 % correct delay estimates and a two-microphone array displays erroneous estimates of up to 2 samples off the true delay.

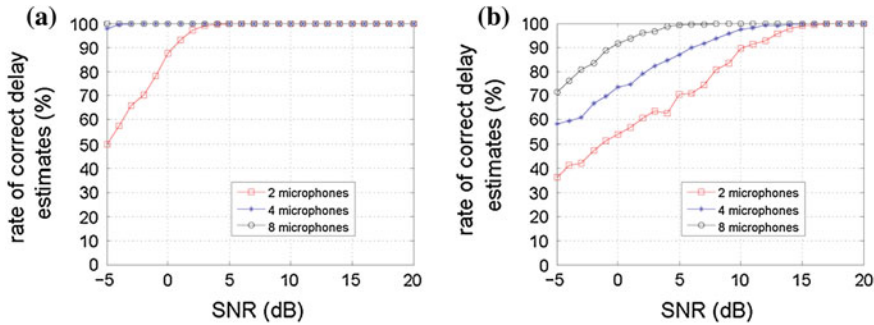
Figure 3 shows the percentage of correct delay estimates in noisy environments with and without reverberation as a function of the SNR. The robustness of the algorithm to estimate the delay correctly increases with the SNR and the number of microphones in the array. The simulation shows that at a SNR of  $-5$  dB in a reverberant-free environment the percentage of successful delay estimates is only about 50 % for two microphones. However, the rate of correct estimates reaches almost 100 % if an array with four or more microphones is used. In a strongly reverberant environment with wall reflection coefficients of  $r_i = 0.8$  the



**Fig. 1** MCCC cost function  $\det \tilde{R}_m$  as a function of the guessed delay  $m$  for microphone arrays with  $L = 2, 4,$  and  $8$  microphones and with a SNR of 0 dB (The true delay is 3 samples); **a** reverberant-free and **b** reverberant environment ( $r_i = 0.8$ )



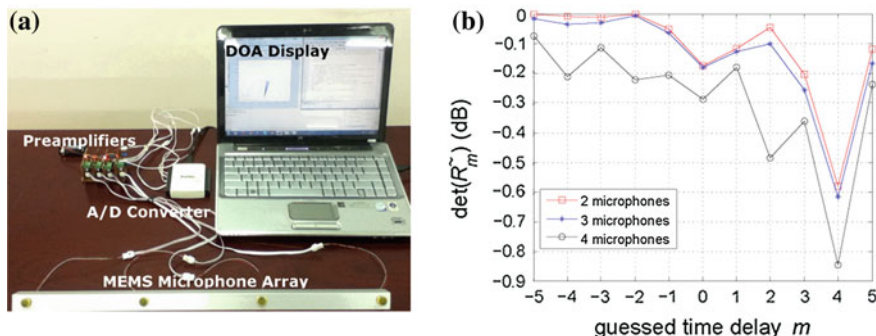
**Fig. 2** Distribution of time delay estimates for repeatedly applying the MCCC algorithm 1,000 times with  $L = 2, 4,$  and  $8$  microphones and with a SNR of  $0$  dB (The true delay is  $3$  samples); **a** reverberant-free and **b** reverberant environment ( $r_i = 0.8$ )



**Fig. 3** Robustness of the MCCC algorithm in terms of percentage of correct delay estimates as a function of the SNR for arrays with  $L = 2, 4,$  and  $8$  microphones; **a** reverberant-free and **b** reverberant environment ( $r_i = 0.8$ )

percentage of correct delay estimates is significantly reduced and the eight-microphone array requires a SNR of better than  $5$  dB to achieve nearly  $100\%$  correct delay estimates.

The simulations confirm the robustness of the MCCC algorithm in estimating the TDOA in both noisy and reverberant environments. A linear equidistantly spaced array of eight microphones performs reliably in strongly reverberant environments with SNRs down to  $5$  dB and with signal durations as short as  $32$  ms.



**Fig. 4** Demonstration system with 4-channel microphone array; **a** Analog MEMS microphone array, preamplifiers, and A/D converter interface to the MCCC algorithm in MATLAB<sup>®</sup> **b** MCCC cost function  $\det(\tilde{R}_m)$  as a function of the guessed delay  $m$  for microphone arrays with  $L = 2, 3,$  and 4 microphones in a reverberant meeting room environment

## 4 Demonstration System

We have set up a demonstration system using a microphone array with analog MEMS microphones and an analog-to-digital (A/D) converter as the interface to the MCCC algorithm in MATLAB<sup>®</sup> (Fig. 4a). The limitations of the A/D converter allow a maximum of four microphone channels at a sample rate of 10.5 kS/s with a resolution of 14 bits. Because of the reduced sample rate compared to the simulation, the distance between the microphones is increased to 17.5 cm. For a sound source at a distance of 2.5 m and at an angle of 35° to the microphone array, and for a SNR of about 10 dB, the cost function is shown in Fig. 4b for microphone arrays with 2, 3, and 4 channels. The estimated delay of  $m = 4$  corresponds to a DOA range between 28° and 47°, which includes the correct DOA of 35°.

## 5 Conclusion

Our investigation confirms that the MCCC algorithm is a suitable candidate for reliable TDOA estimation in real-world environments with a minimum amount of computational cost. The TDOA can be estimated by finding the minimum of the determinant of the cross-correlation coefficient matrix of the aligned microphone array signals.

Results with a 4-channel demonstration system show that the simulated performance is achievable with miniature analog MEMS microphones and an analog-to-digital converter as the interface to the MCCC algorithm in MATLAB<sup>®</sup>. The full hardware implementation will employ an FPGA and a digital MEMS microphone array. The digital MEMS microphones integrate the microphone, amplifier, and A/D converter in a single component, thereby reducing the system

complexity considerably. The system will be smaller, cheaper, and more flexible than conventional analog microphone arrays. In order to evaluate its performance the MEMS microphone arrays will be used to record test data for source localization experiments.

**Acknowledgments** The authors would like to thank Pirmin Rombach and Armin Schober from EPCOS AG for providing free samples of their MEMS microphones.

## References

1. Chen J, Benesty J, Huang Y (2003) Robust time delay estimation exploiting redundancy among multiple microphones. *IEEE Speech Audio Process* 11:549–557
2. Dmochowski JP, Benesty J, Affes S (2007) Broadband MUSIC: Opportunities and challenges for multiple source localization. In: 2007 IEEE Workshop on the applications of signal processing to audio and acoustics, pp 18–21
3. Benesty J, Huang Y, Chen J (2007) Time delay estimation via minimum entropy. *IEEE Signal Process Lett* 14:157–160
4. Huang Y, Benesty J (2002) Adaptive multi-channel least mean square and newton algorithms for blind channel identification. *Signal Process* 82:1127–1138
5. Knapp CH, Carter GC (1976) The generalized correlation method for estimation of time delay. *IEEE Trans Acoust Speech Signal Process* 24:320–327
6. Lehmann EA (2012) Image-source method: MATLAB code implementation. <http://www.eric-lehmann> 10 Mar 2012



# Design and Implementation of a SoPC System for Speech Recognition

Tran Van Hoang, Nguyen Ly Thien Truong, Hoang Trang  
and Xuan-Tu Tran

**Abstract** This paper presents the design of a System on Programmable Chip (SoPC) based on Field Programmable Gate Array (FPGA) for speech recognition in which Mel-Frequency Cepstral Coefficients (MFCC) for speech feature extraction and Vector Quantization (VQ) for recognition are used. The execution speed of the blocks in the speech recognition system is surveyed by calculating the number of clock cycles while executing each block.

**Keywords** Speech recognition · MFCC · VQ · SoPC · FPGA · Nios

## 1 Introduction

Speech recognition system is applied in many application fields such as health care, military, human computer interaction, avionics technicians... [1], especially, the applications which support disabled people to communicate with the world in a better way. For that reason, there are many studies on software/hardware implementation of speech recognition systems for many years. However, because of a large number of accents spoken around the world, there are still many challenges that need further research and development, for example, Vietnamese speech recognition.

---

T. Van Hoang · N. L. T. Truong · H. Trang (✉)  
University of Technology, Vietnam National University, HoChiMinh City, Vietnam  
e-mail: hoangtrang@hcmut.edu.vn

T. Van Hoang  
e-mail: tvhoang@hcmut.edu.vn

N. L. T. Truong  
e-mail: nltruong@hcmut.edu.vn

X.-T. Tran  
VNU University of Engineering and Technology, 144 Xuan Thuy, Hanoi, Vietnam  
e-mail: tutx@vnu.edu.vn

The research on speech recognition going mainly in two directions, namely: the software runs on Personal Computers (PCs) and embedded systems. For the first direction, many studies and software tools have been developed successfully. In particular, the Hidden Markov Model Toolkit (HTK) is a toolkit for building Hidden Markov Models (HMMs) used in speech recognition successfully [2]. There are also many tools running on the PC or smart phone aimed at the control device via speech. For the second direction, embedded systems have many advantages as high performance, convenience, low cost, and great development potential. However, speech recognition research based on embedded systems is more difficult. This paper will present the implementation of a speech recognition system as an embedded system using FPGA technology.

In fact, the implementation of speech recognition systems has been done using FPGA technology in recent years. In paper [3], speech recognition systems are implemented as hardware/software co-design systems using Hidden Markov Model (HMM). This project use Linear Predictive Coding (LPC) method in feature extraction block. So, the recognition accuracy is not high compared with the MFCC method. In paper [4], the MFCC method is applied, but the optimization was not taken into account yet to increase performance.

Another work, presented in [5] and [6], the author proposed an efficient MFCC hardware implementation for feature extraction in speech recognition. However, this work has been done using ASIC technology and therefore less flexible than FPGA based implementations. Other implementations for speech recognition systems can be found at [7–9]. Among these, the work presented in 8 proposes a hardware/software co-design method to tradeoff between the performance and the flexibility of the recognition system while [7] and [9] present FPGA based implementation of the recognition systems. None of them discuss about the optimization method for MFCC algorithm. In our work, the MFCC method is used with some modifications to increase the performance of the system. The whole system has been implemented using Altera FPGA technology to be more flexible.

The paper is organized as follows. The design and implementation of the proposed speech recognition system as a SoPC (System on Programmable Chip) is mentioned and discussed in [Sect. 2](#). [Section 3](#) will show the achieved experimental results. Finally, conclusions and further discussions will be presented in [Sect. 4](#).

## 2 Implementation

In the implementation process, some blocks will be adjusted, modified so that the computing speed of the block can be increased. In this section, we will show some improvements in a few blocks to optimize the computing speed of the block. Evaluated results in terms of the number of clock cycles will be presented in the next section.

## 2.1 Feature Extraction Implementation

### Pre-emphasis

In pre-emphasis block, the coefficient “ $a$ ” has the value from 0.9 to 1. In theory, the normal value of “ $a$ ” is 0.97. But, when we build the system on SoPC, we must choose the value of “ $a$ ” so that the program easy to implement. Thus, the pre-emphasis block will run faster. The value  $a = 1, 15/16, 0.97$  are surveyed about program performance through assessment of pulse clock.

Transfer function of the filter is described by Eq. 1. In the time domain, the relationship between output and input is shown in Eq. 2.

$$H(z) = 1 - a \cdot z^{-1} \tag{1}$$

$$s'_i = s_i - a \cdot s_{i-1} \tag{2}$$

With  $a = 1$ , Eq. 2 will be simplified as:  $s'_i = s_i - s_{i-1}$ .

Advantage of using 15/16 as “ $a$ ” coefficient is expressed in Eq. 3.  $\frac{15}{16}s_{i-1}$  can be realized in binary computation system by shifting  $s_{i-1}$  4 bits to the right. Using this value the multiplication step is simplified to shift and subtract operations.

$$s'_i = s_i - a \cdot s_{i-1}, \quad a = \frac{15}{16} \tag{3}$$

$$s'_i = s_i - \frac{15}{16}s_{i-1} = s_i - \left( s_{i-1} - \frac{1}{16}s_{i-1} \right)$$

### Discrete Fourier Transform (DFT)

In general,  $X(k)$  and  $x(n)$  are the complex numbers. N-point DFT can be calculated as follows:

$$X_R(k) = \sum_{n=0}^{N-1} \left[ x_R(n) \cos \frac{2\pi kn}{N} + x_I(n) \sin \frac{2\pi kn}{N} \right], \quad k = 0, 1, 2, \dots, N - 1. \tag{4}$$

$$X_I(k) = - \sum_{n=0}^{N-1} \left[ x_R(n) \sin \frac{2\pi kn}{N} - x_I(n) \cos \frac{2\pi kn}{N} \right], \quad k = 0, 1, 2, \dots, N - 1. \tag{5}$$

If DFT transformation uses two Eqs. 4 and 5 to calculate, it costs  $2N^2$  trigonometric calculations,  $4N^2$  real multiplications, and  $4N(N - 1)$  additions. This shows that when the direct calculation using the DFT formula above arises large computational cost, it will slow speed program execution. Therefore, in this case we use the Fast Fourier Transform (FFT) algorithm instead. In addition, by using the look-up table of coefficients cosine, sine also increases the computing speed of the program.

### Magnitude computation

If using the conventional formula for calculating the complex amplitude as Eq. 6, then the calculation will be very slow speed, thereby reducing the speed of program execution.

$$M = \sqrt{I^2 + Q^2} \quad (6)$$

Therefore, the estimation algorithm is applied. This algorithm calculates very fast amplitude of a complex number almost exact compared to the normal range by taking the square root operation. For complex number  $I + jQ$ , amplitude estimation algorithm as follows:

$$M \approx \alpha \cdot \max\{|I|, |Q|\} + \beta \cdot \min\{|I|, |Q|\} \quad (7)$$

In this system, we use  $\alpha$  as 1 and  $\beta$  as 1/4. This approach reduces the number of calculations with acceptable error.

### Mel frequency filter bank

The  $k$ th of power coefficient of the  $n$ th frame is calculated by the Eq. 8 as

$$S'_{nk} = \sum_j S_{nj} \cdot FC_{kj}, \quad k = 0, 1, \dots, K \quad (8)$$

where,  $K$  is the number of the filters.  $S_{nj}$  is the  $j$ th point of the  $n$ th frame's spectrum, and  $FC_{kj}$  is the  $j$ th coefficient of the  $k$ th filter. When implementing the speech recognition system on SoPC, the rectangular filter bank is used in the new algorithm instead of the triangular filter bank. So, the Eq. 8 becomes

$$S'_{nk} = \sum_j S_{nj} \cdot FC_{kj}, \quad FC_{kj} = 0 \text{ or } 1 \quad (9)$$

The rectangular filters are proposed to be used instead of the triangular filters because the output characteristic of a rectangular filter is either a "1" or a "0", the multiply and sum operations can be simplified to simple "add" and "no add" operations. No multiplication step is required in the proposed approach.

## 2.2 Training and Recognition Implementation

In this work, training process is done by using Vector Quantization. Codebook size of 128 is considered. K-Mean algorithm is used for training codebook. First,  $M$  vectors in the  $L$  vectors are randomly chosen for training. The second step, for each training vector  $v$ , we find the codeword in the current codebook vectors closest distance this vector and we assign it belongs to the group of the codeword. The third step, for each group, codeword is updated by using the average of all training vectors in this group. Repeat steps 2 and 3 until quantum error smaller than threshold value.

In recognition process, the input speech sample is extracted the feature by the MFCC algorithm first. Then the feature vectors are calculated to find the VQ distortion for each codebook. The word having smallest distortion is the word which needs to be identified.

### ***2.3 SoPC Implementation***

The proposed speech recognition system has been intently implemented on Altera FPGAs for high performance. In this system, Nios II Processor is used as the most important component of the system, a processor to execute programs of the system. All compiled C program is stored in the SDRAM. Flash memory is used to store the parameters of the codebook after training. The ADC interface is the part connected to the Audio Codec WM8731 chip. This chip is responsible for data sampling of voices speak into the microphone. LCD is used to show the implementation of the program, the recognition results will be also displayed on LCD. In particular, the Interval Timer is used to calculate the number of the pulse clocks when executing each block.

## **3 Experimental Results**

As mentioned above, we use Interval Timer to survey the program execution speed of each functional block in speech recognition system. The input speech samples are used for system input is 2,400 samples. The clock of the system is 50 MHz.

### ***3.1 Feature Extraction***

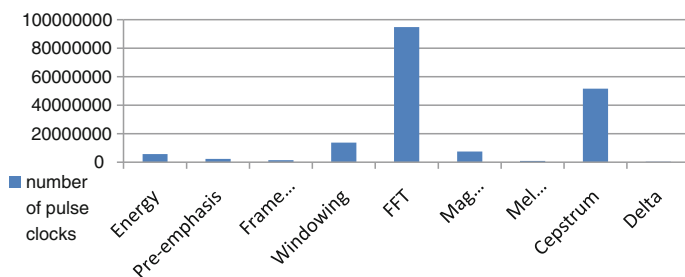
The number of clock cycles for executing each block in feature extraction are presented in Table 1.

The pre-emphasis block with  $a = 1$  is executed fastest. The value  $a = 15/16$  in the pre-emphasis block run faster than the pre-emphasis block with  $a = 0.97$ . In the magnitude computation step, the estimation algorithm calculates amplitude of a complex number much faster than the normal algorithm by taking the square root operation. By using the rectangle filters to replace the triangle filters, the program execution speed of the Mel Frequency Filter Bank block is increased 46 times.

In Fig. 1, the program execution speed of all blocks in MFCC based feature extraction is shown. The FFT block is the slowest, requires 94,874,620 clock cycles to complete the given input samples.

**Table 1** Obtained results of program execution speed by coefficient “a”

Block	Algorithm/parameter	Clock cycles ( <i>cycles</i> )
Pre_emphasis	$a = 1$	2,078,463
	$a = 15/16$	2,155,870
	$a = 0.97$	2,156,018
FFT/DFT	FFT	94,874,620
	DFT	365,586,715
Magnitude_computation	Estimation amplitude	7,463,640
	Accuracy amplitude	80,412,716
Mel-filter-bank	Rectangle filters	418,427
	Triangle filters	19,317,411



**Fig. 1** Program execution speed of the blocks in MFCC based feature extraction

### 3.2 Vector Quantization

With codebook size of 128, Vector Quantization is used in the recognition step. So, it costs 531,067,721 clock cycles.

### 3.3 Recognition Accuracy

The whole recognition system with proposed architectures, parameters as stated above has the recognition accuracy of 88 %, in which 7,416 utterances recorded from male and female adults in three regions of the North, Middle, and South of Vietnam are used.

## 4 Conclusion

In this paper, we propose efficient architectures and design choices for each part in MFCC-HMM-based speech recognition system to improve the processing speed.

The determination of design choices are based on the easiness in implementation and experimental results of whole system. The whole system is built on FPGA, verified by testing with 7,416 utterances which are recorded from male and female adults in three regions of North, Middle and South of Vietnam with a recognition accuracy of 88 %.

## References

1. Lawrence R, Bing-Hwang J (1993) Fundamentals of speech recognition. Prentice Hall PTR, Upper Saddle River
2. Thomas H, Gunnar E, Dan K, Gareth M, Julian O, Dave O, Dan P, Valtcho V, Phil W, Steve Y (1995–2002) The hidden markov model toolkit (HTK) book (for HTK version 3.2.1). Cambridge University. <http://htk.eng.cam.ac.uk/>
3. Amudha V, Venkataramani B, Vinoth kumar R, Ravishankar S (2009) Software/Hardware co-design of HMM based isolated digit recognition system. *J Compu* 4:(2)154–159
4. Zhou H, Han X (2009) Design and implementation of speech recognition system based on field programmable gate array. *Mod Appl Sci* 3(8):106–111
5. Han W, Chan C-F, Choy C-S, Pun K-P (2006) An efficient MFCC extraction method in speech recognition. In: Proceedings of the 2006 IEEE international symposium on circuits and systems (ISCAS). Greece, pp 145–148
6. Han W (2006) A speech recognition IC with an efficient MFCC extraction algorithm and multi-mixture models. Doctor of philosophy thesis, the Chinese University of Hong Kong
7. Pan S-T, Lai C-C, Tsai B-Y (2011) The implementation of speech recognition systems on FPGA—based embedded systems with SOC architecture. *Int J Innovative Comput Inf Control* 7(10):5939–5946
8. Cheng O, Abdulla W, Salci Z (2011) Hardware-Software co-design of automatic speech recognition system for embedded real-time applications. In: Proceedings of the IEEE transactions on industrial electronics, pp 850–859
9. Ge Z, Yin J, Liu, Q, Yang C (2011) A real-time speech recognition system based on the implementation of FPGA. In: Cross strait quad-regional radio science and wireless technology conference (CSQRWC), pp 1375–1378

# Erratum to: Design of a Reliable In-Vehicle Network Using ZigBee Communication

Sunny Ro, Kyung-Jung Lee and Hyun-Sik Ahn

**Erratum to:**  
**Chapter “Design of a Reliable In-Vehicle Network Using ZigBee Communication” in: J. J. (Jong Hyuk) Park et al. (eds.), *Multimedia and Ubiquitous Engineering*, DOI [10.1007/978-94-007-6738-6\\_96](https://doi.org/10.1007/978-94-007-6738-6_96)**

Below given Acknowledgement should be added.

## **Acknowledgment**

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2012-H0301-12-2007) supervised by the NIPA (National IT Industry Promotion Agency) and also supported by the MKE (The Ministry of Knowledge Economy), Korea, under the CITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0401-13-1008) supervised by the NIPA (National IT Industry Promotion Agency).

---

The online version of the original chapter can be found under  
DOI [10.1007/978-94-007-6738-6\\_96](https://doi.org/10.1007/978-94-007-6738-6_96)

---

S. Ro · K.-J. Lee · H.-S. Ahn (✉)  
Department of Electronics Engineering, Kookmin University, Jeongneung-dong,  
Seongbuk-gu, Seoul, Korea  
e-mail: ahs@kookmin.ac.kr

S. Ro  
e-mail: sunyda88@nate.com

K.-J. Lee  
e-mail: streizin@nate.com



# Index

## A

Ahmed, Mohamed A., 841  
Ahn, Hyun-Sik, 769, 777  
Ahn, Sang-Ho, 367  
Aikebaier, Ailixier, 477  
Akrouit, Belhassen, 43  
An, Ming-Shou, 1181  
Arpasilp, Dissakan, 879

## B

Bae, Changseok, 763  
Bae, Nam-Jin, 83, 359, 699  
Baek, Mi Ran, 359  
Baek, Miran, 83  
Baek, Seung Jun, 1047  
Bandara, K., 849  
Bao, Vo Nguyen Quoc, 919  
Bednar, Slavomir, 747  
Bien, Franklin, 1065  
Bok, Junyeong, 801  
Bovet, Jérôme, 93  
Byun, Sangmun, 35

## C

Chang, Hsuan-pu, 665  
Chang, Jae-woo, 277, 345  
Chang, Jaewoo, 337  
Chang, Wen-Chih, 649  
Chao, Yi-Hsiang, 143  
Chen, Nan, 887  
Chen, Yung Hui, 633  
Cheng, H. D., 285  
Chien, Dao-Ngoc, 1169  
Cho, Do-Eun, 57  
Cho, Kyunryong, 691  
Cho, Sang Bock, 943

Cho, Seongsoo, 575, 621  
Cho, Yong-Yun, 83, 359, 691, 699  
Choeysuwan, Patcharaporn, 935  
Choi, Dojin, 51  
Choi, Hyo Sub, 467, 1137  
Choi, Kiyong, 1161  
Choi, KwangHee, 511  
Choi, Miae, 35  
Chong, Kil To, 1031, 1047  
Choomchuay, Somsak, 935  
Chul, Son Kwang, 575, 621  
Chun, Se-Hak, 353  
Chun, Soon-Yong, 1005, 1073, 1081  
Chung, Gi-Soo, 545  
Chung, Tai-Myoung, 485  
Chung, Yeon-ho, 849  
Chung, Young-Suk, 127  
Claesen, Luc, 177

## D

Dang, Hong Quan, 983  
Deepunya, Chairak, 871  
Dinh, Khanh Quoc, 997  
Do, Binh V., 817  
Do, Quoc Trinh, 919  
Doanh, Nguyen Thi Hong, 895  
Dou, Qiang, 449  
Duc Dung, Do, 911  
Duc, Nguyen Tuan, 895  
Duc-Tan, Tran, 911  
Duc-Tuyen, Ta, 911

## E

Enokido, Tomoya, 477  
Eom, Jung-Ho, 485

**F**

Feng, Junliang, 193

**G**

Ganduulga, 1031  
 Gaspar, Stefan, 713  
 Gia, Tuan Anh Nguyen, 1013  
 Giang, Nguyen Linh, 1121, 1129  
 Gil, Joon-Min, 511, 521, 529  
 Gu, Bongen, 51  
 Gu, Junzhong, 193

**H**

Hai, Nguyen Dai, 1121, 1129  
 Han, Jong Wook, 293, 377  
 Han, Kyeong-Soo, 1153  
 Han, Kyuseung, 1161  
 Han, Sewon, 827  
 He, Yi-Jun, 731  
 Hennebert, Jean, 93  
 Hiep, Pham Thanh, 857  
 Hoang, Ngoc-Bach, 1099  
 Hoang, Nguyen Huy, 857  
 Hoang, Tran Van, 1197  
 Hong, Bonghwa, 569, 575  
 Hong, Dong Pyo, 1031  
 Hong, Seungtae, 337  
 Hsu, Victoria, 673  
 Huang, Fengyun, 683  
 Hung, Pham-Viet, 1169  
 HuuTo, Pham, 863  
 Hyun, DongLim, 415, 423, 459

**J**

Jang, Min-Ki, 321  
 Jang, Miyoung, 277  
 Jang, Mi-Young, 345  
 Jeon, Byeungwoo, 949, 997  
 Jeon, Yong-Hee, 505  
 Jeong, Seungdo, 1153  
 Jheng, Ming-Ren, 649  
 Ji, Un-Ho, 1081  
 Jo, Hyun-ho, 1113  
 Jo, Manhwee, 1161  
 Joo, Haejong, 569  
 Jozef, Zajac, 707  
 Jun, Kyungkoo, 77  
 Jung, SeokWon, 1039

**K**

Kang, Byoung-Doo, 367  
 Kang, Dae-Seong, 1181  
 Kang, Dong-oh, 763  
 Kang, Hee-Jun, 1099  
 Kang, M. G., 1147  
 Kao, Bruce C., 633, 641  
 Keikhosrokiani, Pantea, 785  
 Khac, Duy Huynh, 1013  
 Khang, Nguyen-Van, 1169  
 Kim, ByeongSu, 313, 387  
 Kim, Byung-Seo, 27, 287  
 Kim, Chang-Geol, 561  
 Kim, Chung Hyeok, 621  
 Kim, Do-Hyeun, 887  
 Kim, Dong W., 593, 597  
 Kim, Dong-Hyun, 157, 627  
 Kim, Dongkyun, 529  
 Kim, EunGil, 305, 415, 423, 459  
 Kim, Eun-Young, 353  
 Kim, Hag-Tae, 1031  
 Kim, Heung-Shik, 367  
 Kim, Hong Gean, 359  
 Kim, HongGeun, 83, 691  
 Kim, Hoon, 103  
 Kim, Hyeong-Il, 345  
 Kim, Hyoung-il, 277  
 Kim, HyunGon, 1039  
 Kim, Hyunsung, 203, 253,  
     269, 927  
 Kim, I. K., 1147  
 Kim, Intaek, 983  
 Kim, Jae-O, 769  
 Kim, Jeong Ah, 431  
 Kim, Jin-Mook, 585, 627  
 Kim, Jong-Dae, 163, 321  
 Kim, Jong-Ho, 367  
 Kim, JongHoon, 305, 313, 387, 415,  
     423, 459, 493  
 Kim, JongJin, 415  
 Kim, Jong-Seok, 537, 553  
 Kim, Jun Kyo, 431  
 Kim, Kisuk, 691  
 Kim, Mi-Hye, 537, 553, 585  
 Kim, MinSoo, 1039  
 Kim, Sang-Chul, 1091  
 Kim, Sang-Kyoon, 367  
 Kim, Sangsoo, 569  
 Kim, Seungcheon, 903  
 Kim, Seung-Hae, 529  
 Kim, Si Jung, 57

Kim, Siwan, 227  
 Kim, Sungho, 975  
 Kim, Sung-Hwan, 485  
 Kim, SungWan, 305  
 Kim, Tae-Eun, 957, 967  
 Kim, Taeho, 111, 1057  
 Kim, TaeHun, 313, 387  
 Kim, Tae Hyung, 359  
 Kim, Taehyung, 83, 699  
 Kim, Taejin, 227  
 Kim, Won-Tae, 1107  
 Kim, Yongguk, 835  
 Kim, Yong-Kyun, 377  
 Kim, Young-Chon, 841  
 Kim, Yu-Seop, 163, 321  
 Klein, Udo, 1189  
 Kohno, Ryuji, 857  
 Kong, Fei, 211  
 Kong, Ki-Sik, 511  
 Koo, Yong Seo, 989  
 Koonchiang, Krisada, 879  
 Kresman, Ray, 219  
 Kun, She, 169  
 Kwak, Jae Chang, 989  
 Kwak, Yoonsik, 35, 51  
 Kweon, In So, 975

## L

Lee, Chan-Su, 1005  
 Lee, DaeWon, 521  
 Lee, Deok Gyu, 293, 377  
 Lee, Duck Jin, 1047  
 Lee, Gangin, 185  
 Lee, Hankyu, 1153  
 Lee, Ho-Dong, 597  
 Lee, Hyeong-Ok, 537, 553  
 Lee, Hyunjo, 337  
 Lee, J. S., 1147  
 Lee, Jae Kyu, 1137  
 Lee, Jeongsam, 35  
 Lee, Jeongyong, 35  
 Lee, JiHwon, 493  
 Lee, Jong Kwan, 219  
 Lee, Jooyi, 1107  
 Lee, Kilhung, 739  
 Lee, K. T., 1147  
 Lee, Kwang Yeob, 989  
 Lee, Kwanyong, 157  
 Lee, Kyung-Jung, 777  
 Lee, Myeong Bae, 83, 359  
 Lee, Sang-Heon, 1005  
 Lee, Sang Yub, 467, 1137  
 Lee, Seon-oh, 1113

Lee, Tae-Gyu, 545  
 Lee, WonBong, 611  
 Lee, Yunho, 621  
 Lem, Jeongbin, 35  
 Li, Gen, 449  
 Li, Man, 21  
 Lim, Hye-Youn, 1181  
 Lim, JungEun, 1073  
 Lim, KyooSeob, 601, 611  
 Linh, Mai, 809  
 Lisowska, Agnieszka, 3  
 Liu, Jing, 13  
 Liu, Rui, 285  
 Lo, Shou-Chih, 329  
 Lu, Pingjing, 449

## M

Ma, Gunil, 227  
 Maeng, Ji Chan, 1057, 1107  
 Mahdi, Walid, 43  
 Marton, David, 747  
 Masada, Tomonari, 129  
 Matus, Cuma, 707  
 Michal, Hatala, 707  
 Modrak, Vladimir, 747  
 Moldovyan, Nikolay A., 817  
 Moon, Chan-Woo, 769  
 Moon, Jinyoung, 763  
 Mtonga, Kambombo, 203  
 Mustaffa, Norlia, 785

## N

Nakatsuka, Ryo, 243  
 Nam, Jung-hak, 1113  
 Nam, Kyoung-Min, 163  
 Ngoc Nguyen, T. T., 1065  
 Nguyen, Bao Ngoc, 809  
 Nguyen, B. D., 863  
 Nguyen, Binh Duong, 809  
 Nguyen, Dinh Uyen, 809  
 Nguyen, Minh H., 817  
 Nguyen, Pham Minh Luan, 943  
 Nguyen, Viet Anh, 949  
 Ni, Rongrong, 285  
 Niroopan, P., 849  
 Noitubtim, Musleemin, 871

## O

Oh, JungCheol, 493  
 Okuda, Kenji, 243  
 Otsubo, Nobuto, 235

**P**

Pan, Yun, 177  
 Park, Chang-Woo, 83  
 Park, Chan-Young, 163, 321  
 Park, DongGook, 699  
 Park, Duck Keun, 1137  
 Park, Hyeyoung, 157  
 Park, James J., 521  
 Park, Jang Woo, 359, 691, 699  
 Park, Ji-Hoon, 27  
 Park, Jong-Hyun, 575  
 Park, Jong-Wook, 593, 597  
 Park, Koo-Rock, 627  
 Park, Min-Woo, 485  
 Park, Sang Hyun, 467, 1137  
 Park, Sang-June, 1181  
 Park, Seong Ryoung, 359  
 Park, Sung-Wook, 593, 597  
 Park, Tae Ryong, 989  
 Pasko, Jan, 713  
 Peyls, Alexander, 177  
 Pham, Kien T., 863  
 Pham, Phuong Minh, 997  
 Phan, Cuong H., 1021  
 Promwong, Sathaporn, 871, 879  
 Pyun, Gwangbum, 121

**Q**

Qiao, Fei, 13

**R**

Raybourn, Tracey, 219  
 Ren, Yizhi, 211  
 Ro, Sunny, 777  
 Ryang, Heungmo, 113, 135  
 Ryu, Eun-kyung, 1113  
 Ryu, Heung-Gyoon, 801, 835  
 Ryu, Minsoo, 1057, 1107

**S**

Sakamoto, Syunya, 243  
 Sarwar, Muhammad Imran, 785  
 Seo, BoHyeok, 1073  
 Seo, JaeHyun, 1039  
 Shin, Chang-Sun, 83, 359, 691, 699  
 Shin, Jin, 723  
 Shin, Jin-Sup, 1181  
 Sim, Dong-gyu, 1113  
 Skurzok, Dawid, 151  
 Soh, YoungSung, 187  
 Song, Byeonghun, 1107

Song, Byung-Seop, 561  
 Song, Hye-Jeong, 321  
 Song, Hye-Jung, 163  
 Song, Seokil, 35  
 Su, Tran Van, 809  
 Sun, Caixia, 449  
 Sung, Mankyu, 405  
 Syed, Ikram, 103

**T**

Takasu, Atsuhiko, 129  
 Takizawa, Makoto, 477  
 Tang, Dianhua, 269  
 Taniguchi, Hideo, 235  
 Tran, Phuoc Vinh, 1013  
 Tran, Van-Su, 863  
 Tran, Xuan Nam, 919  
 Tran, Xuan-Tu, 1197  
 Trang, Hoang, 1197  
 Truong, Nguyen Ly Thien, 1197  
 Truong, Tuyen T. T., 1021

**U**

Uemura, Shinichiro, 235

**V**

Vallent, Thokozani Felix, 252  
 Van, Tram, 863  
 Võ, Trình Quốc, 1189  
 Vu, Duc Hiep, 919

**W**

Wang, Gicheol, 529  
 Wang, Te-Hua, 657  
 Wang, Yimu, 177  
 Wang, Yixing, 261  
 Wei, Qi, 13  
 Weng, Martin M., 641  
 Won, So-Min, 585  
 Woo, Y. J., 1147  
 Wu, Shih-Wei, 649

**X**

Xiaobo, Song, 169  
 Xiong, Hao, 731  
 Xu, Jian, 211  
 Xu, Jinli, 683  
 Xu, Ming, 211, 261  
 Xu, Wei, 683

**Y**

Yamakami, Toshihiko, [69](#), [395](#), [439](#)  
Yamauchi, Toshihiro, [235](#), [243](#)  
Yan, Xiaolang, [177](#)  
Yang, Eun-Suk, [163](#)  
Yang, Haomiao, [203](#), [269](#)  
Yang, Hsuan-Che, [649](#)  
Yang, Huazhong, [13](#)  
Yang, Won-Hyuk, [841](#)  
Yang, YoungHoon, [415](#)  
Yeo, Sang-Soo, [57](#)  
Yi, Chuho, [1153](#)  
Yi, Hyunyi, [227](#)  
Yi, Jeong Hyun, [227](#)  
Yi, Soo-Yeong, [723](#), [755](#)  
Yin, Haibin, [683](#)  
Yiu, Siu Ming, [731](#)  
Yoon, Eun-Jun, [203](#)  
Yoon, Hyunmin, [1057](#), [1107](#)  
Yoon, Min, [277](#), [345](#)

Yu, Ti-Hsin, [329](#)

Yu, Xiaoqi, [731](#)

Yun, Unil, [113](#), [121](#), [135](#), [185](#)

**Z**

Zakaria, Nasriah, [785](#)

Zhang, Ganglei, [21](#)

Zhang, Haiping, [211](#)

Zhang, Ying, [449](#)

Zhang, Zheng, [1047](#)

Zhang, Zhitong, [285](#)

Zhao, Yao, [285](#)

Zheng, Ning, [211](#), [261](#)

Zhiying, Tan, [169](#)

Zhou, Hongwei, [449](#)

Zhou, Zili, [193](#)

Ziólko, Bartosz, [151](#)