

Ajith Abraham
Jaime Lloret Mauri
John F. Buford
Junichi Suzuki
Sabu M. Thampi (Eds.)

Communications in Computer and Information Science

191

Advances in Computing and Communications

First International Conference, ACC 2011
Kochi, India, July 2011
Proceedings, Part II

Part 2

 Springer

Ajith Abraham Jaime Lloret Mauri
John F. Buford Junichi Suzuki
Sabu M. Thampi (Eds.)

Advances in Computing and Communications

First International Conference, ACC 2011
Kochi, India, July 22-24, 2011
Proceedings, Part II

Volume Editors

Ajith Abraham
Machine Intelligence Research Labs (MIR Labs)
Auburn, WA, USA
E-mail: ajith.abraham@ieee.org

Jaime Lloret Mauri
Polytechnic University of Valencia
Valencia, Spain
E-mail: jlloret@dcom.upv.es

John F. Buford
Avaya Labs Research
Basking Ridge, NJ, USA
E-mail: john.buford@gmail.com

Junichi Suzuki
University of Massachusetts
Boston, MA, USA
E-mail: jxs@acm.org

Sabu M. Thampi
Rajagiri School of Engineering and Technology
Kochi, India
E-mail: smthampi@acm.org

ISSN 1865-0929
ISBN 978-3-642-22713-4
DOI 10.1007/978-3-642-22714-1
Springer Heidelberg Dordrecht London New York

e-ISSN 1865-0937
e-ISBN 978-3-642-22714-1

Library of Congress Control Number: Applied for

CR Subject Classification (1998): I.2, C.2, H.3, H.4, K.6.5, J.1

© Springer-Verlag Berlin Heidelberg 2011

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

The use of general descriptive names, registered names, trademarks, 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.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The First International Conference on Advances in Computing and Communications (ACC 2011) was held in Kochi during July 22–24, 2011. ACC 2011 was organized by Rajagiri School of Engineering & Technology (RSET) in association with the Association of Computing Machinery (ACM)- SIGWEB, Machine Intelligence Research Labs (MIR Labs), International Society for Computers and Their Applications, Inc. (ISCA), All India Council for Technical Education (AICTE), Indira Gandhi National Open University (IGNOU), Kerala State Council for Science, Technology and Environment (KSCSTE), Computer Society of India (CSI)- Div IV and Cochin Chapter, The Institution of Electronics and Telecommunication Engineers (IETE), The Institution of Engineers (India) and Project Management Institute (PMI), Trivandrum, Kerala Chapter. Established in 2001, RSET is a premier professional institution striving for holistic excellence in education to mould young, vibrant engineers.

ACC 2011 was a three-day conference which provided an opportunity to bring together students, researchers and practitioners from both academia and industry. ACC 2011 was focused on advances in computing and communications and it attracted many local and international delegates, presenting a balanced mixture of intellects from the East and from the West. ACC 2011 received 592 research papers from 38 countries including Albania, Algeria, Bangladesh, Brazil, Canada, Colombia, Cyprus, Czech Republic, Denmark, Ecuador, Egypt, France, Germany, India, Indonesia, Iran, Ireland, Italy, Korea, Kuwait, Malaysia, Morocco, New Zealand, P.R. China, Pakistan, Rwanda, Saudi Arabia, Singapore, South Africa, Spain, Sri Lanka, Sweden, Taiwan, The Netherlands, Tunisia, UK, and USA. This clearly reflects the truly international stature of ACC 2011. All papers were rigorously reviewed internationally by an expert technical review committee comprising more than 300 members. The conference had a peer-reviewed program of technical sessions, workshops, tutorials, and demonstration sessions.

There were several people that deserve appreciation and gratitude for helping in the realization of this conference. We would like to thank the Program Committee members and additional reviewers for their hard work in reviewing papers carefully and rigorously. After careful discussions, the Program Committee selected 234 papers (acceptance rate: 39.53%) for presentation at the conference. We would also like to thank the authors for having revised their papers to address the comments and suggestions by the referees.

The conference program was enriched by the outstanding invited talks by Ajith Abraham, Subir Saha, Narayan C. Debnath, Abhijit Mitra, K. Chandra Sekaran, K. Subramanian, Sudip Misra, K.R. Srivathsan, Jaydip Sen, Joyati Debnath and Junichi Suzuki. We believe that ACC 2011 delivered a high-quality, stimulating and enlightening technical program. The tutorials covered topics of

great interest to the cyber forensics and cloud computing communities. The tutorial by Avinash Srinivasan provided an overview of the forensically important artifacts left behind on a MAC computer. In his tutorial on “Network Forensics,” Bhadran provided an introduction to network forensics, packet capture and analysis techniques, and a discussion on various RNA tools. The tutorial on Next-Generation Cloud Computing by Pethuru Raj focused on enabling technologies in cloud computing.

The ACC 2011 conference program also included five workshops: International Workshop on Multimedia Streaming (MultiStreams 2011), Second International Workshop on Trust Management in P2P Systems (IWTMP2PS 2011), International Workshop on Cloud Computing: Architecture, Algorithms and Applications (CloudComp 2011), International Workshop on Identity: Security, Management and Applications (ID2011) and International Workshop on Applications of Signal Processing (I-WASP 2011). We thank all the workshop organizers as well as the Workshop Chair, El-Sayed El-Alfy, for their accomplishment to bring out prosperous workshops. We would like to express our gratitude to the Tutorial Chairs Patrick Seeling, Jaydeep Sen, K.S. Mathew, and Roksana Boreli and Demo Chairs Amitava Mukherjee, Bhadran V.K., and Janardhanan P.S. for their timely expertise in reviewing the proposals. Moreover, we thank Publication Chairs Pruet Boonma, Sajid Hussain and Hiroshi Wada for their kind help in editing the proceedings. The large participation in ACC2011 would not have been possible without the Publicity Co-chairs Victor Govindaswamy, Arun Saha and Biju Paul.

The proceedings of ACC 2011 are organized into four volumes. We hope that you will find these proceedings to be a valuable resource in your professional, research, and educational activities whether you are a student, academic, researcher, or a practicing professional.

July 2011

Ajith Abraham
Jaime Lloret Mauri
John F. Buford
Junichi Suzuki
Sabu M. Thampi

Organization

ACC 2011 was jointly organized by the Department of Computer Science and Engineering and Department of Information Technology, Rajagiri School of Engineering and Technology (RSET), Kochi, India, in cooperation with ACM/SIGWEB.

Organizing Committee

Chief Patrons

Fr. Jose Alex CMI	Manager, RSET
Fr. Antony Kariyil CMI	Director, RSET

Patron

J. Isaac, Principal	RSET
---------------------	------

Advisory Committee

A. Krishna Menon	RSET
A.C. Mathai	RSET
Fr. Varghese Panthaloorkaran	RSET
Karthikeyan Chittayil	RSET
Vinod Kumar, P.B.	RSET
Biju Abraham	
Narayamparambil	RSET
Kuttyamma A.J.	RSET
Asha Panicker	RSET
K. Rajendra Varmah	RSET
P.R. Madhava Panicker	RSET
Liza Annie Joseph	RSET
Varkey Philip	RSET
Fr. Joel George Pullolil	RSET
R. Ajayakumar Varma	KSCSTE
K. Poullose Jacob	Cochin University of Science & Technology
H.R. Mohan, Chairman	Div IV, Computer Society of India (CSI)
Soman S.P., Chairman	Computer Society of India (CSI), Cochin Chapter
S. Radhakrishnan, Chairman	Kerala State Centre, The Institution of Engineers (India)

Steering Committee

John F. Buford	Avaya Labs Research, USA
Rajkumar Buyya	University of Melbourne, Australia
Mukesh Singhai	University of Kentucky, USA
John Strassner	Pohang University of Science and Technology, Republic of Korea
Junichi Suzuki	University of Massachusetts, Boston, USA
Ramakrishna Kappagantu	IEEE India Council
Achuthsankar S. Nair	Centre for Bioinformatics, Trivandrum, India

Conference Chair

Sabu M. Thampi	Rajagiri School of Engineering and Technology, India
----------------	---

ACC 2011 Program Committee Chairs

General Co-chairs

Ajith Abraham	Machine Intelligence Research Labs, Europe
Chandra Sekaran K.	National Institute of Technology Karnataka, India
Waleed W. Smari	University of Dayton, Ohio, USA

Program Co-chairs

Jaime Lloret Mauri	Polytechnic University of Valencia, Spain
Thorsten Strufe	Darmstadt University of Technology, Germany
Gregorio Martinez	University of Murcia, Spain

Special Sessions and Workshops Co-chairs

El-Sayed El-Alfy	King Fahd University of Petroleum and Minerals, Saudi Arabia
Silvio Bortoleto	Positivo University, Brazil

Tutorial Co-chairs

Patrick Seeling	University of Wisconsin - Stevens Point, USA
Jaydeep Sen	Tata Consultancy Services, Calcutta, India
K.S. Mathew	Rajagiri School of Engineering and Technology, India
Roksana Boreli	National ICT Australia Ltd., Australia

Demo Co-chairs

Amitava Mukherjee
Bhadran V.K.

IBM Global Business Services, India
Centre for Development of Advanced
Computing, Trivandrum, India

Janardhanan P.S.

Rajagiri School of Engineering and Technology,
India

Publicity Co-chairs

Victor Govindaswamy
Arun Saha
Biju Paul

Texas A&M University, USA
Fujitsu Network Communications, USA
Rajagiri School of Engineering and Technology,
India

Publication Co-chairs

Pruet Boonma
Sajid Hussain
Hiroshi Wada

Chiang Mai University, Thailand
Fisk University, USA
University of New South Wales, Australia

ACC 2011 Technical Program Committee

A. Hafid

Network Research Lab, University of Montreal,
Canada

Abdallah Shami

The University of Western Ontario, Canada

Abdelhafid Abouaissa

University of Haute Alsace, France

Abdelmalik Bachir

Imperial College London, UK

Abdelouahid Derhab

CERIST, Algeria

Abhijit Mitra

Indian Institute of Technology Guwahati, India

Adão Silva

University of Aveiro, Portugal

Adel Ali

University Technology Malaysia

Ahmed Mehaoua

University of Paris Descartes, France

Ai-Chun Pang

National Taiwan University, Taiwan

Ajay Gupta

Western Michigan University, USA

Alberto Dainotti

University of Naples "Federico II", Italy

Alessandro Leonardi

University of Catania, Italy

Alex Galis

University College London, UK

Alexey Vinel

Saint Petersburg Institute, Russia

Ali Abedi

University of Maine, USA

Alicia Triviño Cabrera

Universidad de Málaga, Spain

Alireza Behbahani

University of California, Irvine, USA

Alois Ferscha

University of Linz, Austria

Al-Sakib Khan Pathan

International Islamic University, Malaysia

Amar Prakash Azad

INRIA, France

Amirhossein Alimohammad

University of Alberta, Canada

Amit Agarwal

Indian Institute of Technology, Roorkee, India

Amitava Mukherjee	IBM Global Business Services, India
Anand Prasad	NEC Corporation, Japan
Andreas Maeder	NEC Laboratories Europe, Germany
Ankur Gupta	Model Institute of Engineering and Technology, India
Antonio Coronato	ICAR-CNR, Naples, Italy
Antonio Pescapé	University of Naples Federico II, Italy
António Rodrigues	IT / Instituto Superior Técnico, Portugal
Anura P. Jayasumana	Colorado State University, USA
Arnab Bhattacharya	Indian Institute of Technology, Kanpur, India
Arun Saha	Fujitsu Network Communications, USA
Arvind Swaminathan	Qualcomm, USA
Ashley Thomas	Secureworks Inc., USA
Ashraf Elnagar	Sharjah University, UAE
Ashraf Mahmoud	KFUPM, Saudi Arabia
Ashwani Singh	Navtel Systems, France
Athanasios Vasilakos	University of Western Macedonia, Greece
Atilio Gameiro	Telecommunications Institute/Aveiro University, Portugal
Aydin Sezgin	Ulm University, Germany
Ayman Assra	McGill University, Canada
Aytac Azgin	Georgia Institute of Technology, USA
B. Sundar Rajan	Indian Institute of Science, India
Babu A.V.	National Institute of Technology, Calicut, India
Babu B.V.	BITS-Pilani, Rajasthan, India
Babu Raj E.	Sun College of Engineering and Technology, India
Balagangadhar G. Bathula	Columbia University, USA
Borhanuddin Mohd. Ali	Universiti Putra Malaysia
Brijendra Kumar Joshi	Military College, Indore, India
Bruno Crispo	Università di Trento, Italy
C.-F. Cheng	National Chiao Tung University, Taiwan
Chang Wu Yu	Chung Hua University, Taiwan
Charalampos Tsimenidis	Newcastle University, UK
Chih-Cheng Tseng	National Ilan University, Taiwan
Chi-Hsiang Yeh	Queen's University, Canada
Chitra Babu	SSN College of Engineering, Chennai, India
Chittaranjan Hota	BITS Hyderabad Campus, India
Chonho Lee	Nanyang Technological University, Singapore
Christian Callegari	University of Pisa, Italy
Christos Chrysoulas	Technological Educational Institute, Greece
Chuan-Ching Sue	National Cheng Kung University, Taiwan
Chung Shue Chen	TREC, INRIA, France

Chun-I. Fan	National Sun Yat-sen University, Taiwan
Chutima Prommak	Suranaree University of Technology, Thailand
Dali Wei	Jiangsu Tianze Infoindustry Company Ltd, P.R. China
Danda B. Rawat	Old Dominion University, USA
Daniele Tarchi	University of Bologna, Italy
Davide Adami	CNIT Pisa Research Unit, University of Pisa, Italy
Deepak Garg	Thapar University, India
Demin Wang	Microsoft Inc., USA
Dennis Pfisterer	University of Lübeck, Germany
Deyun Gao	Beijing Jiaotong University, P.R. China
Dharma Agrawal	University of Cincinnati, USA
Dhiman Barman	Juniper Networks, USA
Di Jin	General Motors, USA
Dimitrios Katsaros	University of Thessaly, Greece
Dimitrios Vergados	National Technical University of Athens, Greece
Dirk Pesch	Cork Institute of Technology, Ireland
Djamel Sadok	Federal University of Pernambuco, Brazil
Eduardo Cerqueira	Federal University of Para (UFPA), Brazil
Eduardo Souto	Federal University of Amazonas, Brazil
Edward Au	Huawei Technologies, P.R. China
Egemen Cetinkaya	University of Kansas, USA
Elizabeth Sherly	IIITM-Kerala, India
El-Sayed El-Alfy	King Fahd University, Saudi Arabia
Emad A. Felemban	Umm Al Qura University, Saudi Arabia
Eric Renault	TELECOM & Management SudParis, France
Errol Lloyd	University of Delaware, USA
Ertan Onur	Delft University of Technology, The Netherlands
Faouzi Bader	CTTC, Spain
Faouzi Kamoun	WTS, UAE
Fernando Velez	University of Beira Interior, Portugal
Filipe Cardoso	ESTSetubal/Polytechnic Institute of Setubal, Portugal
Florian Doetzer	ASKON ConsultingGroup, Germany
Francesco Quaglia	Sapienza Università di Roma, Italy
Francine Krief	University of Bordeaux, France
Frank Yeong-Sung Lin	National Taiwan University, Taiwan
Gianluigi Ferrari	University of Parma, Italy
Giuseppe Ruggeri	University "Mediterranea" of Reggio Calabria, Italy
Grzegorz Danilewicz	Poznan University of Technology, Poland
Guang-Hua Yang	The University of Hong Kong, Hong Kong
Guo Bin	Institut Telecom SudParis, France

Hadi Otrok	Khalifa University, UAE
Hamid Mcheick	Université du Québec à Chicoutimi, Canada
Harry Skianis	University of the Aegean, Greece
Hicham Khalife	ENSEIRB-LaBRI, France
Himal Suraweera	Singapore University of Technology and Design, Singapore
Hiroshi Wada	University of New South Wales, Australia
Hong-Hsu Yen	Shih-Hsin University, Taiwan
Hongli Xu	University of Science and Technology of China, P.R. China
Houcine Hassan	Technical University of Valencia, Spain
Hsuan-Jung Su	National Taiwan University, Taiwan
Huaiyu Dai	NC State University, USA
Huey-Ing Liu	Fu-Jen Catholic University, Taiwan
Hung-Keng Pung	National University of Singapore
Hung-Yu Wei	NTU, Taiwan
Ian Glover	University of Strathclyde, UK
Ian Wells	Swansea Metropolitan University, UK
Ibrahim Develi	Erciyes University, Turkey
Ibrahim El rube	AAST, Egypt
Ibrahim Habib	City University of New York, USA
Ibrahim Korpeoglu	Bilkent University, Turkey
Ilja Radusch	Technische Universität Berlin, Germany
Ilka Miloucheva	Media Technology Research, Germany
Imad Elhajj	American University of Beirut, Lebanon
Ivan Ganchev	University of Limerick, Ireland
Iwan Adhicandra	The University of Pisa, Italy
Jalel Ben-othman	University of Versailles, France
Jane-Hwa Huang	National Chi Nan University, Taiwan
Jaydeep Sen	Tata Consultancy Services, Calcutta, India
Jiankun Hu	RMIT University, Australia
Jie Yang	Cisco Systems, USA
Jiping Xiong	Zhejiang Normal University of China
José de Souza	Federal University of Ceará, Brazil
Jose Moreira	IBM T.J. Watson Research Center, USA
Ju Wang	Virginia State University, USA
Juan-Carlos Cano	Technical University of Valencia, Spain
Judith Kelner	Federal University of Pernambuco, Brazil
Julien Laganier	Juniper Networks Inc., USA
Jussi Haapola	University of Oulu, Finland
K. Komathy	Easwari Engineering College, Chennai, India
Ka Lok Hung	The Hong Kong University, Hong Kong
Ka Lok Man	Xi'an Jiaotong-Liverpool University, China
Kaddar Lamia	University of Versailles Saint Quentin, France
Kainam Thomas	Hong Kong Polytechnic University

Kais Mnif	High Institute of Electronics and Communications of Sfax, Tunisia
Kang Yong Lee	ETRI, Korea
Katia Bortoleto	Positivo University, Brazil
Kejie Lu	University of Puerto Rico at Mayaguez, USA
Kemal Tepe	University of Windsor, Canada
Khalifa Hettak	Communications Research Centre (CRC), Canada
Khushboo Shah	Altusystems Corp, USA
Kotecha K.	Institute of Technology, Nirma University, India
Kpatcha Bayarou	Fraunhofer Institute, Germany
Kumar Padmanabh	General Motors, India
Kyriakos Manousakis	Telcordia Technologies, USA
Kyung Sup Kwak	Inha University, Korea
Li Zhao	Microsoft Corporation, USA
Li-Chun Wang	National Chiao Tung University, Taiwan
Lin Du	Technicolor Research and Innovation Beijing, P.R. China
Liza A. Latiff	University Technology Malaysia
Luca Scalia	University of Palermo, Italy
M Ayoub Khan	C-DAC, Noida, India
Maaruf Ali	Oxford Brookes University, UK
Madhu Kumar S.D.	National Institute of Technology, Calicut, India
Madhu Nair	University of Kerala, India
Madhumita Chatterjee	Indian Institute of Technology Bombay, India
Mahamod Ismail	Universiti Kebangsaan Malaysia
Mahmoud Al-Qutayri	Khalifa University, UAE
Manimaran Govindarasu	Iowa State University, USA
Marcelo Segatto	Federal University of Esp�rito Santo, France
Maria Ganzha	University of Gdansk, Poland
Marilia Curado	University of Coimbra, Portugal
Mario Fanelli	DEIS, University of Bologna, Italy
Mariofanna Milanova	University of Arkansas at Little Rock, USA
Mariusz Glabowski	Poznan University of Technology, Poland
Mariusz Zal	Poznan University of Technology, Poland
Masato Saito	University of the Ryukyus, Japan
Massimiliano Comisso	University of Trieste, Italy
Massimiliano Laddomada	Texas A&M University-Texarkana, USA
Matthias R. Brust	University of Central Florida, USA
Mehrzad Biguesh	Queen's University, Canada
Michael Alexander	Scaledinfra Technologies GmbH, Austria
Michael Hempel	University of Nebraska - Lincoln, USA
Michael Lauer	Vanille-Media, Germany
Ming Xia	NICT, Japan
Ming Xiao	Royal Institute of Technology, Sweden
Mohamed Ali Kaafar	INRIA, France

Mohamed Cheriet	Ecole de Technologie Superieure, Canada
Mohamed Eltoweissy	Pacific Northwest National Laboratory, USA
Mohamed Hamdi	Carthage University, Tunisia
Mohamed Moustafa	Akhbar El Yom Academy, Egypt
Mohammad Banat	Jordan University of Science and Technology, Jordan
Mohammad Hayajneh	UAEU, UAE
Mohammed Misbahuddin	C-DAC, India
Mustafa Badaroglu	IMEC, Belgium
Naceur Malouch	Université Pierre et Marie Curie, France
Nakjung Choi, Alcatel-Lucent	Bell-Labs, Seoul, Korea
Namje Park	Jeju University, South Korea
Natarajan Meghanathan	Jackson State University, USA
Neeli Prasad	Center for TeleInFrastructure (CTIF), Denmark
Nen-Fu Huang	National Tsing Hua University, Taiwan
Nikola Zogovic	University of Belgrade, Serbia
Nikolaos Pantazis	Technological Educational Institution of Athens, Greece
Nilanjan Banerjee	IBM Research, India
Niloy Ganguly	Indian Institute of Technology, Kharagpur, India
Pablo Corral González	University Miguel Hernández, Spain
Patrick Seeling	University of Wisconsin - Stevens Point, USA
Paulo R.L. Gondim	University of Brasília, Brazil
Peter Bertok	Royal Melbourne Institute of Technology (RMIT), Australia
Phan Cong-Vinh	London South Bank University, UK
Pingyi Fan	Tsinghua University, P.R. China
Piotr Zwierzykowski	Poznan University of Technology, Poland
Pascal Lorenz	University of Haute Alsace, France
Pruet Boonma	Chiang Mai University, Thailand
Punam Bedi	University of Delhi, India
Qinghai Gao	Atheros Communications Inc., USA
Rahul Khanna	Intel, USA
Rajendra Akerkar	Western Norway Research Institute, Norway
Raul Santos	University of Colima, Mexico
Ravishankar Iyer	Intel Corp, USA
Regina Araujo	Federal University of Sao Carlos, Brazil
Renjie Huang	Washington State University, USA
Ricardo Lent	Imperial College London, UK
Rio G. L. D'Souza	St. Joseph Engineering College, Mangalore, India
Roberto Pagliari	University of California, Irvine, USA
Roberto Verdone	WiLab, University of Bologna, Italy
Roksana Boreli	National ICT Australia Ltd., Australia

Ronny Yongho Kim	Kyungil University, Korea
Ruay-Shiung Chang	National Dong Hwa University, Taiwan
Ruidong Li	NICT, Japan
S. Ali Ghorashi	Shahid Beheshti University, Iran
Sahar Ghazal	University of Versailles, France
Said Souhli	Ericsson, Sweden
Sajid Hussain	Fisk University, USA
Salah Bourennane	Ecole Centrale Marseille, France
Salman Abdul Moiz	CDAC, Bangalore, India
Sameh Elnikety	Microsoft Research, USA
Sanjay H.A.	Nitte Meenakshi Institute, Bangalore, India
Sathish Rajasekhar	RMIT University, Australia
Sergey Andreev	Tampere University of Technology, Finland
Seshan Srirangarajan	Nanyang Technological University, Singapore
Seyed (Reza) Zekavat	Michigan Technological University, USA
Sghaier Guizani	UAE University, UAE
Shancang Li	School of Engineering, Swansea University, UK
Shi Xiao	Nanyang Technological University, Singapore
Siby Abraham	University of Mumbai, India
Silvio Bortoleto	Positivo University, Brazil
Simon Pietro Romano	University of Naples Federico II, Italy
Somayajulu D. V. L. N.	National Institute of Technology Warangal, India
Song Guo	The University of British Columbia, Canada
Song Lin	University of California, Riverside, USA
Soumya Sen	University of Pennsylvania, USA
Stefano Ferretti	University of Bologna, Italy
Stefano Giordano	University of Pisa, Italy
Stefano Pesic	Cisco Systems, Italy
Stefano Tomasin	University of Padova, Italy
Stefanos Gritzalis	University of the Aegean, Greece
Steven Gordon	Thammasat University, Thailand
Suat Ozdemir	Gazi University, Turkey
Subir Saha	Nokia Siemens Networks, India
Subramanian K.	Advanced Center for Informatics and Innovative Learning, IGNOU, India
Sudarshan T.S.B.	Amrita Vishwa Vidyapeetham, Bangalore, India
Sugam Sharma	Iowa State University, USA
Surekha Mariam Varghese	M.A. College of Engineering, India
T. Aaron Gulliver	University of Victoria, Canada
Tao Jiang	Huazhong University of Science and Technology, P.R. China
Tarek Bejaoui	Mediatron Lab., Carthage University, Tunisia
Tarun Joshi	University of Cincinnati, USA
Theodore Stergiou	Intracom Telecom, UK

Thienne Johnson	University of Arizona, USA
Thomas Chen	Swansea University, UK
Tsern-Huei Lee	National Chiao Tung University, Taiwan
Usman Javaid	Vodafone Group, UK
Vamsi Paruchuri	University of Central Arkansas, USA
Vana Kalogeraki	University of California, Riverside, USA
Vehbi Cagri Gungor	Bahcesehir University, Turkey
Velmurugan Ayyadurai	University of Surrey, UK
Vicent Cholvi	Universitat Jaume I, Spain
Victor Govindaswamy	Texas A&M University, USA
Vijaya Kumar B.P.	Reva Institute of Technology and Management, Bangalore, India
Viji E Chenthamarakshan	IBM T.J. Watson Research Center in New York, USA
Vino D.S. Kingston	Hewlett-Packard, USA
Vinod Chandra S.S.	College of Engineering Thiruvananthapuram, India
Vivek Jain	Robert Bosch LLC, USA
Vivek Singh	Banaras Hindu University, India
Vladimir Kropotov	D-Link Russia, Russia
Wael M El-Medany	University of Bahrain, Kingdom of Bahrain
Waslon Lopes	UFCEG - Federal University of Campina Grande, Brazil
Wei Yu	Towson University, USA
Wei-Chieh Ke	National Tsing Hua University, Taiwan
Wendong Xiao	Institute for Infocomm Research, Singapore
Xiang-Gen Xia	University of Delaware, USA
Xiaodong Wang	Qualcomm, USA
Xiaoguang Niu	Wuhan University, P.R. China
Xiaoqi Jia	Institute of Software, Chinese Academy of Sciences, P.R. China
Xinbing Wang	Shanghai Jiaotong University, P.R. China
Xu Shao	Institute for Infocomm Research, Singapore
Xueping Wang	Fudan University, P.R. China
Yacine Atif	UAE University, UAE
Yali Liu	University of California, Davis, USA
Yang Li	Chinese Academy of Sciences, P.R. China
Yassine Bouslimani	University of Moncton, Canada
Ye Zhu	Cleveland State University, USA
Yi Zhou	Texas A&M University, USA
Yifan Yu	France Telecom R&D Beijing, P.R. China
Yong Wang	University of Nebraska-Lincoln, USA
Youngseok Lee	Chungnam National University, Korea
Youssef SAID	Tunisie Telecom/Sys'Com Lab, ENIT, Tunisia
Yuan-Cheng Lai	Information Management, NTUST, Taiwan
Yuh-Ren Tsai	National Tsing Hua University, Taiwan

Yu-Kai Huang	Quanta Research Institute, Taiwan
Yusuf Ozturk	San Diego State University, USA
Zaher Aghbari	University of Sharjah, UAE
Zbigniew Dziong	University of Quebec, Canada
Zhang Jin	Beijing Normal University, P.R. China
Zhenghao Zhang	Florida State University, USA
Zhenzhen Ye	iBasis, Inc., USA
Zhihua Cui	Taiyuan University of Science and Technology, China
Zhili Sun	University of Surrey, UK
Zhong Zhou	University of Connecticut, USA
Zia Saquib	C-DAC, Mumbai, India

ACC 2011 Additional Reviewers

Akshay Vashist	Telcordia Technologies, USA
Alessandro Testa	University of Naples Federico II, Italy
Amitava	Academy of Technology, India
Ammar Rashid	Auckland University of Technology, New Zealand
Anand	MITS, India
Bjoern W. Schuller	Technical University, Germany
Chi-Ming Wong	Jinwen University of Science and Technology, Taiwan
Danish Faizan	NIC-INDIA, India
Fatos Xhafa	UPC, Barcelona Tech, Spain
Hooman Tahayori	Ryerson University, Canada
John Jose	IIT Madras, India
Jyoti Singh	Academy of Technology, India
Koushik	West Bengal University of Technology, India
Long Zheng	University of Aizu, Japan
Manpreet Singh	M.M. Engineering College, India
Maria Striki	Telcordia Technologies, Piscataway, USA
Mohamad Zoinol Abidin	Universiti Teknikal Malaysia Melaka, Malaysia
Mohamed Dahmane	University of Montreal, Canada
Mohd Helmy Abd Wahab	Universiti Tun Hussein Onn Malaysia, Malaysia
Mohd Riduan Bin Ahmad	Universiti Teknikal Malaysia Melaka, Malaysia
Mohd Sadiq	Jamia Millia Islamia, India
Mudhakar Srivatsa	IBM T.J. Watson Research Center, USA
Nan Yang	CSIRO, Australia
Nurulnadwan Aziz Aziz	Universiti Teknologi MARA, Malaysia

Pooya Taheri

R.C. Wang

Roman Yampolskiy

Shuang Tian

Syed Abbas Ali

Velayutham

Yeong-Luh Ueng

University of Alberta, Canada

NTTU, Taiwan

University of Louisville, USA

The University of Sydney, Australia

Ajman University of Science & Technology,
UAE

Adhiparasakthi Engineering College,
Melmaruvathur, India

National Tsing Hua University, Taiwan

International Workshop on Identity: Security, Management and Applications (ID 2011)

General Chairs

Paul Rodrigues
(CTO, WSS, India) Hindustan University, India
H.R. Vishwakarma
(Secretary, Computer Society of India) VIT University, India

Program Chairs

P. Krishna Reddy
Sundar K.S. IIIT, Hyderabad, India
Education & Research, Infosys Technologies Limited, India
Srinivasa Ragavan
S. Venkatachalam Intel Inc, USA
Jawaharlal Nehru Technological University, India

Organizing Chair

Madhan Kumar Srinivasan
Education & Research, Infosys Technologies Limited, India

Organizing Co-chairs

Abhi Saran
Anireddy Niranjana Reddy
Revathy Madhan Kumar London South Bank University, UK
University of Glamorgan, UK
Education & Research, Infosys Technologies Limited, India

Technical Program Committee

Arjan Durresi
Arun Sivanandham
Avinash Srinivasan
Bezawada Bruhadeshwar
Bhaskara Reddy AV
Bipin Indurkha Indiana University Purdue University
Indianapolis, USA
Infosys Technologies Limited, India
Bloomsburg University, USA
IIIT, Hyderabad, India
Infosys Technologies Limited, India
IIIT, Hyderabad, India

C. Sunil Kumar	Jawaharlal Nehru Technological University, India
Chandrabali Karmakar	Infosys Technologies Limited, India
Farooq Anjum	On-Ramp Wireless, USA
Gudipati Kalyan Kumar	Excellence India, India
Hamid Sharif	University of Nebraska-Lincoln, USA
Hui Chen	Virginia State University, USA
Jie Li	University of Tsukuba, Japan
Kalaiselvam	Infineon Technologies, Germany
Lau Lung	UFSC, Brazil
Lukas Ruf	Consecom AG, Switzerland
Manik Lal Das	Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), India
Manimaran Govindarasu	Iowa State University, USA
Narendra Ahuja	University of Illinois, USA
Omar	University of Jordan, Jordan
Pradeep Kumar T.S.	Infosys Technologies Limited, India
Pradeepa	Wipro Technologies, India
Rajiv Tripathi	NIT, Allahabad, India
Rakesh Chithuluri	Oracle, India
Sanjay Chaudhary	Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), India
Santosh Pasuladi	Jawaharlal Nehru Technological University, India
Satheesh Kumar Varma	IIIT, Pune, India
Saurabh Barjatiya	IIIT, Hyderabad, India
Sreekumar Vobugari	Education & Research, Infosys Technologies Limited, India
Suthershan Vairavel	CTS, India
Tarun Rao	Infosys Technologies Limited, India
Thomas Little	Boston University, USA
Tim Strayer	BBN Technologies, USA
V. Balamurugan	IBM, India
Vasudeva Varma	IIIT, Hyderabad, India
Vinod Babu	Giesecke & Devrient, Germany
Yonghe Liu	UT Arlington, USA

International Workshop on Applications of Signal Processing

(I-WASP 2011)

Workshop Organizers

Jaison Jacob	Rajagiri School of Engineering and Technology, India
Sreeraj K.P.	Rajagiri School of Engineering and Technology, India
Rithu James	Rajagiri School of Engineering and Technology, India

Technical Program Committee

A. Vinod	NTU, Singapore
Aggelos Katsaggelos	Northwestern University, USA
Bing Li	University of Virginia, USA
Carlos Gonzalez	University of Castilla-La Mancha, Spain
Damon Chandler	Oklahoma State University, USA
Egon L. van den Broek	University of Twente, The Netherlands
Feng Wu	Microsoft Research Asia, P.R. China
Hakan Johansson	University of Linköping, Sweden
Joaquim Filipe	EST-Setubal, Portugal
Lotfi Senahdji	Université de Rennes 1, France
Reyer Zwiggelkaar	Aberystwyth University, UK
Xianghua Xie	Swansea University, UK
Yoshikazu Miyanaga	Hokkaido University, Japan

International Workshop on Cloud Computing: Architecture, Algorithms and Applications (CloudComp 2011)

Workshop Organizers

Binu A.	Cochin University of Science and Technology, India
Biju Paul	Rajagiri School of Engineering and Technology, India
Sabu M. Thampi	Rajagiri School of Engineering and Technology, India

Technical Program Committee

Antonio Puliafito	University of Messina, Italy
Bob Callaway	IBM, USA
Chee Shin Yeo	Institute of High-Performance Computing, Singapore
Chin-Sean Sum	National Institute of Information and Communications Technology, Japan
Ching-Hsien Hsu	Chung Hua University, Taiwan
Drissa Houatra	Orange Labs, France
Deepak Unnikrishnan	University of Massachusetts, USA
Jie Song	Northeastern University, P.R. China
Salah Sharieh	McMaster University, Canada
Francesco Longo	Università di Messina, Italy
Fabienne Anhalt	Ecole Normale Supérieure de Lyon-INRIA, France
Gaurav Somani	LNMIIT, Jaipur, India
Hailong Guan	Shanghai Jiao Tong University, P.R. China
Hongbo Jiang	Huazhong University of Science and Technology, P.R. China
Hongkai Xiong	Shanghai Jiao Tong University, P.R. China
Hui Zhang	Nec Laboratories America, USA
Itai Zilbershtein	Avaya, Israel
Jens Nimis	University of Applied Sciences, Germany
Jie Song	Software College, Northeastern University, China

Jorge Carapinha	PT Inovação S.A. Telecom Group, Portugal
Junyi Wang	National Institute of Information and Communications Technology, Japan
K. Chandra Sekaran	NITK, India
Kai Zheng	IBM China Research Lab, P.R. China
Krishna Sankar	Cisco Systems, USA
Laurent Amanton	Havre University, France
Luca Caviglione	National Research Council (CNR), Italy
Lukas Ruf	Consecom AG, Switzerland
Massimiliano Rak	Second University of Naples, Italy
Pallab Datta	IBM Almaden Research Center, USA
Pascale Vicat-Blanc Primet	INRIA, France
Prabu Dorairaj	NetApp Inc, India
Shivani Sud	Intel Labs, USA
Shuicheng Yan	National University of Singapore, Singapore
Siani Pearson	HP Labs, UK
Simon Koo	University of San Diego, USA
Srikumar Venugopal	UNSW, Australia
Stephan Kopf	University of Mannheim, Germany
Thomas Sandholm	Hewlett-Packard Laboratories, USA
Umberto Villano	University of Sannio, Italy
Vipin Chaudhary	University at Buffalo, USA
Yaozu Dong	Intel Corporation, P.R. China
Zhou Lan	National Institute of Information and Communications Technology, Japan

International Workshop on Multimedia Streaming (MultiStreams 2011)

Program Chairs

Pascal Lorenz	University of Haute Alsace, France
Fan Ye	IBM T.J. Watson Research Center, USA
Trung Q. Duong	Blekinge Institute of Technology, Sweden

Technical Program Committee

Guangjie Han	Hohai University, P.R. China
Alex Canovas	Polytechnic University of Valencia, Spain
Brent Lagesse	Oak Ridge National Laboratory, USA
Chung Shue Chen	INRIA-ENS, France
Debasis Giri	Haldia Institute of Technology, India
Mario Montagud	Universidad Politécnic de Valencia, Spain
Doreen Miriam	Anna University, India
Duduku V. Viswacheda	University Malaysia Sabah, Malaysia
Elsa Macías López	University of Las Palmas de Gran Canaria, Spain
Eugénia Bernardino	Polytechnic Institute of Leiria, Portugal
Fernando Boronat	Instituto de Investigación para la Gestión Integrada de Zonas Costeras, Spain
Jen-Wen Ding	National Kaohsiung University of Applied Sciences, Taiwan
Joel Rodrigues IT	University of Beira Interior, Portugal
Jo-Yew Tham	A*STAR Institute for Infocomm Research, Singapore
Marcelo Atenas	Universidad Politecnica de Valencia, Spain
Jorge Bernabé	University of Murcia, Poland
Bao Vo Nguyen	Posts and Telecommunications Institute of Technology, Vietnam
Hans-Juergen Zepernick	Blekinge Institute of Technology, Sweden
Jose Maria Alcaraz Calero	University of Murcia, Spain
Juan Marin Perez	University of Murcia, Spain
Lei Shu	Osaka University, Japan
Lexing Xie	The Australian National University, Australia
Marc Gilg	University of Haute-Alsace, France
Miguel Garcia	Polytechnic University of Valencia, Spain
Mohd Riduan Bin Ahmad	Universiti Teknikal Malaysia, Malaysia

Phan Cong-Vinh

Alvaro Suárez-Sarmiento

Song Guo

Tin-Yu Wu

Zhangbing Zhou

Zuqing Zhu

Juan M. Sánchez

Choong Seon Hong

London South Bank University, UK

University of Las Palmas de Gran Canaria,
Spain

University of British Columbia, Canada

Tamkang University, Taiwan

Institut Telecom & Management SudParis,
France

Cisco System, USA

University of Extremadura, Spain

Kyung Hee University, Korea

Second International Workshop on Trust Management in P2P Systems (IWTMP2PS 2011)

Program Chairs

Visvasuresh Victor

Govindaswamy

Jack Hu

Sabu M. Thampi

Texas A&M University-Texarkana, USA

Microsoft, USA

Rajagiri School of Engineering and Technology,
India

Technical Program Committee

Haiguang

Ioannis Anagnostopoulos

Farag Azzedin

Fudan University, P.R. China

University of the Aegean, Greece

King Fahd University of Petroleum & Minerals,
Saudi Arabia

Roksana Boreli

Yann Busnel

Juan-Carlos Cano

Phan Cong-Vinh

Jianguo Ding

Markus Fiedler

Deepak Garg

Felix Gomez Marmol

Paulo Gondim

Steven Gordon

Ankur Gupta

National ICT Australia, Australia

University of Nantes, France

Universidad Politecnica de Valencia, Spain

London South Bank University, UK

University of Luxembourg, Luxemburg

Blekinge Institute of Technology, Sweden

Thapar University, Patiala, India

University of Murcia, Spain

Universidade de Brasilia, Brazil

Thammasat University, Thailand

Model Institute of Engineering and Technology,
India

Houcine Hassan

Yifeng He

Michael Hempel

Salman Abdul Moiz

Guimin Huang

Universidad Politecnica de Valencia, Spain

Ryerson University, Canada

University of Nebraska-Lincoln, USA

CDAC, India

Guilin University of Electronic Technology,
P.R. China

Renjie Huang

Benoit Hudzia

Helge Janicke

Washington State University, USA

SAP Research, UK

De Montfort University, UK

Mohamed Ali Kaafar	INRIA, France
Eleni Koutrouli	National University of Athens, Greece
Stefan Kraxberger	Graz University of Technology, Austria
Jonathan Loo	Middlesex University, UK
Marjan Naderan	Amirkabir University of Technology, Iran
Lourdes Penalver	Valencia Polytechnic University, Spain
Elvira Popescu	UCV, Romania
Guangzhi Qu	Oakland University, USA
Aneel Rahim	COMSATS Institute of Information Technology, Pakistan
Yonglin Ren	SITE, University of Ottawa, Canada
Andreas Riener	University of Linz, Austria
Samir Saklikar	RSA, Security Division of EMC, India
Thomas Schmidt	HAW Hamburg (DE), Germany
Fangyang Shen	Northern New Mexico College, USA
Thorsten Strufe	TU Darmstadt, Germany
Sudarshan Tsb	Amrita School of Engineering, India
Demin Wang	Microsoft, USA
Fatos Xhafa	UPC, Barcelona, Spain
Jiping Xiong	Zhejiang Normal University, P.R. China
Chang Wu Yu	Chung Hua University, Taiwan

Table of Contents – Part II

Database and Information Systems

Balancing between Utility and Privacy for k-Anonymity	1
<i>Korra Sathya Babu and Sanjay Kumar Jena</i>	
Evaluation of Approaches for Modeling of Security in Data Warehouses	9
<i>Krishna Khajaria and Manoj Kumar</i>	
Content Based Compression for Quicx System	19
<i>Radha Senthilkumar, C. Lingeshwaraa, and A. Kannan</i>	

Distributed Software Development

NL-Based Automated Software Requirements Elicitation and Specification	30
<i>Ashfa Umer, Imran Sarwar Bajwa, and M. Asif Naeem</i>	
Automatic Interface Generation between Incompatible Intellectual Properties (IPs) from UML Models	40
<i>Fateh Boutekkouk, Zakaria Tolba, and Mustapha Okab</i>	
Deadlock Prevention in Distributed Object Oriented Systems	48
<i>V. Geetha and N. Sreenath</i>	
Identification of Error Prone Classes for Fault Prediction Using Object Oriented Metrics	58
<i>Puneet Mittal, Satwinder Singh, and K.S. Kahlon</i>	
An Automated Tool for Computing Object Oriented Metrics Using XML	69
<i>N. Kayarvizhy and S. Kanmani</i>	
Traceability Matrix for Regression Testing in Distributed Software Development	80
<i>B. Athira and Philip Samuel</i>	
Testing Agent-Oriented Software by Measuring Agent's Property Attributes	88
<i>N. Sivakumar, K. Vivekanandan, and S. Sandhya</i>	

Human Computer Interaction and Interface

Classifier Feature Extraction Techniques for Face Recognition System under Variable Illumination Conditions	99
<i>Sneha G. Gondane, M. Dhivya, and D. Shyam</i>	
Bispectrum Analysis of EEG in Estimation of Hand Movement	109
<i>Aditya Saikia and Shyamanta M. Hazarika</i>	
Wavelet Selection for EMG Based Grasp Recognition through CWT	119
<i>Aditya Saikia, Nayan M. Kakoty, and Shyamanta M. Hazarika</i>	
Information Visualization for Tourist and Travelling in Indonesia	130
<i>Adityo Ashari Wirjono, Ricky Lincoln Z.S., William, and Dewi Agushinta R.</i>	
The Smart Goal Monitoring System	138
<i>Dewi Agushinta R., Bima Shakti Ramadhan Utomo, Denny Satria, Jennifer Sabrina Karla Karamoy, and Nuniek Nur Sahaya</i>	
Web Based Virtual Agent for Tourism Guide in Indonesia	146
<i>Kezia Velda Roberta, Lulu Mawaddah Wisudawati, Muhammad Razi, and Dewi Agushinta R.</i>	
Local Feature or Mel Frequency Cepstral Coefficients - Which One is Better for MLN-Based Bangla Speech Recognition?	154
<i>Foyzul Hassan, Mohammed Rokibul Alam Kotwal, Md. Mostafizur Rahman, Mohammad Nasiruddin, Md. Abdul Latif, and Mohammad Nurul Huda</i>	
Power Optimization Techniques for Segmented Digital Displays	162
<i>Rohit Agrawal, C. Sasi Kumar, and Darshan Moodgal</i>	
Language Independent Icon-Based Interface for Accessing Internet	172
<i>Santa Maiti, Debasis Samanta, Satya Ranjan Das, and Monalisa Sarma</i>	
Contribution of Oral Periphery on Visual Speech Intelligibility	183
<i>Preety Singh, Deepika Gupta, V. Laxmi, and M.S. Gaur</i>	

ICT

Geo-Spatial Pattern Determination for SNAP Eligibility in Iowa Using GIS	191
<i>Sugam Sharma, U.S. Tim, Shashi Gadia, and Patrick Smith</i>	
Project Management Model for e-Governance in the Context of Kerala State	201
<i>Anu Paul and Varghese Paul</i>	

ICT Its Role in e-Governance and Rural Development	210
<i>Deka Ganesh Chandra and Dutta Borah Malaya</i>	

Enhancing Sustainability of Software: A Case-Study with Monitoring Software for MGNREGS in India	223
<i>C.K. Raju and Ashok Mishra</i>	

Internet and Web Computing

Proficient Discovery of Service in Event Driven Service Oriented Architecture	234
<i>P. Dharanyadevi, P. Dhavachelvan, S.K.V. Jayakumar, R. Baskaran, and V.S.K. Venkatachalapathy</i>	

Web User Session Clustering Using Modified K-Means Algorithm	243
<i>G. Poornalatha and Prakash S. Raghavendra</i>	

FOL-Mine – A More Efficient Method for Mining Web Access Pattern	253
<i>A. Rajimol and G. Raju</i>	

Semantic Association Mining on Spatial Patterns in Medical Images	263
<i>S. Saritha and G. SanthoshKumar</i>	

FCHC: A Social Semantic Focused Crawler	273
<i>Anjali Thukral, Varun Mendiratta, Abhishek Behl, Hema Banati, and Punam Bedi</i>	

A Dynamic Seller Selection Model for an Agent Mediated e-Market	284
<i>Vibha Gaur and Neeraj Kumar Sharma</i>	

A Modified Ontology Based Personalized Search Engine Using Bond Energy Algorithm	296
<i>Bhaskara Rao Boddu and Valli Kumari Vatsavayi</i>	

A Client Perceived Performance Evaluation of Web Servers	307
<i>Ash Mohammad Abbas and Ravindra Kumar</i>	

Enhanced Quality of Experience through IVR Mashup to Access Same Service Multiple Operator Services	317
<i>Imran Ahmed and Sunil Kumar Koppurapu</i>	

Information Content Based Semantic Similarity Approaches for Multiple Biomedical Ontologies	327
<i>K. Saruladha, G. Aghila, and A. Bhuvaneswary</i>	

Taking Project Tiger to the Classroom: A Virtual Lab Case Study	337
<i>Harilal Parasuram, Bipin Nair, Krishnashree Achuthan, and Shyam Diwakar</i>	

Green Communications through Network Redesign	349
<i>Sami J. Habib, Paulvanna N. Marimuthu, and Naser Zaeri</i>	
Unsupervised Modified Adaptive Floating Search Feature Selection	358
<i>D. Devakumari and K. Thangavel</i>	
Fast and Efficient Mining of Web Access Sequences Using Prefix Based Minimized Trees	366
<i>M. Thilagu and R. Nadarajan</i>	

Mobile Computing

Scalable, High Throughput LDPC Decoder for WiMAX (802.16e) Applications	374
<i>Muhammad Awais, Ashwani Singh, and Guido Masera</i>	
Unique Mechanism of Selection of Traffic Flow Templates for Mobility IP Protocols Using Multihoming and IP Flow Mobility on the NGMN	386
<i>Gustavo Jiménez and Yezid Donoso</i>	
Elliptic Curve Cryptography for Smart Phone OS	397
<i>Sharmishta Desai, R.K. Bedi, B.N. Jagdale, and V.M. Wadhai</i>	
An Improved Secure Authentication Protocol for WiMAX with Formal Verification	407
<i>Anjani Kumar Rai, Shivendu Mishra, and Pramod Narayan Tripathi</i>	
Secured Fault Tolerant Mobile Computing	417
<i>Suparna Biswas and Sarmistha Neogy</i>	
A Survey of Virtualization on Mobiles	430
<i>Suneeta Chawla, Apurv Nigam, Pankaj Doke, and Sanjay Kimbahune</i>	
Mobile Peer to Peer Spontaneous and Real-Time Social Networking	442
<i>Abhishek Varshney and Mohammed Abdul Qadeer</i>	
Analysis of a Traffic Classification Scheme for QoS Provisioning over MANETs	452
<i>Chhagan Lal, V. Laxmi, and M.S. Gaur</i>	

Multi Agent Systems

Modeling and Verification of Chess Game Using NuSMV	460
<i>Vikram Saralaya, J.K. Kishore, Sateesh Reddy, Radhika M. Pai, and Sanjay Singh</i>	

SMMAG: SNMP-Based MPLS-TE Management Using Mobile Agents	471
<i>Muhammad Tahir, Dominique Gaiti, and Majid Iqbal Khan</i>	

Multimedia and Video Systems

Face Detection and Eye Localization in Video by 3D Unconstrained Filter and Neural Network	480
<i>Pradipta K. Banerjee, Jayanta K. Chandra, and Asit K. Datta</i>	
Secret Image Sharing Using Steganography with Different Cover Images	490
<i>Noopa Jagadeesh, Aishwarya Nandakumar, P. Harmya, and S.S. Anju</i>	
A Secure Data Hiding Scheme Based on Combined Steganography and Visual Cryptography Methods	498
<i>Aishwarya Nandakumar, P. Harmya, Noopa Jagadeesh, and S.S. Anju</i>	
Cognitive Environment for Pervasive Learners	506
<i>Sattvik Sharma, R. Sreevathsan, M.V.V.N.S. Srikanth, C. Harshith, and T. Gireesh Kumar</i>	
A Robust Background Subtraction Approach Based on Daubechies Complex Wavelet Transform	516
<i>Anand Singh Jalal and Vrijendra Singh</i>	
File System Level Circularity Requirement	525
<i>Mukhtar Azeem, Majid Iqbal Khan, and Arfan Nazir</i>	
An Adaptive Steganographic Method for Color Images Based on LSB Substitution and Pixel Value Differencing	535
<i>Azzat A. Al-Sadi and El-Sayed M. El-Alfy</i>	

Parallel and Distributed Algorithms

Communication Aware Co-scheduling for Parallel Job Scheduling in Cluster Computing	545
<i>A. Neela Madheswari and R.S.D. Wahida Banu</i>	
Shared Resource Allocation Using Token Based Control Strategy in Augmented Ring Networks	555
<i>Rajendra Prasath</i>	
An Algorithmic Approach to Minimize the Conflicts in an Optical Multistage Interconnection Network	568
<i>Ved Prakash Bhardwaj, Nitin, and Vipin Tyagi</i>	

An Efficient Methodology for Realization of Parallel FFT for Large Data Set	577
<i>Peter Joseph Basil Morris, Saikat Roy Chowdhury, and Debasish Deb</i>	
A Novel Approach for Adaptive Data Gathering in Sensor Networks by Dynamic Spanning Tree Switching	585
<i>Suchetana Chakraborty and Sushanta Karmakar</i>	
Hardware Efficient Root-Raised-Cosine Pulse Shaping Filter for DVB-S2 Receivers.	595
<i>Vikas Agarwal, Pansoo Kim, Deock-Gil Oh, and Do-Seob Ahn</i>	
Security, Trust and Privacy	
Security Analysis of Multimodal Biometric Systems against Spoof Attacks	604
<i>Zahid Akhtar and Sandeep Kale</i>	
A Novel Copyright Protection Scheme Using Visual Cryptography	612
<i>Amitava Nag, Jyoti Prakash Singh, Sushanta Biswas, D. Sarkar, and Partha Pratim Sarkar</i>	
A Weighted Location Based LSB Image Steganography Technique	620
<i>Amitava Nag, Jyoti Prakash Singh, Srabani Khan, Saswati Ghosh, Sushanta Biswas, D. Sarkar, and Partha Pratim Sarkar</i>	
Comments on ID-Based Client Authentication with Key Agreement Protocol on ECC for Mobile Client-Server Environment	628
<i>SK Hafizul Islam and G.P. Biswas</i>	
Covariance Based Steganography Using DCT	636
<i>N. Sathisha, K. Suresh Babu, K.B. Raja, K.R. Venugopal, and L.M. Patnaik</i>	
An Efficient Algorithm to Enable Login into Secure Systems Using Mouse Gestures	648
<i>Usha Banerjee and A. Swaminathan</i>	
Intrusion Detection by Pipelined Approach Using Conditional Random Fields and Optimization Using SVM	656
<i>R. Jayaprakash and V. Uma</i>	
A Flow-Level Taxonomy and Prevalence of Brute Force Attacks	666
<i>Jan Vykopal</i>	
Multi Application User Profiling for Masquerade Attack Detection	676
<i>Hamed Saljooghinejad and Wilson Naik Rathore</i>	

A Novel Technique for Defeating Virtual Keyboards - Exploiting Insecure Features of Modern Browsers	685
<i>Tanusha S. Nadkarni, Radhesh Mohandas, and Alwyn R. Pais</i>	
SQL Injection Disclosure Using BLAH Algorithm	693
<i>Justy Jameson and K.K. Sherly</i>	
Author Index	703

Balancing between Utility and Privacy for k-Anonymity

Korra Sathya Babu and Sanjay Kumar Jena

Department of Computer Science and Engineering, NIT Rourkela, India
{ksathyababu, skjena}@nitrkl.ac.in

Abstract. Organizations need to anonymize the data before releasing them so that data mining cannot predict private information. It's the duty of every organization to ensure privacy of its stakeholders. There is a tradeoff between privacy and utility of the released data. Many methods have been proposed earlier for correlating between the released data and the actual data. All of them use the information theoretic measures. Various methods have been proposed to tackle the privacy preservation problem like Anonymization and perturbation; but the natural consequence of privacy preservation is information loss. The loss of specific information about certain individuals may affect the data quality and in extreme case the data may become completely useless. There are methods like cryptography which completely anonymize the dataset and which renders the dataset useless making the utility of the data is completely lost. One needs to protect the private information and preserve the data utility as much as possible. The objective of this paper is to find an optimum balance between privacy and utility while publishing dataset of any organization. Privacy preservation is hard requirement that must be satisfied and utility is the measure to be optimized. One of the methods for preserving privacy is k-Anonymization which also preserves privacy to a good extent. Many other methods also were proposed after k- Anonymity, but they are impractical. The balancing point will vary from dataset to dataset and the choice of Quasi-identifier sensitive attribute and number of records.

Keywords: Data Mining, Clustering, k-Anonymity, Privacy, Utility.

1 Introduction

Organizations are bound to publish data for survey and research purposes. At the same time they need to ensure the privacy of its stakeholders. Unfortunately the infrastructure that supports data mining shows data mining as a threat to privacy. Availability of complete data in the data warehouse poses the problem of misusing the data. Also publishing summaries of census data carries risk of violating privacy. The solution for misuse data problem falls into two categories [1], data perturbation or secure multiparty computation. Data perturbation is the process of distorting the data before publication. Data perturbation includes distributed data partitioning (horizontal or vertical), Data Swapping, Generalization, Suppression, Randomization. Secure Multiparty Computation (SMC) refers to secure computation of a function with

distributed inputs. Methods such as two party comparison, secure circuit evaluation and secure sum can be used as SMC.

Anonymized Published data need to ensure certain degree of utility. When utility is guaranteed then privacy is hampered and vice versa. There is a need for balancing between utility and privacy of the data. This work deals with an experimental study of balancing for utility and privacy when data is released.

2 Related Work

Privacy has become a major concern to be checked while publishing the data. Anonymization of the data is to be done so that individual information cannot be inferred. The first of the kind of Anonymization algorithm proposed is k -Anonymity [2].

2.1 k -Anonymity

A table T satisfies k -Anonymity if for every tuple $t \in T$ there exists $k-1$ other tuples $t_{i1}, t_{i2}, \dots, t_{i(k-1)} \in T$ such that $t_{i1}[C] = t_{i2}[C] = \dots = t_{i(k-1)}[C]$ for all $C \in$ Quasi Identifier (QI). Other models like l -diversity [3] and t -closeness [4] were proposed to overcome the attacks of k -Anonymity but they are impractical and only lies on theoretical framework.

2.1 One Pass k - Means Algorithm (OKA)

This algorithm was proposed by Jun-Lin and Meng-Cheng in 2008 [5]. It is derived from the standard k -means algorithm and runs within one iteration. This algorithm has two stages first is the clustering stage and second is the adjustment stage.

Clustering Stage. Clustering stage proceeds by sorting all the records and then randomly picking N records as seeds to build clusters. Then for each record r remaining in the dataset, algorithm checks to find the cluster o for which this record is closest and assigns the record to the cluster and updates its centroid. The difference between the traditional k -means algorithm and OKA is that in OKA whenever a record is added to the cluster its centroid is updated thus improving the assignments in future and the centroid represents the real centre of the cluster. In OKA the records are first sorted according to the quasi-identifiers thus making sure that similar tuples are assigned to the same cluster.

Adjustment Stage. In the clustering stage the clusters that are formed can contain more than k tuples and there can be some clusters containing less than k tuples, therefore when these clusters are anonymized will not satisfy condition for k -anonymity. These clusters need to be resized to contain at least k tuples. The goal of this adjustment stage is to make the clusters contain at least k records, while minimizing the information loss. This algorithm first removes the extra tuples from the clusters and then assigns those tuples to the clusters having less than k tuples. The removed tuples are farthest from the centroid of the cluster and while assigning the tuples to the clusters it checks the cluster which is closest to the tuple before assigning

it, thus minimizing the information loss. If no cluster contains less than k tuples and some records are left they are assigned to this respective closest clusters.

3 Proposed Approach for Choosing k

Organizations cannot refrain from publishing information of the stakeholders. They need to publish information by the guidelines of the privacy policy. If the privacy policy is permitting to publish the information, then they need to be anonymized beforehand. The most practical method is the k -anonymity model. Choosing the value of k is a difficult task because it is dataset dependant. The method proposed here has two steps. First construct the clusters and second is to run the clusters through the Enhanced_Privacy Algorithm.

With the availability of the QIs in the dataset, attacks may be possible on k -anonymity. QI are covert channels for the attacks on k -anonymity. It is advisable even to make the QIs as generalized as possible. The following approach is proposed for lessening the attacks.

Once the table T is organized into clusters with k tuples using OKA, apply generalization hierarchy of QI on the clusters to form a privacy preserving k -anonymized table. This algorithm uses the output of OKA. The generalization hierarchy which is made should be complete which can map all possible values of the attribute to a single value. The time complexity of the algorithm is $O(n)$.

Algorithm: Enhanced_Privacy

Input: an adjusted partitioning $P = \{P_1, P_2, \dots, P_K\}$ of T and a generalization hierarchy for attributes

Output: Privacy enhanced k -anonymized table T'

1. *for* each Partition P_i of T do
2. *for* each QI in P_i do
3. *if* attribute values for partition P_i are not same do
4. use Generalization hierarchy to generalize
5. *if* attribute values for partition P_i are not same do
6. *goto* step 4
7. *end If*
8. *end If*
9. *end of For*
10. *end of For*

Parameters to check utility and privacy depend from data to data. To determine the utility of the dataset, Decision stump algorithm for classification is used. Decision stump is a machine learning model consisting of a single-level decision tree with a categorical or numeric class label. The results produced by weka [6] show percentage of tuples that can be correctly classified using the algorithm.

To determine the extent of privacy preserved by the dataset we counted the number of attributes whose values are completely suppressed.

$$\% \text{ of Privacy} = \left(\frac{\text{No. of Suppressed Values}}{\text{No. of QI}} \right) * 100$$

4 Experimental Results

The experiment was conducted on adult database from UCI machine learning repository which comprise of 45,222 instances with known values. It contains numerical as well as categorical attributes. It contains 15 attributes with its properties [7]. The algorithms were implemented in java and executed on a workstation with Intel Dual Core Processor, 1.80 GHz and 1.00 GB of RAM on Window XP SP2 platform.

Clustering of the database is done using WEKA. The K-means clustering algorithm was deployed for clustering. Out of the total 15 attributes of the dataset only 5 (age, education, native-country, race and workclass) were considered as quasi-identifiers and the rest as sensitive attributes. Generalization was performed on the clustered dataset. These generalization hierarchies are used for anonymizing evenly clustered data.

For age which is a numerical attribute, mean of all the tuple values was taken. The other attributes are generalized as shown in the figures Fig. 1 to Fig. 4.

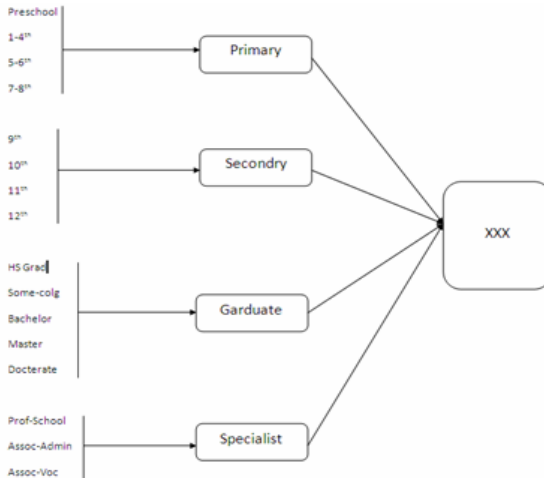


Fig. 1. Generalization hierarchy for education

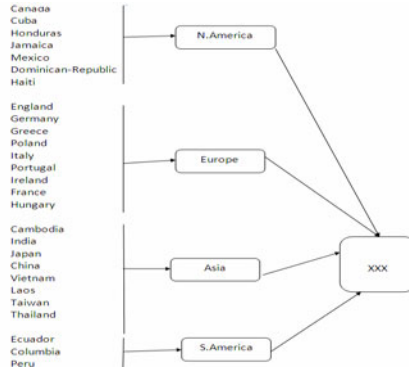


Fig. 2. Generalization hierarchy for native-country

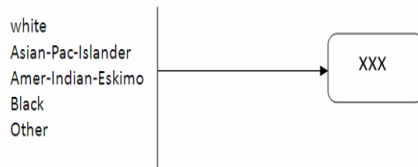


Fig. 3. Generalization hierarchy for race

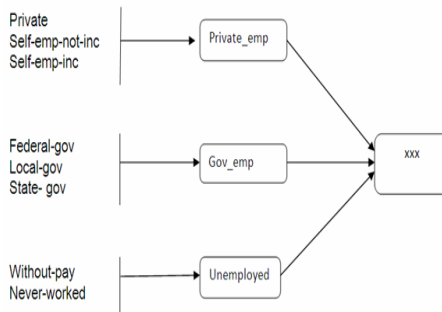


Fig. 4. Generalization hierarchy for workflow

Three experiments were conducted taking different parameters.

4.1 Experiment 1

In the first experiment six attributes (age, education, marital status, occupation, race and native-country) were considered for our analysis. Randomly 1000 tuples from the dataset are selected for anonymization to determine how utility varies with privacy. Age, education, race and country are considered as quasi-identifiers and other two as

sensitive attributes. Clusters were formed using WEKA according to the value of k . As described in section 2.1, the clusters produced may contain less than k tuples, thus an adjustment was done so that each cluster contains at least k tuples.

After adjusting the clusters, k -anonymization is performed based on the generalization hierarchy. The k -anonymization algorithm based on OKA was run to generalize the adjusted clusters.

For evaluating utility, we performed the classification mining on the k -anonymized dataset. Classification was performed by using weka software considering native-country as classification variable. Percentage of correctly classified tuples as the utility of the dataset was taken various values of k . Fig. 5 shows the results produced by the WEKA on using decision stump algorithm for a 3-anonymized dataset. Privacy was calculated by counting the number of tuples which are generalized to xxx. Privacy and utility was calculated by varying the value of k . The balancing point between utility and privacy is the point where privacy and utility curves intersect or tend to converge. Fig. 6 shows the variation of utility and privacy with k . It clearly follows from the figure that on increasing the value of k privacy provided by the dataset increases but utility decreases. For this sample dataset the balancing point comes between $k=8$ and $k=9$, and utility of the dataset at balancing point is around 60%.

Correctly Classified Instances	846	84.6847 %
Incorrectly Classified Instances	153	15.3153 %

Fig. 5. Classification Result for 3-Anonymized Dataset

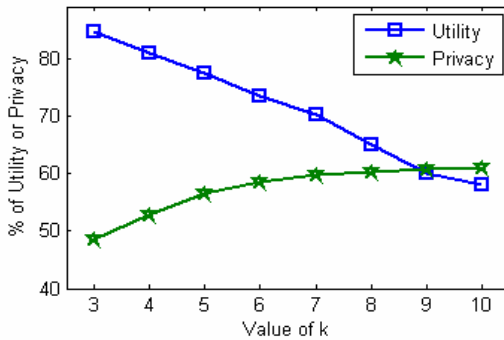


Fig. 6. Variation of Utility and Privacy with Anonymization

4.2 Experiment 2

In the second experiment all attributes were considered for analysis, to study the effect of more number of attributes on the privacy and the utility of the k -anonymized dataset. Again randomly selected 1000 tuples from the dataset for anonymization to determine how utility varies with privacy. Age, work class, education, race and native-country are considered as quasi-identifiers and all other attributes as sensitive

attributes. Similar steps were followed as in experiment 1 to study the variation of utility and privacy on varying k value.

Fig. 7 shows the variation of utility and privacy with k . For this sample dataset the balancing point comes between $k=11$ and $k=12$, and utility of the dataset at balancing point is around 52%. Thus on increasing the number of quasi-identifiers considered for analysis the balancing point falls down and values of k at which balance is achieved increases.

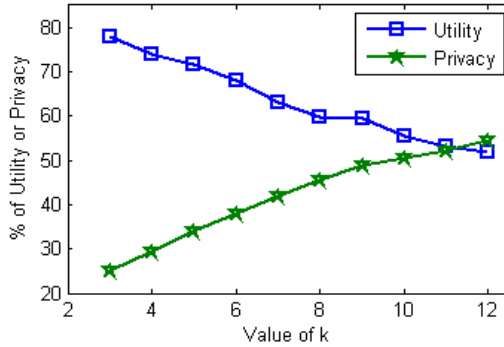


Fig. 7. Variation of Utility and privacy with Anonymization

4.2 Experiment 3

In this experiment 3000 tuples were taken from the adult dataset considering all the attributes, to study the effect of more number of tuples on the privacy and the utility of the k -anonymized dataset. Age, work class, education, race and native-country are considered as quasi-identifiers and all other attributes as sensitive attributes. Similar steps were followed as in experiment 1 to study the variation of utility and privacy on varying k value and shown in Fig. 8. For this sample dataset the balancing point comes between $k=10$ and $k=11$, and utility of the dataset at balancing point is around 50%.

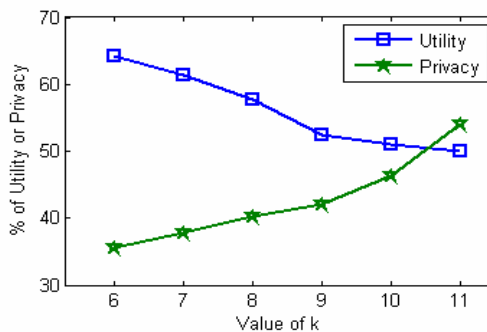


Fig. 8. Variation of Utility and Privacy with anonymization

5 Conclusion

There is a tradeoff between privacy and utility. On conducting the experiments we found that the balancing point between utility and privacy depends on the dataset and value of k cannot be generalized for all datasets such that utility and privacy are balanced. On varying the number of sensitive attributes in a dataset the balancing point varies. If the number of quasi-identifiers increases balancing point moves down and balance between utility and privacy occurs at a higher value of k . Thus if a dataset contains more number of quasi-identifiers then the utility as well as privacy attained at balancing point will be less than the dataset having fewer quasi-identifiers.

The number of tuples in the dataset has an effect on the balancing point. Increase in the number of tuple slightly shifts the balancing point and the value of k for which balancing occurs. Thus an approximate prediction can be made on the balancing point for huge dataset by conducting experiment on a sample dataset.

References

1. Vaidya, J., Clifton, C.W., Zhu, Y.M.: Privacy Preserving Data Mining, ch. 1. Springer, New York (2006)
2. Sweeney, L.: k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5), 557–570 (2002)
3. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-Diversity: Privacy beyond k-anonymity. *ACM Trans. Knowledge Discovery of Data* 1(3) (March 2007)
4. Li, N., et al.: t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In: *Proceedings of IEEE 23rd ICDE*, pp. 106–115 (April 2007)
5. Lin, J.-L., Wei, M.-C.: An Efficient Clustering Method for k-Anonymization. In: *Proceedings of the International Workshop on Privacy and Anonymity in Information Society*, vol. 331, pp. 46–50 (2008)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1) (2009)
7. UCI Repository of machine learning databases, University of California, Irvine, <http://archive.ics.uci.edu/ml/>

Evaluation of Approaches for Modeling of Security in Data Warehouses

Krishna Khajaria and Manoj Kumar

Department of Computer Science & Engineering
Ambedkar Institute of Technology, Delhi, India

Abstract. A Data Warehouse (DW) is a complex system that facilitates decision makers for the fulfillment of strategic, informational and decisional needs by extracting and integrating data from heterogeneous sources. They are complex due to the kind of problems they are built to solve. Due to the sensitive data contained in the DW it is important to assure the security of the DWs from the early stages of the life cycle of its development starting from requirements analysis to implementation and maintenance so that unauthorized attempts cannot access the DW. Traditionally, security was not considered as an important element in modeling of DWs. This survey paper gives a review of the DW security attempts at various abstraction levels like requirements, conceptual, logical, and physical modeling. Further, evaluation of various approaches for modeling security in DWs has been presented. This may help the designer, while selecting the good approach for DW design considering the security aspect as well.

Keywords: Data warehouse, Security modeling, Abstraction levels.

1 Introduction

Data Warehouses (DWs) are complex systems that facilitate decisions makers by combining and integrating data from heterogeneous sources. Multidimensional (MD) modeling forms an approach for storing data in DWs. Representing information in n-dimensions helps users to analyze the data in a way they think. MD modeling represents information into facts and dimensions. Facts are items of interest for the organization and dimensions represent the context in which facts are analyzed [43]. Due to the crucial information stored in the DWs it becomes important to ascertain the security of the information stored in the DW repository. According to Devbandu [1], security in the context of DWs is considered as an important requirement, which must be considered not as a separate aspect but as an element present in all the stages of DW developmental cycle starting from requirements to implementation and maintenance [1].

Security controls specified for Online Transactional Processing (OLTP) systems cannot be used for DWs because in operational systems, security controls are specified on tables, rows, columns etc. while in DWs, a large number of different users with different needs access the contents of the DWs where multidimensionality

is the basis [43, 18]. Many attempts were made traditionally to capture security in DWs but none considers it in the whole developmental cycle [18, 19]. On the other hand, research attempts were also made to capture security aspects of DWs at various abstraction levels i.e. Requirements modeling, Conceptual modeling, Logical modeling and Physical modeling and finally the generation of secure code for implementation on a specific platform [2].

This paper presents a survey of the literature work done by various authors in the recent past in the modeling of security in DWs at different abstraction levels and further various approaches for modeling of DW security have been evaluated. This may help the designer while selecting the good approach for DW design considering security aspects as well.

The rest of the paper is organized as follows: Section 2 discusses the related work consisting of approaches to DW design and security aspects in DWs. Section 3 enlists the modeling approaches to DW design considering DW security and Section 4 presents an evaluation of various approaches for modeling of DW security. Section 5 concludes the discussion and presents open issues.

2 Related Literature Review

A Data Warehouse (DW) is a complex system that fulfills the strategic, informational and decisional needs of an organization by extracting and integrating data from heterogeneous sources [4] into a multidimensional model. The input to DWs comes from various operational sources. The development of a DW system is also different from the development of a conventional operational system. Operational system has to fulfill the application specific business needs, whereas, the design of DW not only involves the information requirements of the user but the structure of the underlying source system. So, the development of a DW is complex and requires an appropriate lifecycle to be followed [2]. In [26] a comparison of the data/ supply-driven, requirements/user-driven and goal-driven methodologies for DW development using a case study is explained. The data/supply- driven approach starts with the purpose of reengineering existing operational systems schemas into the DW schema. The requirements/ user-driven approach rely on the business user for the information needs. The goal driven approach requires that the goals of the organization are to be taken into account when determining the information needs of the organization. One of the multidimensional design methods for generating a logical schema from Entity Relationship (ER) diagrams is given in [22]. Winter and Strauch [23] classified the multidimensional modeling methods in a supply-driven, demand-driven and a hybrid framework. They define them as follows:

Supply-driven approaches: Also known as data-driven, start from a detailed analysis of the data sources to determine the multidimensional concepts in a reengineering process.

Demand-driven approaches: Also known as requirement-driven or goal-driven or user driven, focus on determining the user multidimensional requirements and later map these requirements on data sources.

Hybrid/ Mixed approaches: To avoid the drawbacks of the individual approaches, literature proposed to combine the best of the above two approaches in order to design the DW from the data sources keeping in mind the end-user requirements.

Prat et al. [33] proposed a multidimensional (MD) model by following a demand-driven framework to derive DW relational schema (i.e. logical). One of the most cited MD design methods is presented in [22]. This approach generates a logical schema from Entity-Relationship (ER) diagrams, and it may produce MD schemas in terms of relational databases or MD arrays. A hybrid approach to derive logical schemas from SER (Structured Entity Relationship) diagrams, an extension of ER is given in [24]. In [25] a requirement-driven method is used to derive MD schemas in MD normal form (MNF). The authors argue that the design process must comprise four sequential phases i.e. requirements elicitation, conceptual, logical and physical design) like any classical database design process. A method is proposed in [27] to develop MD schemas from ER models. Bonifati et al. [28] presented a hybrid semi-automatic approach consisting of three basic steps: a demand-driven, a supply-driven and a third stage to conciliate the two first steps. One of the first methods for automating part of the design process by combining both supply and demand driven approaches is introduced in [29]. In [30] a semi-automatic supply-driven approach to derive logical schemas from XML schemas is presented. A supply-driven method of DW design from relational databases is given in [31]. A hybrid approach is presented in [32] to derive the conceptual MD schema by an agent-oriented method based on the i^* framework. A method is provided in [33] to derive the conceptual, logical and physical schema of the DWs according to the three abstraction levels recommended by ANSI / X3 / SPARC. In [34] a semi-automatic hybrid approach for obtaining the conceptual schema from user requirements and verification of its correctness against the data sources by means of Query/View/ Transformation (QVT) relations is presented. This approach work over relational sources and requirements are expressed in the i^* framework. An automatic supply-driven method that derives logical schemas from ER models for automatically identifying facts from ER diagrams by means of the connection topology value (CTV) is proposed in [35]. None of the approaches mentioned above considered security as an important aspect in the design, by providing designers with models which specifies security aspects. Table 1 gives the categorization of various design approaches discussed above.

As in any other system, in DWs also requirements guarantee that the system built meets the end-user requirements and must also consider the underlying data sources of the organization (i) to guarantee that the DW must be populated from data in the organization data sources, and (ii) to allow analytical capabilities for the end user. Currently, several methods for supporting the DW modeling task have been provided. However, they suffer from some significant drawbacks. In short, requirement-driven approaches do not consider the data sources as alternate evidence of analysis whereas data-driven approaches discover information for analysis from the data sources. Most methods agree on the opportunity for distinguishing between a phase of conceptual design (deriving an implementation-independent conceptual schema for the DW according to the chosen conceptual model starting from the user requirements and the

Table 1. Comparison of DW design approaches from literature

S. No.	Method	Supply/data	Demand/ requirements	Hybrid/ Mixed	Automatic/ Semi-automatic
1.	[33]		√		Not Applicable
2.	[22]			√	NA
3.	[24]			√	NA
4.	[25]		√		NA
5.	[28]			√	Semi-Automatic
6.	[29]			√	Semi- Automatic
7.	[23]		√		NA
8.	[30]	√			Semi- Automatic
9.	[31]	√			NA
10.	[32]			√	NA
11.	[33]			√	NA
12.	[34]			√	Semi- Automatic
13.	[35]	√			Automatic

structure of the data sources already available), logical design (takes the conceptual schema as input and creates a corresponding logical schema on the chosen platform by considering constraints. Some methods rely on physical design specifically related to the tools used for implementation. In some a phase of requirements analysis is separately considered [2].

In addition to normal DW functionalities, DWs require powerful security features. In [36] Landwehr defines the term secure as: “se” means without or “apart from” and “cure” means “to care for” or to be concerned about. There are many definitions of the primary requirements of security, the classical requirements of security are summarized by the abbreviation CIA, an acronym for Confidentiality, Integrity and Availability. All other security requirements such as non-repudiation, authentication, authorization, access controls etc can be traced back to these three basic properties. Confidentiality is defined as the absence of unauthorized disclosure of information. Integrity is defined as the absence of the unauthorized modification of information and availability as readiness for service when needed. Security must be implemented in DWs in an end-to-end manner. The real goal is to protect the data in the DW. Before data is loaded into the DW it is extracted, transformed, cleaned and prepared for loading. During this process, data is to be protected to the same standards as in the DW [36]. When users query data from DWs, user access security becomes an issue so that unauthorized attempts cannot be made. Security considerations must be considered at all layers of the DW system involved. A DW is not secure unless the underlying operating system is well secured and network security is adequately addressed [16]. A List of technologies are presented in [17] that have to be integrated to build a secure DW. This list comprises of secure heterogeneous database integration, statistical databases, secure data modeling, secure metadata management,

secure access methods and indexing, secure query processing, secure database administration, general database security, and secure high performance database management. As this survey focuses on DW security modeling, we'll consider only secure data modeling here. Secure data modeling is an essential task for building a DW. We need to integrate the MD data models with secure data models. Securing a DW requires much care and concentration of the efforts because in DW, the information is stored at one location only which poses much risk for the organization like unauthorized access to DW data through a remote query. Second the access to DW may lead to performance degradations. A standard method for setting up a secure DW doesn't exist. These security capabilities must be incorporated in this complex environment to maintain the confidentiality and integrity of the crucial information stored in the DWs. As users can discover the vast amount of information stored within DWs with various information delivery capabilities like complex queries, MD analysis, trends, statistics etc so it is essential to specify the security requirements from the early stages of the DW developmental cycle from requirements analysis to implementation and then to maintenance but not in an isolated manner [1] after the construction of DW.

A complete DW design method should span the three abstraction levels namely requirements, conceptual, logical and physical for modeling operational databases [15]. In the next section we'll give an overview of the modeling approaches for DW Security.

3 Overview of Modeling Approaches for DW Security

DW security modeling is the process of building an abstract model for DW security that is to be stored in the DW. This model is a representation of reality or a part of reality. A proposal that attempts to integrate security in DW based on metadata, is presented in [18]. Bhargava [42] also focuses on DW security. He talked about replication control, user-profile based security, anonymity. In [19] emphasis of DW security is based on view security. Both [27, 28] improves security in acquisition, storage and access attempts. Attempts to integrate security into the conceptual modeling such as [37, 38] were carried out but these operations only deals with OLAP (Online Analytical Processing) tools. The first approach to model security and MD model together with the extension of the UML (Unified Modeling Language) [21] to model secure DWs at the conceptual level [5] is given in [39]. Security issues related with the conceptual modeling are listed in [2] like a need of a reliable and flexible security model that considers all the components of a DW like data sources, ETL process, method for transforming security rules at various abstraction levels and to represent authorization rules in a common language for resolving conflicts and to represent a complete set of the levels of hierarchy of users playing different roles. The authors also focus on the difficulty of integrating the security requirements after the DW is built and deployed. They emphasizes to capture the security requirements at

the beginning of the life cycle and to make it a part of the system design. Some of the prominent approaches/methods for modeling security in DW are reviewed from literature. Till now no attempts were made to integrate security in the whole developmental cycle of DWs. An access control and audit (ACA) model is proposed in [20], specifically designed for DWs to define security constraints in early stages of the development lifecycle. By using an MDA (Model Driven Architecture) approach the authors consider security issues in all stages of the development process by automatically transform models at upper abstraction level towards logical models over a relational or MD approach and finally obtain secure code from the Platform Specific Model (PSM). In [21] authors uses an UML 2.0/OCL (Object Constraint Language) profile for representing the main security aspects in the conceptual modeling of DWs. UML extension [41] contains the necessary stereotypes, tagged values and constraints for a complete and powerful secure MD modeling. These new elements allows to specify security aspects such as security levels on data, compartments and user roles on the main elements of a MD modeling such as facts, dimensions and classification hierarchies. Modeling of both, information and QoS requirements (non-functional requirements) like security as an explicit stage in the development of a DW is presented in [4]. The authors pointed out that DW which satisfies the needs of the users, will be obtained if QoS requirements are modeled together with information requirements (functional), from the early stages in requirements analysis stage. This paper has focused on security modeling using i* framework. In [6] the authors presented an approach, based on UML, to represent main access control and audit rules in the conceptual modeling of data warehouses from the very early stages of a data warehouse project and enforce them in further design steps. In [40] an extension of the relational metamodel of the Common Warehouse Metamodel (CWM) is presented to represent security and audit measures in the logical modeling of data warehouse from the security measures considered at the conceptual level so as to reduce the semantic gap between the two.

4 Evaluation of DW Security Methods

Evaluation of some of the most prominent approaches for modeling DW security based on various subjective/ objective factors as given in table 2. Many types of security constraints are considered like Sensitive Information Assignment Rules (SIAR), Authorization Rules (AUR's), Audit Rules (AR), confidentiality rules etc. An MDA architecture towards secure development of DWs with a MD view towards OLAP tools for automatically developing secure MD logical models from conceptual models is applied in [12, 13] but both approaches were unable to include complex security rules. In [14] a framework for identifying users accessing the DW for improving the access control of DW by combining features from two traditional security models first by encrypting the data before it is stored in the DW and secondly accessing the DW using user level access controls to maintain confidentiality.

Table 2. Comparison of DW security modelling methodologies

S. No.	Citation	Approach/ method used	Model used	Technique used	Security considered at	Transformation Between the models	Permitted security rules	Validation	Traceability of rules	Development Time	Portability
1.	[20] 2008	MDA	ACA	UML extension	all abstraction levels	Possible till the generation of secure code	<ul style="list-style-type: none"> • SIAR • AUR • AR 	Yes	No	long	Not reached
2.	[7] 2006	Extension of UML 2.0	UML Metamodel	UML 2.0	conceptual level	At conceptual only using OCL.	<ul style="list-style-type: none"> • confidentiality 	No	No	long	Not reached
3.	[4] 2008	MDA	i* models	i* modeling framework	Requirements analysis	No	<ul style="list-style-type: none"> • security requirements 	No	No	NA	NA
4.	[40] 2006	Relational metamodel	Common Warehouse Metamodel (CWM)	Extension of relational metamodel	Logical level	security rules at conceptual level are represented at logical level	<ul style="list-style-type: none"> • security rules • audit rules 	Yes	No	short	Not reached
5.	[9,10] 2007	MDA	ACA	Framework based on MDA & QVT	Logical level	Using QVT	<ul style="list-style-type: none"> • security rules • audit rules 	Yes	Yes	Short	Reached
6.	[8] 2006	Extension of UML 2.0	CWM	UML Packages	conceptual level	No	<ul style="list-style-type: none"> • security rules • audit rules 	No	No	Short	Not reached
7.	[11] 2007	MDA	CWM	UML extension	all abstraction levels	Possible till the generation of secure code	<ul style="list-style-type: none"> • security rules • audit rules 	Yes	Yes	Short	Reached

5 Conclusions

The paper presents a comparative study of various approaches for modeling of security in the context of DWs. Having reviewed the various approaches, we point out that security may be modeled at various abstraction levels like requirements modeling, conceptual modeling, logical modeling, and the physical modeling and also there must be a transformation between the various levels of abstraction. We have presented a comparative evaluation of different prominent approaches for modeling of DW security. From the study it is concluded that security is important at all the stages of the developmental cycle of DWs from requirements analysis to implementation and maintenance. Thus, modeling of security will have a strong focus on the design and maintenance of DW.

References

1. Devbandu, P., Stubblebine, S.: Software Engineering for Security: a roadmap. In: Finkelstein, A. (ed.) *The Future of Software Engineering*, pp. 227–239. ACM Press, New York (2000)
2. Rizzi, S., Abello, A., Lechtenborger, J., Trujillo, J.: Research in data warehouse modeling and design: dead or alive? In: *DOLAP*, pp. 3–10 (2006)
3. Chung, L., do Prado Leite, J.C.S.: On Non-Functional Requirements in Software Engineering. In: Borgida, A.T., Chaudhri, V.K., Giorgini, P., Yu, E.S. (eds.) *Conceptual Modeling: Foundations and Applications*. LNCS, vol. 5600, pp. 363–379. Springer, Heidelberg (2009)
4. Soler, E., Stefanov, V., Mazon, N.J.: Towards Comprehensive Requirement Analysis for Data Warehouses: Considering Security Requirements, pp. 104–111. IEEE, Los Alamitos (2008)
5. Fernández-Medina, E., Trujillo, J., Villarroel, R., Piattini, M.: Extending UML for Designing Secure Data Warehouses. In: Atzeni, P., Chu, W., Lu, H., Zhou, S., Ling, T.-W. (eds.) *ER 2004*. LNCS, vol. 3288, pp. 217–230. Springer, Heidelberg (2004)
6. Medina, E.F., Trujillo, J., Villarroel, R., Piattini, M.: Access Control and Audit model for the multidimensional modeling of data warehouses. *Decision Support Systems* 42, 1270–1289 (2006)
7. Villarroel, R., Medina, E.F., Piattini, M.: A UML 2.0/ OCL Extension for designing Secure Data Warehouses. *Journal of Research and Practice in Information Technology* 38(1), 31–43 (2006)
8. Villarroel, R., Soler, E., Fernández-Medina, E., Trujillo, J., Piattini, M.: Using UML Packages for Designing Secure Data Warehouses. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) *ICCSA 2006*. LNCS, vol. 3982, pp. 1024–1034. Springer, Heidelberg (2006)
9. Soler, E., Villarroel, R., Trujillo, J., Medina, E.F., Piattini, M.: A Framework for the Development of secure Data Warehouses based on MDA and QVT. In: *ARES*. IEEE, Los Alamitos (2007)
10. Soler, E., Villarroel, R., Trujillo, J., Medina, E.F., Piattini, M.: A set of QVT relations to transform PIM to PSM in the Design of Secure Data Warehouses. In: *ARES*. IEEE, Los Alamitos (2007)

11. Soler, E., Villarroel, R., Trujillo, J., Medina, E.F., Piattini, M.: Application of QVT for the Development of Secure Data Warehouses: A case study. In: ARES. IEEE, Los Alamitos (2007)
12. Blanco, C., Guzman, I.G.R., Medina, E.F., Trujillo, J., Piattini, M.: Applying an MDA-based approach to consider security rules in the development of secure DWs. In: ARES, pp. 528–238 (2009)
13. Blanco, C., Guzman, I.G.R., Medina, E.F., Trujillo, J., Piattini, M.: Including security rules in a MDA approach for secure DWs. In: ARES, pp. 528–238 (2009)
14. Ahmad, S., Ahmad, R.: An Improved Security Framework for Data Warehouse: A Hybrid Approach, pp. 1586–1590. IEEE, Los Alamitos (2010)
15. Golfarelli, M., Rizzi, S.: A Methodological framework for Data Warehouse design. In: Proceedings of the First International Workshop on Data Warehousing and OLAP (1998)
16. Weipl, E.R.: Security in Data Warehouses
17. Thuraisingham, B., Kantarciogiu, M., Iyer, S.: Extended RBAC-Based design and Implementation for a Secure Data Warehouse (2007)
18. Katie, N., Quirchmayr, G., Schifer, J., Stoba, M., Tjoa, M.: A Prototype Model for Data Warehouse Security based on Metadata. In: DEXA 1998. IEEE Computer Society, Vienna (1998)
19. Rosenthal, A., Sciore, E.: View Security as the basis for data warehouse security. In: 2nd International Workshop on Design and Management of Data Warehouse, Sweden (2000)
20. Blanco, C., Guzman, I.G.R., Medina, E.F., Trujillo, J., Piattini, M.: Automatic generation of secure Multidimensional Code for Data Warehouses: An MDA Approach, In OTM 2008, Part II, LNCS 5332. In: Chung, S. (ed.) OTM 2008, Part II. LNCS, vol. 5332, pp. 1052–1068. Springer, Heidelberg (2008)
21. Luján-Mora, S., Trujillo, J., Song, I.-Y.: Extending the UML for Multidimensional Modeling. In: Jézéquel, J.-M., Hussmann, H., Cook, S. (eds.) UML 2002. LNCS, vol. 2460, pp. 265–276. Springer, Heidelberg (2002)
22. Cabibbo, L., Torlone, R.: A Logical Approach to Multidimensional Databases. In: Schek, H.-J., Salter, F., Ramos, I., Alonso, G. (eds.) EDBT 1998. LNCS, vol. 1377, pp. 183–197. Springer, Heidelberg (1998)
23. Winter, R., Strauch, B.: A method for Demand-driven Information Requirements Analysis in DW Projects. In: Proc. of 36th Annual Hawaii Int. Conf. on System Sciences, pp. 231–239. IEEE, Los Alamitos (2003)
24. Bohnlein, M., Ende, U.: Deriving Initial Data Warehouse Structure from the Conceptual Data Models of the Underlying Operation Information Systems. In: Proc. of 2nd Int. Workshop on Data Warehousing and OLAP, pp. 15–21. ACM, New York (1999)
25. Husemann, B., Lechtenborger, J., Vossen, G.: Conceptual Data Warehouse Modeling. In: Proc. of 2nd Int. Workshop on Design and Management of Data Warehouses (2000), <http://CEUR-WS.org>
26. List, B., Bruckner, R.M., Machaczek, K., Schiefer, J.: A Comparison of Data Warehouse Development Methodologies Case Study of the Process Warehouse. In: Hameurlain, A., Cicchetti, R., Traunmüller, R. (eds.) DEXA 2002. LNCS, vol. 2453, pp. 203–215. Springer, Heidelberg (2002)
27. Moody, D.L., Kortink, M.K.: From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design. In: Proc. of 2nd Int. Workshop on Design and Management of Data Warehouses (2000), <http://CEUR-WS.org>
28. Bonifati, A., Cattaneo, A., Ceri, S., Fuggetta, A., Paraboschi, S.: Designing Data Marts for Data Warehouses. ACM Trans. Softw. Eng. Methodol. 10(4), 452–483 (2001)

29. Phipps, C., Davis, K.C.: Automating Data Warehouse Conceptual Schema Design and Evaluation. In: Proc. of 4th Int. Workshop on Design and Management of Data Warehouses, vol. 58, pp. 23–32 (2002), <http://CEUR-WS.org>
30. Vrdoljak, B., Banek, M., Rizzi, S.: Designing Web Warehouses from XML Schemas. In: Kambayashi, Y., Mohania, M., Wöß, W. (eds.) DaWaK 2003. LNCS, vol. 2737, pp. 89–98. Springer, Heidelberg (2003)
31. Jensen, M.R., Holmgren, T., Pedersen, T.B.: Discovering Multidimensional Structure in Relational Data. In: Kambayashi, Y., Mohania, M., Wöß, W. (eds.) DaWaK 2004. LNCS, vol. 3181, pp. 138–148. Springer, Heidelberg (2004)
32. Giorgini, P., Rizzi, S., Garzetti, M.: Goal-oriented Requirement Analysis for Data Warehouse Design. In: Proc. of 8th Int. Workshop on Data Warehousing and OLAP, pp. 47–56. ACM Press, New York (2005)
33. Prat, N., Akoka, J., Comyn-Wattiau, I.: A UML-based Data Warehouse Design Method. *Decision Support Systems* 42(3), 1449–1473 (2006)
34. Mazon, J.N., Trujillo, J., Lechtenborger, J.: Reconciling Requirement-Driven Data Warehouses with Data Sources Via Multidimensional Normal Forms. *Data & Knowledge Engineering* 23(3), 725–751 (2007)
35. Song, I.Y., Khare, R., Dai, B.: SAMSTAR: A Semi-Automated Lexical Method for Generating STAR Schemas from an ER Diagram. In: Proc. of the 10th Int Workshop on Data Warehousing and OLAP, pp. 9–16. ACM Press, New York (2007)
36. Landwehr, C.E.: Computer security. *Int. Journal of Information Security* 1(1), 13 (2001)
37. Jajodia, S., Wijesekera, D.: Securing OLAP data cubes against privacy breaches. In: Proc. IEEE Symp. on Security and Privacy, pp. 161–178 (2004)
38. Priebe, T., Pernul, G.: A pragmatic approach to conceptual modeling of OLAP security. In: Proc. ER, pp. 311–324 (2000)
39. Fernandez-Medina, E., Trujillo, J., Villarroel, R., Piattini, M.: Extending UML for designing secure data warehouses. In: *Decision Support Systems* (2006)
40. Soler, E., Villarroel, R., Trujillo, J., Medina, E.F., Piattini, M.: Representing Security and audit rules for data warehouse at logical level by using a Common Warehouse Metamodel. In: Proc. of the 1st Int. Conf. on Availability, Reliability and Security. IEEE, Los Alamitos (2006)
41. Medina, E.F., Trujillo, J., Villarroel, R., Piattini, M.: Developing secure data warehouses with the UML extension, pp. 826–856 (2006), doi:10.1016/j.is.2006.07.003
42. Bhargava, B.: Security in Data Warehousing (Invited Talk). In: *Proceedings of the 3rd Data Warehousing and Knowledge Discovery* (2000)
43. Inmon, H.: *Building the Data Warehouse*, 3rd edn. John Wiley & Sons, USA (2002)

Content Based Compression for Quicx System

Radha Senthilkumar, C. Lingeswarara, and A. Kannan

Department of Information Technology, Anna University
Chennai, India

Abstract. XML is widely deployed over the Internet for a prevalent tasks, including configuration files, protocols, and web services. In XML, simple messages can be quite large due to verbose syntax. The two major constraints in XML application is its size and querying efficiency. This paper proposes a scheme to increase the compression ratio of QUICX (Query and Update support for Indexed and Compressed XML) system by applying a bitmask based compression algorithm instead of using LZW compression. The proposed algorithm carries over the data containers present in the system for increasing the compression ratio. In conventional LZW compression, the repetitive contents are replaced by corresponding code that is available in the dictionary. On encountering new contents, corresponding code is created and entered in to the dictionary and to the compressed file. The dictionary size is increased for the file containing more non-repetitive words. This drawback is averted using bitmask based compression and the results are drawn showing the efficiency of the proposed system.

Keywords: XML compression, QUICX, Bit-mask compression, LZW compression.

1 Introduction

XML has become the de facto standard for data exchange. However, its flexibility and portability are gained at the cost of substantially inflated data, which is a consequence of using repeated tags to describe data. This hinders the use of XML in both data exchange and data archiving. The redundancy often present in XML data provides opportunities for compression. In recent years, many XML compressors have been proposed to solve this data inflation problem. Earlier XML compressors make use of the similarities between the semantically related XML data to eliminate data redundancy so that a good compression ratio is always guaranteed. However, in this approach the compressed data is not directly usable; a full chunk of data must be first decompressed in order to process the imposed queries. However, some applications, in particular those frequently querying compressed XML documents, cannot afford to fully decompress the entire document during query evaluation, as the penalty to query performance would be prohibitive. Instead, decompression must be carefully applied on the minimal amount of data needed for each query. Some XML conscious queryable compressors XGrind, XMLZip, XPress and XML Skeleton Compression support direct querying on compressed data, but only at the expense of the

compression ratio, thus the XML size problem is not satisfactorily resolved. XML query language such as XPath and XQuery needs an efficient way to query the structure of XML documents, like XML document could be indexed in advance, as opposed to querying on the fly. QUICX(Query and Update support for Indexed and Compressed XML) system[11] support better compression ratio, efficient query processing by means of indexing and efficient update management. Here overall aim is to achieve a good compression ratio than existing QUICX with LZW and to support efficient evaluation of queries over the compressed data without prior decompression.

2 Related Work

XML is well known for its irregular structure and its verbosity nature. One of the interesting applications for file based XML is web searching. In this application, if each web server manages its own data in the form of XML and transmits it through the network, the storage and the network bandwidth are wasted due to above mentioned issues. Compression has long been recognized as a useful means to improve the performance of large XML databases. In this section we examine standard XML compressors for increasing compression ratio and querying efficiency over compressed XML data. XMILL [7] an efficient compressor for XML data was designed to minimize the size of compressed XML data. It is the first XML conscious compressor technique. In XMILL, we need to decompress whole compressed XML data before processing imposed query on compressed XML. This burden on query processing made a research line to concentrate on querying over compressed XML without decompressing whole compressed file. XGRIND[5], A query-friendly XML compressor and their contribution towards this paper is to support querying over compressed XML data and to achieve better compression ratio. The compression ratio delivered by XGrind is much worse than that of XMILL and it takes much longer time than that of XMILL, since it requires parsing the input XML document twice during the compression process. XPRESS[6], supports querying over compressed XML data and achieved better compression ratio than XGRIND but when compared with XMILL, compression ratio and compressing time yielded by XPRESS is inefficient. QUICX provides a more compact structure for storing XML and querying XML data with the help of indexing with the data. QUICX consider the similar structure of XML for all record for better compression. First step in QUICX is to split up all features from XML, such as structure schema, metadata for tracking the occurrence of tag and content. For each tag, system maintains each container for data and compresses each container using dictionary based LZW compression. LZW Compression technique is one of the most standard and widely used compression techniques. LZW starts with initializing the dictionary with all the possible single character and corresponding code. Then Scan the input string, until it finds the longest string not in the dictionary. Once it found the longest string, insert it in the dictionary with the corresponding code and enter the code for the string (length one less than longest one) in the compressed code. LZW compression is the best technique for files containing more repetitive data and is not suite for the file containing more non-repetitive data; dictionary size occupies more space, if the file contains more non-repetitive data's and at decompression, time taken for regenerating same dictionary

generate at compression phase is time consuming. In proposed compression technique, this inefficiency is eliminated by recording minimum difference between the original and dictionary content in the compressed file. Further querying efficiency over the compressed content is increased by keeping the semantic information about XML file after compression by applying node and content indexing.

3 Proposed Concept

This paper put forwards a new Bitmask based compression that improves the compression ratio of dictionary based compression techniques by considering mismatches in data. The primary motive of this concept is to determine the difference between consecutive bit positions present in the data sequence of dictionary and actual data. Once this difference is determined, the value is stored along with the corresponding location into a compressed file instead of the original content. The compression ratio is given by the number of bits changed in dictionary with respect to the number of bits changed in actual data. The compression ratio also depends on certain factors such as word length, dictionary size and the number of bitmasks used. The relationship between the word size and direct matches is inversely proportional to each other. Another factor taken into consideration while implementing compression is the size of dictionary. A dictionary with larger size and numerous match words needs to be replaced with dictionary index, but at the cost of increased index size in bits. The extensive use of bitmasks results in more compressed words and requires supplementary bits to encode the bitmask information. In QUICX system [11], separate container holds all XML tag data, which gives support for applying content based compression. In new content based compression, text, numeric and mixed data are considered separately by applying different procedure for compression. Individual dictionary have been created for each container present in the QUICX system during the compression of each container. In new content based compression separate compression format is maintained for text, numeric and mixed contents and each format is different from one another in terms of size in bits. General format is represented in Fig 1.

Decision bit	Dictionary index	Bitmask location	Bitmask value
--------------	------------------	------------------	---------------

Fig. 1. Frame Format for Compressed File

Fig 2 represents the numeric data frame format for direct match in dictionary and Fig 3 represents the numeric data frame format for single mismatch from dictionary.

In both the above frame formats, first two bits are called as decision bits, next 4 bits are used for dictionary index. In addition to this, numeric data frame format for single mismatch from dictionary contains 5 more bits, out of which three bits are allocated for bitmask location and last two bits are allocated for bitmask value. Decision bits are used to indicate, whether the compressed content will contain direct

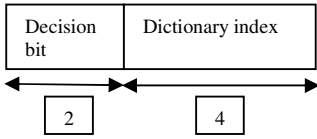


Fig. 2. Frame format for direct

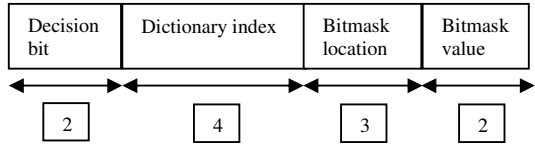


Fig. 3. Frame format for single mismatch from dictionary

dictionary index or dictionary index with corresponding bitmask location and bitmask value. To indicate dictionary index, 4 bits are allocated, which allows 16 dictionary entries for compressing numeric data. Details regarding decision bits are listed in the Table 1.

Table 1. Decision bit values and purposes for numeric data

Values	Purpose	Number of bits
00	Dictionary index alone	6
01	Dictionary index plus bitmask location and bitmask value	11
11	End of particular string	2

The difference between the frame format of text content and numeric content is the number of bits allocated for mentioning dictionary indexes. Fig 4 represents the frame format for text content with direct match in dictionary content, Fig 5 represents the frame format for text content with single mismatch from the dictionary content, and Fig 6 represents the frame format for text content with dual mismatch from the dictionary content in new content based compression. This dual mismatch is also allowed in addition to single mismatch in text content when the dictionary entry exceeds the threshold limit. Here we fix the dictionary threshold limit as 25 for handling the critical situation during the compression.

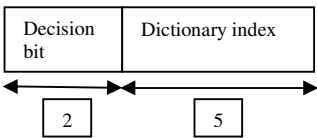


Fig. 4. Frame format for direct match in dictionary

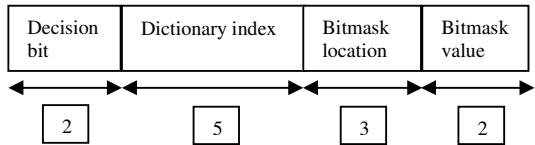


Fig. 5. Frame format for single mismatch from dictionary

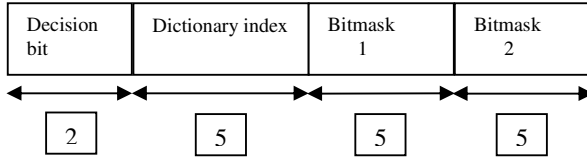


Fig. 6. Frame format for dual mismatch from dictionary

For compressing text content in new bitmask based compression 5 bits are allocated for mentioning dictionary index. It allows 32 dictionary entries for successfully compressing text content. Details regarding decision bits for text data compression are listed in table 2.

Table 2. Decision bit values and purposes for text data

Values	Purpose	Number of bits
00	Dictionary index alone	7
01	Dictionary index plus bitmask location and bitmask value	12
10	Dictionary index plus two bitmask location and value	17
11	End of particular string	2

In compressing mixed content, we consider characters, numbers and all special characters to allocate number of bits for dictionary index. For compressing mixed content, number of entries in dictionary will be more when compared to text and numeric content. Frame format for compressing mixed content is similar to text content except the size. Here 7 bits are allocated for dictionary index and it allows 128 dictionary entries for successfully compressing mixed contents. Similar to text data compression, here also dual mismatch from dictionary is allowed only after the dictionary entries reach the certain threshold limit.

Section 3.1, 3.2 and 3.3 presents the pseudo code for compressing numeric, text and mixed content respectively, in which variables named `decision_dic`, `decision_mis1`, `decision_mis2`, `decision_end` denote the decision bit values. Variables named `original_data` and `compress_data` denote the current data from the original file and the corresponding compressed data. Function “difference” accepts the `dictionary_index` and `original_data` as input parameters and returns the count of bits changed on comparing the `dictionary_data` and `original_data`. Function “bitmask_location” accepts the `dictionary_index` and `original_data` as input parameters and returns the location value. This location value indicates the changed bit location in `dictionary_data` when compare it with `original_data`. Function “bitmask_value” accepts the `dictionary_index` and `original_data` as input parameters and returns the bitmask value which is added to the `dictionary_data` at the `bitmask_location` for getting the original data. Variable named `diff_value` contain the difference value between the

original data and all the dictionary entries present in the dictionary. Variable named `min_value` contains the minimum `diff_value`. In numeric data compression current original is entered in to the dictionary if the `min_value` is greater than one.

3.1 Pseudo Code for Compressing Numeric Content

```

Input:      Meta-table,
               Container file which is to be compressed
Output:    Compressed container file and corresponding dictionary

Begin
String decision_dic = "00", decision_mis1 = "01", decision_end = "11"
while(NOT EOF)
  char ch = getchar();
  while(ch != '>')
    string original_data = ch;
    ch = getchar();
  end while
  append 'X' to the string if string length is odd
  for(each pair of character from the string)
    begin
      if(dictionary is empty)
        insert the string in to the dictionary
        string compress_data = decision_dic + dictionary_index
      else
        for each(i in dictionary entries)
          diff_value[i] = difference(original_data,dictionary_data)
          min_value = minimum(diff_value)
          if(min_value > 1)
            insert the string in to the dictionary
            compress_data = decision_dic +dictionary_index
          else
            find the dictionary index of min_value
            location = bitmask_location(dictionary_index, string );
            value = bitmask_value(dictionary_index, string);
            compress_data = decision_mis1+dictionary_index + location + value
          end if
        end if
      end
    end
    compress_data = decision_end
  end while
End

```

3.2 Pseudo Code for Compressing Text Content

The pseudo code for compressing text content is similar to numeric content. In addition to single mismatch, dual mismatch is also allowed based on the variable `min_value` and dictionary count. Single mismatch is allowed when `min_value` is equal to one and dictionary count is less than 25. Dual mismatch is allowed when `min_value` is equal to two and dictionary count is greater than 25.

3.3 Pseudo Code for Compressing Mixed Content

The pseudo code for compressing mixed contents is also similar to numeric content and text content. Here single mismatch is allowed when min_value is equal to one and dictionary count is less than 75. Dual mismatch is allowed when min_value is equal to two and dictionary count is greater than 75.

3.4 General Compression Steps

Get the input string from container where symbol '>' as a delimiter. Consider the sample container in Fig 7.

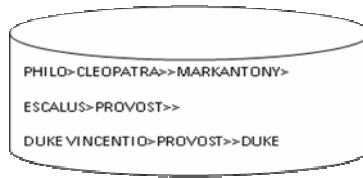


Fig. 7. Sample containers in QUICX

Separate the string in to pairs and handle each pair separately for all the strings.

PHILOX - PH IL OX

For each pair - PH,

1. Convert in to binary streams and it occupies 16 bit for each pair of data before compression
2. If dictionary contains no data, insert the 16 bit input in to the dictionary and output the index value in the dictionary as the compressed data for that pair.
3. If not, compare the 16 bit current data with all data in the dictionary and record the difference by means of XOR operations.

3.4.1 Comparison Steps

1. Split the 16 bit binary data in to 4 pairs (each pair containing FOUR bits).
2. Do (current data (XOR)Dictionary data) as shown in Fig.8 and result is stored temporarily for finding the minimum difference dictionary content

Next pair - IL

Binary form - 0100 0000 0100 1000

(Dic) 0101 (XOR) (curr) 0100	(Dic) 0000 (XOR) (curr) 0000	(Dic) 0100 (XOR) (curr) 0100	(Dic) 1000 (XOR) (Curr) 1000
<u>0001</u>	0000	0000	0000

Fig. 8. Comparison procedure

After comparing with all contents in dictionary, find the minimum bit difference dictionary index and output the indexes with location and bitmask as the compressed output for the current pair.

4 Performance Result

In this paper, the performance analysis of proposed bit mask based compression is carried out by extensive set of experiments. All experiments were run on a Windows 7 machine with Intel Core 2 Duo, 2.2 GHz CPU and 3 GB main memory. The results of the compression performance of proposed scheme are compared with standard XML conscious compressors like XPRESS, XQZIP, RFX and QUICX. Three standard sets are used for evaluation, which includes Line-item and Shakespeare. The compression ratio is defined as:

Compression ratio = $(\text{Original container size} - \text{Compressed container size}) / \text{Original container size} * 100\%$

Table 4 represents the performance measure of QUICX system with LZW compression. Result obtained for QUICX with LZW is reasonable when compare to other queriable XML conscious compressor. Last column in table 3 shows the compression ratio for three standard datasets. It is calculated based upon above mentioned formula.

Table 3. Performance measure of QUICX with LZW compression

Dataset	original size	Compressed size	Compression ratio
Lineitem	30.7	8.28	73.02
Mondial	1.74	0.307	82.35
Shakespeare	7.52	2.79	62.89

As we know QUICX is one among queriable XML compressor, it allows the query processor to processes the query over the compressed containers with the help of indexing. Indexes are created for particular XML file by using uncompressed container and Meta data available in the QUICX system. Once the query is given to the query processor for processing, it will find out the intended record number by using the indexes. Then query processor can directly decompress the intended record number instead of decompressing all the compressed records. It will reduce the query processing time tremendously. But in LZW compression, we need to regenerate the dictionary at the decompression time by using the compressed content. It increases the query processing time even though we find out the intended record number quickly. To overcome this degradation in the QUICX system, we are replacing the LZW compression by new bit mask based compression. In new content based compression we are maintaining same dictionary in decompression process, which is created

during the compression time. So the query processor can timely decompress the intended record by using the same dictionary generated at the time of compression. Major challenge in maintaining static dictionary is, it will increase the size of compressed file. In new bit mask based compression we are maintaining reasonable compression ratio for all standard datasets by having static dictionary which is created during the compression.

4.1 Effect of Different Dictionary Size for Particular Content Type

We carried out set of experiments to explore the effect of using different dictionary size on standard datasets. We choose three representative documents: line-item (which has lot of numeric content) and Shakespeare (which has dominated mixed content) and Mondial (which has moderate text and numeric content). For mixed content, 8 bits are allowed for denoting the dictionary index i.e., 512 dictionary entries. But when we apply this same bit format to text and numeric contents, number of dictionary entries entered are low and the compression ratio obtained for this format is not up to the expected level. It decreases the efficiency of new bit mask compression scheme. In order to get the original efficiency of new bitmask based compression scheme, we reduce the bit size from 8 to 4 for numeric content and from 6 to 5 for text content to adopt the low dictionary entries. Reason for low dictionary entries in numeric and text content is the number of possible combination for any text and numeric content is low, when compared to mixed content. Mixed content includes text, numeric and all special characters.

Table 4. Compression efficiency for line-item dataset

Dataset Name:	Lineitem.XML
Original File Size:	30.7 MB
Containers from QUICX	7.83 MB
Compressed Size:	5.62 MB
Ratio:	.8169
% of Efficiency:	81.69%

Table 5. Compression efficiency for Shakespeare dataset

Dataset Name:	shakespeare.XML
Original File Size:	7.52 MB
Containers from QUICX	4.74 MB
Compressed Size:	3.28 MB
Ratio:	.5638
% of Efficiency:	56.38%

Table 6. Compression efficiency for Mondial dataset

Dataset Name:	Mondial.XML
Original File Size:	1.74 MB
Containers from QUICKX	0.455 MB
Compressed Size:	0.3 MB
Ratio:	.8275
% of Efficiency:	82.75%

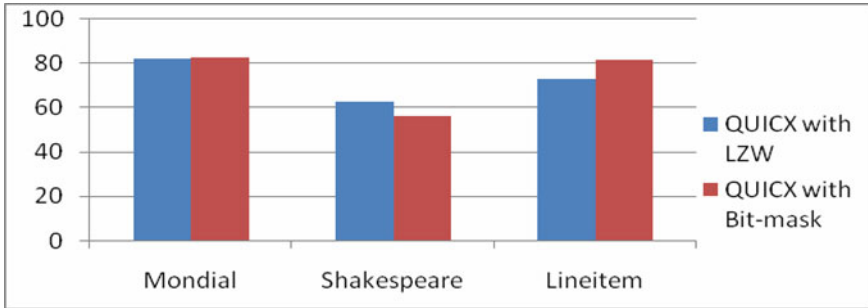
**Fig. 9.** QUICKX with LZW compression VS QUICKX with Bit-Mask based compression

Table 4, Table 5 and Table 6, represents the performance of new proposed algorithm in terms of compression ratio and it shows reasonable compression ratio than other queriable XML compressor. Table 4 shows the compression ratio and compression efficiency for standard dataset line-item, which contains more numeric and text content than mixed content. Four bits are allowed for dictionary indexes to compress the containers in the line-item containing numeric content. Five bits are allocated for dictionary indexes to compress the containers containing text content. Similarly table 5 shows the compression ratio and compression efficiency for standard Shakespeare dataset. This contains dominated mixed content than text and numeric content. Here 7 bits are allowed for dictionary indexes to compress the containers containing mixed content. Fig 9 shows the graphical representation of performance comparison between QUICKX with LZW compression and QUICKX with Bit-Mask based compression.

5 Conclusion

QUICKX provides a flexible and compact framework for storage; retrieval and update mechanism of XML file dynamically. LZW compression in QUICKX system support better compression for the file containing more number of repeated words, but it fails to provide for the file containing more different words. The proposed Bitmask based compression yields better compression ratio for standard benchmark dataset such as Line item, Shakespeare and Mondial. The querying efficiency can be improved by

applying Indexing technique and by using dictionaries present in compressed containers are left as future work.

References

- [1] Arion, A., Bonifati, A., Costa, G., D'Aguanno, S., Manolescu, I., Pugliese, A.: Efficient query evaluation over compressed XML data. In: Hwang, J., Christodoulakis, S., Plexousakis, D., Christophides, V., Koubarakis, M., Böhm, K. (eds.) EDBT 2004. LNCS, vol. 2992, pp. 200–218. Springer, Heidelberg (2004)
- [2] Cheney, J.: Compressing XML with multiplexed hierarchical PPM models. In: Proceedings of the IEEE Data Compression Conference, pp. 163–172 (2000)
- [3] Murthy, C., Mishra, P.: Lossless Compression using Efficient Encoding of Bitmasks. In: IEEE Computer Society Annual Symposium on VLSI (2009)
- [4] Cheng, J., Ng, W.: XQzip: Querying compressed XML using structural indexing. In: Hwang, J., Christodoulakis, S., Plexousakis, D., Christophides, V., Koubarakis, M., Böhm, K. (eds.) EDBT 2004. LNCS, vol. 2992, pp. 219–236. Springer, Heidelberg (2004)
- [5] Haritsa, J.R., Tolani, P.M.: XGRIND: A query-friendly XML compressor. In: IEEE Proceedings of the 18th International Conference on Data Engineering (2002)
- [6] Min, J.K., Park, M.J., Chung, C.W.: XPRESS: A queriable compression for XML data. In: Proceedings of the ACM SIGMOD International Conference on Management of Data (2003)
- [7] Liefke, H., Suciu, D.: XMill: An efficient compressor for XML data. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 153–164 (2000)
- [8] Murthy, C., Mishra, P.: Bitmask-based control word compression for NISC architectures. In: Proceedings of ACM Great Lakes Symposium on VLSI, GLSVLSI (2009)
- [9] Grimsno, N.: Faster Path indexes for search in XML data. In: Proc.19 Australasian Database Conference (2008)
- [10] Seong, S., Mishra, P.: Bitmask-based code compression for embedded systems. IEEE Trans. CAD 27(4), 673–685 (2008)
- [11] Senthilkumar, R., Kannan, A.: Query and Update support for Indexed and Compressed XML (QUICX). In: WIMo/CoNeCo 2011. CCIS, vol. 162, pp. 414–428 (2011)

NL-Based Automated Software Requirements Elicitation and Specification

Ashfa Umber¹, Imran Sarwar Bajwa¹, and M. Asif Naeem²

¹Department of Computer Science & IT, The Islamia University of Bahawalpur, Pakistan
ashfaumber@yahoo.com, imran.sarwar@iub.edu.pk

²Department of Computer Science, University of Auckland, Auckland, New Zealand
mnae006@aucklanduni.ac.nz

Abstract. This paper presents a novel approach to automate the process of software requirements elicitation and specification. The software requirements elicitation is perhaps the most important phase of software development as a small error at this stage can result in absurd software designs and implementations. The automation of the initial phase (such as requirement elicitation) phase can also contribute to a long standing challenge of automated software development. The presented approach is based on Semantic of Business Vocabulary and Rules (SBVR), an OMG's recent standard. We have also developed a prototype tool SR-Elicitor (an Eclipse plugin), which can be used by software engineers to record and automatically transform the natural language software requirements to SBVR software requirements specification. The major contribution of the presented research is to demonstrate the potential of SBVR based approach, implemented in a prototype tool, proposed to improve the process of requirements elicitation and specification.

Keywords: Requirements Elicitation, Requirement Engineering, Requirements Specification, Natural Language Processing.

1 Introduction

Requirement engineering is a well-known software engineering discipline involving gathering, articulating and verifying the software requirements specifications (SRS). Requirement elicitation is the key phase of software requirement engineering as only the correct, complete, and unambiguous software requirements can result in correct, consistent and fault-tolerant software models [1]. A natural language (NL) is typically used to specify software requirements. However, the software requirements specified in English can be ambiguous and inconsistent due to inherent syntactic ambiguities and semantic inconsistencies [2]. The ambiguous SRS can not only result in conflicting and absurd software models but also complex to machine process.

In this paper, we report a novel approach to automatically translate the English SRS to SBVR (Semantic Business Vocabulary and Rules) [4] representation. The SBVR representation not only generate accurate and consistent software models but also machine process-able as SBVR has a pure mathematical foundation [4].

The presented approach works as the software engineer inputs a piece of English SRS and our approach transforms to SBVR based SRS. A multi-step procedure is adopted for NL to SBVR transition; firstly, the input English text is lexically, syntactically and semantically parsed and then SBVR vocabulary is extracted. Finally, the SBVR vocabulary is used to generate a SBVR rule representation of NL SRS.

The remaining paper is structured into the following sections: Section 2 states preliminaries of the presented research. Section 3 presents the framework of SR-Elicitor. Section 4 presents a case study and the results with performance evaluation are discussed in section 5. Finally, the paper is concluded to discuss the future work.

2 Preliminaries

2.1 Semantic Business Vocabulary and Rules (SBVR)

SBVR [4] is a recently introduced standard by OMG. Using SBVR, requirement can be captured in NL. The SBVR representation is simple to machine process as SBVR is based on formal logic. SBVR can produce SBVR business vocabulary, SBVR business rules and SBVR business facts in particular business domain.

SBVR Business Vocabulary. A business vocabulary [4] (section: 8.1) consists of all the specific terms and definitions of concepts used by an organization or community in course of business. In SBVR, there are four key elements:

- An object type is a general concept that exhibits a set of characteristics to distinguishes that object type from all other object types” [4] e.g. robot, user, etc.
- In SBVR, an individual noun is a qualified noun that corresponds to only one object [4] e.g. ‘Robby’, a famous robot.
- A characteristic is an abstraction of a property of an object [4] e.g. name of robot is Robby, here name is characteristic.
- A verb concept is a verb in English sentences e.g. *orders*.
- A fact type specifies relationships among noun concepts e.g. car has wheels.

SBVR Business Rules. A SBVR business rule is a formal representation ‘Under business jurisdiction’ [4]. Each SBVR business rule is based on at least one fact type. The SBVR rules can be a structural rule [4] used to define an organization’s setup or a behavioural rule [4] used to express the conduct of a business entity.

2.2 SBVR Based Controlled Representation

SBVR was originally presented to assist business people in creating clear and unambiguous business policies and rules in their native language [4]. The following characteristics of SBVR can help in generating a controlled representation of English:

Rule-Based Conceptual Formalization. SBVR standard provides a rule-based conceptual formalization that can be employed to generate a syntactically formal

representation of English. SBVR proposes the use of vocabulary (concepts, terms, etc.) for conceptual modeling. Furthermore, vocabulary can be employed to capture expressions in the form of formal logic structures. These features make SBVR well suited for describing software requirements to implement software models.

Natural Language Semantic Formulation. SBVR is typically proposed for business modeling in NL. However, we are using the formal logic based nature of SBVR to semantically formulate the English software requirements statements. In SBVR 1.0, a collection of semantic formulations (such as atomic formulation, instantiate formulation, logical formulation, quantification, and modal formulation) are proposed to make English statements semantically controlled and restricted.

SBVR Formal Notation. Structured English is one of the possible SBVR notations, given in SBVR 1.0 document, Annex C [4], is applied by prefixing rule keywords in a SBVR rules. The other possible SBVR notation is Rulespeak, given in SBVR 1.0 document, Annex F [4], uses mix-fixing keywords in propositions. SBVR formal notations help in expressing propositions with equivalent semantics that can be captured and formally represented as logical formulations.

3 The SR-Elicitor

This section briefly explains the used approach in ER-Elicitor for transforming English text to SBVR representation. The Figure 1 highlights the used approach:

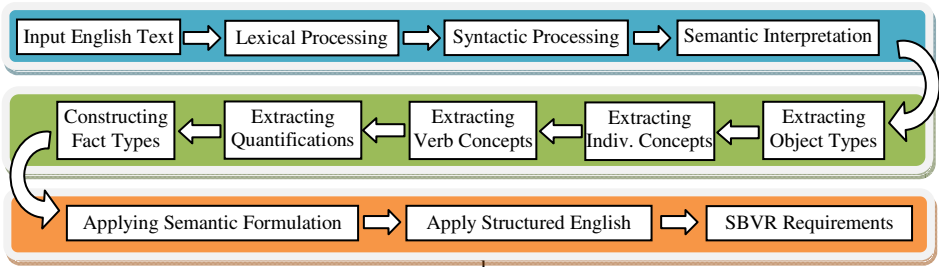


Fig. 1. A Framework for automated transition of English to SBVR requirements

3.1 Parsing NL Software Requirement Text

The first phase of SR-Elicitor is NL parsing that involves a number of processing units (organized in a pipelined architecture) to process complex English statements. The NL parsing phase processes the English text as following:

Lexical Processing. The NL parsing starts with the lexical processing of a plain text file containing English SRS. The lexical phase comprises following four sub-phases:

Tokenization. The lexical processing initiates with the tokenization of the input English text e.g. “A library can issue books to students.” is tokenized as [The] [belt] [conveys] [the] [parts] [towards] [the] [vision] [system] [.]

Sentence Splitting. The tokenized text is further processed to identify the margins of a sentence and each sentence is separately stored in an arraylist.

Parts-of-Speech (POS) Tagging. The tokenized text is further passed to Stanford parts-of- speech (POS) [7] tagger v3.0 to identify the basic POS tags e.g. [The/DT] [belt/NN] [conveys/VBZ] [the/DT] [parts/NNS] [towards/IN] [the/DT] [vision/NN] [system/NN]. The Stanford POS tagger v3.0 can identify 44 POS tags.

Morphological Analysis: After POS tagging, the input text is morphologically processed to separate the suffixes possibly attached to the nouns and verbs [10] e.g. a verb “applies” is analyzed as “convey+s” and a noun “parts” is analyzed as “part+s”.

Syntactic and Semantic Interpretation. We have used an enhanced version of a rule-based bottom-up parser for the syntactic analyze of the input text used in [11]. English grammar rules are base of used parser. The text is syntactically analyzed and a parse tree is generated for further semantic processing. In semantic interpretation phase, role labeling [12] is performed. The desired role labels are actors (nouns used in subject part), co-actor (additional actors conjuncted with ‘and’), action (action verb), thematic object (nouns used in object part), and a beneficiary (nouns used in adverb part) if exists, shown in figure 2. These roles assist in identifying SBVR vocabulary and exported as an xml file.



Fig. 2. Semantic interpretation of English text

3.2 Extracting SBVR Vocabulary

In this phase, the basic SBVR elements e.g. noun concept, individual concept, object type, verb concepts, etc are identified from the English input that is preprocess in the previous phase. The extraction of various SBVR elements is described below:

Extracting Object Types: All common nouns (actors, co-actors, thematic objects, or beneficiaries) are represented as the object types/ general concept [4] (see figure 3) e.g. belt, user, cup, etc. In conceptual modelling, the object types are mapped to classes.

Extracting Individual Concepts: All proper nouns (actors, co-actors, thematic objects, or beneficiaries) are represented as the individual concepts [4] (see figure 3).

Extracting Fact Types: The auxiliary and action verbs are represented as verb concepts [4] (section: 8.1.1) (see figure 3). To constructing a fact types [4] (section: 8.1.1), the combination of an object type/individual concept + verb forms a unary fact

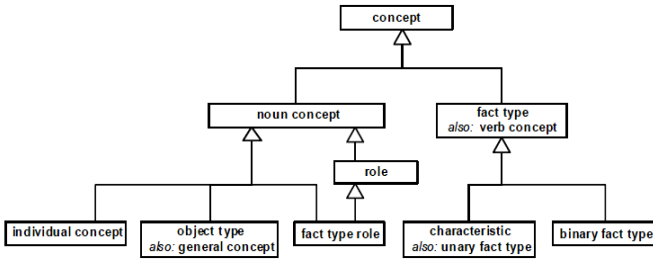


Fig. 3. An extract of the SBVR metamodel: concepts ([4] figure 8.1)

type e.g. “vision system senses”. Similarly, the combination of an object type/individual concept + verb + object type forms a binary fact type e.g. belt conveys part is a binary fact type.

Extracting Characteristics: In English, the characteristic [4] (section: 11.1.2) or attributes are typically represented using *is-property-of* fact type e.g. “name is-property-of customer”. Moreover, the use of possessed nouns (i.e. pre-fixed by *s* or post-fixed by *of*) e.g. student’s age or age of student is also characteristic.

Extracting Quantifications: The key-words such as “Each” or “All” represent SBVR universal quantifications [4] (section: 9.2.6). All indefinite articles (*a* and *an*), plural nouns (prefixed with *s*) and cardinal numbers (2 or two) represent SBVR non-universal quantifications [4] (section: 9.2.6).

Extracting Associative Fact Types: The associative fact types [4] (section 11.1.5.1) are identified by associative or pragmatic relations in English text. In English, the binary fact types are typical examples of associative fact types e.g. “The belt conveys the parts”. In this example, there is a binary association in belt and parts concepts. This association is one-to-many as ‘parts’ concept is plural. In conceptual modeling of SBVR, associative fact types are mapped to associations.

Extracting Partitive Fact Type: The partitive fact types [4] (section 11.1.5.1) are identified by extracting structures such as “*is-part-of*”, “*included-in*” or “*belong-to*” e.g. “The user puts two-kinds-of parts, dish and cup”. Here ‘parts’ is generalized form of ‘dish’ and ‘cup’. In conceptual modeling of SBVR, categorization fact types are mapped to aggregations.

Extracting Categorization Fact Types: The categorization fact types [4] (section 11.1.5.2) are identified by extracting structures such as “*is-category-of*” or “*is-type-of*”, “*is-kind-of*” e.g. “The user puts two-kinds-of parts, dish and cup”. Here ‘parts’ is generalized form of ‘dish’ and ‘cup’. In conceptual modeling of SBVR, categorization fact types are mapped to generalizations. All the extracted information shown in figure 4 is stored in an arraylist for further analysis.

3.3 Generating SBVR Rules

In this phase, a SBVR representation such as SBVR rule is generated from the SBVR vocabulary in previous phase. SBVR rule is generated in three phases as following:

Extracting SBVR Requirements. To generate a rule from an English statement, it is primarily analyzed that it is a structural requirement or a behavioural requirement. Following mapping rules are used to classify a constraint type.

Extracting Structural Requirements: The use of auxiliary verbs such as ‘can’, ‘may’, etc is identified to classify co requirement as a structural requirement. The sentences representing state e.g. “Robby is a robot” or possession e.g. “robot *has* two arms” can be categorized as structural requirements.

Extracting Behavioural Requirements: The use of auxiliary verbs such as ‘should’, ‘must’ are identified to classify requirement as a behavioural rule. Moreover, the use of action verb can be categorized as a behavioural rule e.g. “robot *picks up* parts”.

Applying Semantic Formulation. A set of semantic formulations are applied to each fact type to construct a SBVR rule:

Logical Formulation: A SBVR rule can be composed of multiple fact types using logical operators [4] e.g. AND, OR, NOT, implies, etc. For logical formulation, the tokens ‘not’ or ‘no’ are mapped to negation ($\neg a$). Similarly, the tokens ‘that’ and ‘and’ are mapped to conjunction ($a \wedge b$). The token ‘or’ is mapped to disjunction ($a \vee b$) and the tokens ‘imply’, ‘suggest’, ‘if’, ‘infer’ are mapped to implication ($a \Rightarrow b$).

Quantification: Quantification [4] is used to specify the scope of a concept. Quantifications are applied by mapping tokens like “more than” or “greater than” to at least n quantification; token “less than” is mapped to at most n quantification and token “equal to” or a positive statement is mapped to exactly n quantification.

Modal Formulation: Modal formulation [4] specifies seriousness of a constraint. Modal verbs e.g. ‘can’ and ‘may’ are mapped to possibility formulation to represent a structural requirement and the modal verbs ‘should’, ‘must’ or verb concept “have to” are mapped to obligation formulation to represent a behavioural requirement.

Applying Structured English Notation. The last step in generation of a SBVR is application of the Structured English notation: The object types are underlined e.g. student; the verb concepts are italicized e.g. *should be*; the SBVR keywords are bolded e.g. **at most**; the individual concepts are double underlined e.g. Patron. The characteristics are also italicized but with different colour: e.g. *name*.

4 A Case Study

To demonstrate the potential of our tool SR-Elicitor, a small case study is discussed from the domain of online ordering systems Cafeteria Ordering System (COS): that was online available at: [16]. Following is the problem statement of the case study:

“The system shall let a Patron, who is logged into the Cafeteria Ordering System, place an order for one or more meals. The system shall confirm that the Patron is registered for payroll deduction to place an order. If the Patron is not registered for payroll deduction, the system shall give the Patron options to register now and continue placing an order, to place an order for pickup in the cafeteria, or to exit from the COS. The system shall prompt the Patron for the meal date. If the meal date is the current date and the current time is after the order cutoff time, the system shall inform the patron that it’s too late to place an order for today. The Patron may either change the meal date or cancel the order. The Patron shall specify whether the order is to be picked up or delivered. If the order is to be delivered and there are still available delivery times for the meal date, the Patron shall provide a valid delivery location.”

The problem statement of the case study was given as input (NL specification) to the SR-Elicitor tool. The tool parses and semantically interprets English text and extracts the SBVR vocabulary from the case study as shown in table 1:

Table 1. SBVR vocabulary generated from English text

<i>Category</i>	<i>Count</i>	<i>Details</i>
Object Types	05	system, order, payroll, date, time
Verb Concepts	14	let, log, place, confirm, register; pick_up, exit, inform, change, cancel, specify, pick, deliver, provide
Individual Concepts	03	Cafeteria_Ordering_System, Patron, COS
Characteristics	04	meal_date, cutoff_time, delivery_time, delivery_location
Quantifications	08	Universal (01), At least n (07)
Unary Fact Types	05	Patron <i>registers</i> , , Patron change, order <i>picked</i> , order <i>delivered</i> , Patron provide
Associative Fact Types	08	Patron <i>logged</i> into Cafeteria_Ordering_System, Patron <i>place</i> order, system <i>confirm</i> Patron, Patron <i>registered</i> for payroll, Patron <i>pickup</i> order, system <i>prompt</i> Patron, System <i>inform</i> Patron, Patron <i>exit</i> from COS, Patron <i>cancel</i> order, Patron <i>specify</i> order,
Partitive fact Types	00	
Categorization Fact Types	00	

Here, Cafeteria_Ordering_System and COS are synonyms of each other but not picked but our system and these are specified as separate individual concepts. One object type has not been picked that cafeteria. Moreover, current date and current time are characteristics but they are picked as object types. In the used case study’s problem statement, there were seven requirements as shown in table 2:

Table 2. SBVR Rule representation of software requirements

<i>Details</i>
It is obligatory that the <u>system shall let</u> , each <u>Patron</u> who <i>is logged</i> into the <u>Cafeteria Ordering System</u> , <u>place at least one order</u> for <u>at least one</u> or more <u>meals</u> .
It is obligatory that the <u>system shall confirm</u> that the ' <u>Patron</u> ' <i>is registered</i> for <u>payroll</u> deduction to <u>place at least one order</u> .
If the <u>Patron</u> <i>is not registered</i> for <u>payroll</u> deduction, It is obligatory that the <u>system shall give</u> the <u>Patron</u> options to <u>register</u> and <u>continue placing at least one order</u> , to <u>place at least one order</u> for <u>pickup</u> in the cafeteria, or to <u>exit</u> from the <u>COS</u> .
It is obligatory the <u>system shall prompt</u> the <u>Patron</u> for the <u>meal date</u> .
If the <u>meal date</u> <i>is</i> the current <u>date</u> and the current <u>time is</u> after the <u>order cutoff time</u> , it is obligatory that the <u>system shall inform</u> the <u>Patron</u> that it's too late to <u>place at least one order</u> for today.
It is possibility that the <u>Patron may change</u> the <u>meal date</u> or <u>cancel</u> the <u>order</u> .
It is obligatory the <u>Patron shall specify</u> whether the <u>order is to be picked or delivered</u> .
If the <u>order is to be delivered</u> and there still <i>are</i> available <u>delivery times</u> for the <u>meal date</u> , it is obligatory that the <u>Patron shall provide at least one valid delivery location</u> .

5 Evaluation

We have done performance evaluation to evaluate that how accurately the English specification of the software requirements has been translated into the SBVR based controlled representation by our tool SR-Elicitor.

There were seven sentences in the used case study problem. The largest sentence was composed of 39 words and the smallest sentence contained 10 words. The average length of all sentences is 24. The major reason to select this case study was to test our tool with the complex examples. The correct, incorrect, and missing SBVR elements are shown in table 2.

Table 3. Results of NL to SBVR Translation by SR-Elicitor

#	Type/Metrics	N_{sample}	$N_{correct}$	$N_{incorrect}$	$N_{missing}$
1	Object Types	05	3	2	1
2	Verb Concepts	14	14	0	0
3	Individual Concepts	02	2	1	0
4	Characteristics	06	4	0	2
5	Quantifications	08	8	0	0
6	Unary Fact Types	05	5	0	0
7	Associative Fact Types	08	8	0	0
8	Partitive fact Types	00	0	0	0
9	Categorization Fact Types	00	0	0	0
	Total	48	44	3	3

In table 3, the average recall for SBVR software requirement specification is calculated 91.66% while average precision is calculated 93.61%. Considering the lengthy input English sentences including complex linguistic structures, the results of this initial performance evaluation are very encouraging and support both the approach adopted in this paper and the potential of this technology in general.

Table 3. Recall and Precision of SR-Elicitor for NL software requirements

<i>Type/Metrics</i>	N_{sample}	$N_{correct}$	$N_{incorrect}$	$N_{missing}$	<i>Rec%</i>	<i>Prec%</i>
Software Requirements	48	44	3	3	91.66	93.61

6 Related Work

A few controlled natural language (CNL) representations are introduced in last two decades such as Attempto Controlled English (ACE) [8], Processable English (PENG) [9], computer Processable Language (CPL) [10], Formalized-English (Martin, 2002) [11], etc. All above languages are human-oriented CNLs [12], while a machine-oriented CNL [13] can be more helpful in modern software modelling practices. Furthermore, the available CNLs are general purpose and not specifically designed for natural language based software requirement specifications.

An automated approach was presented in [16] to generate SBVR representation from English language description. However, English is difficult to machine process and translate to formal languages [15], [17]. SBVR based controlled natural language is not a brand new proposal as it has been previously presented and implemented in a tool RuleXpress [5] but it is specifically designed for business people to express and communicate business rules. The related work shows that currently there is no approach and tool available that can automatically translate natural language software requirements to a CNL representation such as SBVR.

7 Conclusion and Future Work

The primary objective of the paper was to automate the process of software requirement elicitation and specification by overcoming ambiguous nature of natural languages (such as English) and generating a controlled representation. To address this challenge we have present a NL based too SR-Elicitor that is based on an automated approach to parse English software requirement specifications and generated a controlled representation using SBVR. The output of out tool can be used for automated object oriented analysis and design from natural language software requirements. Additionally, our SR-Elicitor provides a higher accuracy as compared to other available NL-based tools.

The future work is to extract the object-oriented information from SBVR specification of software requirements such as classes, instances and their respective attributes, operations, associations, aggregations, and generalizations.

References

- [1] Denger, C., Berry, D.M., Kamsties, E.: Higher Quality Requirements Specifications through Natural Language Patterns. In: Proceedings of IEEE International Conference on Software-Science, Technology & Engineering (SWSTE 2003), pp. 80–85 (2003)
- [2] Ormandjieva, O., Hussain, I., Kosseim, L.: Toward A Text Classification System for the Quality Assessment of Software Requirements written in Natural Language. In: 4th International Workshop on Software Quality Assurance (SOQUA 2007), pp. 39–45 (2007)
- [3] Tobias, K.: Controlled English for Knowledge Representation. Doctoral Thesis. Faculty of Economics, Business Administration and Information Technology of the University of Zurich (2010)
- [4] OMG. Semantics of Business vocabulary and Rules. (SBVR) Standard v.1.0. Object Management Group (2008), <http://www.omg.org/spec/SBVR/1.0/> To insert individual citation into a bibliography in a word-processor, select your preferred citation style below and drag-and-drop it into the document
- [5] Spreuwenberg, S., Healy, K.A.: SBVR's Approach to Controlled Natural Language. In: Fuchs, N.E. (ed.) CNL 2009. LNCS, vol. 5972, pp. 155–169. Springer, Heidelberg (2010)
- [6] Ilieva, M.G., Ormandjieva, O.: Automatic transition of natural language software requirements specification into formal presentation. In: Montoyo, A., Muñoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 392–397. Springer, Heidelberg (2005)
- [7] Toutanova, K., Manning, C.D.: Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In: The Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Hong Kong, pp. 63–70 (2000)
- [8] Fuchs, N.E., Kaljurand, K., Kuhn, T.: Attempto Controlled English for Knowledge Representation. In: Baroglio, C., Bonatti, P.A., Małuszyński, J., Marchiori, M., Polleres, A., Schaffert, S. (eds.) Reasoning Web. LNCS, vol. 5224, pp. 104–124. Springer, Heidelberg (2008)
- [9] White, C., Rolf, S.: An Update on PENG Light. In: Proceedings of ALTA 2009, pp. 80–88 (2009)
- [10] Clark, P., Murray, W.R., Harrison, P., Thompson, J.: Naturalness vs. Predictability: A key debate in controlled languages. In: Fuchs, N.E. (ed.) CNL 2009. LNCS, vol. 5972, pp. 65–81. Springer, Heidelberg (2010)
- [11] Martin, P.: Knowledge representation in CGLF, CGIF, KIF, frame-CG and formalized-english. In: Priss, U., Corbett, D.R., Angelova, G. (eds.) ICCS 2002. LNCS (LNAI), vol. 2393, pp. 77–91. Springer, Heidelberg (2002)
- [12] Schwitter, R.: Controlled Natural Languages for Knowledge Representation. In: Coling 2010. Poster vol., Beijing, pp. 1113–1121 (August 2010)
- [13] Huijsen, W.O.: Controlled Language –An Introduction. In: Proceedings of CLAW 1998, pp. 1–15 (1998)
- [14] Hirschman, L., Thompson, H.S.: Chapter 13 evaluation: Overview of evaluation in speech and natural language processing. In: Survey of the State of the Art in Human Language Technology (1995)
- [15] Bajwa, I.S., Samad, A., Mumtaz, S.: Object Oriented Software modeling Using NLP based Knowledge Extraction. European Journal of Scientific Research 35(01), 22–33 (2009)
- [16] Bajwa, I.S., Lee, Mark, G., Behzad, B.: SBVR Business Rules Generation from Natural Language Specification. In: AAAI Spring Symposium 2011, San Francisco, USA, pp. 2–8 (2011)
- [17] Bajwa, I.S., Choudhary, M.A.: A Rule Based System for Speech Language Context Understanding. Journal of Donghua University 23(6), 39–42 (2006)

Automatic Interface Generation between Incompatible Intellectual Properties (IPs) from UML Models

Fateh Boutekkouk, Zakaria Tolba, and Mustapha Okab

Department of Computer Science, University of Larbi Ben M'hedi,
Route de Constantine, BP 358, Oum El Bouaghi, 04000, Algeria
{fateh_boutekkouk, tolba, okab}@yahoo.fr

Abstract. This paper presents an UML based tool for incompatible hardware Intellectual properties (IPs) integration. Our aim is to provide Systems On Chips (SOCs) designers with a UML based environment for modeling incompatible IPs, automatic generation of interface between IPs, and functional simulation. In our case, each IP is modeled as an UML component with a well defined interface including input and output signals and some attributes. The whole SOC is modeled via UML structure diagram. Memory timing constraints are modeled via UML timing diagrams. Communication protocols for incompatible IPs are modeled via UML Statecharts with hierarchic and concurrent states. From these diagrams, a Finite State Machine with Data path (FSMD) for interface is generated automatically. Functional simulation of the interface is performed by translating the result FSMD to a VHDL code.

Keywords: SOC, IP, UML, FSMD, VHDL.

1 Introduction

Nowadays, System On Chip (SOC) [1] design is becoming more complex and may lead to the non satisfactory of customers requirements and the time to market constraints. To cope with this problem, it seems that Core Based Design (CBD) brings a significant improvement of design in general and to decrease the time to market window in particular [2]. The main idea behind the CBD is to reutilize existing hardware and/or software components with some customization and adaptation.

In the SOC field, designers have considered the reuse of complex hardware and software components (Intellectual Property or simply IP components), already used and tested in previous designs can dramatically decrease the design time [2].

Reuse is essential to master the complexity of SOC design; however it does not come for free. Since most IPs are provided by different vendors, and they have different interface schemes, and different data rates, combining these components is an error-prone task and the most important part of system integration. Designers have to find and evaluate IPs that fit particular needs and the selected IPs must be integrated together to implement the desired SOC functionality. This integration may require some adaptation and customization. The basic goal of an interface synthesis is

to generate interfaces between incompatible components. Data could be transferred at different bit width, operating frequency, data rates. For this reason, researchers in both academia and industry [3, 4, 5, 6, 7, 8, 9] have developed many algorithms and CAD tools to explore, to optimize, and to generate possible interfaces between incompatible IPs. Unfortunately, most of these efforts target notations, models and languages at lower levels of abstractions. Another problem is the fact that software engineers are not very familiar with hardware notations and languages. As a consequence the task of interface generation is usually hard especially for software designers. In this context, we have developed an UML 2.0 [10, 11] tool that permits to both software and hardware SOC designers to model, configure, and link the incompatible IPs graphically. From UML diagrams, an FSMD (Finite State Machine with Data path) modeling the interface is generated automatically. Our work tries to take advantages of the UML 2.0 standard for IPs modeling and interface generation with minimal user inputs exploiting the algorithm proposed in [8]. In its basic form, this algorithm was used to generate the glue logic between two incompatible IPs. Since the system may contain many incompatible IPs, we have to apply the same algorithm for each pair of communicating incompatible IPs.

The rest of this paper is organized as follows: section two is dedicated to related works concerning the synthesis of interface for incompatible protocols. Section three puts light on IPs and their classes. The algorithm of synthesis we have adopted is detailed in section four. Section five discusses our tool with an illustrative example before concluding.

2 Related Work

Here, we try to mention some pertinent works targeting interface generation. In [4], signal transition graph was introduced for protocol specification and the hardware interface is synthesized with asynchronous logic. In [5], the protocol specification is decomposed into five basic operations (data read/write, control read/write, time delay), while the protocol is represented as an ordered set of relations whose execution is guarded by a condition or by a time delay. In [6], the two protocols are described using regular expressions and are translated into corresponding deterministic finite automata. Then interface protocol can be synthesized as an FSM by production computation algorithm. In [7, 8], a novel queue-based interface scheme which is general enough to accommodate any component protocols was proposed. In order to implement the queue-based interface architecture, a canonical model of a queue which can contain various memories was defined on the basis of timing constraints of various memory organizations. An algorithm which generates FSMD model for queue from timing specification of the given memory was developed. From given protocol specifications and clock period of the selected queue, the interface synthesis algorithm generates the FSMD for interface including the queue FSMD.

The main limitation of these approaches is that IPs communication protocols are expressed in low level models and/or programming languages such as waveforms,

VHDL or C language. Another tendency to address the problem of IPs integration is the use of standards for promoting reuse in the design process. Several standards have been proposed. Among these, the Open Core Protocol (OCP) by OCP-IP [9] has gained wide industrial acceptance. However, for existing non OCP compliant IP cores, it is very expensive to customize them to comply with the OCP standard.

3 Intellectual Properties (IPs)

An Intellectual Property or a virtual core (IP) [1, 2] is a reusable software or hardware pre-designed block and maybe delivered by third party companies. Hardware IP components may come in several forms: hard, firm or soft. An IP is *hard*, when all its gates and interconnects are placed and routed. It has the advantage of more predictable estimations of performance, power, and area considering the target technology. But, it is less flexible and therefore less reusable. An IP can be *soft*, with only an RTL (Register Transfer Level) representation. It is available in source code and therefore adaptable to different platforms at the price of less predictable estimations on performance and area. An IP can be *firm*, with an RTL description together with some physical floor planning or placement.

4 Interface Generation Algorithm

In this section, we try to detail the synthesis algorithm for interface between incompatible protocols as proposed in [8]. In [7], the authors proposed a novel queue-based interface scheme, which is general enough to accommodate any component protocols. The interface architecture is basically composed of synchronous system interfaces as shown in Figure 1. The system components (PE1 and PE2) may operate at different frequencies and at different data rates. The interface architecture includes a buffer (FIFO queue) to smoothen the burst data transfer requests and two FSMDs (Finite State Machine with Data path) to queue and un-queue data. In the interface architecture, system components (PE1 and PE2) in Figure 1 are directly connected to its corresponding state machines and will transfer data to other component through the state machines. The state machines are responsible for receiving (sending) data from (to) the corresponding system components and writing (reading) the data to (from) the queues. The interface protocol between state machines and queues will be fixed because the queue interface is predefined. But the interface protocol between system components and state machines will be varied depending on the protocol of system components. A queue can have one or two I/O ports. The queue is implemented with a memory to store large amount of data. The clock period of the queue is frequently less than the memory read access time.

Generally, a queue contains memory to store data internally. The operation of the queue is determined by memory organization and timing [7].

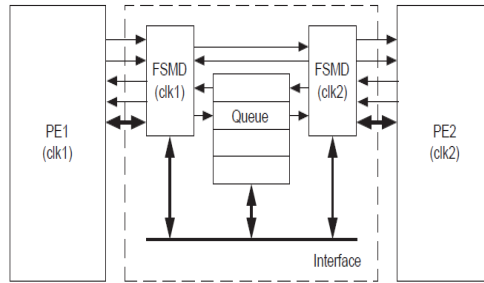


Fig. 1. The interface architecture of two incompatible Components [7]

4.1 Interface Generation Algorithm

Problem definition

Given:

1. Protocol descriptions of two communicating parties (producer and consumer).
2. Bit width and size for the selected memory.
3. Clock period $TQclk$ of the queue.

Determine:

1. FSMDs for state machines.
2. FSMD for the queue.

Conditions: Timing constraints are met.

Algorithm of the figure 2 shows the interface synthesis algorithm from given protocol specifications and clock period of the selected queue. We have applied the same algorithm as proposed in [8] but with two major modifications: firstly, the scheduling of actions over states is achieved by the designer thus the *Schedule()* function is removed from the algorithm. Secondly, the *MakeDual()* function transforms the Statechart (instead of the Protocol sequence graph) of original protocol specification to the corresponding dual Statechart, which can be done by replacing the operators in actions with their duals. The method *GenerateQueue()* will generate the queue based on the selected memory and the clock of the queue (see the generate Queue algorithm). The generated producer interface FSMD, consumer interface FSMD and queue interface FSMD should be collapsed into a single FSMD to obtain interface FSMD. The method *AddFSMD()* will collapse the producer and queue interface FSMDs ($FMSDS_i$, $FSMDQ_i$) into the transducer interface FSMD for the producer ($FSMDT_S$). In the same way, the consumer and queue interface FSMD ($FSMDR_i$, $FSMDQ_i$) will collapse into the transducer interface FSMD for the consumer. Finally we have two FSMDs for transducer: the producer interface FSMD and the consumer interface FSMD in the transducer. For more details on this algorithm, one can refer to [8]. In order to simulate the interface between incompatible IPs, we have to translate its FSMD to a HDL like VHDL [13] or a Software Language. In our case, we choose VHDL. By translating the interface

FSMD to VHDL, we can easily perform simulation and/or synthesis using existing commercial tools such as *ModelSim* simulator [14]. The communication protocols actions are expressed in the C language. The interface FSMD is implemented as a VHDL entity with an architecture comprising of one or more than a process that calls some procedures and functions implementing the queue read and write operations. C actions are also translated to VHDL. UML required and provided interfaces are translated to VHDL input and output ports respectively. Concurrent and hierarchical states are transformed to VHDL processes and procedures respectively.

5 Our Tool

Our tool generates automatically the interface FSMD including the Queue FSMD for each pair of communicating IPs. From Memory Timing diagram, some temporal parameters are extracted to be used for FSMD queue generation. In our case, each IP is modeled via UML 2.0 components with required (input) and provides (output) signals. Furthermore, each IP is parameterized by some parameters such as the HDL (e.g. VHDL, Verilog), the clock period, and the abstraction level of each IP. Regardless of the IP HDL, we assume that all IPs are in RTL level and all IPs communication protocols are modeled via UML Statecharts. The communication protocols actions are expressed in the C language. As an example of application, we have chosen three IPs that are: *ColdFire* processor, *ARM9TDMI* processor, and *TMS320C50* DSP processor [8]. The objective is the generation of interface between these three cores with incompatible communication protocols. Figure 3 shows IPs modeling with their connections and queue configuration between two incompatible IPs. Queue configuration consists in horologe period selection and I/O ports number. Figure 4 shows memory write and read operations timing constraints modeling using timing diagrams. From these diagrams, a set of timing constraints are extracted to generate the queue model. Figure 5 shows the queue FSMD which is generated automatically. Here two types of generic hierarchic states are created: read states (Rxx states) and write states (Wxx states). Figure 6 shows the interface FSMD. The latter is generated automatically and it includes Queue states (Qxx).

```

Algorithm GenerateInterface (  $PSG_s$  ,  $PSG_r$  ,  $T_{Qc1k}$  )
 $PSG_Q$  = Generate_Queue( $T_{Qc1k}$ ); //generate Queue FSMD
 $PSG_{s1}$  = Make_Dual( $PSG_s$ ); //generate the dual of producer
 $PSG_{r1}$  = Make_Dual( $PSG_r$ ); //generate the dual of consumer
 $PSG_{Q1}$  = Make_Dual( $PSG_Q$ ); // generate the dual of queue
For  $i = 1$  to (  $bw_Q$  /  $bw_s$  ) do
    Add_FSMD( $FSMD_{TS}$ ,  $FSMD_{s1}$ ); // add the producer FSMD to interface
    FSMD
End for
Add_FSMD( $FSMD_{TS}$ ,  $FSMD_{Q1}$ ); //add the queue FSMD to producer interface
FSMD
Add_FSMD( $FSMD_{TR}$ ,  $FSMD_{Q1}$ ); //add the queue FSMD to consumer interface
FSMD
For  $i = 1$  to (  $bw_r$  /  $bw_Q$  ) do
    Add_FSMD( $FSMD_{TR}$ ,  $FSMD_{r1}$ ); // add the consumer FSMD to interface
    FSMD

```

Fig. 2. Interface generation algorithm [8]

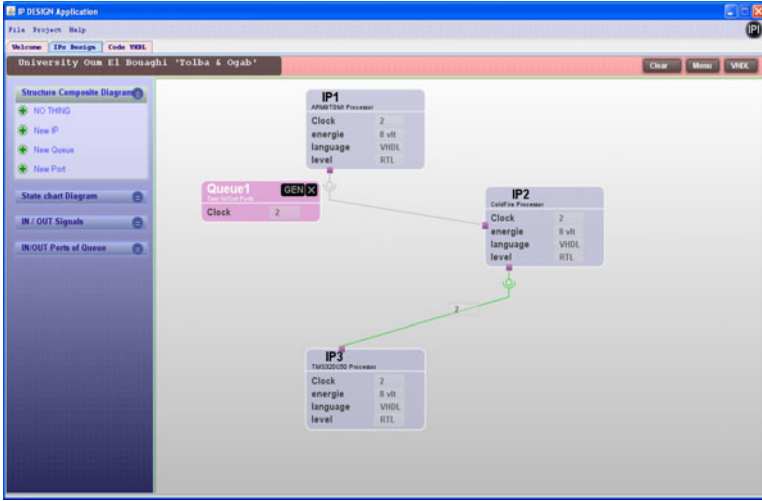


Fig. 3. IPs modeling and Queue configuration between IPs



Fig. 4. Memory write and read operations modeling using Timing diagrams

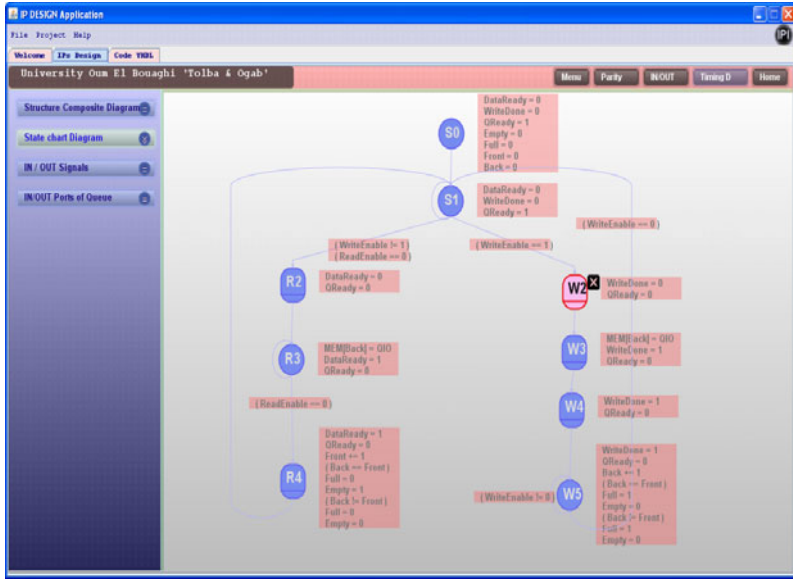


Fig. 5. Queue FSMD generation



Fig. 6. FSMD of the interface

6 Conclusion and Future Work

In this paper, we present our tool for incompatible hardware IPs modeling and interface generation between them using UML 2.0 diagrams. We have exploited UML 2.0 diagrams such as structure diagram for IPs modeling and their connections, timing diagram to model memory read and write operations timing constraints, and Statecharts with hierarchic and concurrent states to model IPs communication protocols and Queue behavior. From the result interface FSM, designers can analyze the whole SOC performance and generate a VHDL code for functional and timing simulations purposes. As perspectives, we plan to integrate a VHDL simulator into our tool and perform functional verification of the generated interface.

References

1. Jerraya, A.A., Wolf, W.: Multiprocessor systems on chip. Morgan Kaufmann publishers, San Francisco (2005)
2. Wagner, F.R., Cesario, W.O., Carro, L., Jerraya, A.A.: Strategies for integration of hardware and software IP components in embedded systems on chip. *VLSI Journal* (2004)
3. Borriello, G., Katz, R.: Synthesis and optimization of interface transducer logic. In: *Proceedings of the International Conference on Computer-Aided Design*, pp. 274–277 (November 1987)
4. Lin, B., Vercauteren, S.: Synthesis of concurrent system interface modules with automatic protocol conversion generation. In: *Proceedings of the International Conference on Computer-Aided Design*, pp. 101–108 (November 1994)
5. Narayan, S., Gajski, D.: Interfacing incompatible protocols using interface process generation. In: *Proceedings of the Design Automation Conference*, pp. 468–473 (June 1995 November 1994)
6. Passerone, R., Rowson, J.A., Sangiovanni-Vincentelli, A.: Automatic synthesis of interfaces between incompatible protocols. In: *Proceedings of the Design Automation Conference*, pp. 8–13 (June 1998)
7. Shin, D., Gajski, D.: Queue Generation Algorithm for Interface Synthesis. Technical Report ICS-TR-02-03, University of California, Irvine (February 2002)
8. Shin, D., Gajski, D.: Interface synthesis from protocol specification. Technical Report CECS-02-13, University of California, Irvine (April 12, 2002)
9. OCP-IP, <http://www.ocp-ip.org>
10. Booch, G., Rumbaugh, J., Jacobson, I.: *Unified Modeling Language User Guide*. Addison-Wesley, Reading (1999)
11. Schattkowsky, T.: UML2.0 Overview and Perspectives in SOC Design. In: *Proceedings of the Design, Automation and Test in Europe (DATE 2005)*, vol. 2 (2005)
12. Gajski, D., Vahid, F., Narayan, S., Gong, J.: *Specification and Design of Embedded Systems*, p. 07632. Prentice Hall, Englewood (1994)
13. IEEE Standard VHDL Language Reference Manual. IEEE, IEEE Std 1076 (2000)
14. ModelSim documentation, <ftp://ftp.xilinx.com/pub/documentation>

Deadlock Prevention in Distributed Object Oriented Systems

V. Geetha¹ and N. Sreenath²

¹ Dept. of Information Technology

² Dept. of Computer Science & Engg

Pondicherry Engineering College

Puducherry – 605014

vgeetha@pec.edu, nsreenath@pec.edu

Abstract. This paper proposes a deadlock prevention algorithm for Distributed Object Oriented Systems (DOOS) based on the popular resource ordering technique. In distributed object oriented system, objects are the resources requested by the transactions. Though resource-ordering technique is not new, novelty of the proposed deadlock prevention algorithm lies in exploiting the relationships among objects present in the domain to do the resource ordering. In this paper a resource ordering technique based on semantics of the object relationships like inheritance, aggregation and association is proposed. A formal model of the resource ordering technique is defined using predicate calculus.

Keywords: DOOS, resource ordering, objects, class relationships.

1 Introduction

Distributed System (DS) is a collection of sites that allows several transactions to execute parallelly. Concurrency Control Techniques (CCT) are applied to maintain consistency of data resources. However, the CCT have the negative effect of resulting in deadlocks. Deadlock prevention is one of the proactive techniques to handle deadlocks. Prevention of deadlocks has the benefit of low runtime cost and better response time. Hac et al. [1] have shown that deadlock prevention algorithms are better than detection algorithms with better performance and response time in distributed systems.

In DOOS, objects play a major role. They have state and behavior. State is defined by the values of attributes or data members defined in the class. Behavior of the class is defined by the methods or member functions. Data from reusable data resources like databases are copied into the object state and member functions operate on them. Classes define the template of objects. Complex relationships among objects are defined using inheritance, aggregation and association relationships. The class diagram describes the objects and their relationships in the domain.

In DOOS, a transaction is made through interfaces. An interface may typically contain one or more operations. The operations are defined and implemented in the

member functions of classes. Hence, transaction is actually a series of calls to member functions of objects.

Riehle and Berzuk [9] [10] have stated that member functions have types and properties. The member functions are classified into (1) *Query method* – for reading the object state (2) *Mutation method* – for modifying the state and (3) *Helper method* – for supporting tasks. Properties specify whether the member function is *primitive* or *composed*, and whether *instance* (object) method or *class* method. Before proposing a resource ordering technique in DOOS, it is worth noting certain points:

1. Garza et al. [13] states that when class level member functions are called, instead of setting individual locks on all objects, a single lock on its class may be set to minimize the lock escalation.
2. When a sub class object is requested, then both sub class object and its corresponding base class object that are mapping to the same record in a database table must also be locked to maintain consistency. However base class objects are independent objects (IO) as they can be accessed on their own. Sub class objects are dependent objects (DO) as they depend on base class objects for consistency as in fig 1. This is applicable for aggregation and association also.

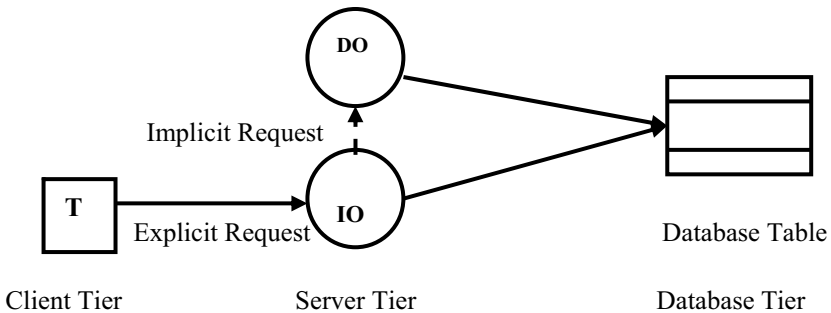


Fig. 1. Locking Dependent Objects along with Independent Objects to maintain Consistency

2 Related Works

Several deadlock prevention algorithms have been defined for DS and DOOS. Reddy and Bhalla [3] have proposed DPA for distributed database systems which eliminate the deadlock by giving higher priority to active transactions. However, it does not consider the case where conflicting transactions require multiple resources and latter transactions already have more resources than earlier transactions. Davidson et al. [4] have proposed AND-OR DPA for concurrent real time systems using resource ordering technique. In this DPA, the interdependency or relationship of data resources among themselves is not addressed. Lewis [7] has proposed DPA for multi threaded environment. Here also, deadlock is prevented by access ordering and effect of mutual dependency of resources is not addressed. Anand and Sethi [8] have proposed DPA for distributed environment by preempting lower priority transaction. Cummins [6] recursively checks for presence of cycle of any size. It does not exploit the

structure of object oriented system defined using class diagrams and does not utilize the object relationships to know the required resources a priori. Hence, our objectives are to propose resource-ordering policy for objects and deadlock prevention algorithm using the proposed resource ordering technique for objects in DOOS. Though resource-ordering technique is not a novel technique, same resource ordering policy cannot be adapted for all systems. It depends on the characteristics of the resources under consideration. [1, 4, 5] are some of the resource-ordering policies adapted for various systems. The novelty in the proposed technique lies in defining the resource ordering policy by exploiting the dependency among objects participating in the domain. The dependency of objects with other objects can be inferred by their relationship with other objects. Thus the proposed scheme prevents deadlock using resource ordering.

The paper is organized as follows. Chapter 3 describes the proposed system with resource ordering technique, formal model of resource ordering and deadlock prevention algorithm. Chapter 4 gives formal proof of our algorithm. Chapter 5 concludes the paper.

3 Proposed Scheme

In DOOS, the reusable data sources i.e. databases are mapped onto the objects. So concurrency control is applied on objects and hence live locks and deadlocks might occur. DOOS follows AND request model, as transactions need all the resources before execution. Then it is ideal to assign smaller resource IDs to independent resources and higher resource IDs to resources dependent on these independent resources. The proposed resource ordering technique is based on the dependency among the objects participating in the system. In DOOS, the dependency among objects is based on their relationship with other objects. Then dependency can be categorized based on

1. Relationship between class and its objects
2. Relationship between objects namely inheritance, aggregation and association.

Based on the constraints in section 1, we will do partial ordering based on the dependencies above and then propose total ordering of all the objects in a system.

3.1 Resource Ordering Technique

1. Partial ordering on a class and its objects:

The transactions can request for resources in two granularities namely one object or all the objects in a class. Simultaneous requests to both the cases are not allowed to maintain consistency. In this situation, two possibilities exist:

Case 1: Objects are assigned lower resource IDs than their class.

Case 2: Classes are assigned lower resource IDs than its objects.

For example, Let A be a class. Let a1, a2 and a3 be its objects.

Let T1 requests A. (i.e. all objects in the class. Hence, the class itself is locked to minimize number of locks)

Let T1 requests {a1, a2, a3}; T2 requests {a1}; T3 requests {a3}.

If case 1 is considered, both T2 and T3 are executed first and T1 is executed afterwards. If case 2 is considered, both T2 and T3 are blocked, while T1 is executed first. Here case 1 improves concurrency and hence improves the throughput of the system. Hence, case 1 is better than case 2. Then partial ordering on objects and classes can be defined as follows.

Rule 1: For all objects O belonging to class C, resource IDs of objects O should be less than resource ID of their class C.

(Note: However since transactions will request either only any one or all objects, ordering among the objects of a class is not necessary)

2. Partial ordering between objects related by inheritance, aggregation and association:

2.1 Partial ordering on base class and its inherited sub classes:

Here also resource IDs can be assigned in two ways:

Case 1: Base Classes are assigned lower resource IDs than their subclasses.

Case 2: Sub classes are assigned lower resource IDs than their base classes.

In inheritance, attributes of base class are included as attributes of subclass.

Example 2:

Let A and B be base classes. Let a1, a2 be instances of A and b1, b2 be instances of B. Let C be a subclass inherited from A and B. Let c1, c2 be instances of C. Let a1 and b1 be associated base class objects for sub class object c1. Similarly, a2 and b2 be associated base class objects for c2.

Let T1 requests {A} = {a1, a2}; T2 requests {B} = {b1, b2}; T3 requests {C}; Then resource set for T3 = {A, B, C} = {a1, a2, b1, b2, c1, c2}.

If case 1 is considered, then T1 and T2 get more priority than T3. Then concurrency is improved in case 1 than case 2.

Rule 2: For all base classes BC, if a sub class SC is inherited from BC, then resource ID of base class BC should be less than the resource ID of subclass SC.

(Note: Since a transaction will request only one of the sub classes at a time, the ordering among sub classes of a parent class is not necessary.)

2.2 Partial ordering in aggregation and association:

From previous example it is obvious that rule 2 can be extended to aggregation and association also. As proof is trivial it is ignored here.

Rule 3: For all composite classes CM, if a component class CC is a part of CM, then resource ID of component class CC should be less than the resource ID of composite class CM.

Rule 4: For all associative class (TC), if an associative class (TC) is associated with associated class (AC), then resource ID of associative class TC should be less than associated class AC.

3. Total resource ordering using 1, 2.1, and 2.2:

In a business domain, objects may have complex relationships by having combination of above relationships. By combining the partial ordering rules proposed earlier, the total ordering of resources can be defined as follows:

Case 1: When a class has inheritance relationship:

Here the ordering needs to be done on base class, its objects, sub class and its objects. This can be done by combining rule 1 and 2 in the previous section.

Rule 5: For all base class objects BCO belonging to BC and all sub class objects SCO belonging to SC, if a sub class SC is inherited from base class BC, then Resource ID (BCO) \prec Resource ID (SCO) \prec Resource ID(BC) \prec Resource ID(SC).

Case 2: When a class has aggregation relationship:

Component objects and composite objects are given lower IDs and their classes are given higher IDs by combining rules 1 and 3.

Rule 6: For all component class objects CCO belonging to CC and all composite class objects CMO belonging to CM, if a component class CC is a part of composite class CM, then Resource ID (CCO) \prec Resource ID (CMO) \prec Resource ID (CM) \prec Resource ID (CM).

Case 3: When a class has association relationship:

Associative objects and associated objects are given lower IDs and their classes are given higher ID by combining rules 1 and 3.

Rule 7: For all associative class objects TCO belonging to TC and all associated class objects ACO belonging to associated class AC, if an associative class TC is associated with associated class AC, then Resource ID (ACO) \prec Resource ID (TCO) \prec Resource ID (AC) \prec Resource ID (TC).

Case 4: When a class has inheritance, association and aggregation relationships:

A class can have all the relationships in any combination. The class diagram representing the classes is partitioned horizontally by various levels. In a class diagram, as the level increases, dependency increases and concurrency decreases. Concurrency decreases because more resources are required for a transaction requesting high-level object or class to maintain consistency.

Rule 8: For all objects OA belonging to A and for all objects OB belonging to B, for all classes A in level i and for all classes B in level j of the class diagram where $i < j$, then Resource ID (OA) \prec Resource ID (OB) \prec Resource ID (A) \prec Resource ID (B).

3.2 Formal Model for Resource Ordering Using Predicate Calculus

Let class diagram representing the domain be the Universal set U . U is a collection of classes C representing the domain. Let O be the collection of objects instantiated from the classes C . Then U can be represented as $U(C(O))$. The classes are related to each other by inheritance, aggregation and association relationships. Let $Rid(X)$ be a function that returns resource id for a resource X . Let Inherit-from (SC, BC) be a predicate that means SC is inherited from BC . Let Part-of (CC, CM) be a predicate that means CC is a part of CM . Let Associated-with (TC, AC) be a predicate that means TC is associated with AC . Let Level (Y_i) be a predicate that means class Y is in level i . Let $LL(i, j)$ be a predicate that means level i is less than level j . Then rules 1- 8 can be written in predicate calculus as follows:

Rule 1: $(O)(C)(O \in C \Rightarrow Rid(O) \prec Rid(C))$.

Rule 2: $(BC)(SC)(\text{Inherit-from}(SC, BC) \Rightarrow Rid(BC) \prec Rid(SC))$.

Rule 3: $(CC)(CM)(\text{Part-of}(CC, CM) \Rightarrow Rid(CC) \prec Rid(CM))$.

Rule 4: $(TC)(AC)(\text{Associated-with}(TC, AC) \Rightarrow Rid(AC) \prec Rid(TC))$.

Rule 5: $(BCO)(BC)(SCO)(SC)((BCO \in BC) \wedge (SCO \in SC) \wedge \text{Inherit-from}(SC, BC) \Rightarrow Rid(BCO) \prec Rid(SCO) \prec Rid(BC) \prec Rid(SC))$.

Rule 6: $(CCO)(CC)(CMO)(CM)((CCO \in CC) \wedge (CMO \in CM) \wedge \text{Part-of}(CC, CM) \Rightarrow Rid(CCO) \prec Rid(CMO) \prec Rid(CC) \prec Rid(CM))$.

Rule 7: $(TCO)(TC)(ACO)(AC)((TCO \in TC) \wedge (ACO \in AC) \wedge \text{Associated-with}(TC, AC) \Rightarrow Rid(ACO) \prec Rid(TCO) \prec Rid(AC) \prec Rid(TC))$.

Rule 8: $(i)(j)[(LL(i, j) \wedge Level(C_i) \wedge Level(C_j)) \rightarrow \{(O_i \in C_i)(O_j \in C_j) \rightarrow Rid(O_i) \prec Rid(O_j) \prec Rid(C_i) \prec Rid(C_j)\}]$.

3.3 Deadlock Prevention Algorithm

The algorithm is based on prevention of circular wait condition. Prevention of circular wait condition is achieved using resource ordering and access ordering. Hence, DPA requires some preprocessing to be done before accepting the client transactions. The preprocessing steps are Global ordering of the resources in the system and deployment of resources to various sites.

As the proposed algorithm is based on resource ordering, it is necessary to order the resources globally to avoid access conflicts. Since objects are the resources and their global relationship is known from the class diagram, the ordering is done on the class diagram. The class diagram is partitioned into levels based on the rules of horizontal partitioning [2]. Figure 2 shows the resource ordering of a sample class diagram. The objects and classes are ordered from level 1 to level 4 using the resource ordering rules in 3.1.

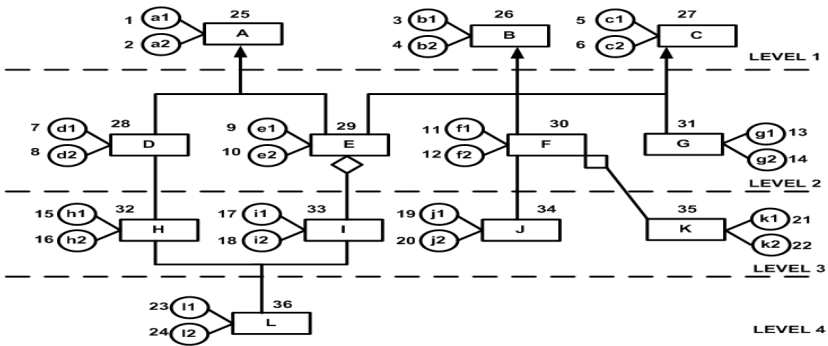


Fig. 2. Sample class diagram with resource ordering and horizontal partitioning

Resource ordering method

1. Let \prec be the initial total order of all the objects. This ordering starts from level 1 up to level n.
2. For every set of objects O_i belonging to classes C_i in level 1, add the objects arbitrarily.
3. For objects in level 2 to level n, order the objects in increasing dependency such that resource id of objects in level i is less than resource id of objects in level j, where $i < j$. Define the function $\text{max_object_Rid}(O)$, to return the maximum of resource ID of all objects O participating in the system which will be the resource id of last object in level n.
4. For all classes in level 1, i.e. independent classes (classes that need not lock any other classes to preserve consistency, like base classes and component classes), add the classes arbitrarily to create the total order such that $\text{max_object_Rid}(O)$ is always less than resource IDs of independent classes i.e. classes in level 1.
5. For classes in level 2 to n, order the classes in increasing dependency, such that resource ID of classes C_i in level i is less than resource IDs of classes C_j , where $i < j$.
6. Let \prec be the smallest transitively – closed order that is compatible with steps 1-5.

By applying the technique in the sample class diagram, the resource ordering is done. In this class diagram, there are four levels. The objects and classes are ordered from level 1 to level 4 using the resource ordering technique. The maximum resource id for objects is 24. The classes in class diagram along with their associated database are horizontally partitioned and assigned to four sites S1 to S4. However the class diagram is duplicated and placed in all the sites for global information.

The list of resources for every transaction is sorted in increasing order of resource IDs. The objective of this request ordering is to enforce the rule that resources of higher IDs can be requested only after obtaining resources of lower IDs. The location of the servers for these distributed requests need to be identified, as the resources are partitioned to several sites.

4 Formal Proof

Theorem 1: The resource ordering \prec done using the proposed resource ordering method is total and meets rules 1-8.

\prec **orders each pair of resources r_i and r_j :** If the resources r_i and r_j are not already ordered, by step 2 and 3 all objects are ordered. Then at least one of them should be a class.

Case 1: r_i is an object and r_j is its parent class (Rule 1)

Steps 2 and 3 orders all objects by their dependency. Without loss of generality, by step 3 and 4, $\text{max_object_Rid}(r_i) \prec \text{Rid}(r_j)$. By transitivity of \prec , it is proved that $r_i \prec r_j$.

Case 2: r_i is a base class and r_j is its sub class (Rule 2)

If r_i is in level i and r_j is in level j , and levels $i < j$, then in general by steps 4 and 5, $\text{rid}(r_i) \prec \text{rid}(r_j)$. By definition of inheritance, base classes will be always in lower level than their subclasses. Hence, it is proved.

Case 3: r_i is a component class and r_j is its composite class (Rule 3)

If r_i is in level i and r_j is in level j , and levels $i < j$, then in general by steps 4 and 5, $\text{rid}(r_i) \prec \text{rid}(r_j)$. By definition of composition, component classes will be always in lower level than their composite classes. Hence, it is proved.

Case 4: r_i is an associative class and r_j is associated class (Rule 4)

If r_i is in level i and r_j is in level j , and levels $i < j$, then in general by steps 4 and 5, $\text{rid}(r_i) \prec \text{rid}(r_j)$. By definition of association, associated classes will be always in lower level than their associated classes. Hence, it is proved.

Case 5: Proof of Rule 5

From case 1 and case 2, by transitivity, rule 5 holds good for all resources r_i and r_j .

Case 6: Proof of Rule 6

From cases 1 and 3, it is inferred that rule 6 holds good for all resources r_i and r_j .

Case 7: Proof of Rule 7

From cases 1 and 4, it is inferred that rule 7 holds good for all resources r_i and r_j .

Case 8: Proof of Rule 8

r_i is in level i , r_j is in level j and $i < j$. By steps 1- 6, from cases 5, 6 and 7, it is proved that they are totally ordered.

Theorem 2: Ordering \prec is well defined, if $r_i \prec r_j$, then $r_j \prec r_i$ does not hold.

Let $\zeta \prec$ be defined as $\{R, E\}$, where resources R represent nodes and E is defined as set of edges linking those resources that are related by the ordering $r_i \prec r_j$, where $r_i, r_j \in R$ and \prec defines E , then it is enough to show that $\zeta \prec$ contains no cycles. The resources can be either objects or classes. All objects r_i and r_j are ordered by step 1 and 2 and edges $r_i \rightarrow r_j$ are added for all objects at these steps. If r_i and r_j are classes related by inheritance and/or composition and/or association, then by steps 4 and 5 they are ordered and $r_i \rightarrow r_j$ are added in this step. So it is clear that so far there are no back

edges $r_j \rightarrow r_i$ added, as the objects and classes are ordered individually. In step 3 objects and classes are ordered by defining \max_object_Rid (O) to be less than the resource IDs for all classes. So, it is clear by transitivity, that all objects and classes are ordered. The edge $r_i \rightarrow r_j$ is added where r_i is an object with maximum resource ID and r_j is a class. Since edges are added for each type of resource exclusively, there will be no cycles.

Theorem 3: Deadlocks are prevented in single request model by eliminating cycles in wait for graph.

The proof for this is given in Holt [11].

5 Conclusion

The proposed scheme has proposed resource ordering technique for objects in distributed object oriented systems. A good Deadlock Prevention Algorithm (DPA) should avoid starvation. Starvation usually arises due to poor access ordering policy. Starvation is categorized and defined as follows:

Starvation in poverty [11]: A resource request made by a transaction is never satisfied there after, while on the other hand the requested resource is assigned to other transactions repeatedly. *Starvation in wealth [12]:* A resource requested by a transaction is never satisfied though it is permanently satisfiable from a particular time instant.

In DPA, access ordering is usually on FIFO basis to ensure fairness in the system. However, strict adherence of FIFO strategy may introduce starvation in wealth, which states that latter transaction which could have been satisfied, is kept waiting, since earlier transaction is waiting. On the other hand, assigning priority introduces starvation in poverty, which is a consequence of expedient scheduling strategy. This makes shortage of resources and makes lower priority transactions permanently blocked. Hence it is required to propose expedite access ordering that alleviates the problem of starvation in poverty and starvation in wealth as future work and conduct simulation experiments using the enhanced proposed scheme and give out the results.

References

- [1] H.X.J., Soo, J.: A Performance comparison of Deadlock Prevention and Detection Algorithms in a Distributed File System. In: Eighth International Conference on Computer & communications (1989)
- [2] Ozsu, M.T., Valduriez, P.: Principles of Distributed Database Systems. Pearson Education, London (1999)
- [3] Reddy, P.K., Bhalla, S.: Deadlock Prevention in Distributed Database Systems. ACM SIGMOD Record 22(3), 40–46 (1993)
- [4] Davidson, S., Lee, I., Wolfe, V.F.: Deadlock Prevention in Concurrent Real time Systems, pp. 305–318. Kluwer Academic Publishers, Real time Systems, Dordrecht (1993)

- [5] Gunther, K.D.: Prevention of Deadlocks in Packet Switched Data Transport Systems. IEEE Transactions on Communications Com-29(4), 512–524 (1981)
- [6] Cummins, F.A.: Distributed Object System with Deadlock Prevention, US Patents No. US 6236995B1 (2001)
- [7] Lewis, R.L.: Preventing Deadlocks, US Patent No. 0168448 A1 (2008)
- [8] Anand, A., Sethi, M.: Prevention of Deadlock in a Distributed Computing Environment, US Patent No. 0138886A1 (2009)
- [9] Riehle, D., Berzuk, S.P.: Properties of Member Functions in C++, Report (2000)
- [10] Riehle, D., Berzuk, S.P.: Types of Member Functions in C++, Report (2000)
- [11] Holt, R.: Some deadlock properties of Computer Systems. ACM Computing Surveys 4(3), 179–196 (1972)
- [12] Parnas, D.L., Habermann, A.N.: Comment on deadlock prevention method. Communications ACM 15(9), 840–841 (1972)
- [13] Garza, J.F., Kim, W.: Transaction management in an object oriented database system. In: ACM SIGMOD Int'l Conference, Management Data (1987)

Identification of Error Prone Classes for Fault Prediction Using Object Oriented Metrics

Puneet Mittal¹, Satwinder Singh¹, and K.S. Kahlon²

¹ BBSBEC, Fatehgarh Sahib, Punjab, India

² GNDU, Amritsar, Punjab, India

Abstract. Various studies have found that software metrics can predict class error proneness. However their study is focused on the relationship between class error proneness and software metrics during the development phase of software projects not in system's post-release evolution. This study is focused on the three releases of Javassist- open source java based software. This paper describes how we calculated the object-oriented metrics to illustrate error-proneness detection. Using Findbugs we collected errors in the post-release system and applied logistic regression to find that some metrics can predict the class error proneness in post release evolution of system. We also calculated model's accuracy by applying one model on other version's data.

Keywords: Object oriented metrics, Class error-proneness, Javassist, Findbugs, JColumbus, Together tool.

1 Introduction

It is the dream of every developer to develop error-free software. But inspite of undergoing software testing, walkthroughs and inspections; post maintenance of software is required. Few errors still remain in software that can make the life of software developer hell. It is very difficult to do changes after the software has been released. But if we know the probable areas where error can be located, the problem can be solved and the software testers can be helped to locate the errors easily and effectively. Software metrics are one way to measure the software and can be used for locating the errors. One goal of software metrics is to identify and ensure the essential parameters that affect software development. Software metrics provide quantitative basis for development and validation of models of software development process. Metrics can be used to improve software productivity and quality.

In this paper, we describe how we calculated the object-oriented metrics for error-proneness detection from source code of open source software, Javassist[12]. We employed statistical methods (i.e. logistic regression) for predicting error proneness of code.

2 Literature Survey

Various software metrics have been proposed by various researchers in different paradigms. Various studies have sought to analyze the connection between object-oriented metrics and code quality ([16], [9], [2]).

Chidamber et al. [5] developed and implemented a new set of software metrics for OO designs. These metrics were based on measurement theory and also reflect the viewpoints of experienced OO software developers. He gave set of six OO metrics (WMC, DIT, RFC, LCOM, NOC, and CBO) where WMC, NOC and DIT metrics reflect class hierarchy; CBO and RFC metric reflect class coupling and LCOM reflects cohesion. **Basili et al. [2]** collected data about faults found in object oriented classes. Based on these data, they verified how much fault-proneness is influenced by internal (e.g., size, cohesion) and external (e.g., coupling) design characteristics of OO classes. From their results, five out of the six CK OO metrics appear to be useful to predict class fault-proneness during the high- and low-level design phases of the life-cycle. **Chidamber et al. [6]** investigated the relationship between the CK metrics and various quality factors: software productivity, rework effort, and design effort. The study also showed that the WMC, RFC, CBO metrics were highly correlated. Therefore, **Chidamber et al.** did not include these three variables in the regression analysis to avoid generating coefficient estimates that would be difficult to interpret. The study concluded that there were associations between the high CBO metric value and lower productivity, more rework, and greater design effort. In another study, **Wilkie and Kitchenham [18]** validated the relationship between the CBO metric and change ripple effect in a commercial multimedia conferencing system. The study showed that the CBO metric identified the most change-prone classes, but not the classes that were most exposed to change ripple effect. **Cartwright and Shepperd [4]** also investigated the relationship between a subset of CK metrics in a real-time system. The study showed that the parts of the system that used inheritance were three times more error prone than the parts that did not use inheritance. **Subramanyam and Krishnan [16]** validated the WMC, CBO, and DIT metrics as predictors of the error counts in a class in a business-to-consumer commerce system. Their results indicated that the CK metrics can predict error counts. They examined the effect of the size along with the WMC, CBO, and DIT values on the faults by using multivariate regression analysis. They concluded that the size was a good predictor in both languages, but WMC and CBO could be validated only for C++. **Alshayeb and Li [1]** conducted a study on the relationship between some OO metrics and the changes in the source code in two client-server systems and three Java Development Kit (JDK) releases. Three of the CK metrics (WMC, DIT, and LCOM) and three of the Li metrics (NLM, CTA, and CTM) were validated. They found that the OO metrics were effective to predict design effort and source lines of code added, changed, and deleted in short-cycled agile process (client-server systems); however, the metrics were ineffective predictors of those variables in long-cycled framework evolution process. **Olague et al. [14]** indicated that the CK and QMOOD OO class

metrics suites are useful in developing quality classification models to predict defects in both traditional and highly iterative, or agile, software development processes for both initial delivery and for multiple, sequential releases.

The above studies suggest that various OO metrics can predict the error proneness during development but still their usability is dubious. It seems to be difficult to predict the class error proneness in post-release system as system has already passed through various quality tests and very few exceptional errors remain behind.

In this study we want to know whether software metrics can predict error proneness of class in post-release system. Our study is based on open source java project Javassist- byte code manipulator. Source code of Javassist is freely available online.

3 Data Collection

We collected errors for three releases of Javassist project version 2.4, 2.6 and 3.0 using the FindBugs[7] tool. FindBugs is a program which uses static analysis to look for bugs in Java code. FindBugs gives list of probable bugs or errors along with their package name, class name and method name. After that we collected metrics for all the three versions of Javassist using the JColumbus[8] and Together[15] tool. Together tool is a plugin in eclipse for finding the metrics. Metrics used in study are NM, NA, NOA, NOC, DIT, CBO, RFC, LCOM5, WMC, and TCC (all listed in appendix). JColumbus could find 9 metrics NM, NA, NOA, NOC, DIT, CBO, RFC, LCOM5, WMC and for TCC we used Together tool. The object oriented metrics were extracted with the help of tool. Next we associated errors with each class in metrics list. Each class was marked erroneous if atleast one error was found and not erroneous if no error was found.

3.1 The Descriptive Statistics of Data

Table 1 shows the distribution of errors and summarizes the number of classes that had errors, the number of error prone classes, the number of classes that did not have errors and the total number of classes considered in the study. Tables 2–4 summarize the metrics descriptive statistics.

Table 1. Distribution of errors based on the error categories

Error	Javassist 2.4	Javassist 2.6	Javassist 3.0
Total no. of errors	43	67	78
No. of error prone classes	25	36	46
No. of classes without error	168	160	192
Total classes	193	196	238

Table 2. Javassist2.4

Metrics	Mean	Standard Deviation	MIN	MAX	Percentile		
					25%	50%	75%
NM	16.82383	22.30751	0	145	4	10	22
NA	22.1399	61.80418	0	248	1	2	4
NOA	0.865285	1.021889	0	5	0	1	1
NOC	0.450777	1.464479	0	10	0	0	0
DIT	0.787565	0.854797	0	4	0	1	1
CBO	5.756477	6.471715	0	30	2	3	8
RFC	16.93782	23.48351	0	147	4	9	19
LCOM5	5.518135	7.891909	0	53	1	3	6
WMC	18.50259	38.17983	0	330	3	6	18
TCC	19.54839	30.74617	0	100	0	0	33

Table 3. Javassist2.6

Metrics	Mean	Standard Deviation	MIN	MAX	Percentile		
					25%	50%	75%
NM	17.16837	22.79467	0	147	4	10.5	22.25
NA	22.95408	63.04279	0	249	1	2	4
NOA	0.872449	1.017315	0	5	0	1	1
NOC	0.454082	1.479062	0	10	0	0	0
DIT	0.795918	0.852859	0	4	0	1	1
CBO	5.821429	6.569842	0	30	2	3	8
RFC	17.35385	24.17682	0	151	4	9	19
LCOM5	5.607143	8.01369	0	55	1.75	3	6
WMC	18.79592	39.13224	0	349	4	6	18
TCC	18.89172	29.6024	0	100	0	0	33

Table 4. Javassist3.0

Metrics	Mean	Standard Deviation	MIN	MAX	Percentile		
					25%	50%	75%
NM	17.96639	25.28852	0	158	5	11	22.75
NA	22.64286	63.26699	0	310	1	2	4
NOA	0.890756	1.008737	0	5	0	1	1

Table 4. (continued)

NOC	0.487395	1.751988	0	13	0	0	0
DIT	0.806723	0.829593	0	4	0	1	1
CBO	6.189076	6.982948	0	36	2	4	8
RFC	18.31933	25.77129	0	180	5	10	19.75
LCOM5	5.57563	7.826208	0	71	2	3	6
WMC	20.0084	40.29239	0	376	4	8	19.75
TCC	22.06842	30.8437	0	100	0	4.5	33

We noticed that NOC metrics was zero for 75% of classes and the highest DIT value was 4 and 75 % of classes have only 1 level of inheritance.

To get more insight on the relationship between the classes that have errors and the classes that have no error, we used the error box plot chart for Javassist2.4. The error box plot charts help us in comparing groups (no error/error groups) to draw conclusions that one group is higher (or lower) on average than another. If the means and their confidence intervals do not overlap, then we will find a statistically significant difference between the groups. If the means and their confidence intervals for two groups overlap, then the groups are probably not significantly different from one another in a statistical sense. Error box plot charts for Javassist2.4 in Fig.1 shows the means and 95% confidence intervals for all the metrics. All error charts show that the distribution of the classes with errors has more variability than the no error classes for all metrics. We noticed that NM, CBO, RFC, LCOM5, WMC don't overlap (i.e., significantly different) and the means for the error group are higher than the no error group. The mean for the error group for the DIT metric is lower than the no error group which gives opposite results to our expectations. The means for NA, NOA, NOC, TCC metrics are higher in the error group than the no error group, but the intervals overlap, which mean that they may not be significantly different.

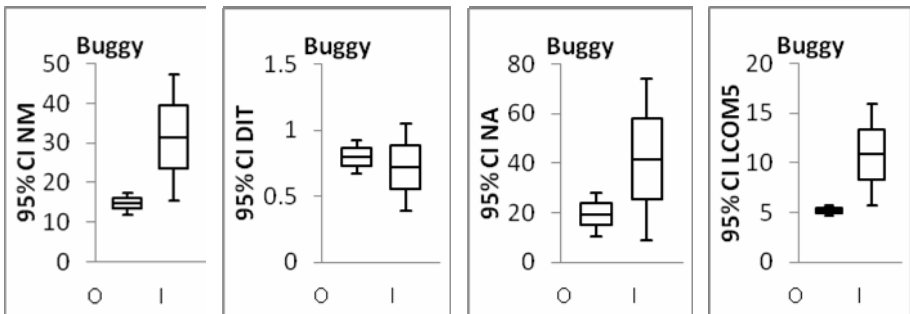


Fig. 1. Means and CI of all metrics

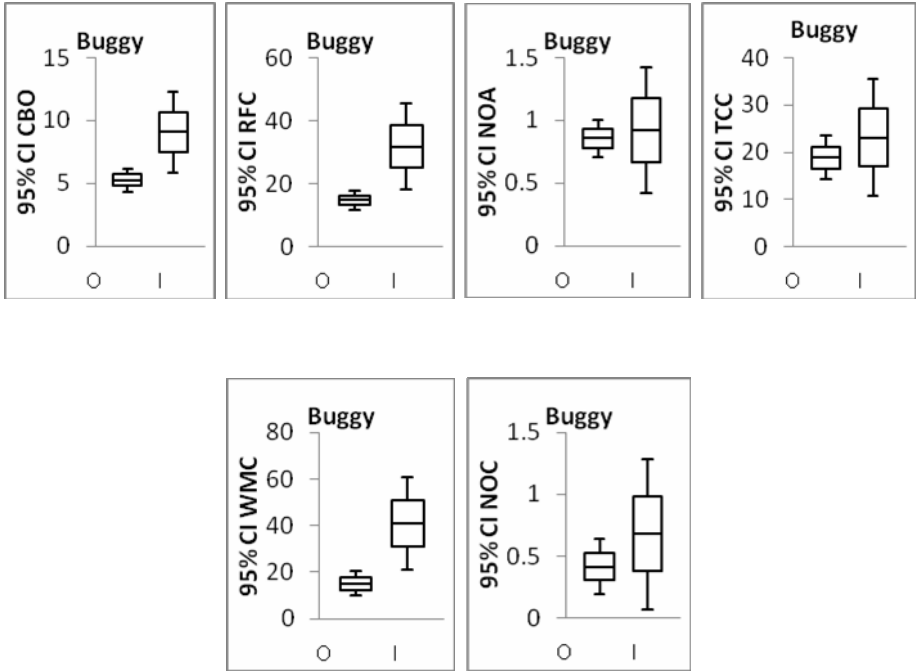


Fig. 1. (continued)

4 The Statistical Models

In this study, we used logistic regression for class error probability. It is used for prediction of the probability of occurrence of an event by fitting data to a logit function. We used the Univariate Binary Regression (UBR) test to examine whether there was any significant association between a metric and the class error proneness. We used 0.05 as the cutoff P-value in both tests [10]. We then combined the significant metrics into one set to build the multivariate prediction models for the error probabilities. We used the multivariate logistic regression (MLR) to predict class error probability. The independent variables were selected for the MLR analysis. Binary dependent variable tells whether class is erroneous or not in MLR model.

The general MLR model is as follows:

$$\pi(Y=1|X_1, X_2, \dots, X_n) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{1}$$

where $g(x) = B_0 + B_1 * X_1 + B_2 * X_2 + \dots + B_n * X_n$ is the logit function; π is the probability of a class being faulty; Y is the dependent variable- it is a binary variable;

X_i ($1 \leq i \leq n$) are the independent variables, which are the OO metrics investigated in this study; B_i ($0 \leq i \leq n$) are the estimated coefficients from maximizing the log-likelihood.

The B_i 's are used to calculate the Odd Ratio ($OR = e^{B_i}$) for each independent variable that contributes to the model. The OR value describes type of association between dependent and independent variable. Positive association ($OR > 1$) means that likelihood of dependent variable changes in same direction as independent variable changes. Negative association ($OR < 1$) means that likelihood of dependent variable changes in opposite direction of independent variable change. The even association ($OR = 1$) means that changes in independent variable has no impact on dependent variable. The interpretation of OR magnitude depends on scale of independent variable. If an independent variable is continuous, such as CBO metric, then “OR = 2” means that for each unit change of CBO metric, the class error probability changes twice as much in the same direction in its scale.

If an independent variable is dichotomous and is coded as one or zero, then “OR = 2” means that the error probability is twice more likely to occur in a class when the independent variable is present than otherwise.

We started analysis with the ten metrics (NM, NA, NOA, NOC, DIT, CBO, RFC, LCOM5, WMC, and TCC). As we progressed through various analysis steps we dropped various metrics based on some intermediate results and continued our study.

4.1 The Metric Validation

We conducted UBR analysis on ten metrics (NM, NA, NOA, NOC, DIT, CBO, RFC, LCOM5, WMC, and TCC) for error proneness. Table 5 shows UBR results for three versions of Javassist. A metric was significant reflector of error proneness if its p -value ≤ 0.05 . NM, CBO, RFC, LCOM5, WMC were significant. The metrics NOA, NOC, DIT and TCC are highly insignificant as their P -value is > 0.05 .

Next we conducted MLR analysis on the metrics selected after the UBR analysis.

Table 5. UBR Analysis

	Javassist2.4		Javassist2.6		Javassist3.0	
Metrics	B-value	P-value	B-value	P-value	B-value	P-value
NM	0.023	0.003	0.022	0.003	0.011	0.038
NA	0.005	0.102	0.005	0.043	0.003	0.182
NOA	0.059	0.774	-0.048	0.798	-0.026	0.874
NOC	0.099	0.410	0.052	0.650	0.109	0.163
DIT	-0.110	0.671	-0.288	0.223	-0.212	0.312
CBO	0.073	0.009	0.076	0.002	0.060	0.004
RFC	0.021	0.003	0.027	0.000	0.018	0.002
LCOM5	0.065	0.002	0.085	0.000	0.066	0.002
WMC	0.011	0.012	0.019	0.001	0.012	0.004
TCC	0.004	0.535	-0.006	0.369	0.002	0.690

4.2 Error Proneness Prediction Model

We applied forward stepwise logistic regression on the set of metrics obtained from table 5. Results are shown in table 6-8. Each table shows the coefficients, standard errors, P-values and Odd Ratio of the metrics that were included in the model.

Table 6. Javassist2.4

	B-value	S.E.	P-value	OR
LCOM5	0.065	0.021	0.002	1.067
Constant	-2.357	0.279	0.000	0.095

Table 7. Javassist2.6

	B-value	S.E.	P-value	OR
LCOM5	0.085	0.023	0.000	1.088
Constant	-2.057	0.252	0.000	0.128

Table 8. Javassist3.0

	B-value	S.E.	P-value	OR
LCOM5	0.066	0.022	0.002	1.068
Constant	-1.842	0.219	0.000	0.159

4.3 Models' Accuracy Evaluation

We learned through the MLR analysis that we could use some metrics to predict class error-proneness in the post-release evolution of Javassist. This strong association suggests some errors in the source code are indeed related to the design structure of the system (at the class level) in the development as well as the post-release evolution phase of a system.

We use the area under Receivable Operating Characteristics (ROC) curve to evaluate the classification accuracy of MLR. This area measures the association between the observed responses and the predicted probabilities from the model application. The ROC curve plots the probability of detecting true-positives (sensitivity²) and false-positives ($1 - \text{specificity}^3$) for an entire range of cutoff points. The area under ROC curve ranges between 0 and 1; it measures the discrimination power of the models. The general rule to evaluate the discrimination is:

- If $0.5 \leq \text{ROC} < 0.6$: no discrimination
- If $0.6 \leq \text{ROC} < 0.7$: poor discrimination
- If $0.7 \leq \text{ROC} < 0.8$: good discrimination
- If $0.8 \leq \text{ROC} < 0.9$: excellent discrimination
- If $0.9 \leq \text{ROC} < 1$: outstanding discrimination

Table 9. Area under ROC curve in MLR models

	ROC area
Javassist2.4	0.669
Javassist2.6	0.681
Javassist3.0	0.589

4.4 Evaluating Models on Successive Releases

Because our goal was to predict the error-prone classes, we experimented with the use of the prediction model in one release to predict the error-prone classes in the future releases. We applied the Javassist2.4 prediction model to the Javassist2.6 and Javassist3.0 data, and we applied the Javassist2.6 prediction model to the Javassist3.0 data. Table 10 shows the results.

Table 10. Application of MLR models on consecutive releases

	ROC area
Applying 2.4 model on 2.6 data	0.681
Applying 2.4 model on 3.0 data	0.589
Applying 2.6 model on 3.0 data	0.589

5 Conclusion

We investigated whether the object-oriented metrics could predict the class error probability in the post-release evolution of Javassist and found that some metrics could still predict class error probability. Table 6-8 gave the model for error prediction. For checking the error prediction accuracy we have formed the ROC curves. ROC curves for MLR model is above 0.5 level which is though poor but can be used for the prediction. And when one model is applied on other version's data, results were satisfactory (Table 10) as it closely matched with the values of predicted model value (Table 9). So, we can conclude that predicted models can work satisfactory for predicting the errors in general.

6 Threats

We have collected errors with the help of open source software FindBugs. We make no claims about errors discovered using software. The metrics data were collected with the help of JColumbus tool which is research software and together tool. We make no claims about the accuracy of these tools but we believe that the metrics data collected were consistent.

Acknowledgements. We are grateful to FrontEndART Software Ltd. [8] for providing us their research tool JColumbus for finding metrics. Special thanks to

Peter Siket of FrontEndART Software Ltd. and Prof. M. Javed Associate Professor Department of Statistics P.A.U. Ludhiana for their support.

References

1. Alshayeb, M., Li, W.: An Empirical Validation of Object-oriented Metrics in Two Iterative Processes. *IEEE Transactions on Software Engineering* 29(11), 1043–1049 (2003)
2. Basili, V., Briand, L., Melo, W.: A Validation of Object-oriented Design Metrics as Quality Indicators. *IEEE Transactions on Software Engineering* 22(10), 751–761 (1996)
3. Bieman, J.M., Kang, B.K.: Cohesion and Reuse in an Object-Oriented System. In: *Proceedings in ACM Symposium on Software Reusability(SSR 1995)*, pp. 259–262 (1995)
4. Cartwright, M., Shepperd, M.: An Empirical Investigation of an Object-oriented Software System. *IEEE Transactions on Software Engineering* 26(7), 786–796 (2000)
5. Chidamber, S., Kemerer, C.F.: A Metrics Suite for Object Oriented Design. *IEEE Transactions on Software Engineering* 20(6), 476–493 (1994)
6. Chidamber, S., Darcy, D., Kemerer, C.: Managerial Use of Metrics for Object Oriented Software: An Exploratory Analysis. *IEEE Transactions on Software Engineering* 24(8), 629–639 (1998)
7. Findbugs™ -Find bugs in java programs (2010), <http://findbugs.sourceforge.net>
8. FrontEndART Software Ltd- Software and source code quality assurance tools and services, <http://www.frontendart.com/>
9. Gyimothy, T., Ferenc, R., Siket, I.: Empirical Validation of Object Oriented Metrics on Open Source Software for Fault Prediction. *IEEE Transactions on Software Engineering* 31(10), 897–910 (2005)
10. Brian, H.-S.: *A Book of Object-Oriented Knowledge*. Prentice-Hall, Englewood Cliffs (1992); ISBN 0-130-59445-8
11. Hosmer, D., Lemeshow, S.: *Applied Logistic Regression*, 2nd edn. Wiley Series in Probability and Statistics. Wiley, Chichester (2000)
12. Javassist (2010), <http://www.csg.is.titech.ac.jp/~chiba/javassist>
13. Lorenz, M., Kidd, J.: *Object-Oriented Software Metrics*. Prentice Hall, Englewood Cliffs (1994)
14. Olague, H., Etzkorn, L.: Empirical Validation of Three Software Metrics Suites to Predict Fault-Proneness of Object-Oriented Classes Developed Using Highly Iterative or Agile Software Development Processes. *IEEE Transactions on Software Engineering* 33(6), 402–419 (2007)
15. *Software Architecture Design, Visual UML & Business Process Modeling – from Borland* (2010), <http://www.borland.com/us/products/together/index.html>
16. Subramanyam, R., Krishnan, M.: Empirical Analysis of CK Metrics for Object-oriented Design Complexity: Implications for Software Defects. *IEEE Transactions on Software Engineering* 29(4), 297–310 (2003)
17. Li, W.: Another Metric Suite for Object Oriented Programming. *The Journal of Systems and Software* 44(2), 155–162 (1998)
18. Wilkie, F., Kitchenham, B.: Coupling Measures and Change Ripple in C++ Application Software. *Journal of System and Software* 52(2–3), 157–164 (2000)

Appendix: OO Metrics Definition

Metric	Description	Definition	References
NM	Number of Methods	Counts the number of methods	[13]
NA	Number of Attributes	Counts the number of attributes	[13]
NOA	Number Of Ancestors	Number of classes that a given class directly or indirectly inherits from.	[17]
NOC	Number Of Children	Number of classes that directly inherit from a given class.	[5]
DIT	Depth of Inheritance	DIT of a class is the length of the longest path from the class to one of its root in the inheritance hierarchy.	[5]
CBO	Coupling Between Objects	A class is coupled to another if the class uses any method or attribute of the other class or directly inherits from it.	[5]
RFC	Response For Class	RFC is the cardinality of the set M of methods of the class (inherited ones are not taken into account) and the set of methods directly invoked by methods in M	[5]
LCOM5	Lack of Cohesion Of Methods 5	Density of accesses to attributes by methods. Consider an undirected graph G, where the vertices are the methods of a class, and there is an edge between two vertices if the corresponding methods use at least one attribute in common or one of them invokes the other. LCOM5 is defined as the number of connected components of G	[10]
WMC	Weighted Methods per Class	Sum of the complexity of all methods for a class	[5]
TCC	Tight Class Cohesion	TCC considers two methods to be connected if they share the use of at least one attribute	[3]

An Automated Tool for Computing Object Oriented Metrics Using XML

N. Kayarvizhy¹ and S. Kanmani²

¹AMC Engineering College, Bangalore, India

²Pondicherry Engineering College, Puducherry, India

Abstract. The importance of object oriented metrics is on the rise and a lot of research is being carried out on various aspects of using object oriented metrics in evaluating the quality attributes of object oriented systems. Metrics computation is an integral step in all these research activities. This demands a quick and easy way to have the metrics computed and presented to the research community. Existing tools fall short as it is not easy to extend them to new metrics or language. This paper focuses on the design of an automated object oriented metrics tool which has a generic framework for computing the metrics. The tool converts the source code developed using a particular object oriented language to a language independent XML format which is then used for computing the required metrics.

Keywords: Object Oriented Metrics, Java, C#, Tool, AMT.

1 Introduction

The metrics for object oriented software systems focus on measurement that can be applied to the class and the design characteristics viz., localization, encapsulation, inheritance, information hiding, polymorphism, messaging and object abstraction. Object oriented metrics are widely used in the life cycle of a software product for assessing various attributes like quality of the design, complexity of the code, prediction of faults and for estimating the effort involved for maintenance. As the popularity of object oriented metrics increased, a large number of metrics were proposed for capturing the various aspects of object oriented programming [2][11][12][13][14]. The number of new metrics keeps increasing continuously with each new related research work. Metrics can either be collected manually or by an automated tool.

Existing tools suffer mainly from two problems. The first is that most existing tools are commercial tools where extensibility becomes an issue [1]. In such tools, the researcher will not have access to the source code for adding his own metric to the list of existing metrics. Free tools are available but are specific to a language or written to satisfy just the need of one particular research topic and hence are not easy to adapt to other metrics. The second problem which is common to all tools is the interpretation of the metrics. Most metric definitions are ambiguous and hence more than one

variant of the same metric exists which have been proposed by different researchers. This results in tools implementing each metric in their own way.

Given these problems, we report on the design and development of an automated metrics computation tool (AMT) that is inherently extensible. We believe that with constantly improving software processes, it is essential that any tool used to gather metrics data should be easy to modify and extend to accommodate both new and evolving metrics. Tools must also be readily adaptable to a variety of programming languages so that a consistent software process may be applied across multiple languages.

The automated tool that is proposed in this paper uses two important mechanisms. The first mechanism aims at providing a language independent representation of the source code using XML. This will ensure that that tool is generic enough to include any object oriented language, even those that are not currently available in use. The second mechanism is the metric collection which uses the standard XML format as input to collect each metric. This metric collector part is easily modifiable and extensible to include new metrics or for changing the way existing metrics are computed. This ensures that all concerns existing in the current tools are addressed.

2 Related Work

The primary objective of gathering software metrics is to provide the capability to predict future software development efforts based on past performance. Automated metrics collection has been attempted and a number of tools have been proposed in the literature.

Most tools available today gather and analyse traditional software metrics. Lincke et al. [1] have done a study of both commercial and free tools available for metric computation. They conducted their experiments on nine metrics using ten different tools on three software systems. They concluded that the values for the same metrics are calculated differently by different tools.

Many researchers have also attempted at creating their own tools. Some of them like Baroni et al. [3,4] adapt existing technologies for the purpose of computing metrics. Baroni et al. use OCL as a means to express metrics. Since OCL is defined to express constraints on UML class diagrams, design-related metrics can be implemented mainly as post-conditions. On the other hand, metrics in which implementation-related data are involved, such as implementation coupling metrics, cannot be directly defined using OCL. Similarly, Harmer and Wilkie [5] define a meta-model in the form of relational database schema, and use SQL to express metric computations. However, for some complex metrics, SQL is mixed with a programming language. The SQL-based approach is also used by Lavazza et al. [6], to collect only UML-based metrics (design level). In the same family, El-Wakil et al. [7] propose the use of XQuery to compute metrics on UML models that are represented in XML documents. Eichberg et al. [8] also developed a framework, called QScope, for measuring software projects. It is built on top of Magellan framework in which all documents of a project are stored as XML documents. XQuery is used as a definition language to express metrics. This approach allows the user to collect metrics on different artefacts with the use of a uniform mechanism, i.e., XQuery.

Mens et al. [9] define an object oriented meta-model as graphs and formalism for metric definition based on graph manipulation. Using this formalism, they define three generic metrics: NodeCount, EdgeCount and PathLength, and a number of complementary higher-order metrics (e.g. ratio, sum, and average). Starting from these generic and higher-order metrics more than 30 object-oriented metrics have been implemented. Marinescu et al. [10] propose a simplified implementation of object-oriented design metrics. In their approach, a new interpreted language, called SAIL, is defined to express metrics.

3 Generic Framework

The proposed generic framework model for calculating metrics is shown in Fig. 1 below. It contains the parser and XML converter which takes the source code as input

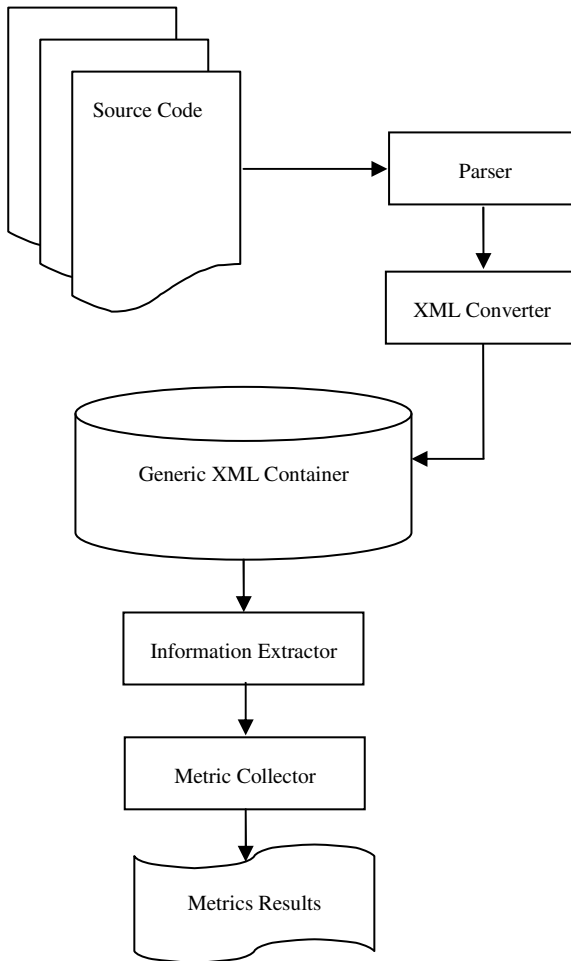


Fig. 1. Generic Framework

and creates the intermediate generic XML format. This is then passed on to the metrics calculator which extracts the required metrics and displays them. The tool was developed in Java and currently supports only software developed in Java and C# languages as these languages are the current popular choice in object oriented languages.

XML has been chosen as the intermediate generic format as it has the following advantages [16]. XML is fully compatible with applications like JAVA, and it can be combined with any application which is capable of processing XML irrespective of the platform it is being used on. XML is an extremely portable language. XML is an extendable language, meaning that we can create our own tags, or use the tags which have already been created. It is a platform independent language. It can be deployed on any network and the application can work along with XML. It can work on any platform and has no boundaries. It is also vendor independent and system independent. While data is being exchanged using XML, there will be no loss of data even between systems that use totally different formats

3.1 Parser

The main function of the parser is to break the source code into its object oriented constructs like classes, methods and attributes. The parser has to be unique for each object oriented language that the tool supports. Parsing functionality for languages is generally available as part of the compiler. Free tools and open source software are also available which can be made use of for implementing the parser. Extending the tool for including languages involves getting the equivalent parser in place for the language

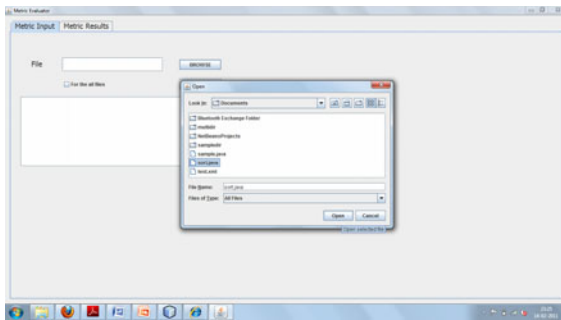


Fig. 2. Parser and XML Converter Module

3.2 XML Converter

Once the parser has done the necessary partitioning of the source code to various constructs, the XML generator then acts on them to provide a generic XML file.

This XML format has standardized tags for all the major constructs along with their property as well. For example the entire information pertaining to a class is under the tag named 'class'. Within this parent tag are the children tags 'attribute' and 'method'. Each tag has properties that further provide additional information on the

```

<?xml version="1.0" encoding="UTF-8" ?>|
- <xjava version="1.1">
- <class name="MyTestClass" public="yes"
  unqualifiedName="MyTestClass">
  <import kind="package">java.util.*</import>
  <extends class="MyBaseClass" />
- <field name="MyTestClass.names"
  protected="yes" unqualifiedName="names">
  <type dimension="0" fullName="java.util.Vector"
    name="java.util.Vector" unqualifiedName="Vector" />
  </field>
- <constructor name="MyTestClass.MyTestClass" public="yes"
  unqualifiedName="MyTestClass">
  </constructor>
- <method name="MyTestClass.isNameAvailable" public="yes"
  unqualifiedName="isNameAvailable">
  <type dimension="0" fullName="boolean" name="boolean"
    unqualifiedName="boolean" />
- <parameter name="name">
  <type dimension="0" fullName="java.lang.String"
    name="java.lang.String" unqualifiedName="String" />
  </parameter>
</method>
- <method name="MyTestClass.countUppercaseLetters"
  public="yes" static="yes"
  unqualifiedName="countUppercaseLetters">
  <type dimension="0" fullName="int" name="int"
    unqualifiedName="int" />
- <parameter name="name">
  <type dimension="0" fullName="java.lang.String"
    name="java.lang.String" unqualifiedName="String" />
  </parameter>
</method>
</class>
</xjava>

```

Fig. 3. Language Independent XML Output

construct. For a class tag the additional information are whether it is ‘public’ and if it ‘inherits’ from other class. An example XML is shown in Fig 3 above.

3.3 Information Extractor from XML

The module has two parts, the first one being an XML parser which parses the generic XML file and takes the necessary information from it. The second part creates a generic data structure which aids easy metric computation. This module is common for all the languages that would be supported by the tool.

For example the classes are arranged based on hierarchy which is used directly for all inheritance based metrics. The ‘class’ tag indicates the main class and ‘extends class’ tag indicates its parent. For example the ‘MyTestClass’ in the example in Fig 3. has ‘MyBaseClass’ as the parent class. This is taken into account in the data structure as follows

$$\text{Nodes}[\text{'MyTestClass'}].\text{parentnode} = \text{Nodes}[\text{'MyBaseClass'}] \quad (1)$$

3.4 Metrics Calculator

The metrics calculator module like the Information Extractor above is common for all languages. It calculates the number based on this information. The metrics calculator can be easily extended to include new metrics. Each metric is calculated in a separate function to maintain clarity.

For example to find the Number of Children metric (NOC) for the class 'MyClass' we just have to do the following where x is the all the nodes

$$\sum \text{Nodes}[\text{'x'}].\text{parentnode} = \text{Nodes}[\text{'MyClass'}], \quad (2)$$

To find the Coupling between Object (CBO), which is the number of usage of other classes' methods and attributes in this class we have to do the following

$$\sum (\text{Nodes}[\text{'MyClass'}].\text{Method}[i].\text{uses}(\text{Nodes}[\text{'x'}].\text{Attribute}[j]) \parallel \text{Nodes}[\text{'MyClass'}].\text{Method}[i].\text{uses}(\text{Nodes}[\text{'x'}].\text{Method}[k])) \quad (3)$$

To find Lack of Cohesion (LCOM), which is the measure of how closely related are the class' methods and attributes we need to the following

$$\begin{aligned} & \sum ((\text{Nodes}[\text{'MyClass'}].\text{Method}[i].\text{hasAttribute}[k]) \&\& (\text{Nodes}[\text{'MyClass'}].\text{Method}[j].\text{hasAttribute}[k])) - \\ & \sum (((\text{Nodes}[\text{'MyClass'}].\text{Method}[i].\text{hasAttribute}[k] \&\& !(\text{Nodes}[\text{'MyClass'}].\text{Method}[j].\text{hasAttribute}[k])) \parallel \\ & (!(\text{Nodes}[\text{'MyClass'}].\text{Method}[i].\text{hasAttribute}[k] \&\& (\text{Nodes}[\text{'MyClass'}].\text{Method}[j].\text{hasAttribute}[k]))) \end{aligned} \quad (4)$$

3.5 Metrics Listing

The metrics are listed in the tool GUI for the convenience of the researcher. Option is also provided to select the class that the researcher is interested in and all the metrics for the class gets displayed. Also generally researchers end up using the metrics elsewhere to model their research. To aid this, the metrics are also provided as tab separated data that can be exported to an excel sheet for example to carry out further processing or mapping. Software metrics can be gathered at many levels of granularity. This tool provides measurement at the class level. Class-level metrics measure the complexity [15] of individual classes contained in the system.

These metrics can be very useful for predicting defect rates and estimating cost [14] and schedule of future development projects. Examples of class-level metrics are depth of inheritance tree and class coupling. Currently the following metrics are calculated by the tool – cohesion, coupling and inheritance metrics. Cohesion metrics describe how well the methods and attributes gel with the class. A poor cohesive class might be required to be split into multiple individual classes. Coupling metrics indicate how pairs of classes are bound to each other. A high binding indicates that the classes depend on each other strongly indicating a bad design and high complexity

overhead. Inheritance metrics indicate the level of depth of reuse attempted. A high inheritance might indicate high reusability but it comes at the cost of high maintenance complexity while a low or no inheritance indicates that there is no attempt at reuse.

Table 1 lists the metrics that have been considered for computation currently. We implemented a variety of existing size/complexity, inheritance, coupling and cohesion metrics. Since the entire information about the source code is captured in the XML format an addition of a new metric computation will only result in modification to the metrics calculator part and thus eases the addition of new metrics

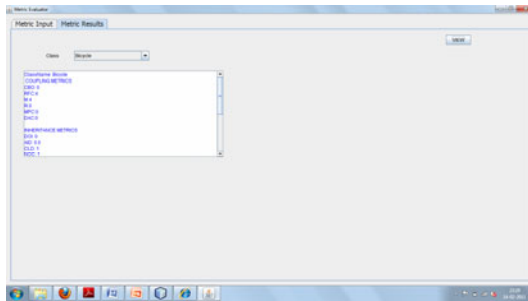


Fig. 4. Metrics Calculator and Display Module

4 Evaluation Results

A team of post graduate students were initially trained well on computing the metrics manually using a set of small programs in Java and C#. They were guided on the steps to be followed for computing each metric. They were monitored regularly and any doubts on their part were clarified.

To evaluate our tools we took two student projects ‘ATM Simulator’ in Java and ‘Student Management System’ in C#. The programs were specifically coded to ensure that the necessary test cases were taken into account for testing each metric comprehensively. The students were then advised to compute by hand all the metrics on these programs. After this the tools were used to obtain the values. The automated tool that we had proposed here were used on both Java and C# programs. Additionally we took two tools – CKJM and JMT to compare the metric values. CKJM is a freely available tool for Java that computes only the metrics from CK metric suite [17].

JMT is another freely available tool for computing metrics from JAVA language that has an option of a Graphical User Interface as well. Both the tools were also computing other traditional metrics in addition to the listed object oriented metrics.

The following table shows the values computed for the Java program given by AMT and the values got by using the tools CKJM and JMT [18]. If a particular metric is not supported by the tool, then it is denoted as ‘X’ in the tables

Table 1. Java Metrics

Metric	AMT	CKJM	JMT
LCOM1	0.38	36.61	X
LCOM2	0.24	X	X
LCOM3	0.31	X	X
CBO	1.92	3.00	29.53
CBO'	2.84	X	X
RFC	3.23	20.61	12.07
MPC	1.85	X	X
DAC	0.00	X	X
DAC'	0.00	X	X
WMC	4.07	5.84	5.76
DIT	0.46	1.69	0.46
AID	0.46	X	X
CLD	0.15	X	X
NOC	0.46	0.46	0.46
NOP	0.46	X	X
NOD	0.46	X	X
NOA	0.46	X	X
NMO	0.31	X	0.46
NMI	0.46	X	1.07
NMA	2.77	X	X

From the Table 1 above, we can infer that the values computed using JMT and CKJM did not match with the values from AMT and also did not match between them. This confirms the notion that the existing tools interpret the same metric differently. For example CKJM and JMT reported extremely different values for CBO. The only metric which had identical values among all the tools was NOC.

The low values on the LCOM metrics shows that the classes in the ATM Simulator program are highly cohesive. The Inheritance metrics also indicate that inheritance has been appropriately used as per the requirement. Coupling is present but the value indicates that it is at a manageable level.

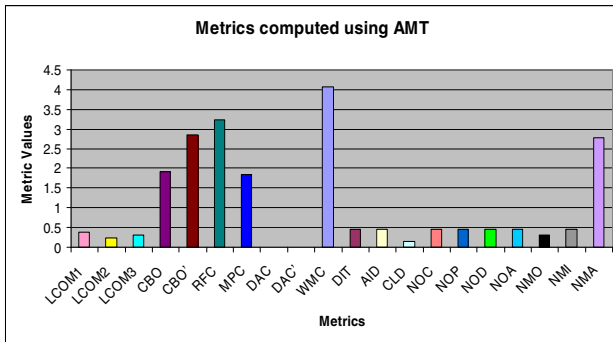


Fig. 5. Java Metric values computed by AMT

Since the AMT tool also supports C#, we have compared AMT with an existing tool NDepend [19]. NDepend is a commercial tool which is available as a limited version for free educational use. It computes some of the object oriented metrics as listed in Table 2 and other traditional metrics. Similar to the approach followed for Java, we have computed the metric values by AMT and by the commercial tool NDepend. The values are listed in the Table 2 below.

Table 2. C# Metrics

Metric	AMT	NDepend
LCOM1	0.31	X
LCOM2	0.21	0.30
LCOM3	0.32	0.40
CBO	33.42	41.85
CBO'	42.75	X
RFC	33.19	X
MPC	22.35	X
DAC	0.00	X
DAC'	0.00	X
WMC	4.42	4.42
DIT	0.00	2.71
AID	0.00	X
CLD	0.00	X
NOC	0.00	0.46
NOP	0.00	X
NOD	0.00	X
NOA	0.00	X
NMO	0.00	X
NMI	0.00	X
NMA	4.42	X

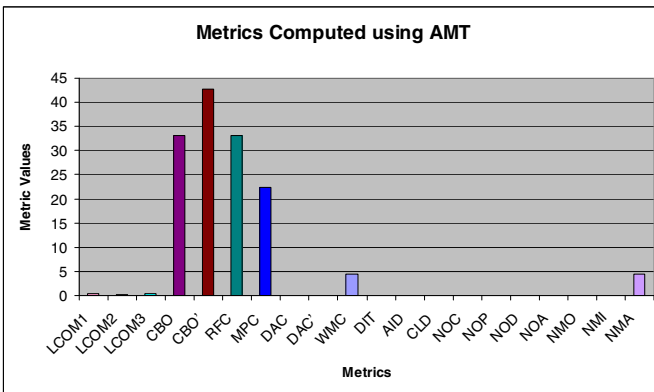


Fig. 6. C# Metric values computed by AMT

For C#, NDepend reported comparable values to that AMT computation. The metrics not available in NDepend are marked as 'X'.

The system exhibits high cohesion which is good. But the coupling is also high indicating that there might be a problem in maintaining the system as the classes interact with each other a lot.

5 Conclusion

Gathering metrics for a software development process should be non-intrusive as it requires a high level of automation for gathering product metrics as well as to support more metrics and implementation languages.

The proposed Automated Tool solves the issues present in the existing tools. It helps researchers to concentrate on their actual research work instead of spending valuable time in designing a tool for each new language or metric.

Our approach enables users to adapt existing metrics to their needs, and extend new metrics with reasonable effort. The parser module for Java and C# is available making the tool ready to use for these languages but as indicated the tool can be easily accommodated for other object oriented languages as well. Although the experiments showed that our approach is feasible and scalable, there is still a room for improvement.

References

1. Rudiger, L., Jonas, L., Welf, L.: Comparing Software Metrics Tools. In: Proceedings of the 2008 International Symposium on Software Testing and Analysis. ACM, New York (2008)
2. Abreu, F.B., Goulao, M., Esteves, R.: Toward the design quality evaluation of object-oriented software systems. In: Proceedings of the 5th International Conference on Software Quality, Austin, Texas, USA (1995)
3. Baroni, A.L., Brito, F.: An OCL-Based formalization of the MOOSE metric suite. In: Proceedings of QUAOOSE, Darmstadt, Germany (2003)
4. Baroni, A.L., Brito, F.: A formal library for aiding metrics extraction. In: International Workshop on Object-Oriented Re-Engineering, Darmstadt, Germany (2003)
5. Harmer, T.J., Wilkie, F.G.: An extensible metrics extraction environment for object-oriented programming languages. In: Proceedings of IEEE International Conference on Software Maintenance, Montreal, Canada (2002)
6. Lavazza, L., Agostini, A.: Automated measurement of UML models: an open toolset approach. *Journal of Object Technology*, 115–134 (2005)
7. Wakil, M.E., Bastawissi, A.E., Boshra, M., Fahmy, A.: A novel approach to formalize and collect object-oriented design-metrics. In: Proceedings of the 9th International Conference on Empirical Assessment in Software Engineering (2005)
8. Eichberg, M., Germanus, D., Mezini, M., Mrokon, L., Schafer, T.: QScope: an open, extensible framework for measuring software projects. In: Proceedings of 10th European Conference on Software Maintenance and Reengineering, CSMR (2006)
9. Mens, T., Lanza, M.: A graph-based metamodel for object-oriented software metrics. *Electronic Notes in Theoretical Computer Science* 72 (2002)

10. Marinescu, C., Marinescu, R., Girba, T.: Towards a simplified implementation of object-oriented design metrics. In: IEEE METRICS, pp. 10–11 (2005)
11. Abreu, F.B., Melo, W.L.: Evaluating the impact of object-oriented design on software quality. In: 3rd International Software Metrics Symposium, Berlin, Germany (1996)
12. Briand, L.C., Daly, J.W., Wust, J.: A unified framework for coupling measurement in object-oriented systems: Technical report ISERN, Fraunhofer Institute for Experimental Software Engineering, Germany (1996)
13. Chidamber, S.R., Kemerer, C.F.: A metrics suite for object-oriented design. IEEE Transactions on Software Engineering, 476–493 (1994)
14. Li, W., Henry, S.: Object-oriented metrics that predict maintainability. Journal of Systems and Software, 111–122 (1993)
15. Henry, S., Selig, C.: Predicting source-code complexity at the design stage. IEEE Software, 36–44 (1990)
16. Ray, E.T.: Learning XML, 2nd edn. O’ Reily Media (2003)
17. Spinellis, D.: ckjm: a tool for calculating Chidamber and Kemerer Java metrics: Technical report, Athens University of Economics and Business, Athens, Greece (2006)
18. Java Metrics Tool,
[http://ivs.cs.uni-magdeburg.de/
sw-eng/agruppe/forschung/tools/](http://ivs.cs.uni-magdeburg.de/sw-eng/agruppe/forschung/tools/)
19. NDepend – C# Tool, <http://www.ndepend.com/>

Traceability Matrix for Regression Testing in Distributed Software Development

B. Athira¹ and Philip Samuel²

¹ Department of Computer Science, Cochin University of Science and Technology
athiramalu@yahoo.com

² Information Technology, School of Engineering, Cochin
University of Science and Technology
philips@cusat.ac.in

Abstract. Distributed software development is a process that is done across many business worksites or locations. Geographically distributed teams and cross-platform functionalities complicated the process of quality management. Software quality is essential to business success. One of the main issues distributed teams face is change management. For large changes, retesting the entire system is complex. So an automated regression testing technique is a must in distributed system. Thus identifying proper test cases is essential in regression testing. In this paper we propose a new method to find out essential test cases using traceability matrix. Traceability matrix correlates both user requirements and functional requirements with test cases of the system. By analyzing the traceability matrix we can ensure that we have covered all the required functionalities of the system. It also ensures that all the modification done to the system is tackled. It leads to the forward as well as backward traceability of the system. Our approach is an effective method to avoid redundant test cases for regression testing.

Keywords: Distributed software development, regression testing, Activity Diagrams, Traceability Matrix.

1 Introduction

Distributed development teams are becoming the norm for today's software projects. In lieu of close physical interaction, distributed teams are faced with challenges of keeping software projects on track. Earlier the term "distributed development" "did not exist. Every team member on a given project was in the same location, often with same offices. But today, the economies of "global village" increasingly motivate the most of resources across time zones and continents. Through acquisitions and mergers, geographically dispersed teams are continually divided and recombined to form projects based on developer availability and talent. Internet-enabled, repository-centric tools are available for key development process.

Although powerful, distributed development raises some challenges. How to verify the software quality and change management are some of these issues. Bug reports,

new feature requests, defect tracking, traceability etc. have to be tackled. Team members often use different testing tools, development processes and technological platforms. This increases the complexity in test management process and subsequently in regression testing. Hence the development of an efficient regression testing technique is essential.

Regression testing is the process of validating modifications introduced in a system are correct and do not adversely affect the unchanged portion of the system. During regression testing the modified elements of the system are first tested. Then the whole system needs to be retested using the existing test suite to have confidence that the modifications did not introduce new faults into the system. Because of the large size of a test suite, system retesting may be very expensive; it may last for days, or even months. One of the issues developers face during retesting of the system is ordering tests for execution. Test prioritization [1], [2], [3] is a technique that tries to address this issue.

In this paper, we propose a traceability matrix for regression test cases. First step is to model the system. Several modeling languages have been developed to model the systems, e.g., State Charts, Extended Finite State Machine (EFSM) [1] and Specification Description Language. System models can be used to generate partial code or to design test cases [9]. Here, we use UML activity diagram to model the system. We plot activity diagram for original and modified system. Then we conduct a diagram change analysis to find out the modified path. Next is to create a traceability matrix which plot test data against each functional specification. Finally we find out test data which is covering the modified part by analyzing the traceability matrix.

The rest of the paper is organized as follows. Section 2 covers some related works in this area. In section 3, plotting activity diagram and diagram change analysis is presented. Traceability matrix is in section 4 and the paper is concluded in section 5.

2 Related Work

In this section, we present some of the previous work done in the field of regression testing. Test prioritization orders tests for execution so that the test cases with the higher priority, based on some criteria, are executed first. Variety of criteria for test prioritization can be used. For example, tests can be ordered to achieve selected code coverage at the fastest rate. In [2] several test prioritization criteria were presented and their influence on the improvement of the rate of fault detection was investigated. An empirical study of the application of several greedy, meta-heuristic and evolutionary search algorithms is presented in [4].

Most of the test prioritization methods [3] are code-based, i.e., information about the source code is used to prioritize tests. Estimate program change based on comparisons of binary representations of code [5]. Other methods include history-based test prioritization [7] test cost and fault severity based test prioritization [6] and model-based test prioritization in which effectiveness of test prioritization with respect to early fault detection [8].

Wu and Offutt [10] presented an analysis model for modeling the server side components and the client server interactions. Wu et. Al. extended their model to cover intercomponent connections between different server side components in [11]. Several techniques for prioritizing test cases and report empirical results measuring the effectiveness of these techniques for improving rate of fault detection in [12]. A Safe regression test selection (RTS) technique has been integrated into a systematic method that monitors distributed code modifications and automates the RTS in [13]. [14] proposes a scenario-based functional regression testing, which is based on end-to-end (E2E) integration test scenarios.

3 System Modeling

This section describes how to model the original system and modified system. Then find out the difference between original and modified model. Most of the previous work described above in the section 2, in this area is code-based. Here in this work we use standard UML modeling, which makes the system independent of different platform. The Unified Modeling Language (UML) is used to specify, visualize, and modify the artifacts of any system. It offers a standard way to visualize a system's architectural blueprints. UML activity diagram is used to model the system. We plot activity diagram for original and modified system. We use node to represent activity node and link to represent transition between two activities. We also annotate the node with information such as input-output parameters of activities.

3.1 An Example

We illustrate with an example of Electricity Bill Payment System. Customers are classified as different slabs according to the range of usage of power. The bill depends on the customer usage. It is as per the following rule of usage, 50ps for below 100 usages, and 60ps for each additional usage of 100 to 200 and 20ps for each additional usage of above 200. The Fig. 1. depicts the activity diagram of this process.

3.2 The Flow of Process

The flow of process is described as follows. The nodes are numbered as activities A1 to A7. A1 receives the customer information. Cond1 checks whether customer's usage type is domestic or commercial. Cond2 checks whether usage is less than or equal to 100 or not for domestic customers. Cond3 checks whether usage is less than or equal to 100 or not for commercial customers. Cond4 checks whether usage is in between 100 and 200 for domestic customers. A2 calculate bill for usage in between 100 and 200. A3 calculate bill for usage above 200. A4 throws a message as commercial customers of usage below 100 are eligible subsidy. A5 calculate 10% discount for commercial customers of usage above 100. Cond5 checks whether the final bill is above 100 or not and A6 give 5% discount for the bill above 100. A7 return the bill.

3.3 Diagram Change Analysis

Suppose a new rule is decided to implement as the range of usage 100-200 changed to 100-150 for domestic customers. That is domestic customers of above 150 usage changed to higher slab. The customers come under 100-150 have to pay 60ps for each additional bill and the customers above 150 have to pay 20ps for each additional bill. The following Fig. 2. shows this modification. The change done to the system is shown in dotted line.

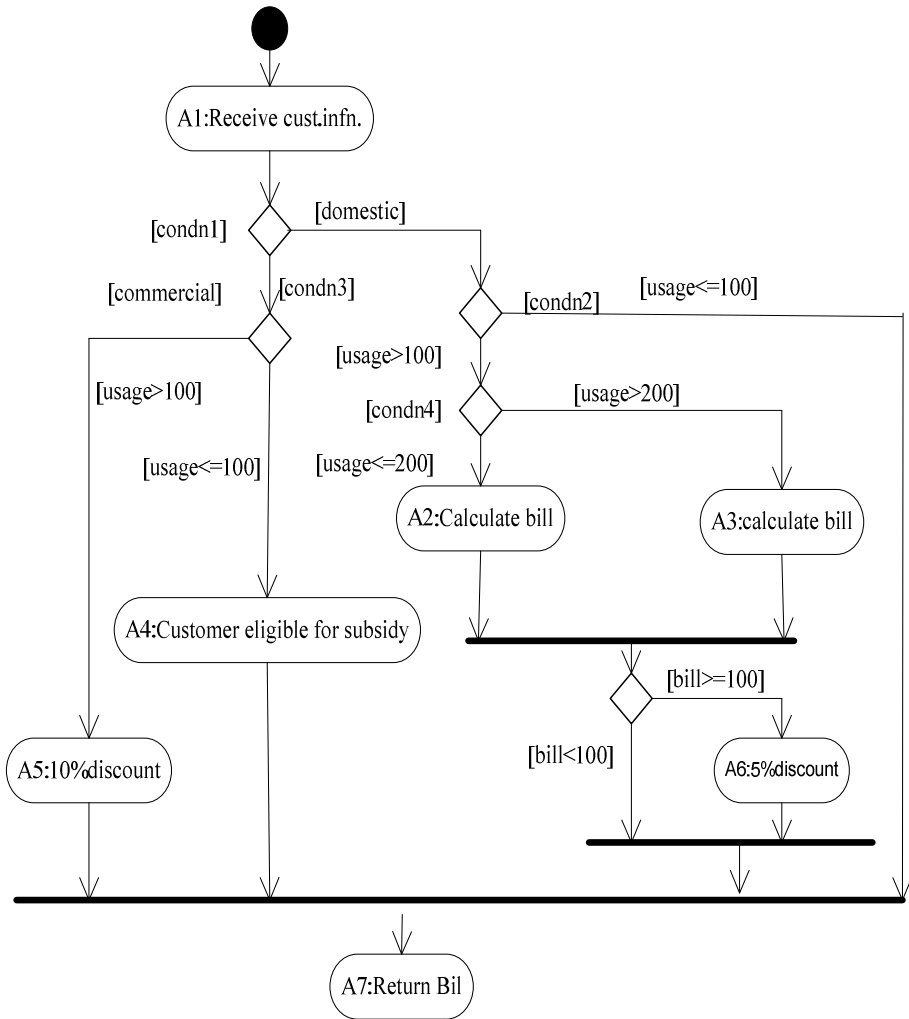


Fig. 1. Activity diagram for original program

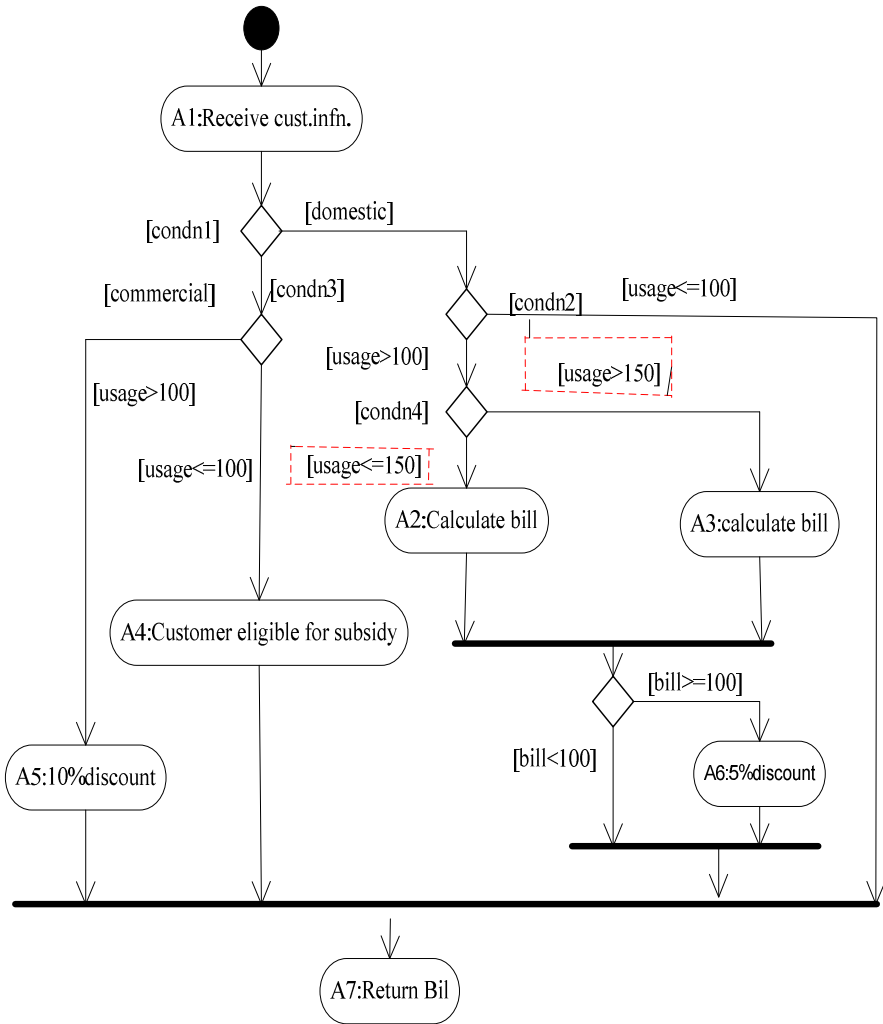


Fig. 2. Activity diagram for modified program

The difference between original program and modified program is find out by diagram change analysis. We import XML document for both the diagrams. XML is used to store and transport data. XML document contains the information about each nodes of activity diagram .By parsing the xml document we get much information about each node. Here we use SAX parser to parse the xml document. SAX parser operates on each piece of XML document sequentially. Thus each control flow path can be retrieved. By comparing this control flow of both the diagrams, we easily get the information about the modified part. In the above mentioned example, by diagram change analysis we get the information that the change is made to the range of usage from 200 to 150.

4 Traceability Matrix

This section describes how to find out the test cases for regression testing using traceability matrix. It also avoids the redundant test cases which are used in the original system. Traceability matrix is a method used to validate the compliance of a product with requirements for that product. It is a document which maps requirements with test cases. The requirements are each listed in a row of the matrix and the columns of the matrix are used to identify how and where each requirement has been addressed. It is often used with high-level requirements and detailed requirements of the software product to the matching parts of high-level design, detailed design, test plan and test cases. It is used to manage change and provide the basis for test planning. By preparing traceability matrix we can ensure that we have covered all the required functionalities of the application in our test cases. It is also easy to track changes if we have a good traceability matrix. Thus it improves the quality of a system.

We use a set of test cases (T1 to T12) to illustrate the challenges in test case prioritization. Following Fig. 3. shows the test case data used for testing the original program.

```

Test case1 (T1) :< 50,"domestic" >
Test case2 (T2) : <60,"commercial" >
Test case3 (T3) : <90,"domestic" >
Test case4 (T4) : <100,"commercial" >
Test case5 (T5) : <100,"domestic" >
Test case6 (T6) : <120,"domestic" >
Test case7 (T7) : <150,"domestic" >
Test case8 (T8) : <170,"commercial" >
Test case9 (T9) : <190,"domestic" >
Test case10 (T10) : <200,"domestic" >
Test case11 (T11) : <200,"commercial" >
Test case12 (T12) : <220,"domestic" >

```

Fig. 3. Test case data

Let us explain the results at cond4, which is modified, for test cases T1 to T12 in Fig. 2. Test cases T1 through T8 and T11 satisfies the cond4 in modified program in the same way as executed in the original program. But test cases T9, T10 and T12 in modified program are not executed in the same way as that of original program. Thus test cases T9, T10 and T12 are covering the modified part. Traceability matrix of the modified program clearly proves this result.

Table 2. shows the traceability matrix of the modified program. It is plotted the test cases against the functional requirement specification. Put cross 'x' against each of the test case to each requirement if that particular test case is checking that particular requirement partially or completely. Functional requirement specification is shown in table Table 1.

Table 1. Functional requirement specification

Requirement ID	Description
Req1	Commercial customers with usage of 100 and below 100
Req2	Commercial customers with usage of above 100
Req3	Domestic customers with usage of 100 and below 100
Req4	Domestic customers with usage of above 100 and below and equal to 150
Req4	Domestic customers with usage of above 150

Table 2. Traceability matrix of modified program

Requirement identifiers	Req1	Req2	Req 3	Req4	Re5
Test cases					
T1			x		
T2	x				
T3			x		
T4	x				
T5			x		
T6				x	
T7				x	
T8		x			
T9					x
T10					x
T11		x			
T12					x

By analyzing the traceability matrix of modified program it is clear that the modified requirements Req4 and Req5 are checked by test cases T6, T7, T9, T10 and T12. But T6 and T7 produces same result as that of original program. Thus test cases T9, T10 and T12 need to be tested earlier than other test case.

5 Conclusion

In this paper, we presented a new approach for selecting regression test cases for distributed software development projects. Remote development of software offers several advantages, but it is also fraught with challenges. How to run tests in different software configuration with multiple platforms , how to select proper test cases and what are the test result for specific build are some among them. Our traceability matrix approach is an effective method to face these challenges. Traceability ensures

quality of the system as well as provide basis for test planning. Requirement change analysis is done easily with our traceability matrix. Since UML is a standard modeling language for concurrent and distributed system, the model which we created is independent of multiple platforms.

References

1. Rothermel, G., Harrold, M.: A safe, Efficient Regression Test Selection Technique. *ACM Transaction on software Engineering & Methodology* 6(2), 173–210 (1997)
2. Rothermel, G., Untch, R., Harrold, M.: IEEE Transactions on Prioritising Test case for regression Testing 27(10), 929–948 (2001)
3. Wong, W., Horgan, J., London, S., Agarwal, H.: A study of effective Regression Testing in Practise. In: *Proc.8th International Conf.on Software Maintenance*, pp. 214–223 (2002)
4. Li, Z., Harman, M., Hierons, R.: Search Algorithms for Regression test case prioritization. *IEEE Tran.Software Engineering* 33(4), 225–237 (2007)
5. Srivastava, A., Thiagarajan, J.: Effectively Prioritising tests in Development Environment. In: *Proc. ACM International Symposium on Software Testing and Analysis, ISSTA 2002* (2002)
6. Elbaum, S., Malishevsky, A., Rothermel, G.: Incorporating varying test costs and fault severities into Test case prioritisation. In: *Proc.23rd International Conference on Software Engineering, ICSE 2001*, pp. 329–338 (2001)
7. Kim, J., Porter, A.: A History-based Test prioritisation Technique for Regression testing in Resource constraint Environment. In: *Proc.24th International Conference on Software Engineering*, pp. 119–129 (2002)
8. Korel, B., Tahat, L.H., Harman, M.: Test Prioritisation using system models. In: *Proc.21st IEEE Intenational Conference on Software Engineering (ICM 2005)*, pp. 559–568 (2005)
9. Vaysburg, B., Tahat, L., Korel, B.: Dependence Analysis in Reduction of requirement based Test suites. In: *Proc.ACM International Symposium on Software Testing and Analysis*, pp. 107–111 (2002)
10. Wu, Y., Offutt, J.: Modeling and Testing Web-based Applications. *GMU ISE Technical ISE-TR-02-08* (2002)
11. Wu, Y., Offutt, J., Duz, X.: Modeling and Testing of Dynamic Aspects of Web Applications. *GMU ISE Technical ISE-TR-04-01* (2004)
12. Rothermel, G., Roland, H., Harrold, C.: Test Case Prioritization: An Empirical Study. *IEEE Transactions on Software Engineering*, 159–182 (2002)
13. Ruth, M., Tu, S.: Towards Automating Regression Test Selection for Web Services. In: *Proc.31st Annual international conference on Computer software and Applications (COMPSAC 2007)*, pp. 729–736 (2007)
14. Tsai, Bai, W., Paul, X., Yu, R.: Scenario-Based Functional Regression Testing. In: *Proceedings of the 25'h Annual IEEE International Computer Software and Applications Conference (COMPSAC)*, Chicago, IL, pp. 496–501 (2001)

Testing Agent-Oriented Software by Measuring Agent's Property Attributes

N. Sivakumar, K. Vivekanandan, and S. Sandhya

Department of Computer Science and Engineering, Pondicherry Engineering College.
Puducherry-605014, India
sivakumar11@pec.edu, k.vivekanandan@pec.edu,
sandhya_38@pec.edu

Abstract. Agent technology has been the subject of extensive discussion and investigation within the research community for several years, but it is perhaps only recently, it has seen any significant degree of exploitation in commercial applications. We could view Agent Oriented Software Engineering (AOSE) as a new programming paradigm that has evolved itself from Object Oriented Software Engineering (OOSE). AOSE has placed greater emphasis on agent characteristics such as the autonomy, learning, interaction, adaptability, reactivity, pro-activity etc. However, no paradigm will be completed if there were no form of measurement (metrics) to determine the efficiency and quality of its application. AOSE needs proper product metrics to enhance its stand as a lasting methodology in software engineering, just like its parent. For this reason, it is important to develop comprehensive measures of excellence to evaluate agent based software. No set of measures defining the overall quality of an agent has been developed to date. Our research is an elucidation to the need for a more encompassing definition of agent quality through metrics. We present a set of metrics that can be easily applied to a design that measures certain software quality characteristics of an agent oriented system. These metrics are “design and product” metrics for Agent Oriented system. They provide information regarding the ability of their design to match software quality. This paper first identifies the quality requirements for an agent system and secondly proposes a tool that has been developed to measure the agent metrics that quantifies and assists quality engineers to decide the efficiency and the degree of quality of Agent-oriented system.

Keywords: Agent Oriented Software Engineering (AOSE), Agent metrics, Agent-oriented System.

1 Introduction

Throughout the history of software engineering, a number of software development paradigm such as procedural, object-oriented and agent oriented were developed. The main objective of these paradigms is to develop quality software that completely satisfies the user needs. Software quality control has spurred the research on software

metric technology. The IEEE standard of Software Engineering Terms defines metric as “a qualitative measure of the degree to which a system, component or process possess a given attribute”. Since the traditional software metrics such as line of code (LOC), cyclomatic complexity etc., aims at the procedure-oriented software development and cannot fulfill the requirement of the object-oriented software, a set of new object oriented software metrics suites such as MOOD (Metrics for Object Oriented Design), C.K or MOOSE (Metrics for Object Oriented Software Engineering) and QMOOD (Quality Model for Object-Oriented Design) were proposed [1]. Research has been conducted on adapting some metrics of procedural and object-oriented software (encapsulation, data hiding, etc.) to evaluate agent-oriented software quality. As the agent characteristics [2][3] such as autonomy[4], pro-activity, reactivity, social ability [4][5], intelligence etc., differs with object characteristics, our goal set out to develop metrics exclusively targeting agent-oriented software. The increasing importance on agent-oriented software development in industries and research field has led to the development of automated tools to support agent-oriented metrics.

Our research focus is to develop an automated agent metrics tool that supports users and managers to measure the attributes [6] of the agent-oriented software and thus evaluate the quality of the software according to the specified hierarchical metric model. This tool would also serve as an aid in Post-Mortem analysis to review software design i.e. the process of looking back at a completed project's design and its development process, in order to identify those aspects where improvements can be made in future projects. Besides analyzing the effectiveness of the design of a particular system, these metrics also aid in measuring the higher level attributes of a software system like flexibility, reusability, readability, maintainability, extensibility, functionality and effectiveness.

2 Proposed Work

The main objective of this paper is to provide an automated software metrics tool to collect agent- oriented metrics for several agent programs and to store the calculated metrics in a database for future reference and analysis. This automated tool significantly improves developer's ability to identify, analyze, fix and improve quality characteristics [7] of agent-oriented software's design and implementation.

2.1 System Description

Java is used as a front-end tool to provide a user-friendly, interactive interface. The agent based projects to be analyzed have been developed using JADE [8] [9] framework and FIPA standards [10]. These projects shouldn't have any syntax errors and the code should be capable of being executed independently. The designer inputs agent oriented projects to be analyzed through the interface after placing the projects in the root directory specified by the tool. A preprocessor is designed to remove all spaces and statements that would not be useful for the purpose of metrics calculation. The result from this preprocessor is then sent to a parser that analyses the program

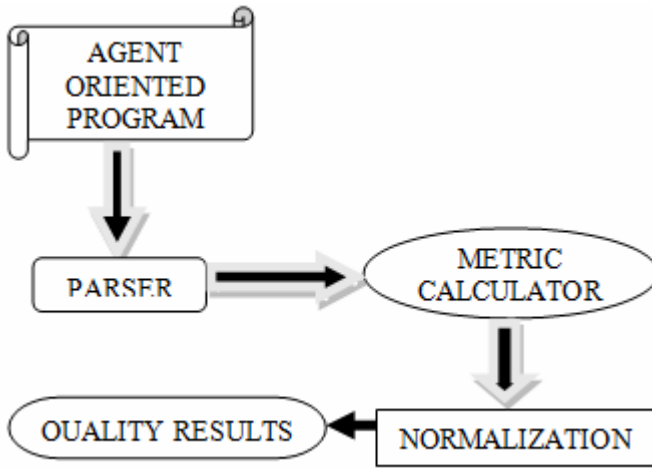


Fig. 1. System Design

and collects information needed for calculating various metrics. The information gathered from the parser is stored in text files for future use by the tool.

3 Implementation

The need to quantify distinctive features of the agent-oriented paradigm gave birth to the need for different metric suites. Though various agent-oriented metrics prevail, we propose to develop a tool that calculates the attributes of agent properties at various levels such as Interaction level, Learning level, Adaptation level, Collaboration level, Performance level, Mobility level, Specialization level and Communication level. Our implementation focuses on developing agent metric calculator tool that determines and collects agent specific metric data according to above mentioned levels. The tool is designed to evaluate metrics that relate to quality of the agent-oriented programs. The calculated metric values are stored in a database for further reference and analysis. The following are the various metric suite parameters calculated using the tool and stored in the database.

3.1 Interaction Level

It expresses the activity of an agent during interaction. Under different situation, agents might react differently with other agents and their environment. A high interaction level might indicate that the agent is able to react to multiple scenarios, thus more intelligent but complex. The lists of agent interaction level metrics are as follows:

Weighted Methods per class (WMC)

This metric measured the number of methods implemented within the agent and the sum of cyclomatic complexities of the methods. Higher value likely to indicate a complicated agent, which exhibit more behaviors depending on the environment parameters, service requests and own goals.

Number of Message type (NMT)

This metric measured the number of different type of agent message that can be resolved or catered by the agent. The more message types an agent could handle, the better it has developed its interaction capability.

$$\text{NMT} = \text{IM} + \text{OM} \quad (1)$$

Where IM and OM is the number of unique incoming and outgoing message type respectively.

3.2 Learning Level

We could determine how much the agent can learn from the system through discovery of data knowledge, the effect upon the agent, time and effort needed to maintain this knowledge and the value of this new knowledge. A strong learning level might indicate that the overall agent learning overhead would be high and a greater unpredictable agent behavior, as the behavior would be more likely to be subject to the new knowledge it has learned. The following are the agent learning level metrics:

Attribute Hiding factor (AHF)

This metric determine the visibilities of the agent knowledge. Poor information encapsulation and easy modification of the agent internal states can directly affect the agent behavior. Easy read access to the agent states may allow others to anticipate the agent action. We can also include the number of methods that are available for other agent to modify the values directly. Methods should not be available to other agents for direct attribute modification or access, this is to preserve the integrity of the knowledge obtained and derived by the agent.

Knowledge Usage (KUG)

Derive from Variable Scope, we count the average number of internal agent attributes used in the decision statements inside the agent methods. Also from the metrics value, we could determine the impact of each variables value on the agent behaviors. Variables which affect more decision making process would have a stronger influence over the agent behavior. Given more of the decision making process uses the internal states, then the agent is said to be greater affected by the learning process and might be less predictable if the values changed frequently.

Knowledge Update (KUP)

Derive from live variables, this metric count the number of statement that will update the variables in the agent. Each variable is dependent on different event occurrence, where the event would change the variable value, thus agent internal states. Some variables might be quite stable and do not changed much.

Variables Density (VD)

We use the number and data type of variables to determine the different internal states of the agents. Large number of internal states required corresponding more updates to maintain the values.

3.3 Adaptation Level

It determines the stability and complexity of the agent implementation. More complex algorithm is needed for the agent to adapt to different environment conditions and ensure its lasting impact on the new location, especially for mobile agent. It is hard to determine if the agent is able to adapt to different environment and continue to prove useful using just metrics computed from the agent code and design. However, we can determine the adaptation level by simulation or observe agent performance in the real system.

Knowledge Usage (KUG) – *Refer the previous KUG*

Knowledge Update (KUP) – *Refer the previous KUP*

Weighted Methods per Class (WMC) – *Refer the previous WMC*

Exception Handling Functionality (EHF)

This metrics measures the quality of exception handling functions that are found within the agent code. We count the exception type that is handled by the agent. High EHF value can indicate that the agent is capable to handle different environment situation more efficiently. However, this also depends on the code within the exception handling function and the type of exception that present in the agent. Some general exception type may be define even if the actual exception are of much specific class and this do not gives a concise design of the agent code.

3.4 Collaboration Level

A high collaboration can classify an agent's roles in the given tasks. Not all collaboration reduce agents effort in accomplish common task, some competitive agents might in fact hinder the process if it has been mistaken as cooperating agents. The following is the agent collaboration level metrics:

Coupling between Objects (CBO)

CBO is the number of other Classes that are coupled to the current one.

3.5 Mobility Level

It considers the efficiency relating to the agent movement. This movement also includes the message transferred by the agent. The following are the agent mobility level metrics:

Agent Executable Size (AES)

A large agent size can cause low mobility efficiency. If the agent size to be transferred is greater than the amount of data that needs to be transferred over to the local host for computation, then it may not be worthwhile to use mobile agents.

Static Message Size (SMS)

Derived from the Function Parameters Density, this metric determine the minimum size of the message that is transmitted by the agents by the number and data type of parameters sent by the message. Bigger message size might indicate more information is transferred to the receiving agent but it could create a larger overhead over the agent system such that the services performed do not justify the resources spent by the message.

Average Message Size (AMS)

AMS measures the influence of the data size of the messages sent by the agent on its communication.

3.6 Specialization Level

Agents with more behavior tend to have more roles specific, leading to lower specialization. A high specialization can lead to high performance. The following are the agent specialization level metrics:

Weighted Methods per Class (WMC) – Refer the previous WMC

Number of Roles (NR)

We measure the number of potential roles that agent must perform. The agent role(s) comes from the design phase of the system. Agents may have dynamic roles allocation or changes during its running course and this requires additional function implemented in order to meet the different roles needs. Each new role does not necessary means a new function is added, some complex roles have to accomplish more goal(s) and required more function to be added.

3.7 Communication Level

The level of conversation may view as the amount of messages that have to be transferred to and from, in order to maintain a meaningful communication link or accomplish some objectives. High communication intensity can affect the flexibility of an agent as it may mean that the agent has spent much of its resources in the handling of incoming request from other agents for its service thus making it harder to modify. It could also mean the agent has much outgoing request to other agents for their services, indicating an excessive coupling design. Agents should have minimal communication as most agents will only interact with the service providing agents and when providing services or detecting and responding to the environment changes. Agents usually communicate with the services yellow page to search for required service and thus do not required to send messages to all other agents in the system for services. The following are the agent communication level metrics:

Response for Message (RFM)

RFM measures the amount of messages that are invoked in response to a message received by the agent.

Average Message Size (AMS) – *Refer the previous AMS*

Incoming Message (IM)

IM measures the relation of incoming messages to agent communication during its lifetime.

Outgoing Message (OM)

OM measures the relationship between direct outgoing messages and agent communication during its lifetime. Higher values could indicate that the agent is dependent on other agents.

4 Result Interpretation

We developed an automated agent metric calculating tool for Agent oriented software. We tested our tool by applying 10 different agent based projects namely P1, P2... and P10. The calculated attribute level agent metrics were tabulated in Table 1. To measure the quality, the measured metrics value will be expressed in the range of 0 and 1 or [0, 1] in short. The process of transforming our index from its value into a range of 0 and 1 is called normalization. In principle, to aggregate a sequence of numbers into range of [0, 1], we need to make them positive and divide with something that is bigger than the nominator. Using this principle, we can make use any in-equality to normalize the value. In our paper the calculated metrics at each level is normalized in the range of 0 and 1 using the following formula,

$$N = d / \text{square root}(d^2+a) \quad (2)$$

Where, 'd' is the similarity between index and 'a' is the actual value. Based on the normalization interval, the quality range is fixed as mentioned in Table 2. The normalized value and their corresponding quality ranges are tabulated in Table 3. Table 3 indicates that the attribute levels of the agents such as learning, collaborating, mobility and communication for all the 10 projects are good as per the FIPA standards, whereas the interaction (P1, P3, P4, P5, P6), adaptation (P1, P2) and specialization (P4, P7) attributes of the agent are moderate and particularly P1 has very poor interaction. Fig.3 and Fig.4 shows the snapshots of our metric calculator tool interface and the output screen respectively.

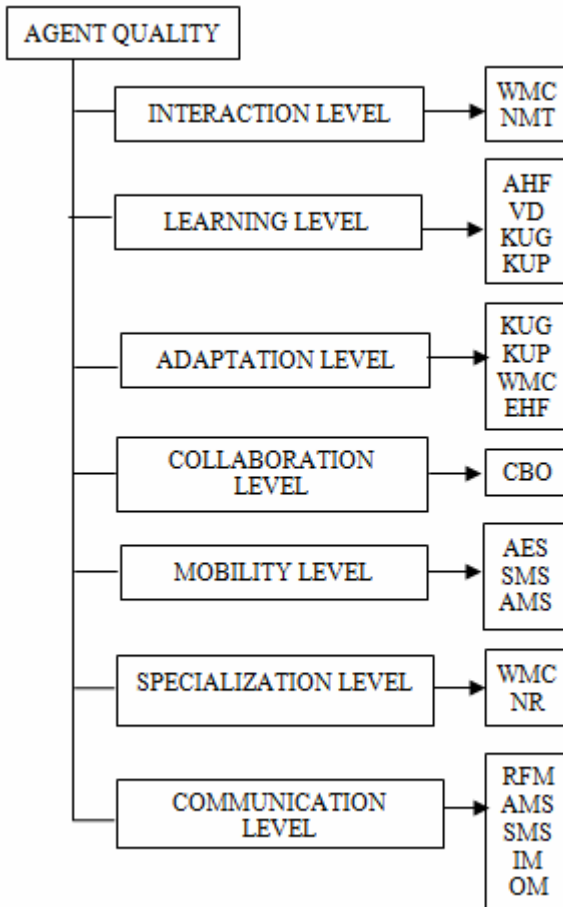


Fig. 2. Agent Quality Model

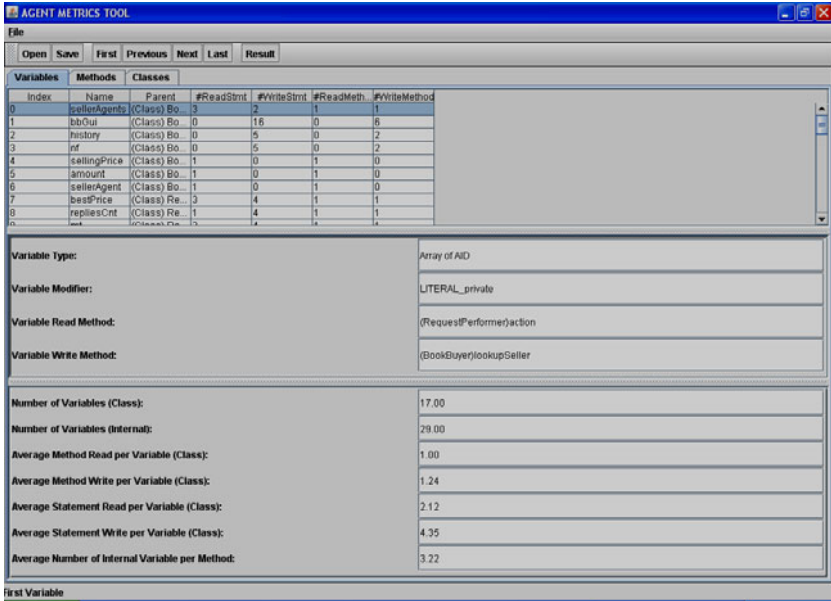


Fig. 3. Snapshot of Agent Metric Tool Calculator Interface



Fig. 4. Snapshot of Agent's Attribute Level Metric Values (Output Screen)

Table 1. Agent attribute values at various levels

Project	Interaction Level		Learning Level			Adaptation Level		Collaboration Level		Mobility Level			Specialization Level		Communication Level		
	WMC	NMT	AHF	VD	KUG	KUP	WMC	EHF	CBO	AES	SMS	AMS	WMC	NR	RFM	IM	OM
P1	0.4	4.0	1.0	7.0	1.1	4.3	0.7	1.3	3.5	9.2	51.0	17.0	0.8	2.8	1.0	3.0	3.0
P2	0.7	6.0	1.0	7.5	1.2	4.5	0.9	1.1	3.1	9.2	51.3	17.12	0.8	3.0	0.9	1.8	1.8
P3	0.4	4.3	0.8	7.2	1.1	4.1	0.8	1.2	3.0	9.0	49.5	16.5	1.0	3.0	1.0	2.0	2.0
P4	0.5	4.5	1.0	7.0	1.2	4.5	0.8	1.4	3.2	9.0	50.5	17.2	0.7	2.8	0.8	1.8	1.7
P5	0.6	5.5	1.0	7.5	1.2	4.5	0.7	1.2	3.4	9.2	5.1	17	0.9	3.0	0.9	1.8	1.8
P6	0.6	5.5	1.0	7.5	1.2	4.5	0.7	1.2	3.4	9.2	5.1	17	0.9	3.0	1.0	2.0	2.0
P7	0.2	5.0	0.9	6.0	1.0	4.2	0.6	1.1	3.6	9.5	52	18.2	0.8	3.0	0.8	1.8	1.8
P8	0.8	6.0	1.0	7.0	0.9	4.5	0.8	1.2	3.4	9.5	51.5	17.12	1.0	3.0	1.0	2.0	2.0
P9	0.8	6.0	1.0	7.0	1.0	4.5	0.8	1.2	3.4	9.2	51.3	17.12	0.8	3.0	1.0	2.0	2.0
P10	0.6	5.5	1.0	7.5	1.2	4.5	0.7	1.2	3.4	9.2	5.1	17	0.9	3.0	0.9	1.8	1.8

Table 2. Quality rating

<i>Normalized value Interval</i>	<i>Rating</i>	<i>Abbreviation</i>
0 - 0.20	Poor	P
0.20 - 0.4	Below Average	BA
0.40 - 0.60	Average	A
0.60 - 0.80	Above Average	AA
0.80 - 1	Good	G

Table 3. Normalized values of metrics at its quality at various levels

Project	Interaction Level		Learning Level		Adaptation Level		Collaboration Level		Mobility Level		Specialization Level		Communication Level	
P1	0.64	AA	0.99	G	0.78	AA	0.95	G	0.98	G	0.85	G	1.00	G
P2	0.90	G	1.00	G	0.73	AA	0.92	G	0.99	G	0.86	G	1.00	G
P3	0.72	AA	0.91	G	0.80	G	0.90	G	0.96	G	0.92	G	1.00	G
P4	0.76	AA	1.00	G	0.91	G	0.93	G	0.96	G	0.76	AA	0.96	G
P5	0.76	AA	0.99	G	0.91	G	0.94	G	0.95	G	0.87	G	0.99	G
P6	0.76	AA	0.99	G	0.91	G	0.94	G	0.95	G	0.87	G	1.00	G
P7	0.17	P	0.93	G	0.49	A	0.97	G	1.00	G	0.71	AA	1.00	G
P8	0.95	G	0.95	G	0.93	G	0.94	G	1.00	G	0.91	G	1.00	G
P9	0.95	G	0.96	G	0.93	G	0.94	G	0.99	G	0.86	G	1.00	G
P10	0.76	AA	0.99	G	0.91	G	0.94	G	0.95	G	0.87	G	0.99	G

5 Conclusion

Agent technology has become an important and meaningful research field in software industry for building complex system. Since Agent Oriented Software Engineering (AOSE) needs proper product metrics to enhance its stand as a lasting methodology in software engineering, we developed an automated agent metric tool that calculates a set of metrics that quantifies and assists quality engineers to decide the efficiency and the degree of quality of an agent-oriented system. Some of the agent attribute levels that lead to specialized metrics for agent-oriented software are Interaction, learning, adaptation, collaboration, performance, mobility, specialization and communication. These agent attribute levels greatly influence the quality of the agent-oriented software. This automated tool significantly improves developer's ability to identify, analyze, fix and improve quality characteristics of agent-oriented software's design and implementation. The availability of an easy-to-use automated tool will encourage designers and developers to measure, make improvements, and develop designs that have high internal quality characteristics. The net result will be a significant increase in agent-oriented software quality and hence better software products.

References

1. Abreau, F.B., Melo, W.: Evaluating the impact of object-oriented design on software quality. In: Proc. 3rd International Software Metrics Symposium. IEEE, Berlin (1996)
2. Jennings, N.R., Wooldridge, M.: Agent-Oriented Software Engineering. In: Proc. of the 2002 ACM Symposium on Applied Computing, Madrid (2002)
3. Foundation for Intelligent Physical Agents, FIPA Communicative Act Library Specification, Geneva, Switzerland (2002)
4. Alonso, F., Fuertes, J.L., Martinex, L., Soza, H.: Evaluating Software Agent Quality: Measuring Social Ability and Autonomy. In: Innovations in Computing Sciences and software Engineering. Springer, Heidelberg (2010)
5. Alonso, F., Fuertes, J.L., Martínez, L., Soza, H.: Measuring the Social Ability of Software Agents. In: Sixth International Conference on Software Engineering Research, Management and Applications (2008)
6. Jang, K.S.: A Preliminary Study & Development of a Metric Analyzer. National University of Singapore (2003)
7. ISO, ISO/IEC 9126-1: Software engineering- Product quality- Part 1: Quality Model, International Standard ISO/IEC 9126-1:2001 (2001)
8. Bellifemine, F., Caire, G., Greenwood, D.: Developing Multiagent Systems with JADE. John Wiley & Sons, Inc., Chichester (2007)
9. JADE, <http://jade.tilab.com>
10. FIPA, <http://www.fipa.org>

Classifier Feature Extraction Techniques for Face Recognition System under Variable Illumination Conditions

Sneha G. Gondane^{*}, M. Dhivya^{**}, and D. Shyam^{*}

Department of Electrical and Electronics Engineering,
Anna University of Technology, Coimbatore-641047
Tamilnadu

sneha_cie@yahoo.com, shyamjoe1987@gmail.com
dhivya.erts@gmail.com

Abstract. Recognition of object under uncontrolled illumination environment is imprecise and vague. A simple image preprocessing chain is taken for precept. Local binary pattern (LBP) is capable of reducing noise levels in background regions. Local ternary patterns (LTP) fragmenting less under noise in uniform regions. Gabor filter acts as a resounding filtering source for local spatial frequencies. Phase congruency is to extract the image in phase as well as in magnitude levels. The result is obtained by the KLDA based classifiers with combination of LBP and Gabor features. The above explained features are obtained from both the input and the data base image. In that the LBP and Gabor features are fused and the distance is calculated. If both the input and database images are same, the face is recognized; otherwise the face is not recognized. The simulation results exemplify the proposed technique for image with different lighting, expressions and structural defects.

Keywords: Face recognition, lightning invariance, local binary patterns, local ternary pattern, phase congruency.

1 Introduction

FACE recognition has created immense opportunity in the field of science and technology to establish a well set platform among the basic standards in this technological world. Computer analysis does not recognize various techniques which can only extract and recognize its features. Face image recognition acquired in its surrounding environment with changes in lighting or pose remains a largely unanswered dilemma. Most of these methods were initially developed with face images collected under relatively well-controlled conditions. In practice, they have complexity in dealing with the range of appearance varying from certain

* PG Student.

** Research Scholar.

approaches that can commonly occur in unconstrained situation due to illumination levels, pose, facial expressions, aging of networks and partial occlusions. This paper focuses mainly on the issue of sturdiness to illumination variations. A face corroboration system capable of implementing images for a portable device should be able to verify a client at any time.

2 Existing System

Recognition system models the structural patterns of an imaging process under difficult lightning conditions. Patterns are being processed in various stages according to their level of intensity. Specifically three main contributions are described in previous works. First Efficient preprocessing chain needed before any feature extraction techniques is widely carried on. Local binary pattern (LBP) improves the face identification and gabor wavelets which is used to filter the noise levels. Improve sturdiness by adding Kernel principal component analysis (KPCA) feature extraction during later stages of analysis. The result is obtained by the KPCA based classifiers. The above explained features are obtained from both the input and the data base image. If both the input and database images are exactly same, then the face is recognized otherwise the face is not recognized. The image with difficult lighting, same expressions can be identified.

2.1 Existing System Flow Diagram

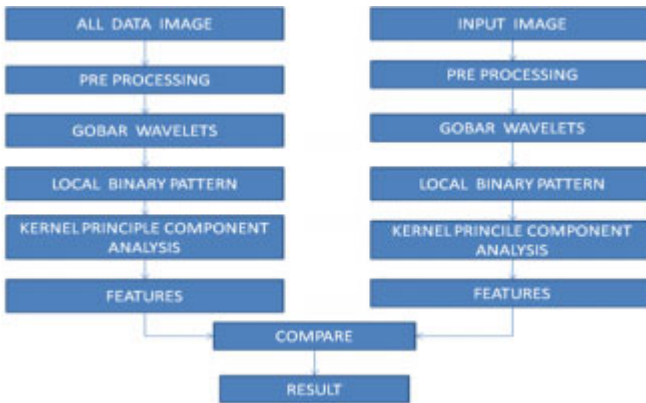


Fig. 1. Flowchart of existing system

2.1.1 Pre Processing

Pre processing is an important stage of face detection system. It is better to present a simple and efficient preprocessing chain which eliminates the structural effects of illumination while still preserving the necessary details that are needed for recognition. Preprocessing stage includes: Gamma correction, Difference of Gaussian filtering, Masking and Contrast Equalization.

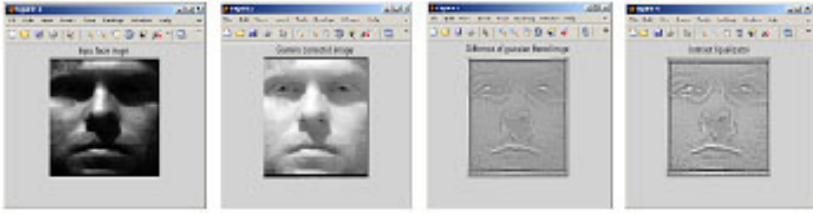


Fig. 2. Various stages of preprocessing: input image; Gamma corrected image; image after DoG filtering; image after contrast normalization

2.1.1.1 Gamma Correction. Gamma Correction improves the structural formation of arrays which is under the region of darkness. Shading effects will not undergo any changes during its illumination stages. Data images are compressed in bright regions and improve the quality level of images being processed. Gamma correction transforms the gray level into object level based on the intensity of light being reflected on the data images.

2.1.1.2 Difference of Gaussian (DOG) Filtering. DOG will suppress the highest spatial frequencies based on the reduction level of both aliasing and noise signals. Decaying of the structural images causes simplification of internal progression of the filtering steps further more. High-pass filtering removes both useful and source information gathered during the structural progression. Level of analysis during simplification process will greatly influence the overall system performance.

2.1.1.3 Masking. Masking is the track of achieving better imaging process in the facial regions (hair style, beard) that are felt to be unconditional or source variables which are needed to be masked should be applied at this point. Masking is the optional process during analysis.

2.1.1.4 Contrast Equalization. The final stage of our preprocessing chain rescales the image intensities to standardize a robust measure of overall contrast or intensity variation. It is important to use a global estimator because the signal indication level typically contains extreme values produced by highlighters, small dark regions such as nostrils, garbage at the image borders, etc.

2.1.2 Local Binary Patterns

LBP method provides very good results, both in terms of speed as well as discrimination performance. Because of the way the texture and shape of images is described, the method seems to be quite robust against face images with different facial expressions under different lightening conditions, image rotation and aging of persons. The middle pixel value is used as a threshold frequency which compares with the eight neighboring channels of a pixel.

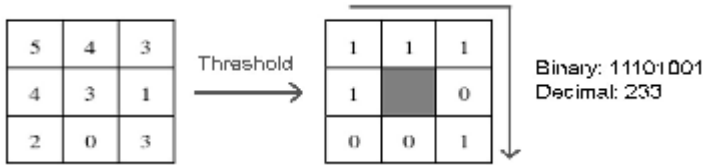


Fig. 3. Illustration of the basic LBP operator

If a neighbor pixel has a higher gray value than the center pixel, then one is assigned to that pixel, else it gets a zero.

2.1.3 Gabor Features

Gabor filter suits the certain level of frequencies which allows a particular band to pass and helps in local spatial frequency distribution. Images are optimally oriented in each part of the object with equal frequency and spatial domains. Gabor feature also encodes the face over broader range of scales. The Gabor feature is a frequency based solution technique under the region of sinusoidal plane wave and orientation phase.

2.1.4 Kernel Principle Component Analysis

Kernel principal component analysis is a combination of local binary patterns and Gabor feature through which the distance is calculated. Using a kernel, the original linear operations of PCA are done by reproducing Kernel Hilbert Space with a non-linear mapping.

3 Proposed System

Face recognition system has created advance level of technology in the field of medical science. Everyday new advancement from the previous stages is carried out for the welfare of the society. When compared to previous works, solution with phase congruency provides precise and accurate results. Existing system is the base for our proposed system. Various stages are being carried out based on the existing system. Local binary patterns and local ternary patterns are being processed with the Gabor filters in order to reduce the noise level in unconditional regions. Gabor filter gives an optimal solution for structural analysis in spatial and frequency domains. Phase congruency is to extract the images in phase as well as in the magnitude levels. The kernel linear discriminant analysis is used to extract the feature using combination of local binary patterns and Gabor features through which distance is calculated. The input image and data base image compared and face is recognized.

3.1 Flowchart for Feature Extraction Techniques

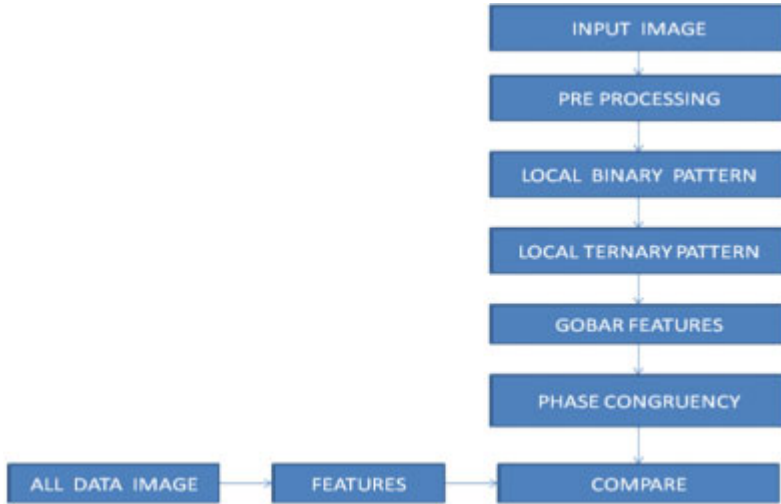


Fig. 4. Flow chart of feature extraction

3.1.1 Local Ternary Pattern

Local ternary patterns reduce the illumination effects. In this we itself choose a threshold value and compare with the neighbor pixels value. If the central value is greater than neighbor pixel value, then it will show as 1. If the central value is in between than neighbor pixel value, then it will show as 0. If the central value is lesser than neighbor pixel value, then it will show as 0.

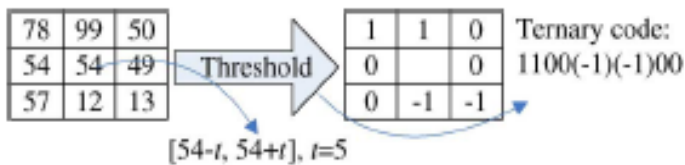


Fig. 5. Illustration of the basic LTP operator

3.1.2 Phase Congruency Features

Gradient-based operators who search for their effective approaches look under the maximum intensity gradient. Intensity gradient will undergo some structural failures during detection and localize a large proportion of features within images. Phase congruency is used to identify and analyze the corners and edges from the

structural images. Unlike the edge detectors, which identify the sharp changes from the database congruency model to detect points of order in the phase spectrum. Phase component is the most important part of phase congruency rather than its magnitude.

4 Simulation Results and Discussion

The simulations are carried out in MATLAB (7.11.0.584). A detailed survey and analysis of previous works is carried out and the simulation parameters are chosen. The Input Image sequence and the Data base image sequence are listed in figure 6 (a) and 6(b) respectively. In table 1, the comparison results are given. It is evident that the proposed technique is less time consuming/time constraint. The classifier extraction technique requires nearly less than half the time consumed by the previous work.

Table 1. Comparison

Parameter	Existing System	Proposed System
Time	40.2015 sec	16.1929 sec
Recognition	The image with difficult lighting, same expressions can be identified	The image is recognized with different lightning, facial expressions, any spot or scratches.
Techniques	Local binary patterns, Gobar features, kernel principle component analysis	Local ternary patterns, phase congruency



Fig. 6 (a). Input images



Fig. 6 (b). Data base images

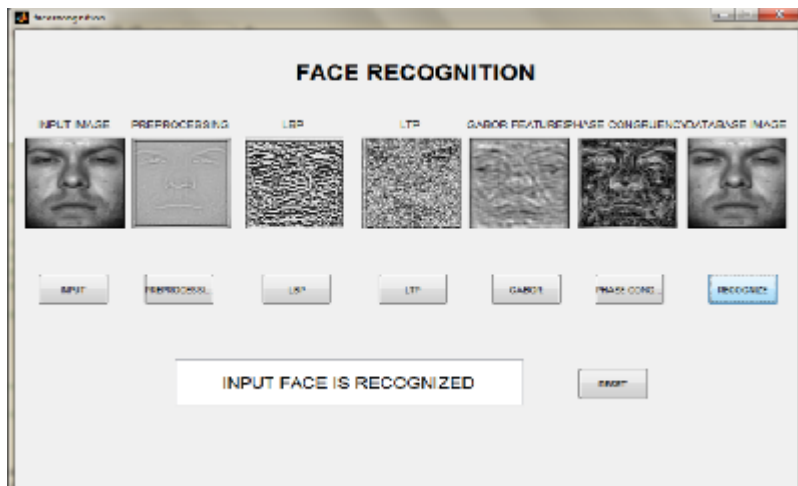


Fig. 7. Face Recognition under blurred image

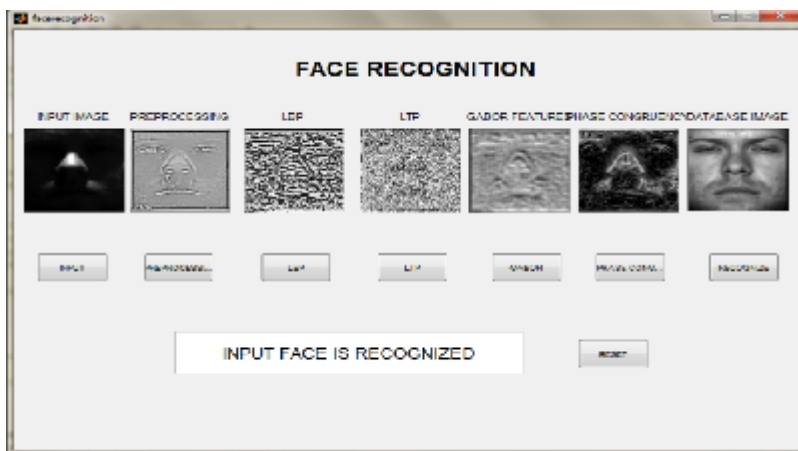


Fig. 8. Face Recognition under darker lighting

In Fig 7, the input and data base image is recognize under difficult lightning conditions when the images are same. In Fig 8, the input and data base image is recognize when the input image is selected under dark lightning. The input image passes through various feature extraction techniques , compare with data base image and recognize the image.

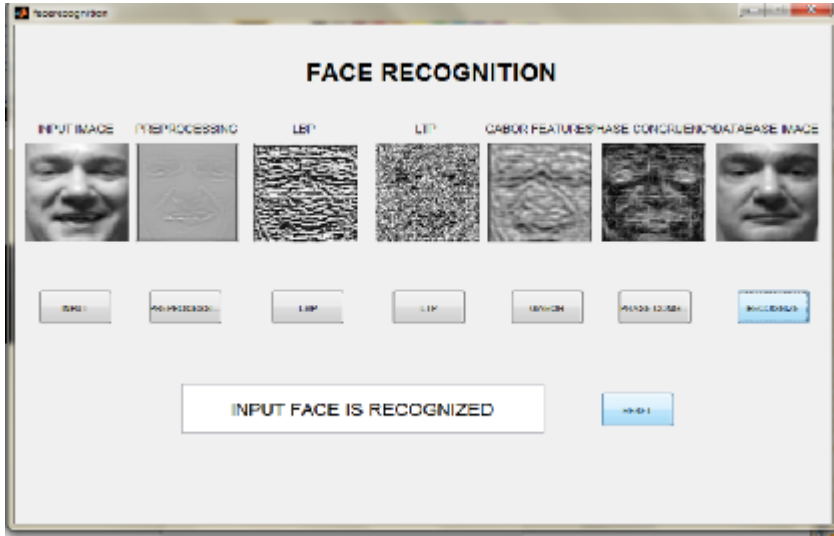


Fig. 9. Face recognition under different expressions

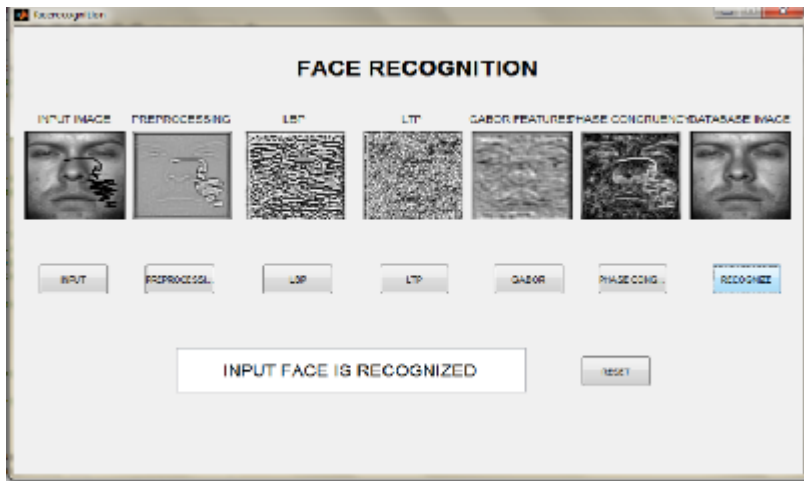


Fig. 10. Face Recognition under structural defects

Fig 9 In this figure the input and data base image is recognize when the input image with expressions under difficult illumination is selected. The input image passes through various feature extraction techniques, compare with data base image and recognize the image. Fig 10: In this figure the input and data base image is recognize when the input image with scratches under difficult illumination is selected. The input image passes through various feature extraction techniques, compare with data base image and recognize the image. Fig 11: In this figure if the input image is selected which is not present in data base image than after comparing input and data base image the image will not recognize.

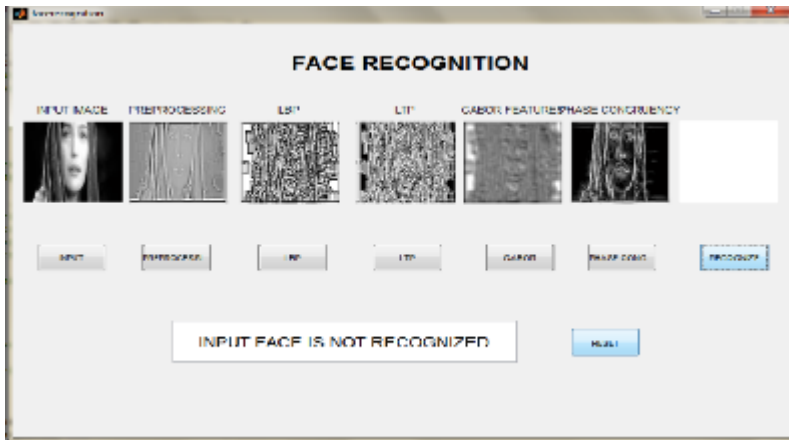


Fig. 11. Face Recognition out of data base

5 Conclusion

In this paper a classifier based feature extraction technique is implemented for face recognition under uncontrolled illumination with different lightning conditions. The result is obtained by the KLDA based classifiers with combination of LBP and Gabor features. Local ternary patterns reduce noise level since using the own threshold value makes the feature to extract and recognize the face, precisely. Phase congruency can map 0 to 360 degrees to 0 to 255 gray values, which help to extract the minute features and is applicable in phase and time domains. Time consumption is lesser in proposed method compared to the previous methods. From the simulated image results the efficiency of the classifier technique is hence shown.

References

1. Tan, X., Triggs, B.: Enhanced Local Texture Feature Sets for Face Recognition under Difficult Lighting Conditions. *IEEE Transactions on Image Processing* 19(6) (June 2010)
2. Kovese, P.: Phase Congruency Detects Corners and Edges. School of Computer Science & Software Engineering The University of Western Australia Crawley, W.A. 6009 pk@csse.uwa.edu.au
3. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition: A Literature Survey. *ACM computing surveys* 35(4), 399–458 (2003)
4. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7) (July 1997)
5. Kernel principal component analysis, http://en.wikipedia.org/wiki/Kernel_analysis from Wikipedia, the free encyclopedia
6. Malik, J., Sainarayanan, G., Dahiya, R.: Corner Detection using Phase Congruency Features. In: *International Conference on Signal and Image Processing (ICSIP)*, December 15-17 (2010)
7. Smith, A.R.: Gamma Correction. Technical Memo (September 9, 1995)
8. Gondane, S.G.: Recognition and Detection System in Human Face Under Variable Illumination Conditions. In: *Third National Conference on Recent Trends in Computation and Signal Held on Amrita School of Engineering*, March 1 (2011)
9. Gupta, R., Mittal, H.P.A.: Robust Order-based Methods for Feature Description, Department of Computer Science and Engg. Indian Institute of Technology Madras, Chennai INDIA-600036
10. Bangalore, N., Young, R., Birch, P., Chatwin, C.: Tracking Moving Objects Using Band pass Filter Enhanced Localization and Automated Initialization of Active Contour Snakes Industrial Informatics Research group. In: *Department of Engineering and Design, University Of Sussex, Brighton*
11. Ding, W., Li, X., Wang, W.: Robust Image Corner Detection Using Local Line Detector and Phase Congruency Model. In: *WASE International conference on Inst. of Electr. Eng., Yanshan Univ., Qinhuangdao, China Information Engineering (ICIE)*, (2010)
12. Ahmadian, A., Mostafa, A.: An efficient texture classification algorithm using Gabor wavelet. *Dept. of Biomedica System & Medical Phys., Tehran Univ. of Medical Sci., Iran Engineering in Medicine and Biology Society 1* (September 17-21, 2003); (Date of Current Version: April 5, 2004)
13. Zhang, C., Guo, K., Yu, G.: An Analysis of Gabor Wavelet Algorithm for Tracking Driver's Feature Point. In: *International Conference Key Lab. of Intell. Transp. Syst. Technol., Nat. Center of ITS Eng. & Technol. Electrical and Control Engineering (ICECE)*, Beijing, China (November 2010)
14. Ekenel, H.K., Fischer, M., Tekeli, E., Stiefelhagen, R.: A Local Binary Pattern Domain Local Appearance Face Recognition. In: *Signal Processing, Communication and Applications Conference* (2008)
15. Verschae, R., Ruiz-del-Solar, J., Correa, M.: Face Recognition in Unconstrained Environments: A Comparative Study, Department of Electrical Engineering, Universidad de Chile

Bispectrum Analysis of EEG in Estimation of Hand Movement

Aditya Saikia and Shyamanta M. Hazarika

School of Engineering, Tezpur University

Tezpur, India

{aditya10, smh}@tezu.ernet.in

Abstract. Bispectrum analysis is presented to analyze electroencephalogram (EEG) signals recorded during two states of motor acts i.e. during *imagination* and *observation* of hand movements. EEG signals are recorded from primary hand areas i.e. from electrode position C3 and C4. This paper emphasizes the nonlinear behavior of EEG signal and we figure out, by bispectrum analysis it is possible to estimate spontaneous rhythm in the EEG during imagination and observation of hand movements. The results show that the location of bispectral peaks in bifrequency are quite different depending on the EEG signals in different motor acts.

Keywords: Electroencephalogram (EEG), Bispectrum.

1 Introduction

To characterize electroencephalogram (EEG) signals, to date, both time and frequency domain analysis have been applied [1,2]. The gold standard technique has been the estimation of power spectral density or power spectrum [3]. In power spectrum estimation, only linear mechanisms governing the process are investigated since the phase between the frequency components are suppressed [25]. However, EEG possesses nonlinear behavior. One of the characteristics of non-linear behavior is that various frequencies *combine* to form a new combination of sum and difference frequencies and there exists quadratic phase coupling that occurs when two waves interact and generate a third wave with a frequency equal to the sum of frequency of the first two wave. Thus power spectrum, which start with an assumption of linearity, gaussianity and minimum phase systems [3] is in-appropriate for EEG signals. Although, conventional power spectrum technique may be able to identify the rhythm, power and the phase of the EEG signal; it is incapable of detecting quadratic nonlinear coupling signatures among the underlying sinusoidal components in the same spectrum. This paper explores the use of a signal processing technique called bispectrum in estimation of spontaneous rhythm in the EEG during imagination and observation of hand movements. The study of bispectrum analysis for EEG signals is motivated by the following:

- a. Bispectrum analysis detect the presence of quadratic phase coupling.
- b. Bispectrum analysis as higher order spectra, retain the phase relations between the spectral components.

In this work we study EEG signals for two states of motor acts i.e. during *imagination* and *observation* of hand movements. This is done in accordance with Porro et al. [17] two kinds of motor acts can be generated by normal subjects: an internal or first person process in which subjects *feel* themselves executing movements, and an external or third-person process involving a *visual* representation of actions. Mu rhythm, overlapping with the alpha-rhythm and the lower part of the beta-rhythm band [16] is generated by sensorimotor cortex [14]. Furthermore, it has been reported that area F5 in the monkey premotor cortex have neurons (termed mirror neurons) that discharge when the monkey performs specific hand actions and also when it observes another individual performing the same or a similar action [11]. Hari et al. [18] were the first to show that when a human subject observes hand actions there is a desynchronization of the premotor cortex; (similar, although weaker) to that occurring during active movements. Also, a number of neuroimaging studies have demonstrated a distinct temporal sequence of brain activations during the observation of hand actions [12,13,15]. The sequence starts in the occipital cortex, moves to the inferior parietal lobule, then to Brocas area and finally the motor cortex. In addition to these studies, Muthukumaraswamy et al. [14] stated that the Mu rhythm particularly influence downstream modulation of primary sensorimotor areas by mirror neuron activity, and the frontal mirror neuron system is the only network in the region of sensorimotor cortex that has been identified as responding to observed hand actions. Based on these empirical studies, we have planned to record EEG signal activity during the imagination and observation of hand movement from primary hand areas i.e. electrode position C3 and C4 as defined by the 10-20 system of electrode placement where electrodes are placed overlying sensorimotor cortex.

The rest of the paper is organized as follows: section 2 covers the background on Bispectrum Analysis. In section 3, we present materials and methods of experimentation. Section 4, discuss the results and we make final comments in section 5.

2 Bispectrum Analysis

2.1 Bispectrum of EEG Signal

Bispectrum have been used for EEG analysis. Bispectrum analysis of EEG was first reported by Huber et.al [5]. Huber et al. [6] in 1997 demonstrated the use of bispectrum analysis for the visual evoked potential. Muthuswamy et al. [4] reported the bispectral analysis of burst patterns in EEG. Barnett et al. [19] studied the interaction of component waves in EEG in awake and asleep subjects. Sigl et al. [7] have used the bispectral parameters of EEG to quantify the depth of anesthesia in a patient. Ning et al. [20] distinguished the different stages of vigilance using the bispectrum of EEG recorded from rat hippocampus.

Marceglia et al. [22] investigate the existence of possible nonlinear interactions between local field potentials (LFPs) rhythm characterizing the output structure of the basal ganglia, the globus pallidus internus, by means of bispectral analysis. Although several studies have established the applicability of bispectrum technique for EEG signal, an attempt to explore the possibility of its application on BCI have been rather limited. Only one published report by Zhong et al [8] exists; who have developed bispectrum based feature extraction method for classification of left /right-hand motor imaginary task.

2.2 Background Theory

The mathematical analysis of bispectrum have been described in a number of papers [3,20,25]. Here only a brief review is presented. The bispectrum is the expectation of three frequencies: two direct frequency components and the third is the conjugate frequency of the sum of those two frequencies of a random signal [3]. Given the Fourier frequency components, $X(f)$, of the signal $X(n)$, the bispectrum, $B(f_1, f_2)$, can be estimated using the Fourier-Stieltjes representation [9,10].

$$B(f_1, f_2) = E \langle X(f_1)X(f_2)X^*(f_1 + f_2) \rangle \quad (1)$$

where $X^*(f)$ is the complex conjugate of $X(f)$ and $E\langle \rangle$ is the statistical expectation operator. $X^*(f_1 + f_2)$ represents the correlation among different frequency in $(f_1 + f_2)$ plane. The bispectrum is a complex quantity and therefore it has magnitude and phase. Therefore, for each bispectrum value, (f_1, f_2) can be represented as a point in a complex space: $\text{Re}[B(f_1, f_2)]$ versus $\text{Im}[B(f_1, f_2)]$, thus defining a vector. Thus phase is determined by the angle between the vector and the positive real axis.

Unlike the power spectrum, which suppress the phase information bispectrum exclusively measures the correlation of phases between the frequency components f_1, f_2 and $(f_1 + f_2)$. Phase information offers the bispectrum an advantage over the power spectrum method. As a matter of fact, bispectrum is the only spectral method which can provide the phase information of signal [24]. Bispectrum analysis, is also able to detect and quantify existence of the significant quadratic phase coupling between the different frequency components of the signal [5]. In EEG signal analysis, frequencies that contributed by quadratic phase coupling, i.e. generated due to quadratic phase is indicated by a peak in the bispectrum at the bifrequency $B(f_1, f_2)$.

3 Materials and Methods

3.1 Subjects

For the experiments, we choose five right-handed subjects between 20 to 25 years of age. Among the five subjects, one is female and four are male subjects. The subjects were all volunteers.



Fig. 1. The experimental setup: A subject, performing tasks, with the Electro-Cap for EEG recording

3.2 EEG Data Acquisition and Representation

An Electro-Cap was used to record from two sites C3 and C4 (defined by the 10-20 system of electrode placement). The reference electrodes were placed on ear lobes. Eye blinks were detected by means of a separate channel of data recorded from two electrodes placed on outer canthus of left and right eye. The EEG signals were sampled at 200Hz with a 0.3-100 Hz band pass filter.

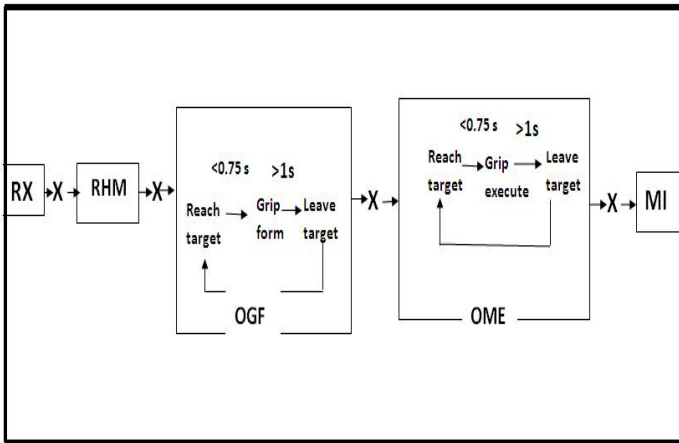


Fig. 2. Experimental tasks sequence and timing

The experimental protocol consist of five different task blocks. Subjects were asked to perform task / observe task on the presentation of an audio cue (both at the start and end of session) and data was recorded for 10 seconds. A audio cue marked the end of the trial. Experimental tasks sequence and timing with

3s inter task period (X) is depicted in Fig. 2. The following five task blocks were observed and subsequently performed during the experiment (in separate runs of 10 trials):

- Relax (RX): subjects were asked to relax with their eyes closed.
- Right hand motor imaginary (RHM): subjects were asked to imagine opening and closing the right hand
- Observation of grip form (OGF): subjects observed right hand of confederate of the experimenter (who seated right to the subject) in a position to perform power grip.
- Motor execution (OME): subjects observed right hand of confederate of the experimenter clinch a bottle i.e execute power grip and return to the rest position.
- Motor imagery (MI): Subjects were asked to imagine right hand movement to clinch the bottle at a rate similar to the previous task without actually doing it.

With 200 Hz sampling rate, each 10 second trial produces 2,000 sample per channel. Further off-line analysis, a bandpass filter (Butterworth IIR, 6th order, lower cut-off frequency 8Hz and upper cut-off 30 Hz) was designed to isolate Mu rhythm. The filtered EEG signals per subject per channel were averaged over the trials. Trials containing eye blinks were discarded.

4 Results

The acquired EEG signals after averaging in time domain were analyzed in bifrequency domain using bispectrum. Bispectrum were computed using a 512 points Fast Fourier Transform (FFT) with Hanning windows and 50% overlap of time segments. In our preliminary investigation, we have shown that bispectrum of EEG during relax state i.e. for task RX exhibits phase coupling between the frequency range 10-15 Hz, that is fall in the alpha range. This is true for all subjects. Fig. 3 shows the 3-D plot of the bispectrum of EEG for the electrode position C3 when subjects are at relax state. It is seen that the contour plot of the bispectrum indicates a peak at location (12 Hz, 12 Hz), i.e. in the alpha range of frequency (Fig. 4).

Fig. 5 to Fig. 8 depicts the ability of bispectrum in detecting non-linear phase coupling between alpha and beta rhythms during imagination (Task RHM and MI) of hand movement for the electrode position C3. In the experiment, we perform two motor imaginary tasks: RHM and MI; before and after visual representation of the motor acts (OGF and OME) respectively. We were interested to know whether visual representation of motor acts make any difference during motor imagination. Without any such difference, RHM and MI would have the same bispectrum. However, we have observed that for hand movement imagination tasks, task RHM shows bispectrum peak at locations [10Hz, 10 Hz] for the electrode position C3. For the electrode position C4, the bispectrum peak at [9Hz, 9Hz]. For task MI, the bispectrum peaks at [11Hz, 10Hz] for C3 and

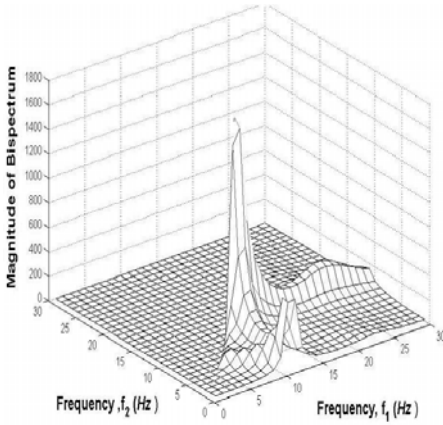


Fig. 3. 3D bispectrum during the relax state (Task-RX)

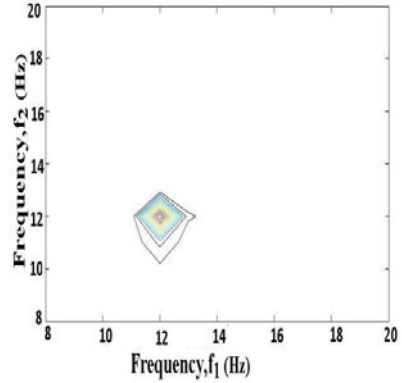


Fig. 4. Contour plot of bispectrum during relax state (Task-RX)

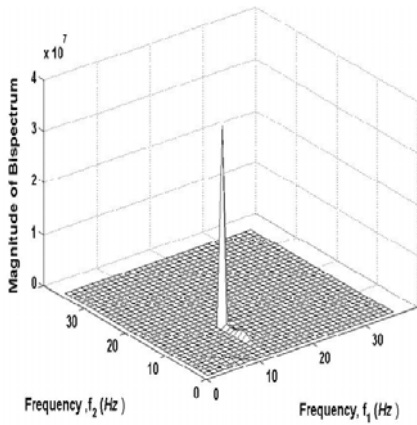


Fig. 5. 3D bispectrum during the Task-RHM (Average over the all subjects)

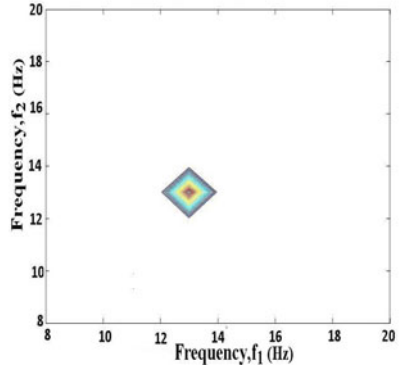


Fig. 6. Contour plot of bispectrum during the Task-RHM

[16Hz, 16Hz] for C4 electrode position. The non-linear interaction of alpha and beta rhythms during tasks, for the electrode position C3 and C4 are summarized in Table 1. Another primary interest of our experiment was to study motor control plan subsequent to visual motor imaginary. We examined the change that occur in the alpha and beta rhythms during the visually guided hand movement representation (OGF and OME).

Table 1. Bispectrum peaks $B(f_1, f_2)$ during five tasks in bifrequency domain

Tasks.	C3	C4
RX	(12,12)Hz	(10,10)Hz
RHM	(10,10)Hz	(9,9)Hz
OGF	(8,8)Hz	(14,14)Hz
OME	(9,8)Hz (16.5,16.5)Hz	(10,10)Hz, (17,17)Hz
MI	(11,10)Hz	(16,16)Hz

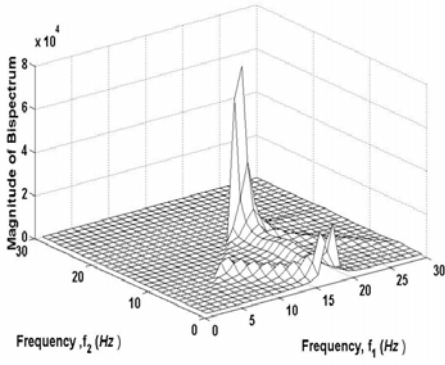


Fig. 7. 3D bispectrum during the Task-MI (Average over the all subjects)

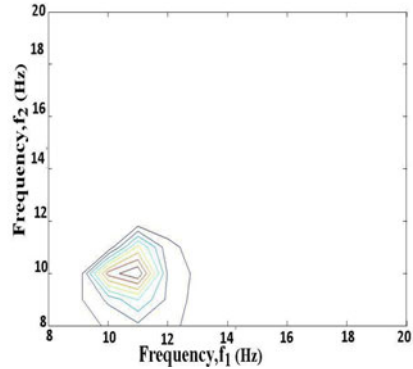


Fig. 8. Contour plot of bispectrum during Task-MI

Fig. 9 shows the contour plot of the bispectrum for the electrode position C3 during the period in which subjects observed another hand (third-person process) touch the object and engaged in a position to perform the grip (OGF) i.e task that represent the action without accompanying motor execution. Bispectrum peak at [8 Hz,8Hz] for the electrode position C3 and for the electrode position C4 shows non-linear interaction at [14Hz,14Hz] as seen in Fig. 10.

Fig. 11 shows the contour plot of the bispectrum for the electrode position C3 during the period in which subjects observed another hand (third-person process) perform hand movement accompanying motor execution (OME). The task indicate peaks at locations [9 Hz, 8 Hz] and [16.5 Hz, 16.5 Hz] for electrode position C3 and at locations [10 Hz, 10 Hz] and [17 Hz, 17 Hz] for electrode position C4 (Fig. 12).

The task OGF which represented motor act with *action possibility* shows bispectrum peaks which are different from the task OME which represented execution of motor act. Bispectrum of EEG signals for the electrode position C3 and C4 is able to distinguish visual representation of motor acts.

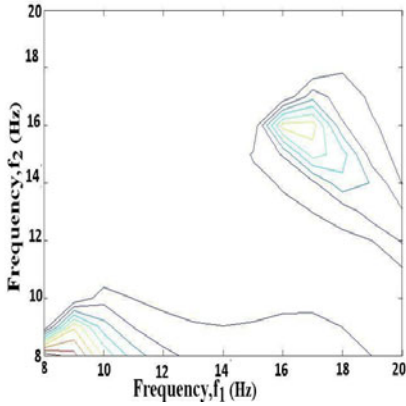


Fig. 9. Contour plot of bispectrum during the Task-OGF for the electrode position C3

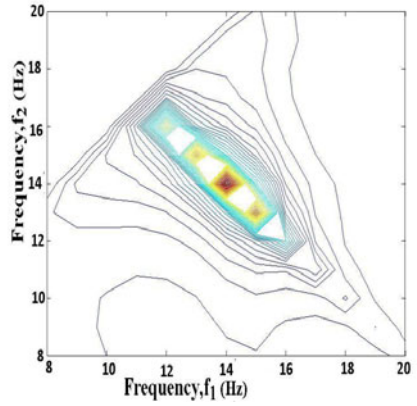


Fig. 10. Contour plot of bispectrum during the Task-OGF for the electrode position C4

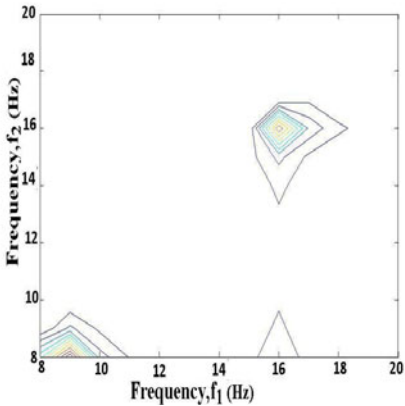


Fig. 11. Contour plot of bispectrum during the Task-OME for the electrode position C3

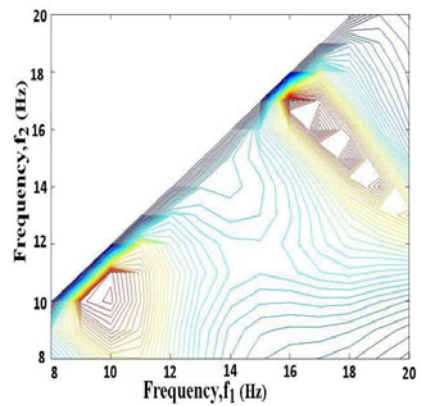


Fig. 12. Contour plot of bispectrum during the Task-OME for the electrode position C4

The results indicate bispectrum analysis of EEG provide an appropriate way to evaluate mental representation during observation and imagination of hand movement. From the results we can draw following conclusions:

- Bispectrum analysis reflects specific frequency content for different mental tasks.
- The location of significant bispectral peaks in bifrequency are quite different depending on the EEG signals in different kinds of mental tasks.

- The different bispectral peaks (for tasks with and without prior visual representation) have reinforced our belief that visual representation of motor acts make difference during motor imagination.

5 Final Comments

The main objective of this research was to analyze EEG signals using bispectrum to evaluate mental representation during observation and imagination of hand movement. Bispectrum peaks in bifrequency domain are different depending on the EEG signals during observation and imagination of hand movement. Even though the present study had a limitation that only five subjects with 10 trials were conducted (and this may not be statistically significant); from the analysis we can clearly see that bispectrum peaks in bifrequency domain are different depending on the EEG signals during observation and imagination of hand movement. This is expected to be extremely useful technique to predict the instant of transition of the mental state and may be useful for neurobionic prosthetics. This is part of ongoing research. The presented results need also to be improved by the calculation of the bispectrum using short time FFT or any other time frequency representation of the EEG signals.

Acknowledgments. We acknowledge the financial support received from DIT, Govt of India through its project *Design and Development of Cost-effective Biosignals Controlled Prosthetic Hand* under Grant No. 1(9)/2008-ME & TMD.

References

1. Pardey, J.: A review of parametric modeling techniques for EEG analysis. *Medical Engineering & Physics* 18(1), 2–11 (1996)
2. Jung, T.P.: Estimating alertness from the EEG Power spectrum. *IEEE Transaction on Biomedical Engineering* 4(1), 60–69 (1997)
3. Mendel, J.M.: Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications. *Proc. IEEE* 79, 278–305 (1991)
4. Muthuswamy, J., Sherman, D.L., Thakor, N.V.: Higher-order spectral analysis of burst patterns in EEG. *IEEE Transactions on Biomedical Engineering* 46(1), 92–99 (1999)
5. Huber, P.J., Kleiner, B., Gasser, T., Dumermuth, G.: Statistical methods for investigating phase relations in stationary stochastic processes. *IEEE Transactions On Audio Electroacoust* AU-19, 78–86 (1971)
6. Huber, P.J., Kleiner, B.: Bispectrum analysis of visuly evoked potentials. *IEEE Engineering In Medicine And Biology* 16(1), 57–63 (1997)
7. Sigl, J.C., Chamoun, N.G.: An introduction to bispectral analysis for the electroencephalogram. *Journal of Clinical Monitoring and Computing* 10(15), 392–404 (1994)
8. Zhou, S., Gan, J.Q., Sepulveda, F.: Classifying mental tasks based on features of higher-order statistics from EEG signals in brain-computer interface. *Information Sciences* 178, 1629–1640 (2008)

9. Kim, Y.C., Powers, E.J.: Digital bispectral analysis and its applications to non-linear wave interactions. *IEEE Transactions on Plasma Science PS-7(2)*, 120–131 (1979)
10. Collis, W.B., White, P.R., Hammond, J.K.: Higher-order spectra: The bispectrum and trispectrum. *Mechanical Systems and Signal Processing* 12(3), 375–394 (1998)
11. Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G.: Action recognition in the premotor cortex. *Brain* 119, 593–609 (1996)
12. Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J., Rizzolatti, G.: Cortical mechanisms of human imitation. *Science* 286, 2526–2528 (1999)
13. Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., Fazio, F.: Localisation of grasp representations in humans by PET: 1. Observation versus execution. *Exp. Brain Res.* 111(2), 246–252 (1996)
14. Muthukumaraswamy, S.D., Johnson, B.W., McNair, N.: Murhythm modulation during observation of an object-directed grasp. *Journal Cogn. Brain Res.* 19, 195–201 (2004)
15. Nishitani, N., Hari, R.: Temporal dynamics of cortical representation for action. *Proc. Natl. Acad. Sci.* 97, 913–918 (2000)
16. Pfurtscheller, G., Neuper, C.: Motor imagery activates primary sensorimotor area in human. *Neuroscience Letters* 239, 65–68 (1997)
17. Porro, C.A., Francescato, M.P., Cettolo, V., Diamond, M.E., Baraldi, P., Zuiani, C., Bazzocchi, M., di Prampero, P.E.: Primary Motor and Sensory Cortex Activation during Motor Performance and Motor Imagery: A Functional Magnetic Resonance Imaging Study. *The Journal of Neuroscience* 16, 7688–7698 (1996)
18. Hari, R., Forss, N., Avikainen, S., Kirveskari, E., Salenius, S., Rizzolatti, G.: Activation of human primary motor cortex during action observation: a neuromagnetic study. *Proc. Natl Acad. Sci.* 95, 15061–15065 (1998)
19. Barnett, T., Johnson, L.C., Naitoh, P., Hicks, N., Nute, C.: Bispectrum analysis of electroencephalogram signals during waking and sleeping. *Science* 172, 401–402 (1971)
20. Ning, T., Bronzino, J.D.: Bispectral analysis of the rat EEG during various vigilance states. *IEEE Transactions On Biomedical Engineering* 36, 497–499 (1989)
21. Niekas, C.L., Raghuvver, M.R.: Bispectrum estimation: A digital signal processing framework. *Proc.IEEE* 75(7), 869–891 (1987)
22. Marceglia, S., et al.: Interaction Between Rhythms in the Human Basal Ganglia:Application of Bispectral Analysis to Local Field Potentials. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 15(5), 483–492 (2007)
23. Andreassi, J.L. (ed.): *Psychophysiology:human behavior and physiological response.* Lawrence Erlbaum Associates Inc., Mahwah (2000)
24. Golush, T.M. (ed.): *Waste management research trends.* Nova Science Publishers, Bombay (2008)
25. Huber, P.J., Kleiner, B., Gasser, T., Dumermuth, G.: Statistical methods for investigating phase relations in stationary stochastic processes. *IEEE Transactions on Audio and Electroacoustics* 19, 78–86 (1971)
26. Raghuvver, M.R., Niekas, C.L.: Bispectrum estimation:A parametric approach. *IEEE Transactions on Acoust, Speech, Signal Processing* 33, 1213–1230 (1985)

Wavelet Selection for EMG Based Grasp Recognition through CWT

Aditya Saikia, Nayan M. Kakoty, and Shyamanta M. Hazarika

School of Engineering, Tezpur University
Tezpur, India
{adity10,nmk,smh}@tezu.ernet.in

Abstract. This paper details a strategy of discriminating grasp types using surface electromyogram (EMG) signals, which has the potential to be applied for controlling advanced prosthesis for extreme upper limb amputees. We have investigated the classification of six basic grasp types used during 70% of daily living activities. The feature vector for EMG based grasp recognition was derived using continuous wavelet transform (CWT). The proper wavelet basis function was selected through computation of entropy of the preprocessed EMG signals and wavelet transform coefficients of six different wavelet families: Gaussian, Daubechies, Morlet, Mayer, Mexicanhat and Symlet. Based on this, Gaussian wavelet function has been concluded to be possessing maximum informations about grasp types. Experimental results have validated our hypothesis that the CWT coefficients having entropy values close to the entropy of preprocessed EMG signals possesses maximum informations about the grasp types. Classification was through one vs. all multi-class support vector machine with linear kernel following preprocessing and maximum voluntary contraction normalization of EMG signals. We have achieved an average recognition rate of 80% (using the Gaussian wavelet function) cross validated through 10-fold cross validation.

1 Introduction

Surface electromyogram (EMG) is the non-invasive electrical recording of muscle activity from the surface. It is closely related to the strength of muscle contraction and the obvious choice for control of prostheses. Many multi fingered hand prosthesis controlled using surface EMG are currently available that are limited to a few hand postures or a simple proportional estimation of force.

Although advanced prostheses have been developed, the control of these devices requires a considerable training and a great attention during grasping activities. Further, control is non intuitive in the sense that user is required to learn to associate muscle remnants actions to unrelated postures of the prosthesis. Towards this end, an intelligent control architecture for prosthetic hands manipulating grasping operations holds promise. This work represents an ongoing research, investigating EMG based grasp types classification for developing an intelligent controller for advanced prostheses.

An extensive list of basic grasp types has been reported by M. R. Cutkosky [21] and C. MacKenzie [26]. We have selected six basic grasp types: power, pinch, precision, oblique, hook and palm-up, used during 70% of daily living activities [11]. Figure 1 shows the grasp types under study. EMG signal classification

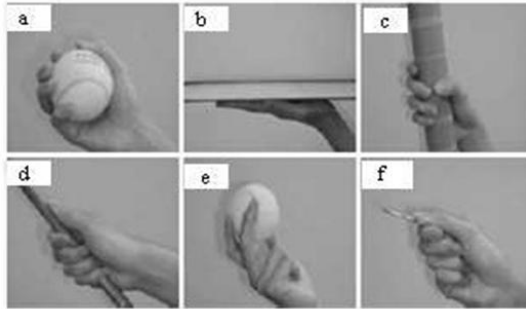


Fig. 1. Grasp Types:(a) power,(b) palm-up, (c) hook, (d)oblique, (e)precision, (f) pinch

to detect various upper-limb movements have been used for many applications including hand prosthesis control and human computer interface for last two decades. Hiraiwa et al. 1989 has classified five finger movements based on single channel EMG using fast Fourier transform analysis [6]. Classification of four grasp modes using principal component analysis and Mahalanobis distance function based on four channel EMG signals has been reported by Vuskovic in 1995 [15]. Nishikawa et al. [12] has reported ten forearm motions discrimination based on two channel forearm electromyogram using real time learning method . Using EMG Fuzzy system i.e ISO-FUZ classification method, Chan et al. has classified four hand functions based on single channel EMG [22]. An accurate and computationally efficient means of classifying surface EMG signal patterns have been the subject of considerable research efforts in the recent years where having effective signal features is crucial for reliable classification [18]. Crawford et al. shows the control of four degrees of freedom (DOF) robotic arm based on seven channel EMG with 90% accuracy using linear kernel support vector machine for classification of eight different hand movements [27]. Neuro-electric interface for replicating virtual joystick and computer keyboard based on eight channel EMG using Hidden Markov Models has been demonstrated by Wheeler and Field in 2003 [10]. Design and development of an underactuated three fingered prosthetic hand has been presented by Massa et al. [14]. A method for EMG pattern classification of hand movements for fourteen Korean characters using self organising feature maps with a 90% recognition rate has been proposed by Eom et al. [19]. Despite serious research in the field of *rehabilitation robotics*, not much has been achieved for grasps manipulation by prostheses. Ferguson and Dunlop in 2002 discussed the development of a system for identification of grasp types with an success rate of 75% to 80% [5]. Recently, classification of three grasps involved in

daily living activities with a success rate of 93-95% (using 10 channel EMG) has been reported by Castellini et al. [23].

In recent years, wavelet transform with its ability to encompass time variations in non periodic signals, time-frequency multi-resolution representations are particularly adapted to analyze EMG signals [7,20,13]. In this paper, we propose to use continuous wavelet transform (CWT) to extract EMG patterns. The work presented in [25,3,8] have demonstrated the success of CWT approach in bio medical signal analysis. The ability of CWT to extract features from the signal is dependent on the appropriate choice of the mother wavelet function. There is no intuitive way to know which base wavelet to choose. We use entropy measure for selection of optimum wavelet function for feature extraction. Aim is to find the wavelet coefficients that possess maximum information about the grasp types from acquired EMG signals. By definition, the entropy is a measure of uncertainty of information in a statistical description of a system. It is relevant to mention here that, Yu et al. [9] used entropy to present the characteristics of EMG signal of finger flexing motions. Al-Nashash et.al [1] applied entropy measure for EEG analysis. EMG classification reported in this paper is through one vs all (OVA) multi-class support vector machine (SVM) with linear kernel function and validated through k-fold cross validation.

The rest of the paper is structured as follows: In section 2 the adopted methodology for EMG acquisition, preprocessing followed by its normalisation is described. Section 3 states the basics of continuous wavelet transform. In section 4 entropy estimation of EMG signals and its wavelet transform coefficients is described. The proposed architecture is detailed in section 5. Section 6 discusses the experimental results. The concluding remarks are made in section 7.

2 Methodology

2.1 EMG Acquisition

Four subjects aged between 25 to 48 years took part in this study. Two pairs of Ag-AgCl electrode were placed on the forearm muscles as tabulated in Table 1. EMG signals were collected while the subjects were performing the six grasp types as shown in Figure 1. The EMG signals were acquired for a period of 400 ms with 5 kHz sampling frequency. Figure 2(a) shows a raw EMG signal. A total of forty eight EMG signals were used in the experiment.

2.2 Pre-processing

Two channel EMG signals generated by six grasp types were typically in the range of 10 Hz to 10 kHz and subjected to further processing. The raw EMG signals were filtered with a 10Hz to 2 kHz band pass filter and notch filtered at 50 Hz to remove unwanted base line drift and the power line interface. The EMG signals obtained after preprocessing is called integrated EMG (IEMG). Figure 2(b) shows a IEMG signal.

Table 1. Placement of EMG Electrode

Electrode Number	Electrode Leads	Specific Muscle
Electrode 1	Lead1	Extensor Digitorum Muscle
	Lead2	Flexor Digitorum Muscle
Electrode 2	Lead1	Flexor Carpi Ulnaris Muscle
	Lead2	Extensor Carpi Radialis Longus Muscle

Normalization

Signals obtained through the surface electrode is easy to get influenced by the skin layer thickness, cross talk from other body signals, change in electrode position and electrode size. To reduce the influence of these physical parameters during each trial of signal collection, signals were normalized with maximum volunteer contraction (MVC) as reference point [17]. The normalized Root Mean Square (RMS) of IEMG signals obtained as follows:

Normalized RMS of EMG =

$$\frac{RMS(IEMG_i) - MinimumRMS(IEMG)}{MaximumRMS(IEMG) - MinimumRMS(IEMG)} \tag{1}$$

where

$IEMG_i = i^{th}$ sample value of IEMG signal.

$RMS(IEMG_i)$ = root mean square value of i^{th} sample of IEMG.

Maximum $RMS(IEMG)$ = maximum RMS value of IEMG.

Figure 2(c) shows a normalized IEMG (nIEMG) signal.

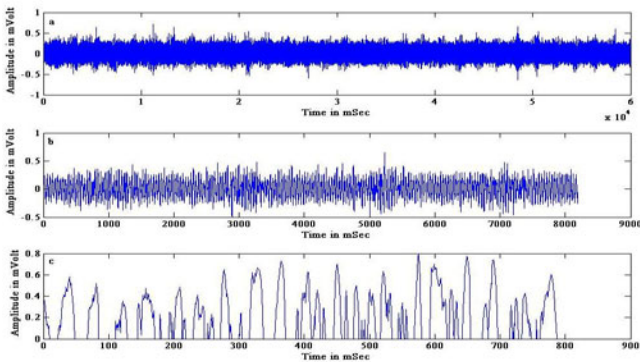


Fig. 2. (a) Raw EMG (b) IEMG (c) nIEMG

3 Continuous Wavelet Transform

The purpose of continuous wavelet transform is to decomposed a signal into localized contributions characterized by scale parameter. The continuous wavelet transform of a signal $f(x)$ is defined as an inner product of the signal and the wavelet bases as follows:

$$W(s, b) = \langle f(x), \psi_{s,b}(x) \rangle \quad (2)$$

Where $\psi_{s,b}(x)$ is referred to as wavelet bases and $W(s,b)$ is referred to as wavelet transform coefficient of signal $f(x)$. The $\psi_{s,b}(x)$ can be formed from a basic wavelet $\psi(x)$ by a series of scaling and shifting operations. The wavelet base is defined as:

$$\psi_{s,b}(x) = 1/\sqrt{s} \cdot \psi((x - b)/s) \quad (3)$$

Where $s > 0$ and b are any real numbers. The variable s indicates the scale of the particular basis function and the variable b specifies its shift operation. Using the wavelet bases in equation (3) the wavelet transform defined in equation (2) can be computed as

$$W(s, b) = 1/\sqrt{s} \cdot \int_{x=-\text{inf}}^{\text{inf}} f(x) \cdot \psi((x - b)/s) \quad (4)$$

The CWT given by equation (4) is the convolution of signal with the wavelet function shifted over the entire signal defined by the wavelet scale [16]. The transform coefficients produce by this process are the correlation of the basis function with the signal.

4 Entropy

CWT based feature extraction for efficient classification of grasp types is dependent on the appropriate choice of the mother wavelet function. We proposed entropy measure to select an optimum wavelet function. The entropy is a measure of uncertainty of information in a statistical description of a system. The entropy H for discrete random variable X is defined as [21]:

$$H(X) = - \sum_i P(X = a_i) \log_2 P(X = a_i) \quad (5)$$

Where i are the possible values of X . To select the optimum wavelet function in representing the EMG signals of grasp types, we compare entropy measure of normalized signal of six class of grasp types with entropy measures of the decomposition coefficients of wavelet functions: Gaussian, Daubechies, Morlet, Mayer, Mexicanhat and Symlet. Figure 3 shows the entropy values of nIEMG signal for six grasp types under study and of wavelet transform coefficients. Based on this the Gaussian wavelet function have been hypothesized to possess the maximum information about the grasp types.

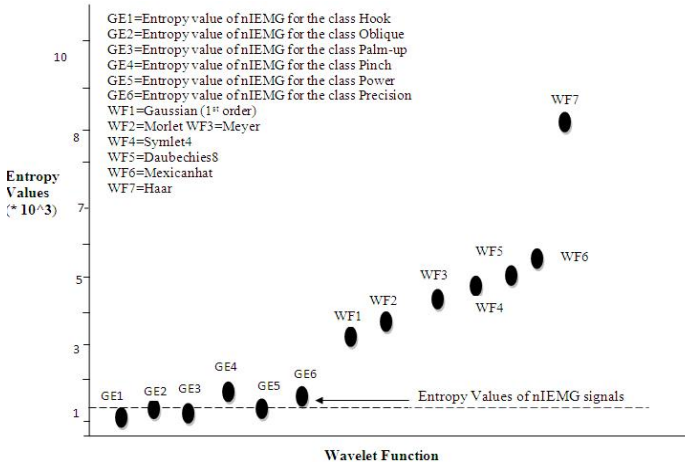


Fig. 3. Entropy values of nIEMG signals and wavelet transform coefficients

5 Proposed Architecture

Figure 4 shows the proposed architecture for EMG based grasp classification through CWT. It comprises of the EMG unit, Normalization unit, Feature Extraction Unit, Entropy Estimation unit and SVM based multi class classifier followed by k-fold cross validation unit.

The raw EMG signals obtained from the subjects for six grasp types were fed into the EMG Unit. The IEMG signal obtained at the output of the EMG Unit were normalized in Normalization Unit. The nIEMG were fed into the CWT based Feature extraction unit. The features extraction unit is followed by the Entropy estimation unit. The Entropy estimation unit works on the basis of the hypothesis discussed in section 4 of this study. The entropy of wavelet transform coefficients obtained through different wavelet families at the output of Feature extraction unit were estimated. The entropy of the wavelet transform coefficients were compared with that of the nIEMG. The wavelet transform coefficient with entropy value closest to that nIEMG were fed into the SVM based multi class classifier. The results obtained from the classifier are cross validated in the k-fold cross validation unit.

5.1 SVM Classifier

The ability of SVM for grasp classification has been studied in [4]. Introduced by Vapnik, SVM try to find the optimal decision boundary by maximizing the margin between the boundaries of different classes [24]. SVMs was originally developed for two-class classification problem. One vs. all (1-vs-A), One vs. One (1-vs-1), Directed Acyclic graph SVM (DAG-SVM) etc. are extended SVM

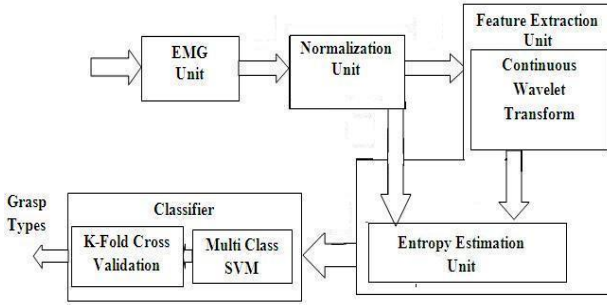


Fig. 4. Proposed EMG based Grasp Classification Architecture

classification method for multiclass. In this study, particularly 1-vs-A method carried out for classification of grasp types.

Suppose that there are n classes of sample to classify, then for 1-vs-A SVM, there will be n classification function required to be constructed. Thus in the 1-vs-A strategy, the i^{th} SVM is trained with all training samples of the i^{th} class with positive labels, and the rest samples with negative labels. In the process of classification, test sample belongs to the class that has the largest value of classification-function of class of $x = \arg \max(w^i)x + b^i$, where w denotes the weight vector and b is the bias term. The classification steps of 1-vs-A SVM for six grasp types are: a) Select all the training samples of the one of the grasp types and label them as +ve samples; label all the training samples of other group types as -ve samples; use all positive and negative samples as input to train a SVM and then, corresponding classification planes are obtained. Label the SVM as SVM1; representing that SVM1 is used for differentiating the grasp type1 from other five grasp types. b) Repeat the previous step for other five grasp types, finally six of the SVMs are obtained as SVM1..SVM6.

The classification performance of the SVM classifier was cross validated through k-fold cross validation. The k-fold cross validation was done with $k=10$.

6 Experimental Result

The estimated entropy of wavelet coefficients obtained through different wavelet functions for six grasp types are shown in table 2. The entropy value of nIEMG signals for six grasp types are shown in table 3.

Hypothesis Validation

The increasing order of entropy values of wavelet transform coefficients results into the following sequence:

$$Haar > Mexicanhat > Daubechies8 > Symlet4 > Meyer > Morlet > Gaussian \quad (6)$$

Table 2. Entropy Measure of Wavelet coefficients

Wavelet Function	Average Entropy Values
Gaussian	$3.07*10^3$
Morlet	$3.46*10^3$
Meyer	$4.22*10^3$
Symlet4	$4.25*10^3$
Daubechies 8	$4.33*10^3$
Mexicanhat	$5.12*10^3$
Haar	$8.60*10^3$

Table 3. Entropy Measure of nIEMG Signals for Grasp Types Under Study

Grasp Types	Entropy Values
Hook	$1.01*10^3$
Oblique	$1.04*10^3$
Palm-up	$1.04*10^3$
Pinch	$1.26*10^3$
Power	$1.04*10^3$
Precision	$1.19*10^3$

The average grasp recognition rate obtained using the different wavelet function coefficient are shown in table 4. The arrangement wavelet functions into an increasing order of grasp recognition rate result into the following sequence:

$$Haar < Mexicanhat < Daubechies8 < Symlet4 < Meyer < Morlet < Gaussian \quad (7)$$

Table 4. Average grasps recognition rate of Wavelet functions

Wavelet Functions Recognition Rate	
Gaussian	80%
Morlet	76.72%
Meyer	74.1%
Symalet 4	73.81%
Daubichies	72.92%
Mexicanhat	70.75%
Haar	62.04%

From sequence (6) and (7), it is clear that wavelet function coefficients having entropy values close to that of the nIEMG signal produces higher recognition rate and wavelet function coefficients having entropy values far from that of nIEMG results into lesser recognition rate. From these results, we hypothesized that *wavelet function coefficients having entropy values close to the entropy values of nIEMG possesses maximum informations about the grasp types*. Based on these experimental results, Gaussian wavelet function coefficients is reported to be possessing maximum information about the grasp types producing an average recognition rate of 80%.

7 Conclusion

A method for recognition of grasp types based on EMG signals is presented. Continuous wavelet decomposition coefficients are used as the feature for

classification. Six basic grasp types have been identified. This is achieved in a single step classification through a multi class linear kernel SVM classifier. We have hypothesized that, CWT coefficients having entropy values close to the entropy of preprocessed EMG signals possesses maximum information about the grasp types. In our study, we achieved an average recognition rate of 80% with Gaussian wavelet decomposition coefficients. It is relevant to mention here that in our previous study [28], we reported that an accuracy of 86% was obtainable with sum of wavelet decomposition coefficients of haar function based feature for a proposed architecture. For same architecture. [28], the present study gave poor accuracy rate of 62% for the same wavelet function. This is because that accuracy rate is heavily influenced by wavelet decomposition coefficients based feature selection. In the present study, EMG data from only four subjects were investigated which may not be statistically significant. Limited numbers of data in each class of grasp types might influence the classification results. Further study is required to explore these influences. While classification of grasp types is the thrust of this paper, the ultimate goal is to develop an intelligent controller for upper limb prostheses discriminating natural grasping operations; this is part of ongoing research.

Acknowledgments. We acknowledge the financial support received from DIT, Govt of India through its project *Design and Development of Cost-effective Biosignals Controlled Prosthetic Hand* under Grant No. 1(9)/2008-ME & TMD.

References

1. Al-Nashash, H.A., Paul, J.S., Thakor, N.V.: Wavelet entropy method for EEG analysis: Application to global brain injury. In: 1st International IEEE EMBS Conference on Neural Engineering, pp. 348–351. IEEE, Los Alamitos (2003)
2. Lee, K.W., He, T., Ilhan, H.T., Linscott, I., Olgin, J.E.: Feature extraction of the atrial fibrillation signal using the continuous wavelet transform. In: Proceedings of 26th EMB, pp. 275–278. IEEE, Los Alamitos (2004)
3. Bitar, A.F., Madi, N., Ramly, E., Saghir, M., Karameh, F.: A portable midi controller using EMG-based individual finger motion classification. In: Proceedings of the BIOCAS, pp. 138–141. IEEE, Los Alamitos (2007)
4. Bitzer, S., Smagt, P.: Grasp recognition from myoelectric signals. In: Proceedings of Australian Conference on Robotics and Automation, pp. 82–84 (2002)
5. Ferguson, S., Dunlop, G.R.: Learning EMG control of a robotics hand: towards active prostheses. In: Proceedings of Robotics and Automation, pp. 2819–2823. IEEE, Los Alamitos (2003)
6. Hiraiwa, A., Shimohara, K., Tokunaga, Y.: EMG pattern analysis and classification by neural network. In: Proceedings of International Conference on Systems, Man and Cybernetics, pp. 711–719. IEEE, Los Alamitos (1989)
7. Jiang, M., Wang, J.J.D., Wang, R.C.: A method of recognizing finger motion using wavelet transform of surface EMG signal. In: Proceedings of EMB, pp. 2672–2674. IEEE, Los Alamitos (2005)

8. Zajdlík, J.: The preliminary design and motion control of a five-fingered prosthetic hand. In: Proceedings of the INES, pp. 202–206. IEEE, Los Alamitos (2006)
9. Cha, K., Yu, K., Shin, H.: Maximum likelihood method for finger motion recognition from sEMG signals. In: Proceeding of ICBME, pp. 452–455. Springer, Heidelberg (2009)
10. Wheeler, K.R., Field, M.: Device control using gesture sensed from EMG. In: Proceedings of International Workshop on Soft Computing in Industrial Applications, pp. 21–26. IEEE, Los Alamitos (2003)
11. Vecchi, F., Micera, S., Carrozza, M.C., Sabatini, A.M., Dario, P.: A sensorized glove for applications in biomechatronics and motor control. In: 6thIFESS Conference (2003)
12. Nishikawa, D., Yokoi, W., Yu, H., Kakazu, Y.: EMG prosthetic hand controller using real-time learning method. In: Proceedings of Intl.Conf.on Systems Man and Cybernetics. IEEE, Los Alamitos (1999)
13. Moshou, I., Hostens, D., Papaioannou, H.: Wavelets and self-organising maps in electromyogram(EMG)analysis. In: Proceedings of ESIT, pp. 186–189 (2002)
14. Massa, B., Roccella, S., Carrozza, M.C., Dario, P.: Design and development of an underactuated prosthetic hand. In: Proceedings of Intl.Conf.on Robotics and Auto. IEEE, Los Alamitos (2002)
15. Vuskovic, M.I., Pozos, A.L., Pozos, R.: Classification of Grasp Modes Based on Electromyographic Patterns of Preshaping Motions. In: Proceedings of Intl.Conf.on Systems, Man and Cybernetics. IEEE, Los Alamitos (1995)
16. Santiago, F.T., Kenemans, J.L., Kok, A.: A comparison of different methods for estimating single-trial P300 latencies. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* 92(2), 107–114 (1994)
17. Reaz, M.B.I., Hussain, M.S.: A non-MVC EMG normalization technique for the trunk musculature: Part I method development. *Journal of Electromyography and Kinesiology* 11(1), 1–9 (2001)
18. Karlsson, S., Gerdle, B.: Mean frequency and signal amplitude of the surface EMG of the quadriceps muscles increase with increasing torque a study using continuous wavelet transform. *Journal of Electromyography and Kinesiology* 11, 131–140 (2001)
19. Eom, K.H., Choi, Y.J., Sirisena, H.: EMG pattern classification using softms for hand signal recognition. *Journal of Soft Computing* 6(6), 436–440 (2002)
20. Englehart, K., Hudgin, B., Parker, P.: A wavelet-based continuous classification scheme for multi-function myoelectric control. *IEEE Trans.on Biomedical Engineering* 48(3), 302–311 (2001)
21. Cutkosky, M.R.: A on grasp choice, grasp models and the design of hands for manufacturing tasks. *IEEE Transactions on Robotics and Automation* 5(3), 269–278 (1989)
22. Chan, F.H.Y., Yang, Y.S., Lam, F.K., Zhang, Y.T., Parker, P.: Fuzzy EMG classification for prosthesis control. *IEEE Transactions on Rehab.Engg.* 8(3), 305–311 (2000)
23. Castellini, C., van der Smagt, P.: Surface EMG in advanced hand prosthetics. *Biological Cybernetics* 100(1), 35–47 (2009)
24. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discov.* 2(2), 121–168 (1998)
25. Bostannov, V.: BCI competition 2003-datasets ib and iib: feature extraction from event related brain potentials with the continuous wavelet transform and t-value scalogram. *IEEE Transactions on Biomedical Engineering* 51(6), 1057–1061 (2004)

26. MacKenzie, C., Iberall, T.: *The Grasping Hand*. North-Holland, Amsterdam (1994)
27. Crawford, B., Miller, K., Shenoy, P., Rao, R.: Real-time classification of electro myographic signals for robotic control, Technical Report, Dept. of Computer Science, University of Washington (March 05, 2005)
28. Kakoty, N.K., Hazarika, S.M.: Classification of Grasp Types through Wavelet Decomposition of EMG Signals. In: *Proceedings of the 2nd International Conference on BioMedical Engineering and Informatics*. IEEE, Los Alamitos (2009)

Information Visualization for Tourist and Travelling in Indonesia

Adityo Ashari Wirjono¹, Z.S. Ricky Lincoln¹,
William¹, and Dewi Agushinta R.²

Informatics Department, Information System
Gunadarma University
Jl. Margonda Raya 100
Depok, Indonesia

¹{adityo.ashari,ricxzone,william_1106}@student.gunadarma.ac.id,
²dewiar@staff.gunadarma.ac.id

Abstract. These days, many developing countries, especially island nations such as Indonesia would try to increase productivity in tourism by trying to improve facilities and services to the tourists and travelers. With so many beauty and diversity in Indonesia, we should see this potential and create a container that can help the people in the world to see the natural beauty found in this country. If this can be realized it will cause a positive impact for both parties, especially governments in the tourist attraction region. However, it is still cannot be realized until today due to the absence of a facility-based of information visualization that can assist travelers in obtaining information about tourist attractions in Indonesia. This paper tries to display a web-based concept of tourism services that can help the tourists in getting the information, location and also the shortest path to reach a tourist destination.

Keywords: Dijkstra Algorithm, GIS, InfoVis, Tourism.

1 Introduction

By looking at the many tourist attractions in Indonesia based on the diversity of cultures, it obvious will attract many tourists both from domestic and abroad to take vacation and travel in Indonesia. Indonesian tourism department is always trying to increase the number of tourists, especially from abroad. But the problem often faced by Indonesia is when the tourists come, they are having a hard time finding a tourist attraction if it does not have guidelines such as tourist maps or other information sources. In 2009, the number of international tourists arriving in Indonesia climbed 3.6Many solutions have been sought by local governments such as the creation and improvement of the guide book that comes with the latest maps. An increasing number of vehicles and imprecise controls of traffic in Jakarta have become major issues that create congestion. The traffic congestion causes loss in productivity, consumes a lot of gasoline, diminishes air quality, creates a variety of safety hazards, often discourages tourism,

and reduces business information [1]. All of these problems are required to be solved so that the traffic congestion can be reduced. There are several different types of solution that have been taken to reduce traffic congestion in Jakarta, such as employs traffic policemen in important traffic points, attempt to lay more pavements to avoid congestion, etc. But with the advent of technology and increment of traffic flow, several approaches with less involvement of human have been taken. Contemporary approaches emphasize better information and control to use the existing infrastructure more efficiently [2]. In contemporary approaches, image processing, computer vision or robot vision, etc are highly recommended. In these types of solutions, involvement of computers provide many promising approaches because information feed through mobile applications or web networks can simply provided. Because of this, we are proposing an innovative method in detecting traffic congestion using mobile application. In Indonesia, we can get the online information of traffic flow on certain location through some websites. One of the websites is <http://lewatmana.com>. In this website, the information of traffic flow will be obtained through cameras that are placed in important traffic points. This information will be updated every two hours. Therefore, the people who are connected to the internet network can use this application and choose an appropriate road to avoid congestion. In this paper, we want to describe mobile-based interaction as a new generation of traffic jam detection system that has tremendous potential to improve decision support system. This application will support real-time maps that can be viewed or accessed and provide the information of the current situation on specific roads in Jakarta. It also presents a better management decision making to the user or the traveler. So, the user not only can avoid traffic congestion but also can choose the appropriate road with a mobile phone. This new and improved application is necessary to develop an innovative traffic jam detection system. This paper is structured as follows. The second section presents the path (graph theory) and the algorithm, Dijkstra's shortest path algorithm, that we used to establish the mobile application. The third section discusses the modeling of mobile-based interaction for detecting traffic congestion. The fourth section shows the final comments and conclusions.

2 Related Works

Nowadays, E-Tourism in the other countries is no longer a new topic. There has been much research and papers discussed about E-Tourism and shortest path algorithm. In this section, we will give a few related paper that has a similarity problem.

Chow proposed a GIS application framework to utilize the Maps Application Programming Interfaces (API) in visualizing and presenting geographic information [3]. This paper produced a web prototype that proved to be effective in providing the users with a dynamic interface for data exploration. With the openness of the Maps API source code and flexibility in adopting open-specification data standards will indicate some of the potential of the Maps APIs in developing new Internet GIS applications.

Dijkstra's algorithm for finding the shortest route tourist attractions in Bali using GIS web form, introduce with the script PHP and MySQL as a database manager [4]. This paper also used the format SVG (Scalable Vector Graphics) to create the map, The advantage of using SVG file format is the quality of images or maps that been generated which will not be diminished or destroyed when the image is enlarged or reduced (scalable).

A method to speed up shortest path computations using graph proposed by Wagner and Wilhalm [5]. This paper describes three geometric speed up techniques for Dijkstra's algorithm. There are goal-directed searches, geometric shortest path containers, and reach based routing. All the three graph-drawing methods that proposed in this paper produce equally good layouts concerning the geometric speed-up techniques. Generating a new force directed and spectral layouts relies non-trivially on parameters that are sometimes hard to optimize.

3 Methodology

Basically, the term Geographic Information System (GIS) is a combination of three main elements: system, information, and geography. Geographic Information System is a computer system that is used to input (capturing), storing, checking, integrating, manipulating, analyzing, and displaying data related to positions on the surface of the earth. Geography Information System can integrate information between the graphics data (spatial) with text data (attributes) that are connected on earth [6].

The main goal of the utilization of Geographic Information System is to facilitate in getting information that has been processed and stored as attributes of a location or object. The data that processed in Geographic Information System is basically composed of spatial data and attribute data in digital form, so analysis can be used is spatial analysis and attribute analysis. Spatial data is data related to the spatial location of the general shape of the map. While attribute data is a data table that serves to explain the existence of various objects as spatial data.

3.1 Collecting and Conversion Data

Collect Data. The first thing we must do is collect the data that needed to make this tourist applications. This application need a tourist map of tourist attractions to be located. Map of the tourist area is taken from a map owned by Google (Google Maps). After that, we need the data from police department about traffic jams or other information related to the traffic, the data from department of transportation about public transportation routes and the latest data needed is from department of tourism [7]. The illustration of the data collecting can be seen in Fig.1.

Convert Maps into Graph. we created an undirect weighted graph based on google maps (see in Fig 2.(A)). The graph uses to find the shortest route for tourists to arrive at their destination place. We use Dijkstra's algorithm to

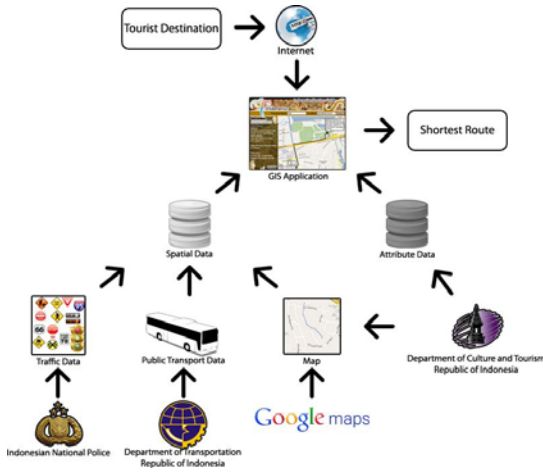


Fig. 1. Collecting Data

search the shortest path from the graph. To create a good graph for this tourist application, a method of drawing graphs that can accelerate the search for the shortest route is required. A high-dimensional embedding method to describe the graph and using a goal-directed search to accelerate the computation of the Dijkstra’s algorithm is used [8]. Edge is a path in the graph representation. The value of each edge is determined from the distance, traffic conditions, and transportation used by tourists. While the vertex is a representation of each tourist places and crossroads that described in Fig.2.(B).

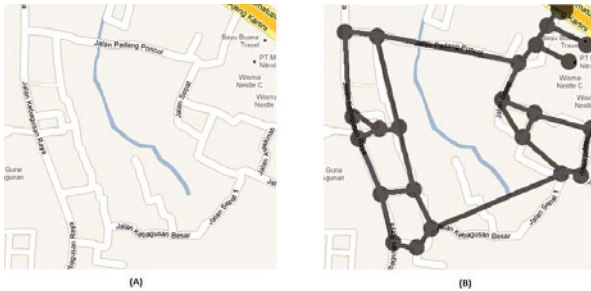


Fig. 2. (A) The map from google.maps.com , (B) The graph from result of conversion map

The key idea of high-dimensional embedding is to draw the graph in a very high dimensional space (usually $d = 50$) and then project the graph to 2D using principal component analysis. As our graphs are weighted, the shortest paths were calculated using these edge weights.

3.2 Find the Shortest Path

After obtaining the graph from the two methods, Dijkstra's algorithm to get the shortest path from the place which the tourist are in to tourist destination place is be used [9]. Dijkstra's algorithm is called the single-source shortest path. Dijkstra's algorithm is one variant of a popular form of algorithms in solving problems that related to the problem of finding a shortest path (a path that has minimum length) from vertex a to vertex z on a weighted graph, where weights are positive numbers.

This algorithm uses a tree diagram to determine the shortest path route. Using a greedy strategy, at each step is taken from the side of the minimum weight that connects a node that has been selected with an other node that has not been selected. The path from initial node to a new node must be the shortest path between all the tracks into nodes that have not been selected.

Dijkstra's algorithm using a time of $O(V * \log V + E)$ where V and E is the number of vertices and edges. The complexity of Dijkstra algorithm is

$$O(n^2)$$

So to find all pairs shortest vertex, the total asymptotic computation time is

$$T(n) = O(n^2) = O(n^3) \quad (1)$$

in other words Dijkstra's algorithm is more beneficial in terms of running time. This is a pseudocode of the Dijkstra's algorithm [9].

```

Procedure Dijkstra (V: set of vertices 1... n {Vertex 1 is the source}
Adj[1..n] of adjacency lists;
EdgeCost(u, w): edge - cost functions;)

Var: sDist[1..n] of path costs from source (vertex 1); {sDist[j] will be equal to
the length of the shortest path to j}
Begin:
Initialize
(Create a virtual set Frontier to store i where sDist[i] is already fully solved)
Create empty Priority Queue New Frontier;
sDist[1]-0; {The distance to the source is zero}

forall vertices w in V - {1} do {no edges have been explored yet}
sDist[w]-∞
end for;

Fill New Frontier with vertices w in V organized by priorities sDist[w];
endInitialize;

repeat
v-DeleteMin(New Frontier); {v is the new closest; sDist[v] is already correct}
forall of the neighbors w in Adj[v] do
if sDist[w]>sDist[v] +EdgeCost(v,w) then
sDist[w]-sDist[v] +EdgeCost(v,w)
update w in New Frontier (with new priority sDist[w])
endif
endfor
until New Frontier is empty
endDijkstra;

```

Fig. 3. Pseudocode of Dijkstra's Algorithm

3.3 Designing an Interactive Map

To get an interactive interface to the tourist application, an easier new map for tourist to find places information is created. First, Google maps were used as the basis of the interactive map, and then we add the data from the department of tourism to the maps. Rebuild the maps using open source software such as Quantum GIS. All the data from the department of tourism were added to the new map on Quantum GIS, automatically the data will be saved on PostgreSQL. So now, the interactive digital maps for the tourist application interface were made. MapServer will make these new digital map to be used in the web server layer. MapServer also provides API (Application Programming Interface) called MapScript, which make the programmer, can modify the map as their want. Chameleon as a product of Open Source that is built with PHP programming language, provides simple access to some features that can only be accessed in MapScript. The function of the Chameleon is a liaison between MapServer and the web server (see fig x). By using the PHP language and with some help of php functions on Chameleon, digital map into a web-based tourist application can be display easily. The next step is to create a web design for tourist applications. The application can help the tourist to search and find the shortest route to their destination (tourist places). For example, when the tourist want to find location of Sea World at Ancol, they only input on search box and then the application will automatically save the position of tourist as beginning node and the destination place as end node. The application can give the information about their destination place and route as described in Fig.4 and Fig.5.



Fig. 4. Web Interface with information of tourist place



Fig. 5. Web Interface with route information

4 Conclusion

Based on the methodology, the using of information visualization can make the tourist easier for searching the tourist attractions in Indonesia. Dijkstra's algorithm is good enough to search the shortest route of a tourist attraction in Indonesia. It is more beneficial in terms of running time. Dijkstra's algorithm combined with the result graph from conversion of maps (high-dimensional embedding and Goal-Directed Search) proved can be useful to find the shortest route for tourists to arrive at their destination place. This GIS application based on web, it will make the tourist easier to get the alternative route (shortest path) to arrive at their destination place.

This paper refers only with web-based applications. For the future this application will be try to implement in other media, such as mobile applications and etc. Online Reservation also will be the next target of this application. This application expects to avoid tourist trouble when they want to buy tickets or lodging around a tourist attraction. So, the tourists cannot just only search or find a way to the tourism object, but they also can make a ticket reservations and lodging for a hotel just by through our application.

References

1. Wikipedia. Tourism in Indonesia (February 2011), <http://en.wikipedia.org/wiki/Tourism-in-Indonesia>
2. Singapore Tourist Guide. Welcome to the Singapore Tourist Guide, <http://www.singaporetouristguide.net/>
3. Edwin Chow, T.: The Potential of Maps APIs for Internet GIS Applications, p. 13. University of Michigan, Department of Earth and Resource Science (2005)

4. Joni, L., Dewi, E.: Pencarian Rute Terpendek Tempat Wisata Di Bali Dengan Menggunakan Algoritma Dijkstra. In: Seminar Nasional Aplikasi Teknologi Informasi 2010 (SNATI 2010), June 19, p. 4 (2010)
5. Wagner, T.W.D.: Drawing Graphs to Speed Up Shortest-Path Computations. p. 9
6. Venia Rachmawati, S., Wirawan, S.: Web Application Of Historical And Cultural Tourism Information Mapping. In: Jakarta. pp. 13 (2010)
7. Prakoso Bhairawa Putera, S. R., Mulatsih, S.: Destinaiton Management Organization (DMO): Parafigma Baru Pengelolaan Pariwisata Daerah Berbasis Teknologi Informasi, June 20, pp. 4 (2009)
8. David Harel, Y.K.: Graph Drawing by High-Dimensional Embedding. *Journal of Graph Algorithms and Applications* 8, 20 (2004)
9. Puthuparampil, M.: Report Dijkstra's Algorithm, 15 (2005)

The Smart Goal Monitoring System

Dewi Agushinta R.¹, Bima Shakti Ramadhan Utomo², Denny Satria²,
Jennifer Sabrina Karla Karamoy², and Nuniek Nur Sahaya²

¹ Information System

dewiar@staff.gunadarma.ac.id

² Informatics

Gunadarma University, Jl. Margonda Raya 100, Depok, Indonesia

{bima_1990,denny_satria,jskk,nunieknursahaya}@student.gunadarma.ac.id


Abstract. In the current era of rapid technology development, many researchers compete each other to make an automated and integrated system. Since soccer is a favorite sport of all ages, a goal monitoring system is very needed, especially goal detection. The goal monitoring system generates fair play and avoids human error on soccer match. It will be very useful to help referee work. The system runs through sensor, image processing, and final decision. Sensor as object reader will activate the camera at many angles. Combining Circle Hough Transform (CHT) with real-time Color Ball Tracking produces a progressive method to process ball detection. The referees use collaboration tool to get the information. Hence, the referees can be collaborated each other to decide a goal on the match better.

Keywords: Computer Support Cooperative Work (CSCW), Goal monitoring system, Soccer.

1 Introduction

Gender is not an obstacle in soccer. Old and young people love this sport. Thousands matches were held, but in fact, there were many mistakes that could not be handled properly. Offside, misconduct, and the goal were the highlighted problems. The referee, who is the key of all decisions, is often less accurate on deciding those issues.

Recently, a less accurate system on soccer happened. A nice goal by Lampard was denied in England's World Cup clash with Germany.

"The ball crashed against the underside of the crossbar before bouncing a yard over the goal-line." 

Jorge Larrionda as the referee failed to see the Lampard's amazing goal. This case attracts us to deeply focus on goal monitoring system. This paper provides a preliminary of the goal monitoring system on soccer match.

Living in the development era of technology makes everything automated and integrated each other. A case likes Lampard's goal can be discussed more

interested by technology point of view. By seeing soccer's growing and high enthusiasm of supporters, this goal monitoring system can be applied well in Indonesia. We will use the concept of Computer Support Cooperative Work (CSCW) to help the referee.

CSCW is a generic term which combines the understanding of the way people work in groups with the enabling technologies of computer networking, and associated hardware, software, services and techniques [2]. It is a part of Human Computer Interaction (HCI) focused on working groups. One of potential applications for CSCW is procedure processing. It can handle paper-based form, and at the same time, the group can access full summary information about status, whereabouts and over-runs [2].

The reason to apply the goal monitoring system, especially goal detection, is to support fair play on soccer match. It will be used for helping referee work. The referees can be collaborated to decide the best decision. Another reason is to decrease human error. It is natural if the referee makes mistakes. With the existence of this system, both player and referee are advantaged.

The system runs through a sensor that is positioned in-line with the goalpost. The sensor functioned as object reader activates the cameras that are positioned at many angles. Once the cameras active, any object will be captured and sent to the main computer. The image processing will generate the captured object into a result image that clearly shows the ball position. It will be sent to referee field as a signal. The referee can decide the best decision through their collaboration tool. This collaboration tool helps the referees to agree the goal. That is why this research related to CSCW. The system can be called as an automated goal detection system, but the decision is on the referee hand. The system illustration is roughly described as in Fig. 1. There are sensor, camera, computer, and referee. All is integrated and provides an automated system.

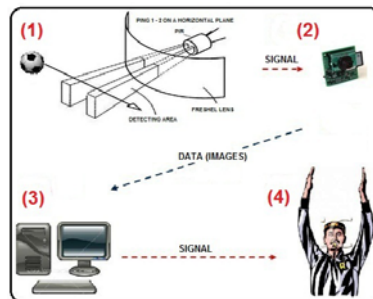


Fig. 1. (1)Sensor Detects Ball, (2)Camera Captures Ball, (3)Image Processing, (4)Final Decision by Referee

Hence, this paper proposes an automated goal detection through sensor and image processing method in detail. This also proposes an integrated system for all referee. How it works will be discussed in detail on methodology section.

2 Related Works

There are many methods or algorithms used for objects detection in soccer which includes players and ball. A software architecture was proposed by Davide *et al.* about ball detection and following based on a stereoscopic vision system. It was able to work in different lighting conditions. The aim was to identify potential arcs in the edge image [3]. The different lighting conditions, such as occlusions, shadows, objects similar to the ball, and real time processing became the important problems. A modified version of the directional Circle Hough Transform with the different lighting was needed. The Circle Hough Transform determined the parameters of a circle when a number of points that felt on the perimeter were known. D’Orazio *et al.* proposed Atherton Algorithm and Modified Atherton Algorithm. Both of them were used for detecting a ball in different lighting condition [4].

The other detection method was from Yu *et al.* It was based on color segmentation and shape analysis in soccer videos. It would detect and locate the players and the ball on the grass playfield. Detecting the shape of an object by using color histogram model was worked to detect the playfield pixels and group them into a playfield region. With the Euclidean distance transform, the players was extracted into skeletons for every foreground blob. Then the transform performed shape analysis [5]. The novel framework by Xinguo *et al.* was not far related with other. Ball candidates were first identified by size, color, and shape, and then these candidates were further verified by trajectory mining with a Kalman filter. It was the most accurate ball detection for broadcast soccer video. The Kalman filter was a tool that can estimate the variables of a wide range of processes [6].

The weakness for each method has been described previously are such as the performance of Atherton Algorithm and Modified Atherton Algorithm. It greatly decreases since the number of points that matches the searched pattern can be very small. Whereas when using the circle detection algorithm, it works very well when a ball passes through in front of the cameras. The ball that comes out of the cameras view always returns a false detection because there is always a peak in the parameter space. Detection based on color segmentation and shape analysis can not detect the ball because the ball’s location must be set manually and requires color histogram models to detect player’s presence.

The methods were widely used in broadcast video. However, many paper [5,6] proposed the detection based on the soccer ball video. This paper uses a little different method. The detection is applied for the images that have been captured by cameras. The cameras are located inside the wicket. Image processing is needed for the ball detection. This paper combines the method of Circle Hough Transform (CHT) and the real-time color ball tracking to detect a ball. There are many weakness in the Circle Hough Transform in terms of background, shadow and accuracy in detecting. To overcome the weakness, we use Circle Hough Transform to detect the ball as a circle first, then get the shape of an object by knowing the distribution of color by using the real-time Color Ball Tracking.

3 Methodology

As discussed before, some cameras are placed inside in the wicket. Each camera has a different capturing time, so there is no delay time to capture the ball. The goal monitoring system has been illustrated clearly in Fig. 1. The sensor has a detecting area to send a signal. If any object passes through the detecting area, sensor will activate the camera. The cameras capture the object immediately and send it directly to the main computer. It will be automatically entered to image processing. The image processing delivers the result detected as a ball. If the result shows the ball crossed the goal line, the system will send a signal to the referees. Every referee has a signal receiving device as collaboration tool to decide the goal.

This goal monitoring system uses 3 steps. The first is grabbing the image, the second is image processing using output from the first step, and the last is final decision system.

3.1 Grabbing the Image

A Passive Infra Red PIR sensor KC7783R is used as switch for the cameras which will capture the pictures of ball [7]. KC7783R PIR is sensor detection functioned normally at 4.7 voltage – 12 volts DC. It will give a high level output between 4.9 to 6 volts. The sensor is positioned in-line with the goalpost. Some cameras will be positioned in some different angles. One camera can be positioned on the corner of the goalpost which leads to the goal line.

The infra red sensor works by detecting a ball passing through the sensor detection area. The sensors will automatically detect the arrival of the ball when the ball past the detection area of the sensor. In other words, if the ball passes the goal line, the sensor will be active. The description of sensor works can be seen in Fig. 2. The camera will be also active as long as the sensor active. Then the camera will grab the image rapidly until the sensor off. The output image will be sent directly to the computer to be processed.

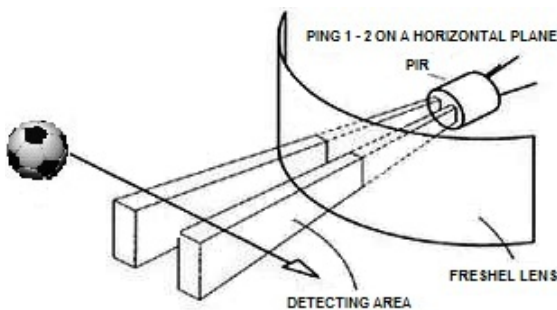


Fig. 2. Sensor System [7]

3.2 Image Processing

Hough Transform. Hough Transform is used for detecting the ball as a circle (amongst other false positives). Hough Transform is also widely used in image analysis, computer vision and digital image processing techniques in terms of extracted features. The purpose of these technique is to find a perfect example of an object in a particular class of shape by the voting procedure.

Circle Hough Transform (CHT) is one part of the Hough Transform method that can retrieve or set on a circular object image. This method transforms the image into the field of Hough.

The algorithm of Circle Hough Transform [8]:

1. Find edges.
2. The hough begin for each edge point.
3. Draw a circle with center in the edge point with radius r and increment all coordinates that the perimeter of the circle passes through in the accumulator.
4. Find one or several maxima in the accumulator and that is the hough end.
5. Map the found parameters (r,a,b) corresponding to the maxima back to the original image.

Each element in the image field is transformed into a circular shape in the form of Hough. From the mapping point edge produced by edge detection, the mapping of the Hough space for each parameter circles through each point edge. The result of Circle Hough Transform with false positives that were reduced can be seen in Fig. 3.

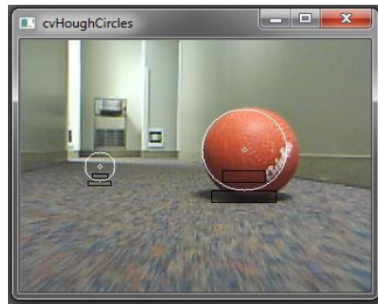


Fig. 3. Result of Circle Hough Transform [8]

There are some weaknesses when using Circle Hough transform algorithm to detect the ball in the soccer field. The backgrounds on the images have a color that is not evenly distributed or have an uneven color. Therefore, a darker section which is not actually a shadow can be clearly detected. The rate of accuracy in detecting the ball as a circle is still quite low. To overcome these problems, a real-time Color Ball Tracking can be used to detect the ball.

Color Ball Tracking. Color Ball Tracking is a system that works to get the shape of an object by knowing the distribution of color. Generally it takes 2 techniques. First is off-line calibration phase the camera's intrinsic parameters and radial distortion. This step purpose is to know the input image distribution color. The approach arranges color balls and acquires a single image. Then there will be image segmentation process to determine the color difference of one object to another. RGB value in each color is converted into one index. The result is shown in Fig. 4.

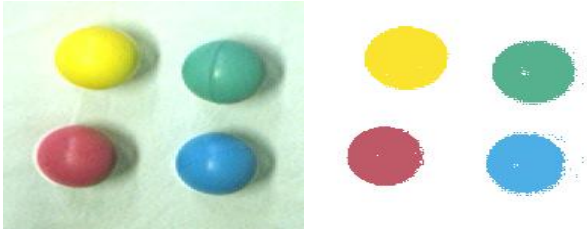


Fig. 4. Input Image and The Result [9]

Then the second step is on-line real-time tracking phase follows where the color classifier is applied to the input images, balls will be detected and 3D positions are returned. Robust estimation of circle parameters and refinement of circle parameters techniques are parts of the input image segmentation. When using these two techniques, we will get the best result of ball detection. The result is shown in Fig. 5.



Fig. 5. Result [9]

3.3 Final Decision System

The system will check the output of the previous step. If the ball pass through the goal line, the system will send the signal to the referee at field using method from Nedad Pejic [10]. The system includes a signaling device and a signal receiving device. Otherwise, this goal monitoring system only needs the signal

receiving device. It includes a radio-frequency receiver and an actuator device. The actuator device is used to provide an indication thereof.

The steps of this method are generating a radio-frequency signal having a predetermined code, receiving the radio-frequency signal, and generation an indication of receipt. If Nedad uses a vibrator device to generate the indication, this paper uses a sound wave device. This device is connected to the smart earpiece of the referee. This smart earpiece is called as referee collaboration tool.

As the ball detected clearly past the goal line, a radio-frequency signal from the main computer would be generated into sound wave. It will be received by the referees through their collaboration tool.

4 Conclusion and Future Work

The advantage of this system is for referee monitoring system. It is expected to help the referee to avoid human error. Hence, the referee can decide the best decision to determine a goal. Since the signal receiving device at final decision step makes all referee connected and receiving a same result just in time, it is a collaboration tool. With the existence of this system, fair play can be well done.

The problem arises when the ball passes through the wicket very fast. The camera can not capture ball at the time. This system could be developed again by using new methods such as detecting the speed of the ball motion. It may be also be able to detect offside and handball violations. In addition, this system uses only one sensor. Hence, it can be developed by adding more sensors to sharpen the accuracy of the sensor, to send a signal which can make the camera faster to detect the ball presence.

References

1. Reporter, S.: World cup 2010: The goal that never was! frank lampard strike against germany denied by referee jorge larrionda (June 2010), <http://www.dailymail.co.uk/sport/worldcup2010/article-1290040/WORLD-CUP-2010-The-goal-Frank-Lampard-strike-Germany-denied-referee-Jorge-Larrionda.html>
2. Wilson, P., Computer, G.B.T.C., Branch, T.A.A.C.: Computer Supported Cooperative Work: an introduction. Intellect (1991)
3. Scaramuzza, D., Pagnottelli, S., Valigi, P.: Ball Detection and Predictive Ball Following Based on a Stereoscopic Vision System. In: CRA, pp. 1561–1566. IEEE, Los Alamitos (2005)
4. d’Orazio, T., Ancona, N., Cicirelli, G., Nitti, M.: A Ball Detection Algorithm for Real Soccer Image Sequences. In: ICPR, pp. I: 210–213 (2002)
5. Huang, Y., Llach, J., Bhagavathy, S.: Players and Ball Detection in Soccer Videos Based on Color Segmentation and Shape Analysis. In: Multimedia Content Analysis and Mining, pp. 416–425 (2007)
6. Yu, X., Tian, Q., Wan, K.W.: A Novel Ball Detection Framework for Real Soccer Video. In: Proc. ICME 2003, vol. II, pp. 265–268 (2003)

7. Electricly. Sensor Passive Infra Red (PIR) pada pintu otomatis (July 2010), <http://electronical-instrument.blogspot.com/2010/07/sensor-passive-infra-red-pir-pada-pintu.html>
8. Simon Just Kjeldgaard Pedersen. Circular Hough Transform (November 2007)
9. Skoral, D., Sedlcek, D., Riege, K.: Real-time Color Ball Tracking for Augmented Reality. In: EGVE Symposium (2008)
10. N.R.O., Nenad, P.: (10241) Foxwood Dr. Method and Device for Indicating a Referee Signal. (6067013) (May 2000)

Web Based Virtual Agent for Tourism Guide in Indonesia

Kezia Velda Roberta¹, Lulu Mawaddah Wisudawati¹, Muhammad Razi¹,
and Dewi Agushinta R.²

¹ Informatics Department

{kezia_velda, lulu_chester91, razi_08}@student.gunadarma.ac.id

² Information System Gunadarma University,
Jl. Margonda Raya 100, Depok, Indonesia
dewiar@staff.gunadarma.ac.id

Abstract. The development of tourism sector in Indonesia is increasing rapidly, judging from the number of local and foreign tourists which is always growing every year. The rapid development of technology is also very influential in the development of this tourism sector, for example using the web to provide information on tourism in Indonesia, where only display text and images and also not interactive. So, an application system of virtual intelligent agent that connects human and computer is created. It makes an intelligent and interactive tour guide. This paper tries to present Smart Indonesian Tourism Agent (SITA) as visual tour guide. This is a web based information system that provides to access location of tourism in Indonesia. This application uses A.L.I.C.E server and Artificial Intelligence Modeling Language (AIML) interpreter. Hence, the information generated in the web can be displayed in text, visualization, image, and the chat box for questions.

Keywords: AIML, SITA, Tourism, Virtual Agent, Web.

1 Introduction

Indonesia is one of the country which has high fascinating tourism attraction spread in all over region from nature tourism until shopping tourism. It is one of the factor that makes tourism sector in Indonesia was improving from year to year. In 2010, the number of tourist improves about 10,79% from 2009 that is 7 million people. Foreign exchange earnings is also improving about 20,63% to be US\$ 7,6 billion [1].

The growth of tourism industry is also impacted by advanced technology, such as using of web site as promotion facility. Information which is displayed on the web sites generally in text and less interesting to read. Therefore, this paper creates a concept of web interactive using virtual agent.

Agents are systems that interact with an environment using sensors to receive perceptual inputs (called percepts) from it, and actuators to act upon it [2]. A agent is a computer system that is situated in some environment, and that is

capable of autonomous action in this environment in order to meet its design objectives. Usually the agent only has partial-actions might not have expected consequences, control systems, software demons [3]. From those definitions, we have two important points. First, an agent has ability to do any task or work. Second, an agent does any task or work for anything or for another people [4] [5]. So, it can be concluded that virtual agent is a human like character in animation form and it can do direct communication with human interactively. This paper prepares that virtual agents use natural human modalities such as speech and gesture. They are capable of real-time perception, cognition and action that allows them to participate in a dynamic social environment.

Smart Tourism agent was created to help the tourist to get the information about tourist attraction in Indonesia. User can do direct interaction to ask about tourist attraction and culture in Indonesia. So, it can makes the tourist feels like communicate with a human tour guide. This paper will be discussed about the concept of web based virtual agent to promote tourism in Indonesia and then will be explained about the smart tourism agent interface.

This paper will be divided to some sections. The explanations start with the introduction. Section 2 explains about related work for the agent concept. Section 3 explains about smart tourism agent architecture. It explains about how agent can communicate with the user and its tools. Interface design for smart tourism agent will be created in section 4. Section 5 explains conclusion and future work.

2 Related Work

Virtual Agent has been widely used for various applications, such as agent for tutoring system. The Application of Intelligent Tutoring System (ITS) into e-learning system that is expected to improve the quality of learning [5]. This paper tried to provide an alternative way to support the creation of intelligent tutoring system, especially the learning of communication systems by presenting the figure of teacher in the form of Virtual Agent Character. This character agent is obtained from Ms. agent. Virtual Agent for e-learning created with Visual Basic programming language. This virtual agent is related part to create Intelligent Tutoring System, especially in presenting an intelligent and communicative virtual teacher. Agent is part of intelligent tutoring system, especially in intelligent and communicative virtual teacher. This paper will develop another virtual agent with Indonesia characteristic to get information a place of tourism in Indonesia using chat box.

An emotion is needed to make real virtual characters. Vallapureddy Rajender Reddy [6], discuss about "Communicating Emotions To Virtual Agents: An Emoticon-Based Approach". This research focuses on using emoticons to communicate emotions in human computer interaction. The use of emoticons has an advantage that it requires no external hardware devices. Emoticons can be used as an added communication channel augmenting natural language input in human computer interaction. A prototype has been developed which consists of a virtual character with natural conversation and appropriate facial expression abilities. The natural language understanding capabilities were extended

to allow for the use of emotions when talking to the agent. A group of users was asked to interact with the agent and asked to submit a feedback pertaining to certain questions. The analysis of the feedback suggests that conversational agents having emotions in them will significantly improve the interaction and believability.

This paper provide facility for user to make direct communication with agent. The work of Marcel Ritschel [7], explain AI Chat Bots and Digital Assistants. This paper explained about how to feed the AIML matching algorithm. So, the system can give the answer based on pattern.

Many papers discuss agents problem with different objectives. This paper tries to create web-based virtual agent with chat box concept that users can interact directly with the agent. Virtual agent is made using real emotion so it looks like a real tour guides by using the background information on AI and AIML method.

3 System Architecture

3.1 Architectural Design of the System

Fig. 1 describes a tourism agent architecture, an agent designed to assist foreign and local tourists as the user to find a place of tourism in Indonesia. Furthermore, users are given a choice of tourist attractions and the system will read it in the database server. After finding it in the database server, then agent will give a feedback to user and user will get information of tourist attractions they want. However, if users find a difficulty or require more information, there is facility to chat with Smart Indonesian Tourism Agent (SITA). The users type their input in the text box and by pressing the enter key or confirm ask button. Next, system will connect to A.L.I.C.E server and patterns will be checked in the AIML interpreter. After the pattern is matched, system will find the answer in database. The agent will response immediately with appropriate face expression.

The system is capable of making a natural conversation with the users. At present the system has just a few patterns to support the conversation. With the increment of the patterns and improving the patterns hierarchy, the system can be made much more effective in the natural conversation. The flow of the conversation with the SITA can be like the example given below [8]:

User : Where is Museum Gajah??

SITA : Museum Gajah is located on Jalan Medan Merdeka Barat no.12, Jakarta Pusat

User : what kind of object i can found there?

SITA : There are ancient stuff from Indonesia such as Ancient Statues, ceramics, textiles.

User : what is the other tourist attraction around Museum Gajah?

SITA : There is Monument National. It's about 1 km from Museum Gajah.

User : ok . thank you

SITA : Your Welcome !!!

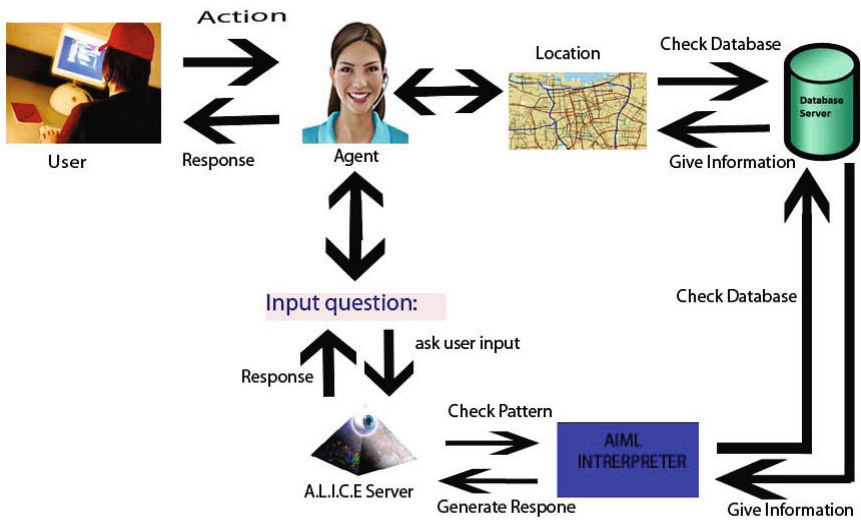


Fig. 1. Smart Tourism Agent System Architecture

3.2 AIML

AIML objects are made up of units called topics and categories, which contain either parsed or unparsed data. Parsed data is made up of characters, some of which form character data, and some of which form AIML elements. AIML elements encapsulate the stimulus-response knowledge contained in the document. Character data within these elements is sometimes parsed by an AIML interpreter, and sometimes left unparsed for later processing by a Responder [9].

A typical AIML formation is made of:

```
<aiml>
<category>
<pattern> </pattern>
<template> </template>
</category>
</aiml>
```

The AIML tag is the root tag, which marks the beginning and end of the AIML document. A category is a top-level element that contains exactly one pattern and exactly one template. A category does not have any attributes. A pattern is an element whose content is a mixed expression. The pattern must always be the first child element of the category. The contents of the pattern are appended to the full match path that is constructed by the AIML interpreter at load time. A template is an element that appears within category elements.

The template must follow the pattern element. The user input is searched for a match in the pattern element. If a match is found the content of the template element is sent as the response to the user.

```
<aiml>
<category>
<pattern>Hello </pattern>
<template> Hi there! How are you? </template>
</category>
</aiml>
```

When the user inputs Hello, the AIML interpreter looks for the pattern which matches the input (in this case the Hello pattern) and then returns the template element content as the response (in this case the user gets the response as Hi there! How are you?).

4 Interface Design

4.1 Flowchart Design

Fig. 2 explains how the user get the information about tourism in Indonesia. When user enter the site, the system will show a message for user to choose what the language that the user wants to use for accessing information. The system will read user choice and connect to database. After that, system replace the language that has been selected by user. Virtual agent will appear as interactive tour guide. The agent will say welcome message and introduce herself. Here, the system will show a map of Indonesia, where the user is free to choose province that would be tourist's destination and system will process data. After that, tourist destination in a smaller area will be shown as user selected. The user selects the tourist attraction which located in that province and agent will explain information about these attractions. In addition, there is chat box facility if users want to know more about the attraction or have difficulty when getting information. If the user input a question in chat box, system will read and directly connect to database then system will give feedback as the answer.

4.2 Website's Interface Design

The implementation flowchart can be illustrated at these interface design (Fig. 3, 4, 5). It is made based on Smart Tourism Agent System Architecture and it is the initial description of the concept are made which will be implemented on tourism sites. At Fig. 3, user can choose one of the province in map or in province list at left side. Then system will bring the user to province's page (fig. 4). In province's page, user can choose one of the tourism attraction and SITA will guide the user to the tourism attraction page (Fig. 5). In here, SITA will

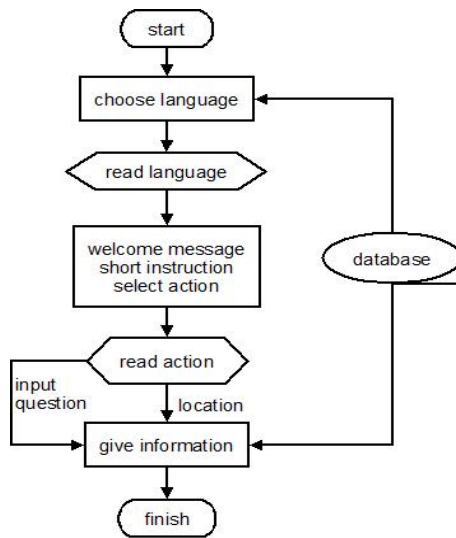


Fig. 2. Flowchart Design

explain about the place and if the user wants to know more, they can input a question in chat box then agent will give the answer. If user already know where to go, they can choose place of interest menu and choose one of the tourism attraction. This web also provide help menu. If user is getting confused, they can go to help menu.

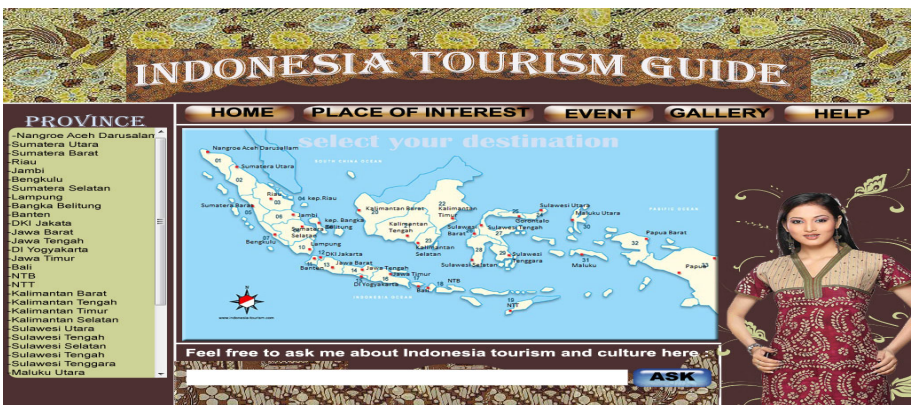


Fig. 3. Interface Design



Fig. 4. Province Page



Fig. 5. Tourist Attraction

5 Conclusion and Future Work

SITA is an intelligent and communicative virtual agent and it is created to give information about tourism in Indonesia. SITA is very helpful for local and foreign tourist to get information of tourist attraction. Beside that, user also can interact directly with the agent if users want to know more about the attraction or have difficulty when getting information. The information generated in the web can be displayed in text, visualization, image, and the chat box for questions.

There are several issues in developing a virtual agent, ie how to display information in 3D shapes and how to make an agent that can move around site. So, it has become more interactive. These several issues need some approached in computer graphic. This approach will be used in the next project.

References

1. William, A.: Wisatawan asing makin betah ke indonesia (February 2011), <http://www.tempointeraktif.com/hg/bisnis/2011/02/01/brk,20110201310549,id.html>
2. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 2nd edn., ch. 2. Prentice Hall, Englewood Cliffs (2003); ISBN 0137903952
3. Wooldridge, M.: Introduction to MultiAgent Systems. John Wiley Sons, Chichester (2002)
4. Wahono, R.S.: Pengantar Software Agent: Teori dan Aplikasi (2003), <http://romisatriawahono.net>
5. Bernard Renaldy Suteja, S.H., Wardoyo, R.: Viirtual agent character untuk mendukung intellegent tutoring system berbasis web. In: Seminar Nasional Aplikasi Teknologi Informasi 2009 (SNATI 2009) (June 2009); ISSN:1907-5022
6. Reddy, V.R.: Communicating emotions to virtual agents: An emoticon-based approach. Master's thesis, International School of New Media, Universitat zu Lubeck, Master of Science in Digital Media (June 2004)
7. Ritschel, M.: Chat Bots and Interaction Design. University of Technology, Sydney (2007)
8. Indonesia, M. N.: Museum nasional, <http://www.museumnasional.or.id/>
9. Wallace, R.S.: Be Your Own Botmaster. ALICE A.I. Foundation, Inc. (2003), www.alice.org

Local Feature or Mel Frequency Cepstral Coefficients - Which One Is Better for MLN-Based Bangla Speech Recognition?

Foyzul Hassan¹, Mohammed Rokibul Alam Kotwal¹, Md. Mostafizur Rahman¹,
Mohammad Nasiruddin², Md. Abdul Latif², and Mohammad Nurul Huda¹

¹ United International University

² The University of Asia Pacific

Dhaka, Bangladesh

foyzul.hassan@gmail.com, rokib_kotwal@yahoo.com,
tuhin_cse@yahoo.com, mohammadnasiruddin@gmail.com,
csefusion@yahoo.com, mnh@cse.uiu.ac.bd

Abstract. This paper discusses the dominance of local features (LFs), as input to the multilayer neural network (MLN), extracted from a Bangla input speech over mel frequency cepstral coefficients (MFCCs). Here, LF-based method comprises three stages: (i) LF extraction from input speech, (ii) phoneme probabilities extraction using MLN from LF and (iii) the hidden Markov model (HMM) based classifier to obtain more accurate phoneme strings. In the experiments on Bangla speech corpus prepared by us, it is observed that the LF-based automatic speech recognition (ASR) system provides higher phoneme correct rate than the MFCC-based system. Moreover, the proposed system requires fewer mixture components in the HMMs.

Keywords: Local Feature; Mel Frequency Cepstral Coefficient; Multilayer Neural Network; Hidden Markov Model; Automatic Speech Recognition.

1 Introduction

Various methods have been proposed for obtaining an accurate automatic speech recognition (ASR) system [1-6]. Although some of them perform adequately, most hidden Markov model (HMM) based methods have several limitations. For example, (a) they require a large number of speech parameters and a large speech corpus to solve coarticulation problems using context-sensitive triphone models, and (b) they require a higher computational cost to achieve an acceptable performance.

On the other hand, many neural network based methods are approached to reduce speech parameters [7-10] and to achieve higher recognition performance. Some of these methods solve coarticulation problems by embedding context dependent input vectors. Most of these methods use mel frequency cepstral coefficients (MFCCs) parameters of speech signal as input vector to multilayer neural network (MLN) [7-10] and provides comparable recognition performance using higher computation cost.

In this paper, we proposed a Bangla phoneme recognition system for an ASR by inputting local features (LFs) instead of MFCCs. The method consists of three stages: (i) the first stage extracts LFs from the Bangla input speech, (ii) the second stage incorporates an MLN to obtain phoneme probabilities for the LFs extracted at first stage and (iii) the final stage embeds an HMM-based classifier to output phoneme strings. This study shows that the LF-based system provides higher phoneme correct rate (PCR) than the system based on MFCCs. For designing the proposed and existing methods, we have prepared a medium size Bangla speech corpus and done the following experiments: (i) MFCC39+MLN+HMM and (ii) LF25+MLN+HMM [Proposed].

The paper is organized as follows: Section 2 shows the preparation of Bangla speech corpus. Section 3 explains the Bangla phonemes and corresponding IPA, while the LF extraction procedure is described in Section 4. The system configuration of the existing phoneme recognition methods with the proposed method is discussed in the Section 5. Experimental setup are provided in Section 6, while experimental results are analyzed in Section 7. Finally, Section 8 draws some conclusion.

2 Bangla Speech Corpus

At present, a real problem to do experiment on Bangla phoneme ASR is the lack of proper Bangla speech corpus. In fact, such a corpus is not available or at least not referenced in any of the existing literature. Therefore, we develop a medium size Bangla speech corpus, which is described below.

Hundred sentences from the Bengali newspaper “Prothom Alo” [11] are uttered by 30 male speakers of different regions of Bangladesh. These sentences (30x100) are used for training corpus (D1). On the other hand, different 100 sentences from the same newspaper uttered by 10 different male speakers (total 1000 sentences) are used as test corpus (D2). All of the speakers are Bangladeshi nationals and native speakers of Bangla. The age of the speakers ranges from 20 to 40 years. We have chosen the speakers from a wide area of Bangladesh: Dhaka (central region), Comilla – Noakhali (East region), Rajshahi (West region), Dinajpur – Rangpur (North-West region), Khulna (South-West region), Mymensingh and Sylhet (North-East region). Though all of them speak in standard Bangla, they are not free from their regional accent.

Recording was done in a quiet room located at United International University (UIU), Dhaka, Bangladesh. A desktop was used to record the voices using a head mounted close-talking microphone. We record the voice in a place, where ceiling fan and air conditioner were switched on and some low level street or corridor noise could be heard.

Jet Audio 7.1.1.3101 software was used to record the voices. The speech was sampled at 16 kHz and quantized to 16 bit stereo coding without any compression and no filter is used on the recorded voice.

3 Bangla Phonemes Schemes

3.1 Bangla Phonemes

The Phonetic inventory of Bangla consists of 14 vowels, including seven nasalized vowels, and 29 consonants. An approximate phonetic scheme in IPA is given in Table 1. In Table 1 (a), only the main 7 vowel sounds are shown, though there exists two more long counterpart of /i/ and /u/, denoted as /i:/ and /u:/, respectively. These two long vowels are seldom pronounced differently than their short counterparts in modern Bangla. There is controversy on the number of Bangla consonants.

Native Bangla words do not allow initial consonant clusters: the maximum syllable structure is CVC (i.e. one vowel flanked by a consonant on each side) [12]. Sanskrit words borrowed into Bangla possess a wide range of clusters, expanding the maximum syllable structure to CCCVC. English or other foreign borrowings add even more cluster types into the Bangla inventory.

Table 1. Bangla phonetic scheme in IPA extracted from http://en.wikipedia.org/wiki/Bengali_phonology

(a) Vowel				(b) Consonants						
	Front	Central	Back		Labial	Dental/Alveolar	Retroflex	Lamino-Postalveolar	Velar	Glottal
				Nasal	m	n			ŋ	
				Plosive	voiceless	p	t	tʃ	k	
						p̪	t̪	tʃ̪	k̪	
Close	i		u		aspirated	pʰ	tʰ	tʃʰ	kʰ	
	i		u			p̪ʰ	t̪ʰ	tʃ̪ʰ	k̪ʰ	
Close-mid	e		o		voiced	b	d	dʒ	g	
	e		o			b̪	d̪	dʒ̪	g̪	
Open-mid	æ		ɔ	murmured	b̃	d̃	dʒ̃	g̃		
	æ		ɔ		b̪̃	d̪̃	dʒ̪̃	g̪̃		
Open		a		Fricative	f	s, z		ʃ		h
		a			f̪	s̪, z̪		ʃ̪		h̪
				Approximant		l				
						l̪				
				Rhotic		r	r̪			
						r̪	r̪̃			

Table 2. Some Bangla words with their orthographic transcriptions and IPA

Bangla Word	English Pronunciation	IPA	Our Symbol
আমরা	AAMRA	/a m r a/	/aa m r ax/
আচরণ	AACHORON	/a tʃ r n/	/aa ch ow r aa n/
আবেদন	ABEDON	/a b æ d n/	/ax b ae d aa n/

3.2 Bangla Words

Table 2 lists some Bangla words with their written forms and the corresponding IPA. From the table, it is shown that the same ‘Av’ (/a/) has different pronunciation based on succeeding phonemes ‘g’, ‘P’ and ‘e’. These pronunciations are sometimes long or short. For long and short ‘Av’ we have used two different phonemes /aa/ and /ax/, respectively. Similarly, we have considered all variations of same phonemes and consequently, found total 51 phonemes excluding beginning and end silence (/sil/) and short pause (/sp/).

4 Local Feature Extraction

At the acoustic feature extraction stage, the input speech is first converted into LFs that represent a variation in spectrum along the time and frequency axes. Two LFs, which are shown in Fig. 1, are then extracted by applying three-point linear regression (LR) along the time (t) and frequency (f) axes on a time spectrum pattern (TS), respectively. Fig. 2 exhibits an example of LFs for an input utterance. After compressing these two LFs with 24 dimensions into LFs with 12 dimensions using discrete cosine transform (DCT), a 25-dimensional (12 Δt , 12 Δf , and ΔP , where P stands for the log power of a raw speech signal) feature vector called LF is extracted.

5 System Configuration

5.1 MFCC-Based System

Fig. 3 shows the phoneme recognition method using MLN [13]. At the acoustic feature extraction stage, input speech is converted into MFCCs of 39 dimensions (12-MFCC, 12- Δ MFCC, 12- $\Delta\Delta$ MFCC, P, ΔP and $\Delta\Delta P$, where P stands for raw energy of the input speech signal). MFCCs are input to an MLN with four layers, including 3 hidden layers, after combining preceding (t-3)-th and succeeding (t+3)-th frames with the current t-th frame. The MLN has 53 output units (total 53 monophones) of phoneme probabilities for the current frame t. The three hidden layers consist of 400, 200 and 100 units, respectively. The MLN is trained by using the standard back-propagation algorithm. This method yields comparable recognition performance.

5.2 Proposed System

Fig. 4 shows the phoneme recognition method using MLN. At the acoustic feature extraction stage, input speech is converted into LFs of 25 dimensions (12 Δt , 12 Δf , and ΔP , where P stands for the log power of a raw speech signal). LFs are input to an MLN with four layers, including 3 hidden layers, after combining preceding (t-3)-th and succeeding (t+3)-th frames with the current t-th frame. The MLN has 53 output units (total 53 monophones) of phoneme probabilities for the current frame t. The architecture and training procedure of MLN are same as Section 5.1.

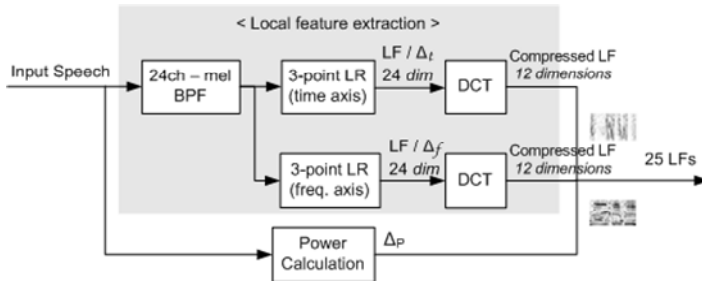


Fig. 1. LFs extraction procedure

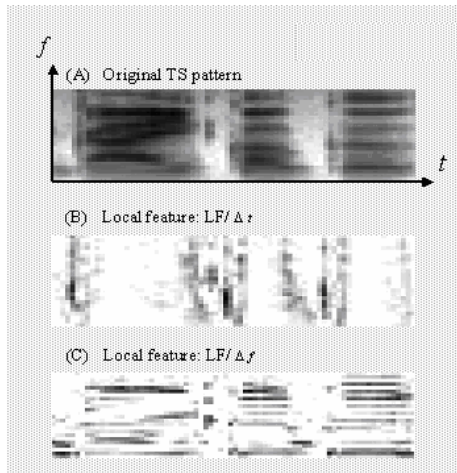


Fig. 2. Examples of LFs

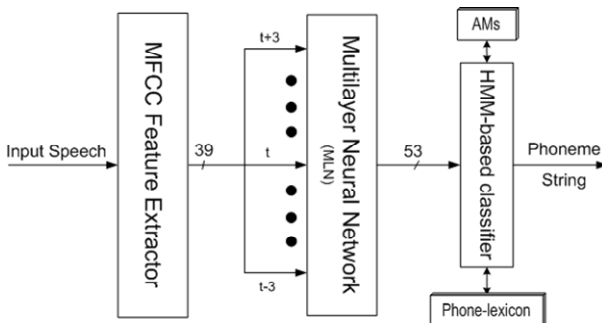


Fig. 3. MFCC-based Phoneme Recognition Method

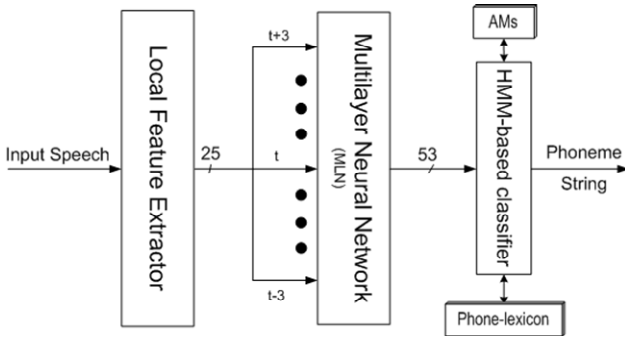


Fig. 4. LF-based Phoneme Recognition Method

6 Experimental Setup

The frame length and frame rate are set to 25 ms and 10 ms (frame shift between two consecutive frames), respectively, to obtain acoustic features (MFCCs) from an input speech. MFCC comprised of 39 dimensional.

For designing an accurate phoneme recognizer, PCR for both D1 and D2 data set are evaluated using an HMM-based classifier. The D1 data set is used to design 53 Bangla monophones HMMs with five states, three loops, and left-to-right models. Input features for the HMM using both the methods is 53 dimensions. In the HMMs, the output probabilities are represented in the form of Gaussian mixtures, and diagonal matrices are used. The mixture components are set to 1, 2, 4, 8 and 16.

In our experiments of the MLN, the non-linear function is a sigmoid from 0 to 1 ($1/(1+\exp(-x))$) for the hidden and output layers.

To evaluate PCR, we have prepared a medium size Bangla speech corpus and done the following experiments for both training (D1) and test (D2) data sets.

- (i) MFCC39+MLN+HMM
- (ii) LF25+MLN+HMM [Proposed].

7 Experimental Results and Discussion

Fig. 5 shows the comparison of PCR of training data set between MFCC39+MLN+HMM and LF25+MLN+HMM systems. It is observed from the figure that LF-based system always provides higher PCR than the conventional MFCC-based method. For an example, at mixture component 16, the LF-based system exhibits 61.07% phoneme correct rate, while 56.27% PCR is obtained by the MFCC-based method.

On the other hand, the PCR the test data (D2) are shown in the Fig. 6 for the investigated methods. The LF-based method outperformed the other methods for the evaluation of PCR. It is noted from mixture component 16 of Fig. 6 that the LF-based system having 55.02% correctness shows its better recognition performance over the MFCC-based method (52.07% PCR).

It is claimed that the proposed method reduces mixture components in HMMs and hence computation time. For an example from the Fig. 6, approximately 50.00% phoneme recognition correctness is obtained by the methods (i) and (ii) at mixture components eight and one, respectively.

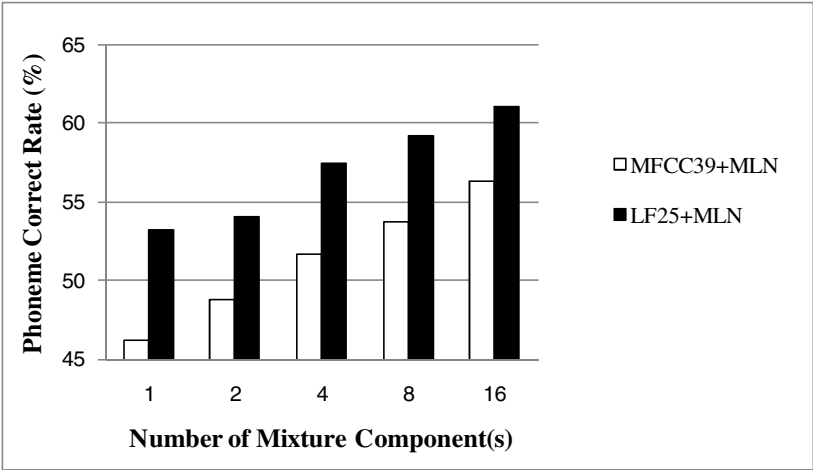


Fig. 5. Phoneme correct rate for training data set, D1

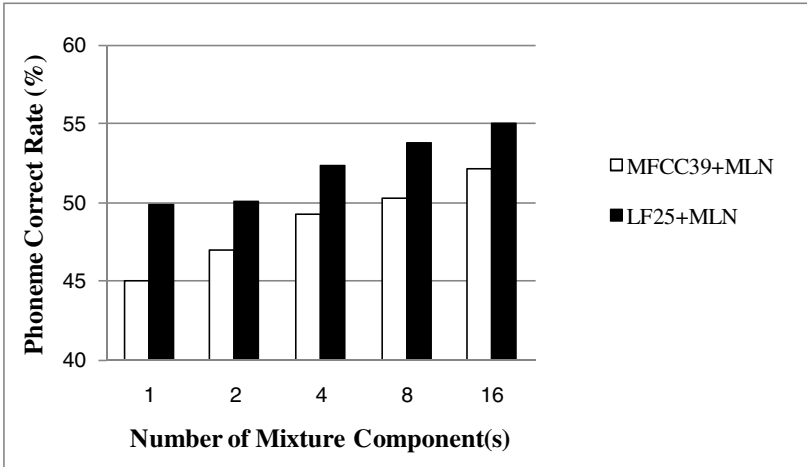


Fig. 6. Phoneme correct rate for test data set, D2

8 Conclusion

In this paper, we proposed a Bangla phoneme recognition method for an ASR using local features. The following conclusions are drawn from the study.

- (i) Our proposed method based on local features show higher phoneme correct rate for both training and test data set.
- (ii) It requires fewer mixture components in HMMs.
- (iii) Moreover, it reduces computation time because of lower dimensional (25 dim) LFs instead of higher dimensional MFCCs (39 dim).

In future, the author would like to do experiments using recurrent neural network (RNN). Moreover, Bangla word recognition using triphone model will be evaluated by the same proposed method in this paper.

References

1. Nakagawa, S., et al.: Noisy Speech Recognition Based on Integration/Selection of Multiple Noise Suppression Methods Using Noise GMMs. *IEICE Trans. Inf & Syst.* . E91-D(3), 411–421 (2008)
2. Jain, P., Hermansky, H., Kingsbury, B.: Distributed Speech Recognition Using Noise-Robust MFCC and TRAPS-Estimated Manner Features. In: *Proc. ICSLP 2002*, vol. I, pp. 473–476 (September 2002)
3. Marcus, J., et al.: Phonetic recognition in a segment-based HMM. In: *Proc. ICASSP* (April 1993)
4. Schwarz, P., et al.: Hierarchical structures of neural networks for phoneme recognition. In: *ICASSP* (2006)
5. Suzuki, H., et al.: Continuous Speech Recognition Based on General Factor Dependent Acoustic Models. *IEICE transactions on information and systems* E88-D(3)(20050301), 410–417 (2005)
6. Matsuda, S., Jitsuhiro, T., Markov, K., Nakamura, S.: Speech recognition system robust to noise and speaking styles. In: *Proc. ICSLP 2004*, vol. IV, pp. 2817–2820 (2004)
7. Kirchoff, K., et al.: Combining acoustic and articulatory feature information for robust speech recognition. *Speech Commun.* 37, 303–319 (2002)
8. Kirchoffs, K.: Robust Speech Recognition Using Articulatory information. Ph.D thesis, University of Bielefeld, Germany (July 1999)
9. Roy, K., Das, D., Ali, M.G.: Development of the speech recognition system using artificial neural network. In: *Proc. 5th International Conference on Computer and Information Technology (ICCIT 2002)*, Dhaka, Bangladesh (2002)
10. Hassan, M.R., Nath, B., Bhuiyan, M.A.: Bengali phoneme recognition: a new approach. In: *Proc. 6th International Conference on Computer and Information Technology (ICCIT 2003)*, Dhaka, Bangladesh (2003)
11. Daily Prothom Alo, <http://www.prothom-alo.com>
12. Masica, C.: *The Indo-Aryan Languages*. Cambridge University Press, Cambridge (1991)
13. Kotwal, M.R.A.: *Neural Network Based Automatic Speech Recognition*. M. Sc. Thesis, United International University, Dhaka, Bangladesh (2010)

Power Optimization Techniques for Segmented Digital Displays

Rohit Agrawal¹, C. Sasi Kumar¹, and Darshan Moodgal²

¹ VIT University, Vellore-632014

Tamil Nadu, India

rohit4849@gmail.com, csasikumar@vit.ac.in

² Indian Institute of Technology Bombay Powai

Mumbai- 400076, Maharashtra, India

darshanmd@gmail.com

Abstract. A number of power optimization techniques for portable electronic systems have been proposed earlier based on instruction-level characterization, compiler optimizations, source-level transformations, memory management schemes at the software level or dynamic power management and voltage scaling at the hardware level. However, reducing the power dissipation in embedded system peripherals namely LCDs, flash memory etc. also helps in improving the battery life. We propose some of these techniques with reference to segmented digital LCDs based on various control parameters like contrast ratio(pixel darkness), frame frequency, multiplexed mode of operation with significant power savings without compromising on the display quality.

Keywords: Segment, refresh rate, power, pixel darkness.

1 Introduction

Portable electronic systems form a huge part of our everyday lives covering a whole variety of devices including mobiles phones, mp3 players, i-pods, personal digital assistants (PDAs) etc. Power consumption has become the most important design challenge for portable devices. These power constraints have led to development of a number of power optimization techniques so as not to compromise upon the System performance and also the Quality of Service.

With reference to embedded system peripherals, the most important ones would definitely be LCDs due to their presence in almost all portable handheld devices PIC16F917 is used as the LCD controller which generates the timing control to drive the panels upto 1 common and 24 segments maximum in static mode and upto 4 commons and 96 segments in multiplexed mode. Here is a brief description of the various power optimization techniques applied to LCDs before.

2 Related Work

Most of the work done so far mainly concentrates on color TFT display systems not only because they are popular in most of the high end applications but also because of being the heaviest power consumers. Some of these methodologies like frame buffer compression using run-length encoding thereby reducing the frame buffer accesses were proposed in [1]. System-level techniques for power reduction in display systems have been recently introduced: variable duty-ratio refresh, dynamic color-depth control, and brightness and contrast shift with backlight luminance dimming [2]. Or other techniques like liquid crystal orientation shift and backlight auto-regulation mentioned in [3]. However dynamic color depth control may lead to image degradation and saves only 10 percent of the total system power. Moreover, some of these techniques are either applicable to only high end displays or to transmissive/trans-reflective LCDs. Another major drawback is the sharing of the resource like frame buffer by both frame buffer compression technique and dynamic color depth so only one of them may be applied at a time making the power saving less efficient.

The power saving techniques proposed here are entirely independent of each other with either no or very little hardware overhead and are primarily meant for reflective and transmissive numeric displays. These types of display panels find use in a variety of modern electronic equipments like calculators, handheld blood glucose meters, gas station pumps, multimeters, digital watches and other measuring instruments.

3 Driving LCD Segments

The liquid crystal is sandwiched between two transparent electrodes namely SEG and COM on the inside of the pieces of glass Fig. 1. The default state of LCD segments is OFF, when no voltage is applied, the segments become transparent and are invisible against the background in the LCD panel. In addition, when the *same* voltage is applied to both a segment line (SEG) and the common backplane (COM), the segment remains off. The segment will only switch ON (i.e., opaque) when a voltage difference is applied between the SEG pin for that segment and the COM plane. As this voltage passes a particular level, known as the threshold voltage, the segment darkens and finally becomes completely opaque. The threshold voltage, which is a percentage of the specified operating voltage of the LCD panel, varies from one LCD to the next.

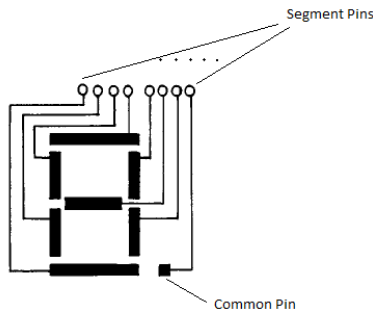


Fig. 1. Single Digit of 7 Segment Display

If a static DC voltage is left across an LCD segment for a long period of time, the segment can become damaged and will no longer switch properly. To avoid this problem, LCD segments are always driven with alternating waveforms to ensure that the overall DC voltage across each segment is always zero, whether the segment is in the ON or OFF state. Particularly, it is the difference between the COM and the SEG pin's alternating waveform that drive the LCD and these waveforms are generated automatically by the controller depending on the LCD bias, respective pixels to be darkened and the mode of operation of the LCD ie either static or multiplexed. Fig. 2 shows static mode of operation, the other mode has been discussed in detail later. Fig. 3 shows the whole hardware setup.

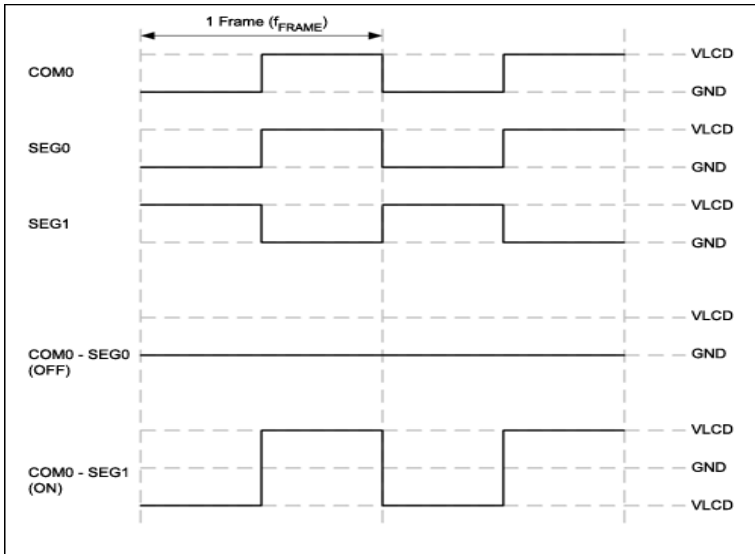


Fig. 2. Driving Waveforms for Static Mode

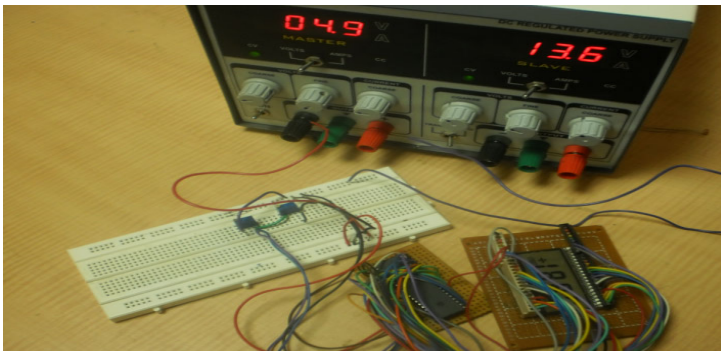


Fig. 3. Hardware Setup

4 Frame Frequency

Refresh rate or the temporal resolution of an LCD is the number of times per second in which the display draws the data it is being given. High-end LCD televisions now feature up to 240 Hz refresh rate. The frequency at which the LCD is driven (known as the frame frequency) varies from one LCD panel to another. The proper value for a given application is usually derived by experimentation on a specific hardware setup. Since the rate at which an LCD segment can change state is limited by the overall capacitance of the segment, the LCD will operate properly only in a specific range of frame frequencies with some particular values dependent on the type of clock source chosen ie either an external crystal or the internal clock source. The frame frequency is also dependent on the prescaler select bits defined in the prescaler registers. The resultant frame frequency is governed by the following formula for the static mode of operation.

$$\text{Clock source}/(4 \times 1 \times (\text{LP} + 1)) \quad (1)$$

Here LP denotes the prescaler bits and Clock Source is $F_{osc}/8192$ where F_{osc} is the internal clock source with a frequency of 8 MHz. As the refresh rate of the display is reduced to the minimum period of an afterimage on the human eye, we start to feel flicker. But at the same time running the LCD at very high frame frequencies does not improve the quality of the image because its effects go unobservable and also leads to unnecessary power wastage. Table 1 and Fig. 4 show the measure of the current consumption with respect to frame frequency.

Table 1.

Frame frequency(Hz)	Current Consumption(mA)
22.199	0.5
24.199	0.53
27.132	0.6
30.51	0.69
34.87	0.78
40.69	0.87
48.828	0.98
61.035	1.18
81.38	1.48
122.07	2.09
244.1473	3.54

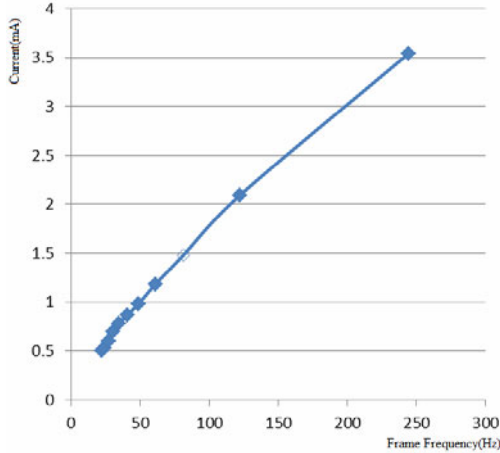


Fig. 4. Current(mA) vs Frequency(Hz)

The after image period of human eye is around 30Hz. The current consumption at 30.51 Hz is .69 mA whereas at 244.1473 Hz is 3.54 mA. For an LCD bias of 5 V, a total of 14.25 mW could be saved operating at 30.51 Hz which is a total of 80.5% that of 244.144 Hz.

5 Contrast Control

This method is based on controlling the darkness of the pixel depending on the ambient light. The contrast depends on the value of the LCD bias ie VLCD. The idea is that when there is sufficient light in the surroundings, pixel darkness can be compromised to save power. This is achieved both manually and automatically with the help of photodetector.

5.1 Manual Control

In this technique, the user can himself decide the contrast of the display by calibrating the resistance of the potentiometer R1 as shown in Fig. 5.

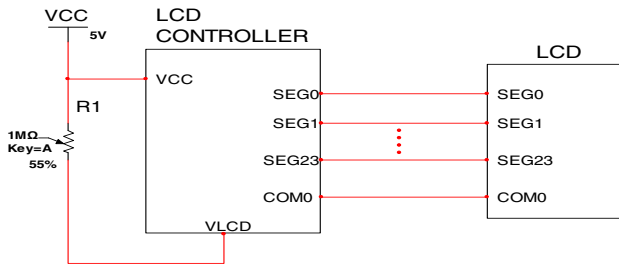


Fig. 5. Manual Control on Contrast

Table 2 and Fig. 6 is a measure of the current consumed with respect to the manually controlled bias voltages.

Table 2.

Voltage(V)	Current(mA)
2.98	0.81
3	0.84
3.05	0.86
3.18	0.92
3.32	0.97
3.62	1.18
3.72	1.22
4.54	2.04
4.58	2.1
4.64	2.22
4.84	2.45

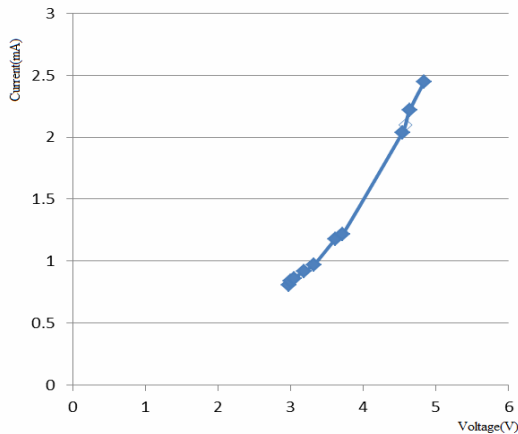


Fig. 6. Current(mA) vs Voltage(V)

An interesting result came out was that the differences in the contrast were insignificant in most of the voltage ranges ie values of VLCD from 4.84 V to around 3.32 V but the differences in the current consumption were quite significant as evident from the above readings. However, if the bias is continued to decrease further, the changes in contrast can be viewed clearly till a value below 2.95 V when the display turns off completely. The point being emphasized here is that it is upon the user that if he finds the display legible enough to read in the ambient light, he may increase the potentiometer resistance (provided in the form of an external knob) suitably to reduce the LCD bias and also the current consumption. A total of 9.338 mW of power could be saved between 4.84 V and 3 V which is around 78.74% that of maximum power with reference frame frequency of around 244 Hz.

5.2 Automatic Control

The LCD bias voltages are being obtained automatically on the basis of the light falling on the photo diode which are within the previous voltage ranges. Here the display became high contrast in darkness and became a little faint in sufficient ambient light, all by itself but always remained readable.

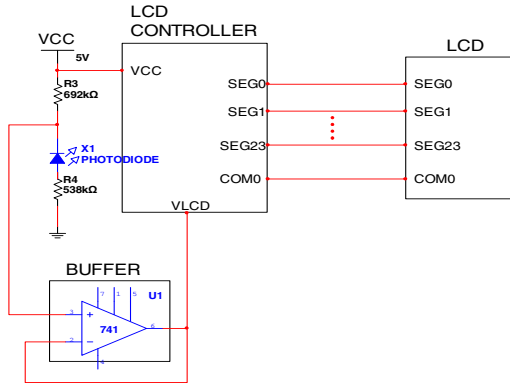


Fig. 7. Automatic Control on Contrast

An operational amplifier was used to prevent the loading effect due to the presence of such high resistances and ensure proper functioning of the system. However, here an additional current consuming branch of that of the two resistors and the photo diode came, but the current consumed in this branch was of the order of micro amperes which was very less compared to the current consumed by the LCD. Table 3 and Fig. 8 draw a comparison between the automatically generated voltage and the current consumed. Power saved between 4.19 V and 2.99 V was around 4.98 mW.

Table 3.

Voltage(V)	Current(mA)
2.99	0.77
3.14	0.86
3.23	0.91
3.37	0.99
3.43	1.05
3.53	1.1
3.63	1.22
3.78	1.29
3.87	1.37
3.94	1.42
4.03	1.52
4.12	1.59
4.19	1.74

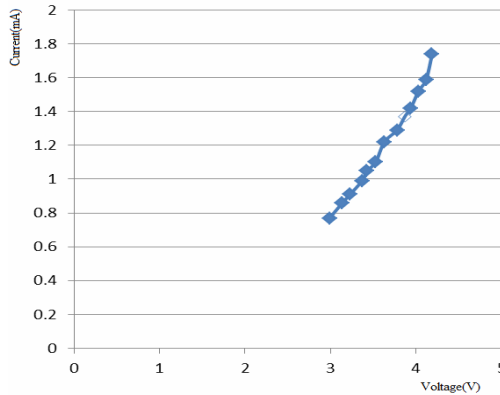


Fig. 8. Current(mA) vs Voltage(V)

6 Multiplexed Mode

When a large number of pixels are needed in a display, it is not technically possible to drive each directly since then each pixel would require independent electrodes. For example, the no. of SEG pins in our controller is 24 so for driving more no. of segments, the display is has to be multiplexed. In a multiplexed display, electrodes on one side of the display are grouped and wired together (typically in columns), and each group gets its own voltage source. On the other side, the electrodes are also grouped (typically in rows), with each group getting a voltage sink. The groups are designed so each pixel has a unique, unshared combination of source and sink. The electronics or the software driving the electronics then turns on sinks in sequence, and drives sources for the pixels of each sink.

The multiplexed mode of operation is primarily used for driving LCDs with large no. of segments but what we suggest is driving LCDs (with lesser no. of segments) also in this mode since it leads to a lot of power saving without much of an effect in display readability. The multiplexed mode is further subdivided into $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$ multiplex depending on the no. of COM pins that the LCD has. But the LCD that we use here consists of 24 segments and only 2 COM pins so only $\frac{1}{2}$ bias multiplexed mode of operation can be used in this particular LCD which means that a total of two voltage levels can be given to the VLCD pins using suitable bias circuitry of Fig. 9.

Here SEG0, SEG6 and SEG3 are driven by VLCD3=5V whereas SEG1, SEG2, SEG4 and SEG5 are driven by VLCD2=2.5V and similarly for the other segments. As it is clear from the Fig. 10, the display is quite readable. Power consumed during the static mode of display at the same frame frequency was 17.7 mW for VLCD=5 V and I=3.54mA. Now for multiplexed mode, power was consumed by both LCD bias voltages.

$$P1(\text{for VLCD3}) = 5 \times 2.25 = 11.25 \text{ mW}$$

$$P2(\text{for VLCD2}) = 2.5 \times .5 = 1.25 \text{ mW}$$

So, the total power consumed was 12.5 mW leading to a power saving of 5.2 mW or around 30% compared to the static mode.

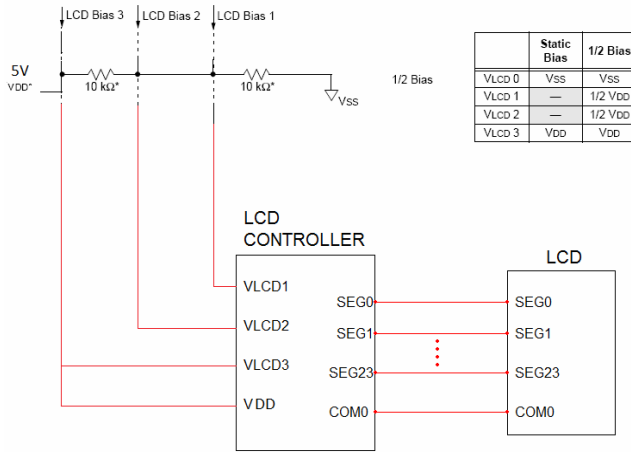


Fig. 9. VLCD bias circuitry

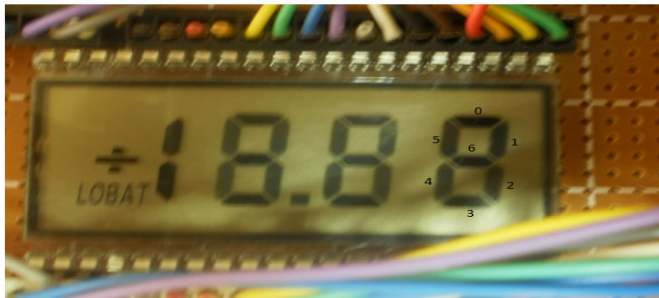


Fig. 10. Multiplexed display

7 Conclusion

A total of around 29 mW of power can be saved by employing the above techniques of optimizing the control parameters like frame frequency, contrast control and mode of display with individual contribution of 14.25 mW, 9.338 mW and 5.2 mW respectively. Most of them are either software based or need very little hardware to be added.

References

1. Shim, H., Pedram, M., Chang, N.: A Compressed Frame to Reduce Display Power Consumption in Mobile Systems. In: Proceedings of ASP DAC (2004)
2. Choi, I., Shim, H., Chang, N.: Low-power color TFT LCD display for hand-held Embedded systems. In: Proceedings of International Symposium on Low Power Electronics and Design, pp. 112–117 (August 2002)

3. Gatti, F., Acquaviva, A., Benini, L., Ricco, B.: Low Power Control Techniques For TFT LCD Displays. In: Proceedings of the 2002 international conference on Compilers, architecture, and synthesis for embedded systems (2002)
4. Pedram, M.: Power Optimization and Management in Embedded Systems. In: Proceedings of the ASP-DAC 2001, Asia and South Pacific (2001)
5. Chen, C., Chen, D., Xu, X., Wen, X., He, L.: A Novel Bias Circuit Design in Low Power LCD Driver. In: Proceedings. 7th International Conference on Solid-State and Integrated Circuits Technology (2004)
6. Microchip: PIC16F917/916/914/913 Data Sheet, Microchip
7. Application note 4039. On: Using the DS89C450 as a Static LCD Display Controller

Language Independent Icon-Based Interface for Accessing Internet

Santa Maiti¹, Debasis Samanta¹, Satya Ranjan Das¹, and Monalisa Sarma²

¹ School of Information Technology, Indian Institute of Technology Kharagpur, India
{santam, satyad, dsamanta}@sit.iitkgp.ernet.in,

² Computer Science and Engineering, National Institute of Technology Durgapur
monalisa_sarma@yahoo.co.in

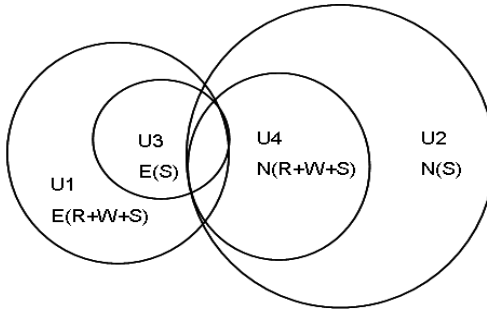
Abstract. With the advancement of the information technology, Internet becomes an essential part in our every sphere of life. However, the benefits of Internet are limited to only educated people who can read and write in English or atleast in their own mother languages. This work addresses this limitation and proposes an icon-based user interface so that low educated people can retrieve information from the Internet. Our proposed interface supports user to compose their queries by means of selecting icons. How such interface can be designed in the context of information retrieval related to tourism domain is addressed in this paper.

Keywords: Icon-based user interface, language independent communication, human-computer interaction, information retrieval.

1 Introduction

At present a huge information repository is maintained in the Internet. According to a recent survey, the indexed web contains at least 14.3 billion pages (March, 2011). Out of this large information repository, 72% of the pages are in English. So, to get the maximum advantages of Internet, a user should be an English literate. Traditionally, literacy is described as the ability to read and write. Though the literacy all over the world is 84%, it is quite a matter of concern for many of the developing countries where approximately 54.74% of the total population is literate. Further, this literacy includes familiarity with their native languages. But, native language literacy is not sufficient because of scarcity of native language support in Web. So, this huge information repository of Internet is confined only within a certain group of literate people.

Presently, English is the most widely published language all over the world. Over 1.8 billion people use English as first, second and foreign language. It is an official language in 52 countries as well as many small colonies and territories. Figure 1 depicts a Venn diagram of language familiarity so far their ability to read (R), write (W), and speak (S) in their native (N) and English (E) languages are concerned. In Figure 1, the users are classified as U1, U2, U3 and U4. User U1 and U4 can read, write and speak English and native language respectively. User



E-English, N-Native Language, R-Read, W-Write, S-Speak

Fig. 1. Language literacy of world population

U3 is a subset of U1 can speak English. Similarly, user U2 (superset of U4) can speak their own native language. So, the users in U2-U4 and a major portion of U3, U4 are our target user. Such people cannot read, write English, mostly used as Internet language. We term such people as under privileged people. Typically these people are rickshaw puller, porter, farmer, shop keeper, gate keeper etc.

An icon can be defined as a small graphics representation of some information. It can also be treated as a language independent faster mechanism of communication, facilitating recognition rather than recall. Recently, icon has been chosen as a primary mode of interaction between man-machine for its expressiveness. Single icon is capable to represent a whole word or a phrase. This work aims to develop a framework of icon-based interaction to Internet so that target users can retrieve their necessary information from web repository. With the proposed interface, a user would fire queries to Internet by selecting some icons. The query would be fed to the search engine and then the search result would be captured by the system. Later it is planned to mine the retrieved result to represent answer in concise text format. Finally, the concise text will be represented into user understandable visual form. The design aspects of icon-based interface and some major issues such as deciding a set of icons that an icon-based interface should have, managing a large set of icons in a limited display area etc. are addressed in this paper.

The organization of the paper is as follows. In Section 2, existing related work on icon-based interface is discussed. Section 3 talks about the proposed methodology of designing icon-based interface. The icon-based interface and its working procedure are discussed in Section 4. Section 5 shows the user evaluation of interface and Section 6 concludes the paper.

2 Related Work

In this section, we survey the existing approaches dealing with icons and implementation of icon based interface in different context. The usefulness of icon in interface is discussed in [15]. Symbols have been used in national and international signposting for public service functions. This includes signs for highways

and airports, electronics and packaging [2]. One actual use of the pictographic representation of data can be found in work on military battlefield displays in which army units on the battlefield are represented on a computer display [11]. Tatomir and Rothkrantz introduced a set of icons for constructing a map representing features such as crossing types and road blocks etc. [16].

The success of iconic interface depends on icon design strategy. The effects of icon design on human-computer interaction is discussed in [7,12] where icon characteristics are investigated to determine the speed and accuracy of icon identification. Icon metaphors, design alternatives, display structures, implementation and a summary of icon design guidelines are addressed in [11]. Abhishek and Basu introduced a disambiguate strategy to remove ambiguities in message generation and interpretation [3]. Gittin classified icon in different ways as by form (static and dynamic), color (monochrome and color) and type. Classification by type is known as icon taxonomy [18]. Different researcher addressed icon taxonomy according to their need although the basic idea is more or less same. Representational (alias - associative, nomic etc.) icons (petrol pump, to represent a petrol pump) were described as an example for a general class of object. Icon that attempt to convey concepts rather than to display the object itself is addressed as abstract (alias - mixed, metaphorical etc.) icon (broken glass to represent fragile). In Arbitrary (alias - key, symbolic etc.) type of icon there is no intuitive connection between the icon and its referent. Presently, icons can be classified on 30 different attributes and sub-options as - detailing, dimension, light-shadow, size, appearance, effects, pixilation etc. Now-a-days various types of icons are found in different computer environments and applications: (1) as a part of an Operating Systems desktop environment like Windows XP, (2) as a part of a specific computer application like in toolbar of Microsoft Word and (3) within Internet websites or other online applications.

Many icon-based interfaces have been implemented for different purpose. Hotel Booking System is designed for booking the hotel room by users with different linguistic backgrounds [19]. For person-to-person communication purpose Pictorial dialogue methods [4], CD-Icon, an iconic language-based on conceptual dependency [5], a graphical chatting program - Visual Messenger [8], language-independent communication using icons on a PDA [9] are developed. Iconic interface application is also designed for children to provide a playful learning environment [17] as the Elephants Memory, Clicker etc. Some icon-based Augmentative and Alternative Communication (AAC) systems are made for quadriplegic people. An AAC iconic system for the people with significant speech and multiple impairments [1], Mobile AAC system for handicapped person [13], Sanyog [6] for people with speech and motor impairments are some significant applications. An icon-based interface for communication in crisis situations on a PDA is developed to alert the people [10]. Optimal audio-visual representations for illiterate users of computers [14] is proposed to help semi-literate users for expressing and understanding information in health domain.

3 Proposed Methodology

The framework of proposed approach is shown in Figure 2. Using the icon-based interface our target user would frame a query related to tourism domain. The query would feed to Google search engine to obtain the web search result against the query. These search result is huge in size and unusable for our user. So this vast result will be mined to extract the concrete result. Finally the concrete result is transformed into iconic form. Using icon as an interaction medium arises some potential problems. Icon ambiguity may lead to major confusion in case of query formation. Still icon medium is selected for our target user as it is language independent and easy to learn. Since, there is innumerable number of queries possible in general and one or more icons are needed to represent each query, it is beyond the scope of this work to plan such a vast icon vocabulary. We therefore limit our work to queries related to tourism domain information only. Tourism domain is selected as the icons of this domain are little bit familiar to user as road signs, traffic symbols, railway symbols etc. Along with this, queries of this domain are definite upto some extent. Our proposed approach includes the following.

- Deciding icon vocabulary
- Maintaining large icon repository
- Organization of icons

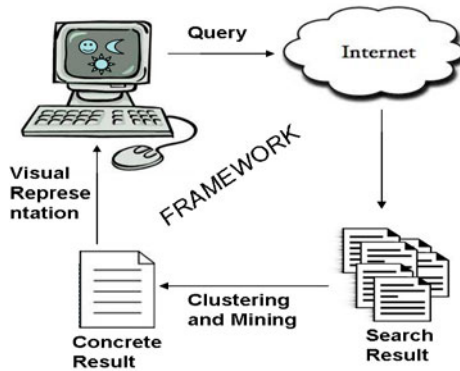


Fig. 2. Language literacy of world population

3.1 Deciding Icon Vocabulary

Towards the design of icon-based interface, our first task is to decide tourism related basic words, corresponding icons and then put them in the interface in a proper way. Icon optimization is important because presence of huge icons in interface may confuse the target user as well as it leads difficulty to find

desired icons from the pile. At the same time, we should ensure to cover the entire domain to generate tourism related queries. Redundancy of similar kind of icons is avoided as it goes against the icon optimization. To build the tourism corpus, basic tourism related words are collected from different tourism related magazines and web-sites. Stopwords (e.g. about, the, in etc.) are frequent in every domain but they are not important. So before calculating weight of each word, 319 stopwords are filtered out from the tourism corpus. Again a word can occur in the corpus in different morphological form (e.g. transport, transports, transported, transporting, transportation etc.) whose stem or root is same. So, the stem form of each word is found out and all words of different morphological forms are mapped to their stem words. Then the unique words of the corpus are identified. The term importance increases proportionally with the number of the times a word appears in the document as well as the occurrence of the word in the corpus. The term count in the given document is simply the number of times a given term appears in that document. This count is normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term t_i within the particular document d_j . Thus, we have the term frequency, defined as follows in Eqn. (1).

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in document d_j and the denominator is the sum of number of occurrences of all terms in document d_j . The document frequency is a measure of the general importance of the term (obtained by dividing the number of documents containing the term by the number of all documents) as shown in Eqn. (2).

$$df_i = \frac{|\{d : t_i \in d\}|}{|D|} \quad (2)$$

with $|D|$ is total number of documents in the corpus. $|\{d : t_i \in d\}|$ is the number of documents where the term t_i appears (that is $n_{i,j} \neq 0$). Then

$$(tf - df)_{i,j} = tf_{i,j} \times df_i \quad (3)$$

The $tf - df$ value for a term will always be greater than or equal to zero. The term rank of tourism corpus is finalized on $tf - df$ weighting. After calculating the $tf - df$ weightage of each term they are sorted to finalize top ranked terms. Frequent queries fired to Google using those keywords are collected. Some supporting query words are added to the main keyword list to complete the query phase. The initial tourism corpus contains 169610 words related to tourism domain. After filtering with stopwords and considering the stem form word the corpus size is reduced to 139052 words. Then 30588 words are identified as unique word. The $tf - df$ value of these unique words are calculated and ranked in decreasing order. Out of top 500 words, 316 words are selected as tourism related important word

as some are implicit to query (e.g. word "list" is implicit in the query "hotel list of Kolkata") and some are synonymous. These words are used to extract frequent queries fired to Google in tourism domain. 127 supporting words are added to the corpus to form the queries. At present the interface contains 235 icons. Later few icons are planned to be reduced considering context sensitive different morphological form of icons. On the other hand, some icons are planned to be incorporated depending on user testing feedback. Presently, we have developed the interface with tourism related icons, available in Internet. Selected icons are modified as necessary. We tried to represent an icon similar to the real object. In some cases the concept is represented where it is difficult to represent a real object. Designing proper icons for tourism domain is beyond the scope of this work.

3.2 Maintaining Large Icon Repository

Presently 235 icons are used in the interface. To manage the icon database, XML document is used as an index to support easy icon retrieval. Different icon attributes like corresponding keyword, storage location (path), position of that icon in the hierarchy are maintained in the XML file. Structure of the XML file is shown below.

```
<?xml version="1.0" encoding="utf-8" ?>
<database>
<icondata xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd=
"http://www.w3.org/2001/XMLSchema" >
  <Identity>d78cc075-014a-49c8-8884-41e72065496c</Identity>
  <pname>where</pname>
  <collection>1</collection>
  <keyword>town</keyword>
  <imagepath>town.jpg</imagepath>
</icondata>
```

3.3 Icon Organization

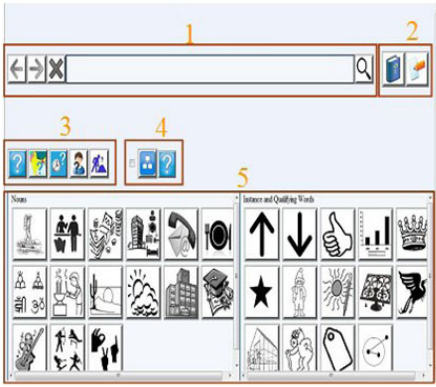
All the icons can not be kept in a single interface, so hierarchy is maintained to organize them. At the primary stage selected icons are categorized under 5 basic 'wh' icons - 'what', 'when', 'where', 'who' and 'how'. Icons related to any place are kept under 'where'. Time, person and verb related icons are kept under 'when', 'who' and 'how' respectively. The rest of the icons are placed under 'what'. Qualifying type of icons is put separately in the interface. With this basic categorization grouping and sub-grouping is also followed (e.g. all types of vehicles are kept under main transport icon). Similar types of icons are placed together. Presently, interface maintains 4 level hierarchies (e.g. *travel* → *transport* → *train* → *ac - 2 - tier*.). Initially, it is not possible for users to find out their desired icons at first chance, but this visual search time would decrease with the familiarity of the system.

4 Icon-Based Interface to Access Internet

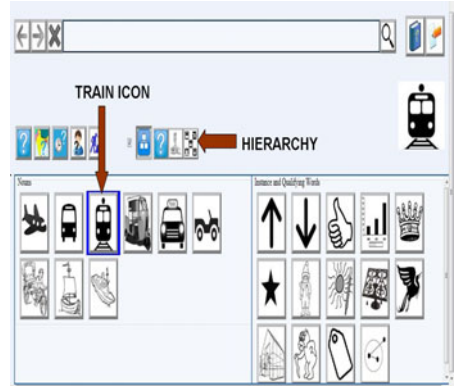
A snapshot of the developed interface is shown in Fig. 3a. The top leftmost panel (Panel 1) contains functional icons (backspace, write back, erase, search) and shows the icon selected by user. This offers easy reversal of actions to the user. Panel 2 contains ‘help’ and ‘feedback’ icons. Basic ‘wh’ icons (‘what’, ‘when’, ‘where’, ‘who’ and ‘how’) are kept in panel 3. Using these ‘wh’ icons a user can see basic icons related to that particular ‘wh’. To go further in the hierarchy Panel 4 is helpful. ‘Hierarchy’ icon enables user to find icons which are kept in hierarchical order. To reduce short-term memory load of user, the hierarchical path user follows is displayed in Panel 4. It also allows user easy state (hierarchy) transition. Panel 5 is the display area of all ‘wh’ icons. User can select any icon displayed in this area by single click. Using this interface the users are able to generate travel related query and feed the query in the search engine using search icon. Google search engine is used in this application for searching purpose. An example of basic tourism related query (“train reservation”) formation is shown in Figure 3b, 3c and 3d. To construct this query the user has to select the ‘hierarchy’ option first. To reach the ‘train’ icon, ‘travel’ and ‘transport’ icons are selected respectively (Figure 3b). To select target icons the ‘hierarchy’ option should turn off. ‘Train’ icon can be selected by single click. Similarly ‘reservation’ icon can be selected from primary interface (Figure 3c). Search result can be obtained by clicking the search button as shown in Figure 3d. The search result is presently in the text form. It would be mined later to extract the concrete result and to present it in a visual form as an output.

5 User Evaluation

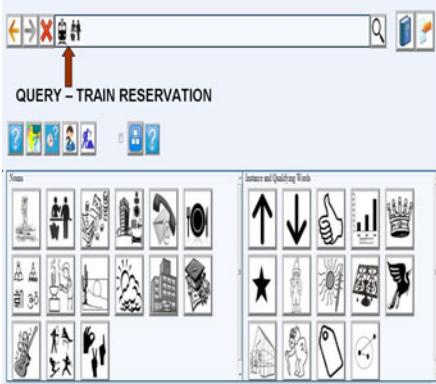
To substantiate the efficiency of the developed interface, we have done user evaluation. Ten users of different backgrounds have participated in our evaluation procedure. The user profiles are summarized in Table 1. As the developed system is totally new for our target users, they are undergone a training followed by testing to check the usability of the interface. In training process the icons are introduced to the user for thrice in some time gap. They are also asked to recognize some icons randomly. Finally, the interface is tested in respect of icon recognisability and query construction ability of user. We decide fifty tourism related queries (given in Table 2) as a benchmark. Participants are asked to generate the queries five each, using the interface. The accuracy of query generation is calculated as the number of icons correctly chosen by the user divided by total number of icons needed to select to form the query in right way. The average accuracy for each user is then calculated as shown in Table 3. From Table 3 we can see that the overall average accuracy of finding search keywords with the proposed icon-based keyboard is 0.758 (approx).



(a) A snapshot of the icon-based interface.



(b) Use of hierarchy.



(c) Query generation.



(d) Search result.

Fig. 3. Icon-based interface

Table 1. User details

User	Background			Language Familiarity	
	Age	Education	Profession	English	Native
S1	22	Class X	Domestic Helper	Low	High
S2	24	Class VI	Gate keeper	Low	Medium
S3	36	Class V	Rickshaw Puller	None	Medium
S4	27	Class IV	Conductor	Low	Low
S5	32	Class V	Farmer	Medium	Medium
S6	43	Class II	Porter	None	Low
S7	19	Class VI	Shop Keeper	None	Medium
S8	32	Class V	Sweeper	None	Medium
S9	39	Class VII	Driver	Low	Medium
S10	20	Class III	Barber	None	Low

Table 2. Tourism related query

Query	Query
1. Nearest station <place>	26. Sea beach in south India
2. Distance <place> and <place>	27. Single room hotel rent in <place>
3. Festivals in <place>	28. Photo gallery of <place>
4. Hotel <place>	29. Climate of <place> in time
5. Transport <place>	30.<place> zoo
6. People culture <place>	31. Buddha stupa in <place>
7. Food <place>	32. Shiva temple in <place>
8. Map <place>	33. <place> lake
9. Weather <place>	34. Botanical garden in <place>
10. Fair in rural India	35. Church in <place>
11. Five star hotel in <place>	36. Forest in <place>
12. Tour package of <place>	37. <place> border
13. Train between <place> and <place>	38. Route from <place> to <place>
14. Adventure sports in <place>	39. Hindu pilgrim in <place>
15. Beaches in <place>	40. Train ticket reservation <place> to <place>
16. Wildlife in <place>	41. Royal museum in <place>
17. Hospital in <place>	42. Reservation status in <place> hotel
18. Five day tour plan <place>	43. Temperature of <place> in <month>
19. Best season to visit <place>	44. Stadium in <place>
20.<place> and <place> flight ticket price	45. Low budget <place> and <place> travel
21. Market in <place>	46. Village craft fair in <place>
22. Arrival time of <place> and <place> plane	47. Fort in <place>
23. Migratory bird watching <place>	48. Sunset in <place> beach
24. Fruit in <place>	49. Tribal festival in <place>
25. Orchid of north <place>	50. <place> college

Table 3. Interface testing result

User	Query	Accuracy	Avg. Accuracy
S1	8, 13, 22, 10, 17	1, 1, 0.8, 0.33, 1	0.826
S2	36, 43, 44, 46, 47	1, 1, 1, .5, 1	0.9
S3	1, 23, 35, 28, 20	0.66, 0.5, 1, 0.5, 1	0.732
S4	3, 11, 16, 21, 26	0.5, 0.75, 1, 0.5, 1	0.75
S5	5, 7, 19, 32, 34	1, 1, 0.75, 1, 0.66	0.882
S6	38, 39, 40, 48, 49	.33, .66, .6, .66, .33	0.516
S7	37, 41, 42, 45, 50	1, .66, .5, .8, 1	0.792
S8	6, 14, 18, 27, 24	0.33, 0.33, 0.8, 0.8, 1	0.652
S9	2, 15, 25, 29, 30	1, 1, 0.66, 1, 1	0.932
S10	4, 9, 12, 31, 33	0.5, 1, 0.66, 0.33, 0.5	0.598

6 Conclusion and Future Work

Developed icon-based interface is a prototype version of the proposed approach. Primary evaluation of interface has been done but more formative evaluation as well as summative evaluation is needed to enhance the user friendliness of

interface. Next work is to mine the search result to point out concrete answer to query. The concrete result would be furnished into iconic form. The interface is also planned to extend so that it can support multiple domain queries. Complete implementation would eventually offer a useful tool to our target users for accessing the Internet.

Acknowledgement

This work is carried out under the project sponsored by Department of Information Technology.

References

1. Albacete, P.L., Chang, S.-K., Polese, G.: Iconic language design for people with significant speech and multiple impairments. In: Mittal, V.O., Yanco, H.A., Aronis, J., Simpson, R.C. (eds.) *Assistive Technology and Artificial Intelligence*. LNCS (LNAI), vol. 1458, p. 12. Springer, Heidelberg (1998)
2. Arnstein, J.: *The international directory of graphic symbols*. Kogan Page (1983)
3. Abhishek, Basu, A.: Disambiguation in ambiguous iconic environments by constraint satisfaction. In: *Proceedings of the Knowledge Base Computer System (KBCS)*, Hyderabad, India (2004)
4. Barker, P.G., Manji, K.A.: Pictorial dialogue methods. *International Journal of Man-Machine Studies* 31, 323–347 (1989)
5. Beardon, C.: Cd-icon: an iconic language based on conceptual dependency. *Digital Creativity* 3, 111–116 (1992)
6. Bhattacharya, S.: An iconic system for multilingual communication for people with speech and motor impairment. Master's thesis, Department of Computer Science and Engineering, IIT Kharagpur (2004)
7. Blankenberger, S., Hahn, K.: Effects of icon design on human-computer interaction. *International Journal of Man-Machine Studies* 35, 363–377 (1991)
8. Chen, A.: *The Visual Messenger Project*. Thesis, University of Washington (2004)
9. Fitrianie, S., Rothkrantz, L.J.M.: Communication in Crisis Situations Using Icon Language. In: *IEEE International Conference on Multimedia and Expo, ICME 2005*, pp. 1370–1373 (2005), doi:10.1109/ICME.2005.1521685
10. Fitrianie, S., Dacu, D., Rothkrantz, L.J.M.: Human communication based on icons in crisis environments. In: Aykin, N. (ed.) *HCI 2007*. LNCS, vol. 4560, pp. 57–66. Springer, Heidelberg (2007)
11. Gittins, D.: Icon-based human-computer interaction. *International Journal of Man-Machine Studies* 24, 519–543 (1986)
12. Isherwood, S.J., McDougall, S.J.P., Curry, M.B.: Icon identification in context: The changing role of icon characteristics with user experience. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49(3), 465–476 (2007)
13. Lee, E., Hwang, E., Hur, T., Woo, Y., Min, H.: A Study on the Predicate Prediction Using Symbols in Augmentative and Alternative Communication System. In: Chittaro, L. (ed.) *Mobile HCI 2003*. LNCS, vol. 2795, pp. 466–470. Springer, Heidelberg (2003)

14. Medhi, I., Prasad, A., Toyama, K.: Optimal audio-visual representations for illiterate users of computers. In: Proceedings of the 16th International Conference on World Wide Web, New York, USA, pp. 873–882 (2007)
15. Rogers, Y.: Icons at the interface: their usefulness. *Interacting with Computers* 1(1), 105–117 (1989)
16. Tatomir, B., Rothkrantz, L.: Crisis management using mobile ad-hoc wireless networks. In: Proceedings of the 2nd International ISCRAM Conference, Brussels, Belgium, pp. 147–149 (2005)
17. Uden, L., Dix, A.: Iconic interfaces for kids on the Internet. In: IFIP World Computer Congress, Beijing, China, pp. 279–286 (2000)
18. Wang, H.-F., Hung, S.-H., Liao, C.-C.: A survey of icon taxonomy used in the interface design. In: Proceedings of the 14th European conference on Cognitive Ergonomics, London, United Kingdom. ACM International Conference Proceeding Series, vol. 250, pp. 203–206 (2007)
19. Yazdani, M., Mealing, S.: Communicating through pictures. *Artificial Intelligence Review* 9(2-3), 205–213 (1995)

Contribution of Oral Periphery on Visual Speech Intelligibility

Preety Singh¹, Deepika Gupta², V. Laxmi³, and M.S. Gaur⁴

Computer Engg. Deptt., Malaviya National Institute of Technology, Jaipur, India
{¹prty Singh,²deepika.guptaa19,³vlgaur,⁴gaurms}@gmail.com

Abstract. Visual speech recognition aims at improving speech recognition for human-computer interaction. Motivated by the cognitive ability of humans to lip-read, visual speech recognition systems take into account the movement of visible speech articulators to classify the spoken word. However, while most of the research has been focussed on lip movement, the contribution of other factors has not been much looked into. This paper studies the effect of the movement of the area around the lips on the accuracy of speech classification. Two sets of visual features are derived: one set corresponds to the parameters from an accurate lip contour while the other feature set takes into account the area around the lips also. The features have been classified using data mining algorithms in WEKA. It is observed from results that features incorporating the area around the lips show an improvement in the performance of machines to recognize the spoken word.

1 Introduction

Visual cues aid automatic speech recognition in adverse audio environment [11]. Visual speech recognition or lipreading is of particular use to people with limited hearing abilities. Extensive research is being done to improve lipreading systems for optimal human-computer interaction. Various studies [10,12,13] use different segmentation techniques to obtain an accurate lip contour. Visual features are derived from the mouth region to form a visual feature vector. These features may be pixel intensities, geometric parameters or fitted curves based on statistical models [10,12,13]. Optical or thermal flow [14] between successive images is also considered for speech recognition. Research has also been done on using visual features extracted from profile views [7].

Humans subconsciously use visual cues from the movement of facial muscles, cheek bones and eyebrows to interpret speech in a challenging audio environment. However, it is difficult to train computers to imitate this cognitive ability of man. Incorporating movement of speech articulators other than lips might increase the efficiency of speech recognition systems for human-computer interface. However, speakers' images carry little information about movements of the facial muscles in the near vicinity of the lips. To our knowledge, no quantitative results have been published that show how the movement of these muscles in the periphery of the oral region correlates with visual speech recognition.

In this paper, we evaluate the contribution of the peripheral oral area on speech classification. Two sets of visual features are used for speech classification. One set is derived from a dilated lip contour that includes the area around the lips. The other set corresponds to a fairly accurate outer lip contour. Comparison of precision accuracy is done to analyze the effect of peripheral oral area on speech recognition.

The paper is organized as follows: Section 2 presents related work in the field of visual speech recognition. Section 3 presents our proposed methodology of determining the lip boundary. Extraction of visual features is discussed in Section 4. The experimental setup is described in Section 5. Section 6 is a discussion of the results and Section 7 concludes the paper.

2 Related Work

Yau et al [17] have extracted dynamic features representing mouth motion. Discrete Stationary Wavelet Transform and moments (Hu, Zernike and geometric) have been used to derive features invariant to translation, rotation and scale. Artificial Neural Network has been used for classification. Recordings from a single speaker were used for experiments. The vocabulary consisted of nine consonants. The paper reports a recognition rate of 85%.

Feng and Wang [3] have used dynamic texture features based on Dual Tree Complex Wavelet Transform to capture lip motion information. The visual dynamic features have been calculated from Canberra distances between lip texture features in adjacent frames. Chinese database is used with 78 isolated words. An accuracy of 79.83% has been reported.

Lee and Park [8] have used the pixel based approach to extract visual features. Principal Component Analysis is applied to determine main modes of variation. A new method, Hybrid Simulated Annealing, has been developed and used for classification. It shows an error rate of 33.4% as compared to 36.1% obtained from the Expected Minimization (EM) algorithm used for training the Hidden Markov Model (HMM) classifier. The experiments have been performed on the digits zero to nine in Korean with 56 speakers and three utterances per speaker.

Zhang et al [18] have used ten utterances of the seven weekdays spoken by ten subjects. Lip dimensions have been derived using a Markov random field framework. With various combinations of the features, visual speech recognition is done for both speaker dependent (S.D.) and speaker independent (S.I.) environment. They have reported a maximum accuracy of 78.28% for S.D. and 48.43% for S.I. visual speech recognition.

Faruque et al [2] have used curve based ASM to extract the lip contour. Twelve different geometric parameters containing four heights in outer contour and eight heights in inner contour along with the eight weights of eigenvectors are used to describe the shape of the lip. Using this feature vector, maximum accuracy obtained is 31.91% for phonetic classification.

3 Proposed Methodology for Visual Speech Recognition

Visual features are to be extracted from the region of interest (ROI). In our paper, we focus on the lips and the surrounding area. the process of extraction of visual features is shown in Figure 1. Segmentation of lips in RGB space is challenging due to the similarity in the colour of skin and lips. Since segmentation of lips is not our primary objective, the lips of the subjects have been painted blue for easy differentiation in the HSV space. To study the effect of the area around the lips on recognition accuracy, we have used two methods to derive the lip contour. The first method extracts the lips, the immediate area around it and the oral cavity. The second method extracts the area bounded by the lips only.

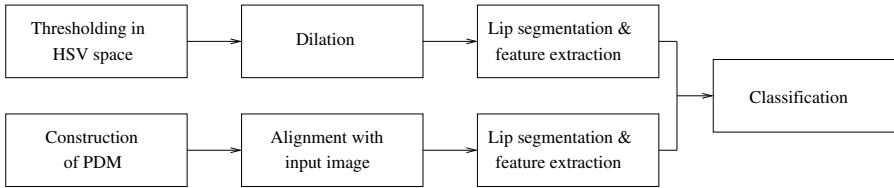


Fig. 1. Proposed Methodology for Visual Speech Recognition

3.1 Visual Feature Extraction by Method 1

In this method [15], the input RGB image is converted into HSV space. Thresholding is done to segment lips from the rest of the face. Morphological operation of dilation is applied to the segmented lips to incorporate the area around the mouth. Application of Sobel edge detector and connectivity analysis yields the boundary of the region of interest, that is, in effect, a dilated contour of the lips (refer Figure 2b). This is a good approximation of the lip boundary containing the peripheral area also.

3.2 Visual Feature Extraction by Method 2

This method makes use of a Point Distribution Model [5] to track the lip contour. A PDM is constructed using ten open mouth images of the speakers. Six key points are located on the lip contour in the training images. Between each pair of key points, twenty points are interpolated to define the lip boundary. This gives us a feature vector of 120 points. The mean of these points for ten images is computed to yield the PDM template which is then placed on the input image. The input image is translated and rotated w.r.t. the PDM template for alignment. The PDM model converges to the lip contour based on intensity analysis. This method gives a fairly accurate lip contour (refer Figure 2c).

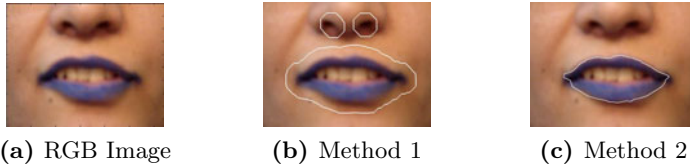
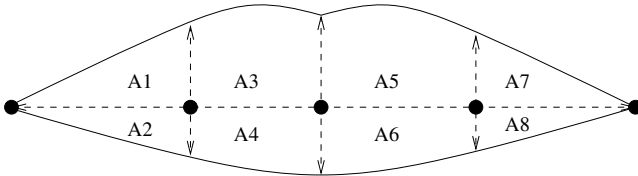


Fig. 2. Region of interest obtained from Methods 1 and 2

4 Extraction of Visual Features

Visual features have been extracted from the lip contour obtained by both the methods. The total lip height (h) and lip width (w) are calculated from the lip boundary. The area within the lip boundary is divided into vertical segments $A_1 \dots A_8$, as shown in Figure 3. It has been shown earlier in [5] that area segments within the lip contour are good contributors to visual speech recognition. The feature set now consists of a total of ten geometric features. This feature vector is used for classification of speech. The derived feature vector is:

$$[h \ w \ A_1 \ A_2 \ A_3 \ A_4 \ A_5 \ A_6 \ A_7 \ A_8]$$



A1..A8 are the areas of the lip segments

Fig. 3. Area segments as visual features

The features extracted from Method 1 is inclusive of the area around the lip region while those extracted by Method 2 gives the actual dimensions of the lip. Comparison of the two feature vectors helps us evaluate the contribution of the additional information on speech classification.

5 Experimental Setup

An audio-visual database was recorded in office environment with moderate illumination. 10 subjects were used for recording. The vocabulary consisted of ten digits viz., *zero* to *nine*. Three utterances per digit were recorded for each speaker. The utterance began and ended with the mouth in the closed position. In our experiments, we assumed that lip detection techniques exist and since this was not our primary objective, the focus of the camera was on the lower

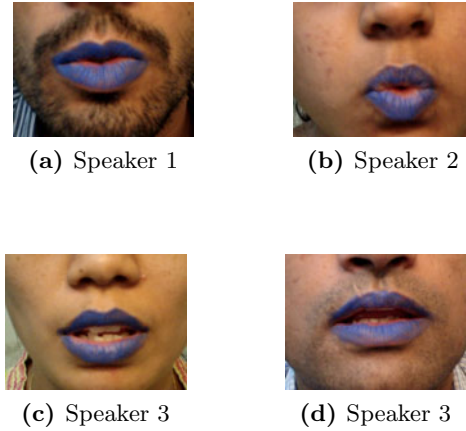


Fig. 4. Example frames of some speakers

part of the speaker's face in a frontal position (Figure 4). A total of 300 speech samples were collected. Each video was processed at 30 frames/second.

Lip boundary was extracted from the input image frames using Method 1 and Method 2, as described in Section 3. The features obtained from the lip contour are the lip height (h), lip width (w) and the area segments $A_1 \cdots A_8$. The features are normalized speaker-wise by dividing each attribute by the largest value in that set. The normalized feature vector $[h \ w \ A_1 \cdots A_8]$ is given as an input to data-mining algorithms in WEKA [16] for classification.

6 Results

Three classifiers in the WEKA toolkit have been used to classify the visual features for speech recognition. These are Random Forest (RF), Multilayer Perceptron (MLP) and Sequential Minimization Optimization (SMO). Random forest is an ensemble classifier consisting of many decision trees. It can handle thousands of input variables without deleting any variable. It can estimate missing data and give good accuracy even when a large proportion of the data is missing. Artificial Neural Networks can solve complicated problems where data computation is a difficult task. Also, due to the parallelism in its structure, it has a high computation rate [6]. They are less dependent on the underlying distribution of classes. A multilayer perceptron (MLP) ANN classifier has been used here. Default values of a single hidden layer, Support Vector Machines have the ability to achieve a globally optimum solution. It has been used in [4] for speech recognition. The sequential minimal optimization (SMO) algorithm has been used for training a support vector classifier. For all classifiers, default values in WEKA have been used.

The k -fold cross-validation method is used for testing with $k = 10$. In k -fold cross-validation, random partitioning of the original sample into k subsamples is

done. From these subsamples, $(k - 1)$ subsamples are used as training data and one subsample is retained for testing the model. This cross-validation process is repeated k times, with each of the k subsamples used as the validation data once. A single estimation is produced by averaging the k results. The advantage of this method is that all observations are used for both training and testing.

6.1 Evaluation Metrics

The metrics used for evaluating the results are True Positive Rate (TPR), False Positive Rate (FPR) and Precision (P).

- **True Positives:** This is the number of speech samples belonging to a particular class correctly classified.
- **False Positives:** It is defined as the number of speech samples falsely classified as belonging to the class being tested.
- **True Positive Rate (TPR):** It is the ratio of true positives to the total number of samples of a particular class. It is equivalent with the sensitivity of a classifier.
- **False Positive Rate (FPR):** This is the ratio of false positives to the total number of samples not belonging to that class.
- **Precision (P):** This is also known as the Positive Predictive Value (PPV). It is defined as the ratio of true positives to the sum of total samples classified as belonging to that class.

Table 1. TPR, FPR and Precision values for Random Forest (RF), Multilayer Perceptron (MLP) and Sequential Minimization Optimization (SMO)

Method	RF			MLP			SMO		
	TPR	FPR	P(%)	TPR	FPR	P(%)	TPR	FPR	P(%)
Method 1	0.54	0.05	54.1	0.36	0.07	35.5	0.27	0.08	26.6
Method 2	0.39	0.07	38.4	0.24	0.09	24	0.17	0.1	14.9
% Increase	–	–	40.89	–	–	47.92	–	–	78.5

6.2 Result Analysis

In our method, we observe from Table 1, that better recognition accuracy is obtained with the features obtained using Method 1 as compared to those obtained from Method 2. Best results are obtained with the Random Forest classifier giving an accuracy of 54.1% with Method 1 and 38.4% using Method 2. There is an average increase of 55.77% in precision values.

In [17], a high recognition rate of 85% has been reported. This could be because recordings from only one speaker have been used and the vocabulary is

also small. Feng and Wang [3] have used a Chinese dataset for experiments to give a recognition rate of 79.83% but these results need to be tested on English words. An accuracy of 48.43% is achieved by Zhang et al [18] for a speaker independent system. A maximum accuracy of 31.91% is reported by Faruque et al [2].

The results obtained by from an accurate lip contour (Method 2) are improved by additional information obtained from features extracted by Method 1. Also, results from our proposed method are comparable to those reported in literature. This suggests that speech recognition performance obtained by conventional features can be improved by incorporating additional information of movement from the surrounding oral area. An increased comprehension of speech by the computer can thus be achieved. The results indicate that just as humans subconsciously take the facial movement into account while lipreading, the computer also tries to imitate the same behaviour and gives better results when additional information is given to it.

7 Conclusion

Generally, visual speech recognition systems focus on features extracted from the lip contour for visual speech classification. Lip segmentation techniques strive to derive accurate lip contours for extraction of features. In our paper, we have included the area around the lips to extract visual features. This is compared with features derived from the actual lip contour. It is observed that addition of the oral periphery improves speech classification. The computational effort involved in extracting a precise lip boundary is not required and the additional area covered by the contour can improve the precision values of the recognition system. However, the results need to be tested on a larger dataset. In future, the authors would like to increase the dataset and also incorporate information from other speech articulators to improve machine speech recognition.

Acknowledgment

The authors are grateful to the Department of Science & Technology, Government of India, for supporting and funding this project.

References

1. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active Shape Models: Their Training and Application. *Computer Vision and Image Understanding* 61(1), 38–59 (1995)
2. Faruque, T.A., Majumdar, A., Rajput, N., Subramaniam, L.V.: Large vocabulary audio-visual speech recognition using active shape models. In: *International Conference on Pattern Recognition*, vol. 3, pp. 106–109 (2000)
3. Feng, X., Wang, W.: DTCWT-based dynamic texture features for visual speech recognition. In: *IEEE Asia Pacific Conference on Circuits and Systems (APCCAS 2008)*, pp. 497–500 (2008)

4. Gordan, M., Kotropoulos, C., Pitas, I.: A Support Vector Machine-Based Dynamic Network for Visual Speech Recognition Applications. *EURASIP Journal on Applied Signal Processing* 2002(11), 1248–1259 (2002)
5. Gupta, D., Singh, P., Laxmi, V., Gaur, M.S.: Comparison of Parametric Visual Features For Speech Recognition. In: *Proceedings of the IEEE International Conference on Network Communication and Computer (ICNCC 2011)*, pp. 432–435 (2011)
6. Kulkarni, A.D.: *Artificial neural networks for image understanding*. Van Nostrand Reinhold, New York (1994)
7. Kumar, K., Chen, T.H., Stern, R.M.: Profile View Lip Reading. In: *International Conference on Acoustics, Speech, and Signal Processing*, pp. IV: 429–432 (2007)
8. Lee, J.S., Park, C.H.: Hybrid Simulated Annealing and Its Application to Optimization of Hidden Markov Models for Visual Speech Recognition. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics* 40(4), 1188–1196 (2010)
9. Liew, A., Leung, S.H., Lau, W.H.: Lip contour extraction from colour images using a deformable model. *Pattern Recognition* 35(12), 2949–2962 (2002)
10. Matthews, I., Cootes, T.F., Bangham, J.A., Cox, S., Harvey, R.: Extraction of Visual Features for Lipreading. *IEEE Trans. Pattern Analysis and Machine Intelligence* 24(2), 198–213 (2002)
11. Neely, K.: Effect of visual factors on the intelligibility of speech. *Journal of the Acoustical Society of America* 28(6), 1275–1277 (1956)
12. Petajan, E., Bischoff, B., Bodoff, D., Brooke, N.M.: An improved automatic lipreading system to enhance speech recognition. In: *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI 1988)*, pp. 19–25 (1988)
13. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W.: Recent Advances in the Automatic Recognition of Audiovisual Speech. *Proceedings of the IEEE* 91(9), 1306–1326 (2003)
14. Saitoh, T., Konishi, R.: Lip Reading using Video and Thermal Images. In: *Proceedings of the International Joint Conference (SICE-ICASE 2006)*, pp. 5011–5015 (2006)
15. Singh, P., Laxmi, V., Gupta, D., Gaur, M.S.: Lipreading Using Gram Feature Vector. In: *Advances in Soft Computing*, vol. 85, pp. 81–88. Springer, Heidelberg (2010)
16. University of Waikato.: Recent Advances in the Automatic Recognition of Audiovisual Speech. Open Source Machine Learning Software WEKA, <http://www.cs.waikato.ac.nz/ml/weka/>
17. Yau, W.C., Weghorn, H., Kumar, D.K.: Visual Speech Recognition and Utterance Segmentation Based on Mouth Movement. In: *9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications*, pp. 7–14 (2007)
18. Zhang, X., Mersereau, R.M., Clements, M., Broun, C.C.: Visual speech feature extraction for improved speech recognition. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1993)*, vol. 2, pp II–II (2002)

Geo-spatial Pattern Determination for SNAP Eligibility in Iowa Using GIS

Sugam Sharma¹, U.S. Tim², Shashi Gadia¹, and Patrick Smith³

¹Dept of Computer Science,

²Dept of ABE,

³Dept of Civil Engg.,

Iowa State University, Ames, Iowa, USA 50011

{sugamsha, tim, gadia, pjsmith}@iastate.edu

Abstract. U. S. Department of Agriculture (USDA) administers SNAP - to assist low income citizens at the federal level, delegating the benefits distribution to the states. To avail the SNAP benefits the target population should be 50% or more low income (185% of federal poverty level) defined.

This work is intended to examine the spatial pattern of eligibility for SNAP in Iowa using GIS technology. Due to unavailability of income data on individual participant, this research employs the census tracts data as the finest spatial granularity. Using GIS technology, the tracts - meeting the income criteria for SNAP eligibility - are rendered on their county map with distinguish rendition pattern, differentiating them from unqualified ones.

Keywords: SNAP, Census Tract, County, Iowa, GIS.

1 Introduction

Supplemental Nutrition Assistance Program (SNAP) [1], previously known as Food Stamp Program (FSP), administrated by the U.S. Department of Agriculture (USDA) [3] is helping the low income or no income population since a long period of time. USDA coordinates with each state, which eventually distributes of the SNAP benefits to their eligible population.

Earlier, FSP used to use colored paper stamps or coupons- brown, blue, and green – each carrying different price value which is one , five and ten dollars respectively. In today's modern world colored paper stamp/coupons are replaced by the cards to buy any food such as soft drinks and confectionery under the food-stamp program. There is no constraint whether to buy nutrition food or not but the food should be prepackaged.

This study targets Iowa as the region of study per-county basis, where we analyze the SNAP eligibility of the population based on the fact that the population is 50% or more low income (185% of federal poverty level). In the study, census tracts are considered as the finest granularity spatially where income data is easily available. The study starts with the rigorous planning about where and how to start. Once the objective is clear, the collection of the current and reliable data which can satisfy the

objective is a great concern thereafter. An exhausted literature review helps to come to a point where we are sure what data is required to complete the study. The data is collected from two different online resources - 1) US census bureau [17], 2) ESRI [18]. The downloaded data is the raw data and requires synthesis and analysis before being used further.

In today's world, to mold the spatial visualization, GIS [4] is considered as one of the most appropriate technology in various descriptions and applications (e.g. natural resource management, environmental science, and community planning etc.). It offers the opportunity for integration of - 1) measures of proximity, 2) connectivity, and 3) density. The potential ability of manipulation of diverse geographic units in analyzing and presenting information, GIS is employed to examine critical application such as environmental and public health issues, analyzing the track changes in neighborhood conditions over time and spatial relationships among socio-economic conditions, measuring the distances how close the gymnasium is from person's home, how far the parks are located from residential area and other parameters that affect the human being in terms of physical activity. The other important utilities involve the simple visualization of complex data, storing and maintaining large amounts of geographic information, and creation of fine and informative maps at various scales.

This study also involves the use of GIS for the mapping of the results and based on the income ratio, the output (at census tract level) is classified into two sections: 1) low-income population, qualified for SNAP, and 2) high-income population and does not need SNAP assistance. In order to provide clear distinction in the representation between the classes, two distinct coloring patterns are used. The study is conducted for all ninety-nine counties in Iowa and only those counties in which the population at the census tracts level (at least in one tract) is eligible for SNAP are collected as the result and due to the space limitation, a few of them are shown as GIS map.

While considerable research work has been done in this area and substantial literature exists but no one has ventured such work for Iowa. The contribution of our research work is three-fold: 1) it could help the state government and intern USDA to collect the finer data regarding SNAP eligible population. Based on these facts, USDA can allocate adequate funds, 2) it could help state and USDA in better decision making to distribute SNAP benefits, 3) it could increase the probability that SNAP benefits reach out to qualified individuals.

The remainder of the paper is structured as follows. Section 2 examines related studies, while section 3 describes SNAP in greater detail. Section 4 is the architecture of this study. The study is conducted in Iowa per-county basis and a few of the counties having SNAP eligible population are visualized in section 5. Section 6 concludes the paper followed by brief outlines of future work.

2 Related Work

Camou explores the theories of policy designing [5] citing the helplessness of a legislative context in policy making understanding at extra governmental and small scale. The study in [6] emphasizes the mobility constraints that low-income population faces, in order to acquire their food, and correlation between these mobility constraints and coping strategies to complete their routine task. In [7] the

authors develop and map, the indices to describe the variation between cost and healthy food availability in London. John et al. [8] examine the residential housing pattern and analyze the effects of using census block groups over census tracts. Sara [9] describes the use of GIS in data management and dissemination and the issues raised in using GIS extensively (e.g. technical, institutional, and political). In [10] the authors took initiative to solve the problem of food desert in Iowa. Kameshwari [11] discusses about the community food assessments (CFAs) which contribute as a significant step in the planning for community food security. Michele brings [12] to the attention, the concerns about the accessibility of affordable and nutrient rich food in feed deserts. In [14] authors attempt to develop an effective GIS method to identify the food deserts in Vermont state, USA.

3 Supplemental Nutrition Assistance Program (SNAP)

Over a long period of time SNAP is commonly known as FSP [15] in United States. In today's modern world, cards - instead of paper stamp/coupons - are used to distribute all the food-stamp benefits. Any prepackaged edible food (regardless of nutritional value), including soft drinks and confectionery, can be purchased using card. Many states has already replaced the use of paper food stamps by EBT card - a specialized debit card system used in late 1900s and issued by private contractor - for public-assistance welfare programs as well. Because of the successful replacement of all paper food stamps by EBT cards, FSP has been renamed to the SNAP in October 2008.

The recipients must have at least near-poverty income in order to qualify for SNAP benefits. As per the statistics available in June 2009, the average monthly benefit per person was \$133.12 and the highest number of US population since 1962 benefited by FSP was 39.68 million, in Feb 2010.

In this study, we consider that our audience is 50% or more low income (185% of federal poverty level) [16] in order to qualify for SNAP program. As the income data at individual level is not available, we used census tracts or census block groups to qualify an area for our programming. GIS analysis and mapping for each SNAP qualified county were created showing which census tracts met the income eligibility threshold.

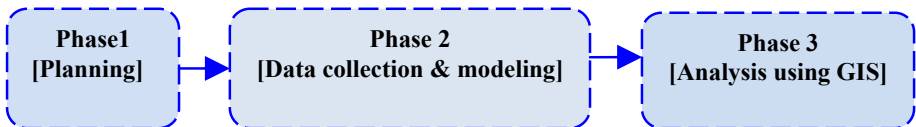


Fig. 1. Architecture

4 Architecture

As shown in Figure 1, the research methodology adopted for this study was divided into three main phases: 1) Planning, 2) Data collection, and 3) Analysis using GIS.

Table 1. Collected data (excel format)

Geography Identifier	Total population	Ratio of income to poverty level; Under 50	Ratio of income to poverty level; 50 to 74	Ratio of income to poverty level; 75 to 99	Ratio of income to poverty level; 1.00 to 1.24	Ratio of income to poverty level; 1.25 to 1.49	Ratio of income to poverty level; 1.50 to 1.74	Ratio of income to poverty level; 1.75 to 1.84	Total Population with ratio less than 1.85	% of population with ratio less than 1.85
19169000100	7086	64	145	31	105	110	160	46	661	9.328252893
19169000200	3696	124	66	31	74	109	95	50	549	14.8538961
19169000300	3284	169	87	116	82	141	39	24	658	20.0365408
19169000400	3036	207	38	92	69	66	87	41	600	19.76284585
19169000500	1426	191	144	67	153	192	156	32	935	65.56802244
19169000600	4621	300	254	188	154	224	153	39	1312	28.39212292
19169000700	3281	893	358	195	180	282	181	25	2116	64.49253276
19169000800	23	0	0	0	0	0	0	0	0	0
19169000900	4115	277	182	140	205	247	142	67	1260	30.61968408
19169001000	4135	575	500	374	361	205	333	167	2515	60.82224909
19169001100	3601	620	376	272	344	297	99	28	2036	56.53985004
19169001200	59	0	0	0	5	5	0	0	10	16.94915254
19169001301	4755	629	392	329	197	243	303	103	2196	46.1829653
19169001302	3561	122	64	73	79	48	122	26	534	14.9957877
19169010100	5986	97	88	118	103	155	183	87	831	13.88239225
19169010200	3964	70	56	96	127	116	82	73	620	15.6407669
19169010300	3996	91	70	65	80	118	120	66	610	15.26526527
19169010400	2881	72	42	29	155	44	122	32	496	17.21624436
19169010500	1961	27	30	24	52	125	146	39	443	22.59051504
19169010600	5087	77	74	108	99	199	155	116	828	16.27678396

4.1 Planning

This phase is the foundation of this study. It involves the rigorous planning about the procedure of the study and generates a set of important questions: what is the objective of this study, what kind of data solves the purpose of this work, what are the resources to collect the data from, whether the collected dataset is a valid and reliable dataset and will produce appropriate results. We conducted literature survey to find the answers of some of the questions regarding data collection and validation. Based on the insight developed after the literature survey, we observe that we needed to have two sets of data for the study: (1) data for calculating SNAP eligibility calculation, and (2) data for SNAP qualified area visualization - shape files.

4.2 Data Collection and Modeling

This phase involves collecting of the data and proper modeling to shape it as an appropriate input for the next phase (refer fig. 1). The desired data is acquired through different online resources. The data for SNAP calculation is downloaded from US census bureau and geospatial data (visualization) is collected from ESRI, considering tract as the finest granularity in both cases.

One can notice that the downloaded data was in Excel spreadsheet form (refer table 1). In order to use it with other dataset in ArcMap [18], the excel format should be converted into a database format (dbase). A database can be joined to attribute tables in ArcMap which will uniquely identify the census tracts geographically in each county.

4.3 Analysis Using GIS

Geographic Information System is widely used in geospatial visualization and is considered as one of the most suitable technologies in geospatial dominated research.

Natural resource management, environmental science, and community planning are such areas which heavily rely on GIS. The plethora use of GIS motivated us to use it in our work as well. This phase involves the tremendous use of GIS. The data (for calculation) downloaded from the above mentioned online resource is in excel from. We do all the necessary calculations in excel sheet, required for SNAP program. Once we are done with the calculation, the data is pretty much ready to be used in arc map provided the excel format is converted into dbase (.dbf) format.

Formula. We assume that N is the total number of population to determine the poverty status. N_x is the total number of population with ratio of income to poverty level x .

$$N_{>1.85} \text{ (total population with ratio to poverty level less 1.85)} = \sum_{i=50}^{1.75 \text{ to } 1.85} N_i.$$

$$\% N_{>1.85} \text{ (percentage of population with ratio to poverty level 1.85)} = \left(\frac{N_{>1.85}}{N} \right) * 100.$$

Though it is possible to convert excel into dbf directly in Microsoft office 2003 but arc catalog can be used to export an excel file into dbf and we adopt later approach due to the unavailability of office 2003. With the help of “join” feature in arc map, the tract qualified for SNAP program are linked with its geography from the shape (.shp format) file and plotted over the map. Once the database and attribute tables are joined we can classify the tracts on the basis of percentage of people who qualify for the SNAP program. From table 1 the last column reads “% of population with ratio less than 1.85”, we classified the data into manual quantities with two groups; less than 50 and 50 or above. Based on this classification the tracts in a county are plotted with two different visualization patterns - the SNAP qualified tracts are plotted by angled hatched lines.

5 Results

As stated above, this study considers Iowa as the study region to determine the spatial pattern eligibility for SNAP. The experiments are run on per county basis and each county is analyzed at the tracts level. Out of all 99 counties in Iowa, a few of those counties are visualized using GIS map, having at least one tract where population is SNAP eligible.

Fig. 2 is the partial visualization for Black Hack county (FIPS [2]-19013), where the tracts - 1, 3, 5, 9, and 23.02 are the qualified tracts. County Dubuque (FIPS-19061), has only one qualified tract (tract ID 1). Johnson county (fig. 3) has three tracts – 11, 16, 21- eligible for SNAP program. Linn county (FIPS-19113) has two tracts - 21, and 27 eligible for SNAP. Polk county (fig. 4) has five qualified tracts -12, 26, 50, 51, and 52. Scott county – FIPS-19163- (fig.5) has five tracts of low-income population, qualified to avail SNAP benefits.

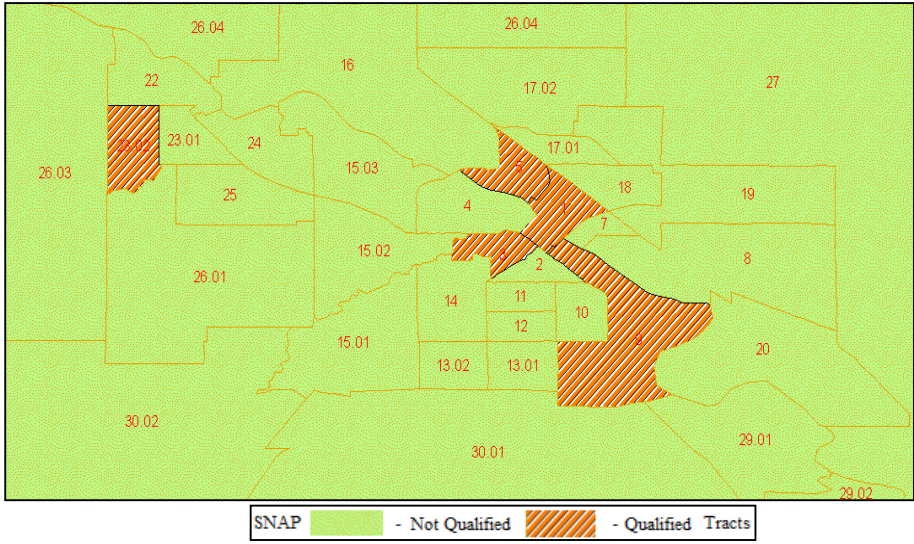


Fig. 2. Black Hawk county (FIPS: 19013)

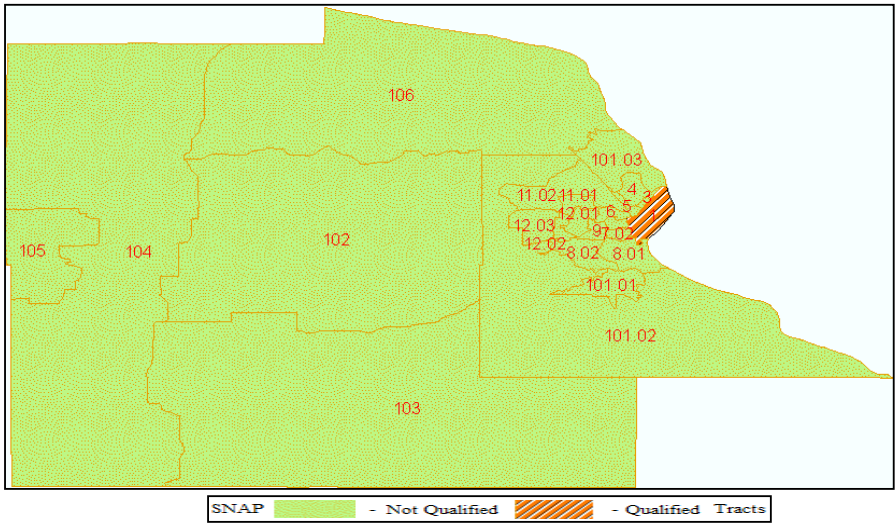


Fig. 3. JOHNSON County (FIPS: 19103)

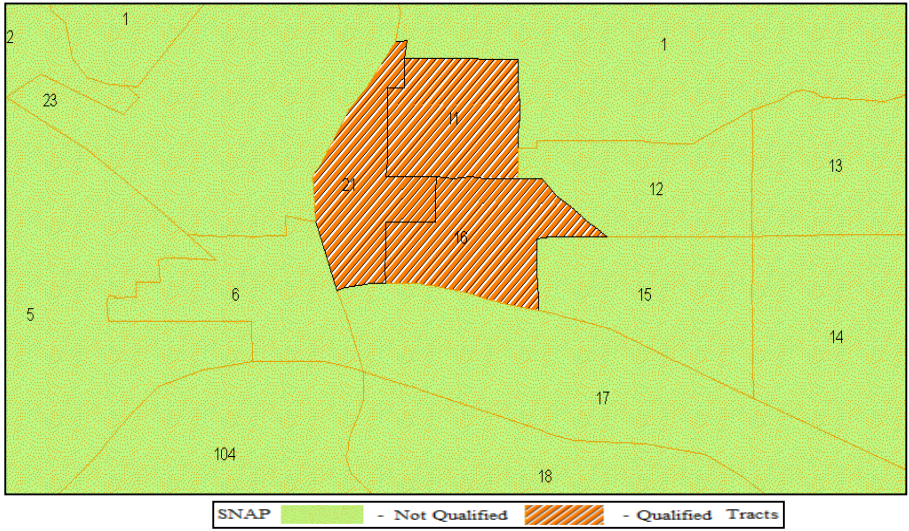


Fig. 4. POLK County (FIPS: 19153)

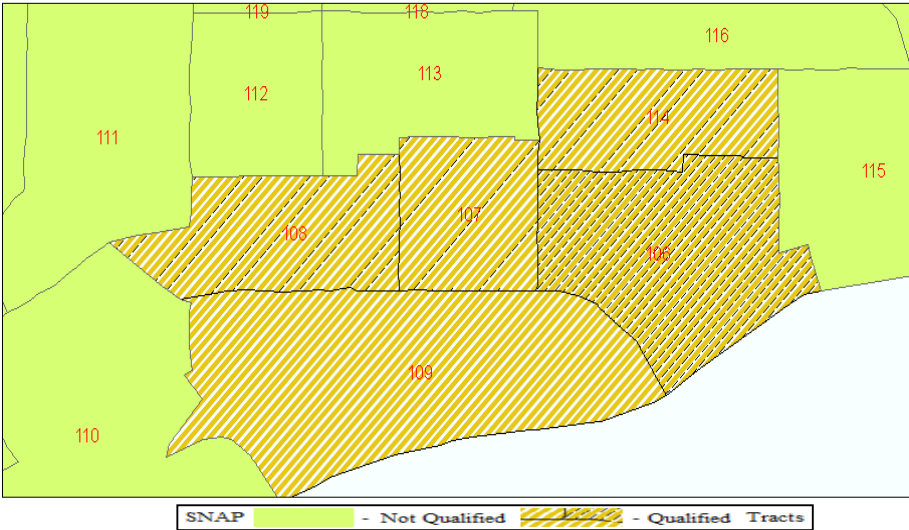


Fig. 5. SCOTT County (FIPS: 19163)

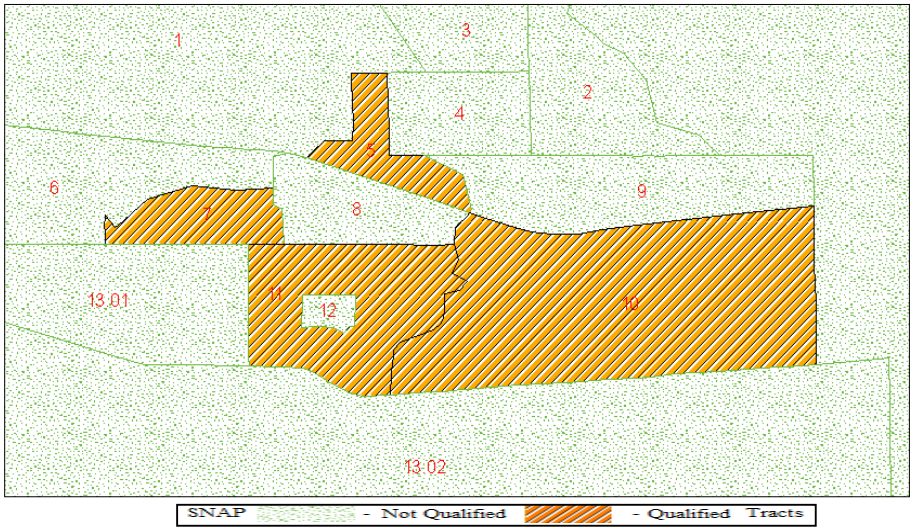


Fig. 6. STORY County (FIPS: 19169)

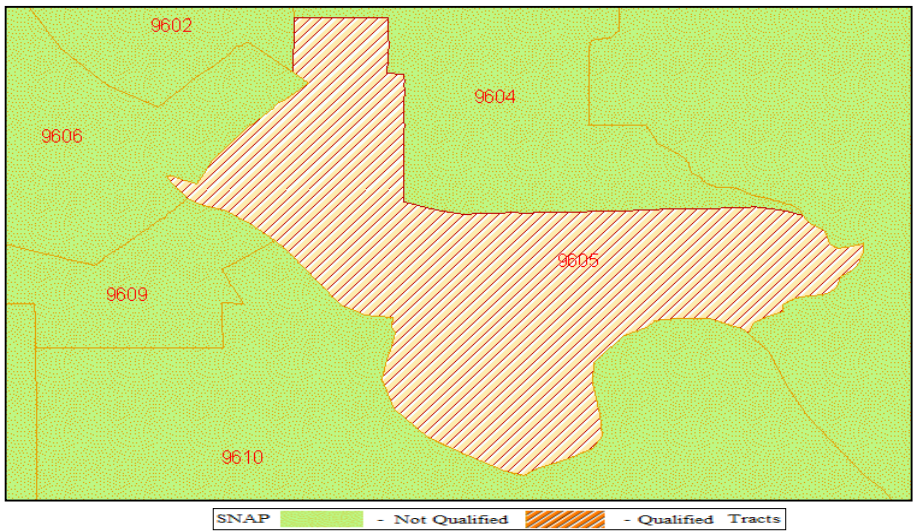


Fig. 7. WOODBURY County (FIPS: 19193)

In Story county- FIPS-19169- (fig. 6) the results are very interesting. There are four qualified tracts: 5, 7, 10, and 11. It can be noticed that tract 12 has rich population, but completely surrounded by tract 11- low income population. There is only one tract: 9605 in which the population is low-income/ no-income and qualified for SNAP program in Wapello county with FIPS-19179. Webster county - FIPS -

19187 - also has only one qualified tract - tract Id as 7. Woodbury county (fig. 7) - FIPS-19193 - has four qualified tracts: 12, 15, 16, and 17.

6 Conclusion and Future Work

We investigated the spatial patterns where the population is 50% or more low income (185% of federal poverty level) and qualified for SNAP assistance benefits. Census tracts were used as the finest spatial granularity in our study and those qualified tracts were rendered on GIS map of their counties.

The future work involves the investigation of other aspects e.g. socioeconomic impact, further analysis of population based on gender, age, etc. in those tracts where the population is qualified for SNAP program.

Acknowledgements

Authors are thankful to Xinyuan Zhao, a Ph.D. student in Computer Science at Iowa State University for his helpful comments to improve this paper.

References

1. <http://www.fns.usda.gov/snap/>
2. http://en.wikipedia.org/wiki/FIPS_county_code
3. United States Department of Agriculture, National Agriculture Statistics Service
4. Alibrandi, et al.: Using GIS to Answer the ‘Whys’ of ‘Where’ in Social Studies. *Social Education* 70(3), 138–143 (2006)
5. Camou, M.: Deservedness in Poor Neighborhoods: A Morality Struggle. Deserving and Entitled: Social Constructions and Public Policy, 197–218 (2005)
6. Clifton, K.: Mobility Strategies and Food Shopping for Low-income Families. *Journal of Planning Education and Research* 23, 402–413 (2004)
7. Donkin, et al.: Mapping Access to Food in a Deprived Area: The Development of Price and Availability Indices. *Public Health Nutrition* 3, 31–38 (1999)
8. Iceland, et al.: The Effects of Using Census Block Groups Instead of Census Tracts When Examining Residential Housing Patterns (2003)
9. McLafferty, S.: The Socialization of GIS. *Cartographica: The International Journal for Geographic Information and Geovisualization* 39(2), 51–53 (2004)
10. Morton, L., et al.: Solving the Problems of Iowa Food Deserts: Food Insecurity and Civic Structure. *Rural Sociology* 70(1), 94–112 (2005)
11. Pothukuchi, K.: Community Food Assessment: A First Step in Planning for Community Food Security. *Journal of Planning Education and Research* 23, 356–377 (2004)
12. Ramasubramanian, L.: Nurturing Community Empowerment: Participatory Decision Making and Community Based Problem Solving Using GIS. In: *Geographic Information Research: Trans- Atlantic Perspectives*, pp. 87–102. Taylor & Francis, Briston (1999)
13. Ver Ploeg, M.: Access to Affordable, Nutritious Food is Limited in ‘Food Deserts’. *Amber Waves*, USDA Economic Research Service (2010)
14. McEntee, J., et al.: Towards the development of a GIS method for identifying rural food deserts: Geographic access in Vermont. *Applied Geography* 30, 165–176 (2010)

15. http://en.wikipedia.org/wiki/Supplemental_Nutrition_Assistance_Program
16. <http://liheap.ncat.org/profiles/povertytables/FY2010/popstate.htm>
17. <http://www.census.gov/>
18. <http://www.esri.com/>

Project Management Model for e-Governance in the Context of Kerala State

Anu Paul¹ and Varghese Paul²

¹ Research Scholar, M G University, Kottayam, Kerala, India
anupaul71@gmail.com

² ToCH Institute of Science and Technology, Arakkunnam, Ernakulam, Kerala, India

Abstract. Government Of India (GOI) acknowledged the pivotal role that Information and Communication Technology (ICT) has played since the last decade, in bringing government services to the doorstep of the people across different segments and geographical locations. Kerala is one of the federal states of India, implementing e-governance for the effective and efficient administration and service to the state. Implementation of e-Governance projects has defined goals like timeliness and cost optimization. But still it is a dream. In this paper we analyzes the reasons and it points the finger to the current project management scenario under Kerala state. The data collection was done by personal interviews. This paper propose a model, which can be functional in a team for e-Governance project management in order to the effective implementation of e-Governance projects and its completion in time while respecting timeliness and cost optimization.

Keywords: e-Governance; Project Management; Service Design Process; Object Oriented Design; Reusability; Interoperability.

1 Introduction

e-Governance is a paradigm shift over the traditional approaches in public administration. It means that rendering of government services and information to the public using electronic means. The aim of e-governance is to bridge the gap between the government and the public by providing effective, interactive and transparent governance [1]. e-Governance implementations are cost and resource intensive with high degree of risk at times. Most often e-Government initiatives suffer delays and encounter failures. A survey of e-government projects revealed that 85 percent are partial or total failure i.e., some of them are unattained goals or abandoned implementation [2]. Lack of internal ownership, absence of vision or strategy, poor project management, inadequate technological infrastructure and obstacles to data interchange etc. are the reasons of these failures [3]. Ample administrative reform for the e-Governance project can make the failure into success.

Government of India (GOI) recognizes that e-Governance is a strategic tool for transforming Governance and improving the quality of services provided by the government to its people. Government of India has approved the National

e-Governance Plan (NeGP) in pursuance of its policy of introducing e-Governance on a massive scale. The state administration initiatives by the guidelines of NeGP have been discussed in section 2 of this paper.

Kerala is one of the federal states of India, implementing many citizen-friendly e-Governance projects [4] [5]. This paper analyses the project management initiatives by Kerala state and is discussed in section 3 and 4. Due to lack of in-house technical manpower, the Kerala State Information Technology Mission (KSITM) have to depend on external agencies to build up, manage, implement and maintain the e-Governance projects. The e-Governance solutions of various departments are similar nature. This paper proposes a project management model which can be functional in a team under KSITM. This has been discussed in section 5. The paper closes in section 6 with description of related works.

2 Programme Management in the State Domain

Programme Management is vital to the successful implementation of any e-Governance Project. It includes developing a project plan which includes defining project goals and objectives; specifying tasks or how goals will be achieved; what resources are need; associating budgets and timeliness for completion and also ensure the plan is being managed according to plan. To facilitate the state administration, three level professional teams were formed by the guidelines of NeGP of India [6].

- Apex Level - Provide strategy direction and oversee the state Programme and ensure inter- departmental co-ordination.
- Programme level - State e-Governance Mission Team (SeMT)
- Project level - Project e-Governance Mission Team (PeMT)

Chief Secretary of the State is the Chairman of State Apex Committee. This committee is responsible for overall guidance, deciding policy level matters and to act as final body for approving all issues relating to project implementation. SeMT would be responsible for program management of the e-Governance initiatives in the State. Various departments of the State Government taking up e-Governance projects would require a full-time dedicated Project e-Governance Mission Team (PeMT). Figure 3 provides representation of the institutional framework at the state government level:

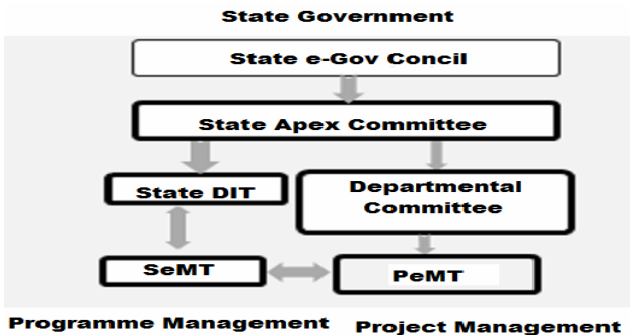


Fig. 1. Institutional framework at the State Government

3 Current Scenario in the Kerala State

The Government of Kerala acknowledges the importance of IT as an instrument for the state's overall development. It is deeply committed to its dissemination as a crucial engine of economic growth and a tool for increasing productivity, speed and transparency in governance and improves quality of life for the common man. There are many project management initiatives by Kerala state, for the implementation of e-Governance projects and is discussed and sum up with the following information. The data collection was done by personal interview with officials of Kerala state e-Governance team.

- Kerala State Information Technology Mission (KSITM) is an autonomous nodal IT implementation agency under Information Technology Department (ITD), Government of Kerala.
- KSITM is responsible for co-ordinating e-Governance initiatives in the Kerala State. It provides managerial support to initiatives of various state departments. It is a team of professionals from the industry and the Government which is headed by the Director, with the Secretary - IT as the Chairman.
- The state level projects of various departments are designed, developed, and implemented independently by the line departments.
- The early stages of project i.e. preparation of SRS (Software Requirement Specification); FRS (Functional Requirement Specification) based on URS (User Requirement Specification); and DPR (Detailed Project Report) are done by any of the government agencies like, Centre For Development of Imaging Technology (C-DIT); National Informatics Centre (NIC) Kerala; The Kerala State Electronics Development Corporation (KELTRON) or any other agencies interested by the line departments.
- This system design and DPR (Detailed Project Report) is approved by the line departments in consultation with the e-Governance team, and then it is outsourced to the vendors for coding. The project e-Mission Team (PeMT) implementing these projects in their departments.
- A large amount of money is spent for each of these independent projects and delays much beyond scheduled time. So the e-Governance goals, timeliness, cost optimization and the interoperability between departments couldn't be achieved.
- To facilitate the state administration in Kerala, three level professional teams were formed by the guidelines of NeGP. Kerala State e-Governance Mission Team (SeMT) starts functioning from August 2009. A seven member team of consultants have appointed from M/s Wipro for a period of one year. Now GOI have provided the Technical people for the SeMT to support the e-governance initiatives in the state. The State Nodal Agency has initiated the process for Project e-Governance Mission Teams (PeMT) formation.
- Some of the present Kerala state government projects related to different departments are: JnNRUM - Local Self Government Department (Corporations); FAST-Transport Department; SPARK - Payroll & Personnel Management System of state government employees; Treasury Information System - Treasuries Department; KVATIS - Commercial Taxes Department; Tetra PDS - Food & Supplies department; e-District - Revenue, Food & Supplies,

Agriculture, Home Department (police & Passport), Local Self Government Department etc.

- The above projects are formulated, designed, developed and implemented independently and so many of the similar processes are duplicated.
- The ICT Infrastructures implemented in the Kerala state are State Data Center (SDC), Kerala State Wide Area Network (KSWAN) and State Service Delivery Gateway (SSDG).

4 Discussion

Computerization of government departments is not a complete e-Governance solution. Different government departments should be linked with each other to exchange of information. Then the decision making process would be quicken and better. The government processes are highly complex and changing very quickly, day by day number of users and services are increasing drastically. The complex nature of e-Government projects raises the challenges in project management [7] [8]. In spite of financial constraints, the Indian state Kerala has made significance achievements in e-Governance. To facilitate the state administration, professional teams were formed by the guidelines of NeGP of India. But there is no common team to manage the software solutions of entire state level projects in the Kerala state. Kerala State Information Technology Mission is responsible for implementing and co-ordinating e-Governance initiatives in the state. It is a team of professionals from the industry and the Government by the choices of the ruled political party. Due to the lack of in-house technical manpower, the KSITM have to depend on external agencies. In a state like Kerala, large number of projects has to be managed at a time, on a continual basis, by this unstable administrative group. This will lead to a big challenge in the programme management activity.

By the analysis of the current project management scenario under the Kerala state, it has been seen that most often e-Government initiatives suffer delays and goes beyond the targeted cost, as the implementation agencies lack guidance in the area of planning and implementation of state level e-Governance projects. There is hardly any business re-engineering process had been attempted for the projects. The changes have been incorporated in a demand driven manner. There are nearly 192 state departments in Kerala. Most of them maintain their websites and are linked with the Kerala state portal. KSITM provides managerial support to the initiatives of these various state departments. The e-Governance projects related to these departments are in similar nature. Currently these projects are formulated, designed, developed and implemented independently by the corresponding line departments. Hence, many of the similar processes are duplicated. A large amount of money is spent for each of these independent projects and delayed much beyond scheduled time. Thus, the e-Governance goals like timeliness, cost optimization and the interoperability between departments couldn't be achieved. If there is a stable, common team with sufficient technical manpower, to manage the software solutions of entire state level projects, many of the components can reuse in projects of various state departments. Thus a timely completion and reduction of cost become a reality.

5 The Project Management Model (PMM)

To overcome the practical issues of the current project management scenario under Kerala state, this paper propose a project management model which can be functional in a team to the effective implementation of e-Governance projects and its completion in time while respecting timeliness and cost optimization.

Kerala State Information Technology Mission is an autonomous nodal IT implementation agency under Information Technology Department of Kerala government. We propose a dedicated team under this nodal to handle the project management of development process of entire state level projects.

The PMM consist of two parts:

- First, a service design process in the suggested team.
- Second part is the reusability of components in e-Governance projects.

This support to develop e-Governance solution with better quality, common architecture, higher reliability and minimum maintenance. Sufficient technical manpower in the suggested team could avoid independent design and development of projects in various departments of the state by the external agencies. However, to incorporate a large number of state government departments and to develop solutions under an umbrella by a single team and a common methodology is very difficult task. Even though, by the proper administration and mindset it is possible.

5.1 The Service Design Process

The project management process is analyzed by the customer needs into service design in any of the e-Government organization, which is responsible for designing, coding and maintaining the web interface as well as the back-end software functions. In e-Governance systems, intelligence generation, intelligence dissemination and organization-wide responsiveness are the three components. Intelligence generation is the collection of customer needs. Intelligence dissemination is the distribution of the collected information within the organization system, web page, database design and other related technical operations. Organization-wide responsiveness is the evaluation of the service design process.

In this paper, we suggest enough technical manpower in the team to manage the entire state level projects. The structure of the team is shows in Figure 2.

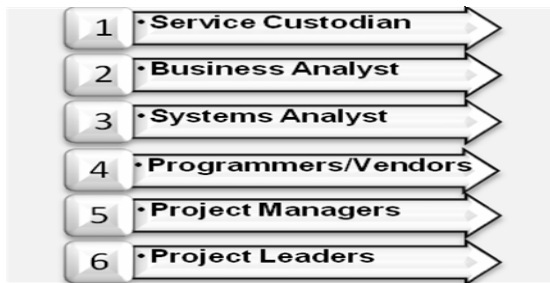


Fig. 2. Employee Level Structure

The suggested team would have the major responsibilities like design and launch of new services online and maintaining all the e-services that have already been launched. We suggest a service design process for the proper management of entire state level projects. In this service design process, custodian would have to approve the process flow designed by the business analyst and also the system design by the system analyst. Different activities like survey, training, interviews, developments, dissemination and research could be carried out by the business analyst for the collection of customer needs. The system analyst could reuse the design, knowledge, components and code of the existing similar systems. The system analyst would have to evaluate and approves the code and then it would go for the approval of the custodian. Project manager is responsible for the overall control and implementation of state government department projects. The project tasks like inputs (resources), outputs (deliverables), ordering (process) and governing mechanisms (control) would have to be managed by the project leaders. This service design process is shown in the figure 3.

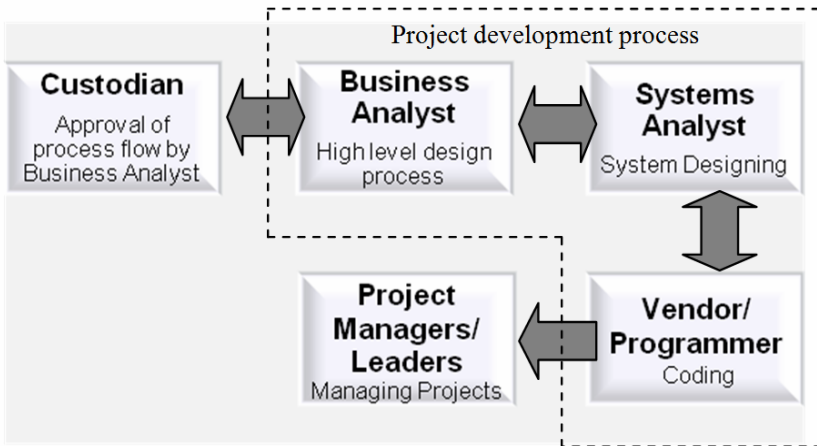


Fig. 3. The service design process

This process will help the timely completion by the proper management of the life cycle of each project.

5.2 Reusability of Components

Most of the processes in e-governance solutions are similar, like, Filling up the form for new application and verifying it for renewal; Documents verification; Fee Payment etc. There are common behaviours among the processes across the state government departments and within the department also. It is expected that 85% of the processes should be similar across different government departments . A similar fraction of the processes can be similar across different government solutions. Thus, it should be possible to reuse the solutions developed for one department to another. By using the Object Oriented Design (OOD) methodology to solve problems with similar

behaviour of e-governance system will help to incorporate the future changes. Reusing the e-governance skill across different governments departments can substantially bring down the cost of developing e-governance solutions. This leads towards not only a code or component reusability but also a reusability of design and knowledge.

The OOD methodology enhances the reuse by the features like; encapsulation of data and functions as objects, specialization by inheritance and delegation through association. In the component-oriented approach, delegation allows building generalized subsystems which can be plugged into new designs with little effort. Specialization is an architectural approach; the system analyst can pre-define hierarchies of generalized object types and then later, develop specialized objects which automatically inherit a basic set of properties. These techniques make reuse easier. This is shown in the figure 4.

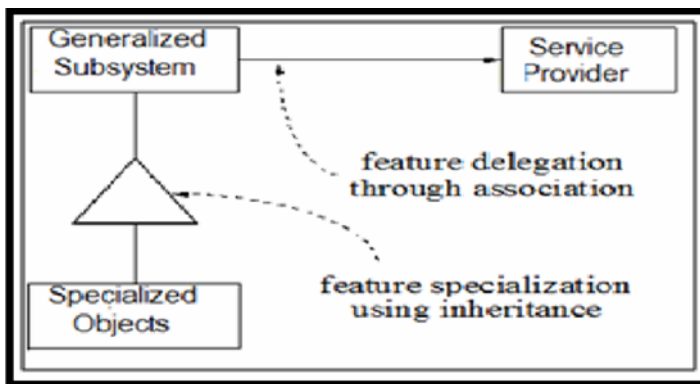


Fig. 4. Reuse in Object-Oriented Design

An e-Governance frame work of existing system could be developed as a partially built subsystem by the basic building blocks and ready to use classes. The specification of class, object interactions and their underlying intents are easy to reuse across different application domain according to the government requirements. OOD gives a concrete base in the software design and provides a robust and economical solution. The software designer can reuse their knowledge to solve similar type of the problems by providing the flexibility to integrate new necessity in the existing e-governance solution with minimum efforts. By building generalized subsystems for the e-Governance projects with similar behavior will help to include the future changes and alterations very easily by the software developers and designers.

6 Related Studies

e-Governance is a term which has recently came into regular use in India. GOI has taken several initiatives to introduce e-Governance [9]. Currently India is in between phase 2 and 3 of e-Governance implementation, i.e., the information are available in

the form of websites and the citizen can download information and can interact with the government. India have to attain the 4th phase, integration [10]. e-Governance is a developing field in India. So that a very few researchers and academia are there in the field of e-Governance in India in computer science area. In that, most of them concentrated on the network security and authentication. There is hardly any research paper or article found out, based on the software engineering perspective in the Indian context. Even though we have been inspired and our concept has been formulated by other researchers in this field.

The first part of our PMM, the service design process of an e-Governance team is developed from the customer orientation model by Jaworski and Kohli (1993) [11] and this model is used in an e-Governance organization of Dubai municipality by Arif, M. (2008) [12]. Our design model is better and suitable for the project management in the Indian context, especially in the Kerala state as per the current status. A well structured service design process is not enough for reducing the development time and cost. The reusability of similar process in e-Governance projects will help to attain the goals. reusability by object orientation is not a new concept, and considerable research and analysis on a wide range of concepts and approaches was available [13] [14] [15] [16] [17] [18] [19]. So the second part of our PMM suggest the reusability of design, knowledge, component and code in e-Governance projects in the state. The independent project development task would not achieve the goal [20]. A common team should be there to manage the entire state level projects. Then only the reusability will become possible. By the analysis of PM scenario under Kerala state, we propose a dedicated team under KSITM. The PMM could be functional in this team. The proposed PMM is a realistic one for the PM of entire state level projects in the context of Indian state kerala, to attain the goals like timeliness and cost optimization.

7 Conclusion

The Government of Kerala is a front-runner in implementing e-Governance. e-Governance project are cost and resource intensive with high degree of risk at times. Project Management is essential for successful e-Governance solutions of the state government projects. Many of the initiatives had completed for the administration of the Kerala state government projects by the guidelines of the NeGP of GOI. Still many of the projects are delayed for its completion and require more money than the target. This paper proposes a team under KSITM to manage the software solution of the state government projects. By applying the suggested service orientation and reuse of components in the projects could achieve goals like, timelines, consistent strategies for cost optimization and integration. The future work would be to achieve interoperability in e-Governance initiatives by the OOD methodology.

Acknowledgment. We wish to thank Mr. K.S Anil Kumar, Former Head of e-Governance in the Kerala State, for his valuable comments and guidance.

References

1. Bagga, R.K., Gupta, P.: Transforming Government: eGovernance Initiatives in India. The ICFAI University Press (2009)
2. Heeks, R.: Most eGovernment-for-Development Projects Fail: How Can Risks be Reduced? iGovernment Working Paper Series, Paperno.14 (2003)
3. Mrinalini, S.: E-Governance in India: Dream or reality? International Journal of Education and Development using Information and Communication Technology (IJEDICT) 3(2), 125–137 (2007)
4. Alex, J., Unnikrishnan, V., et al.: A Model of Human Resource Development for IT Enabled Governance: The Experiences of Information Kerala Mission. A report of Information Kerala Mission, Thiruvananthapuram.
5. Kumar, A.: Secretary, IT, Government of Kerala (2010), e-Governance in Kerala, e-Gov, Asia's First Monthly Magazine on e-Government (September 2010)
6. Bhattacharya, J.: Executive level e-Governance Capacity Building e-Governance Oracle-HP. e-Governance Center of Excellence (2005)
7. Kliem, R.: Risk Management for Business Process Reengineering Projects. Information Systems Management 17(4), 71–73 (2001)
8. Bansal, V., Bhattacharya, J.: E-governance solution for government of Maharashtra. Technology white paper, India Research Lab, IBM (2000)
9. Anu, P.: e-Governance, a Paradigm shift in India. International Journal of Interdisciplinary Studies and Research: Baselius Researcher XI (1), 63–73 (2010)
10. Anu, P., Varghese, P.: The Implementation issues of e-Governance in India. In: Proceedings of the UGC & IETE Sponsored National Seminar on Modern Trends in Electronic Communication & Signal Processing, February 3-4, pp. 43–48. Excel India Publishers, Kerala (2011)
11. Jaworski, B.J., Kohli, A.K.: Market Orientation: Antecedents and Consequences. Journal of Marketing 57(1), 53–70 (1993)
12. Arif, M.: Customer Orientation in e-Government Project Management: a Case Study. Electronic Journal of e-Government 6(1), 1–10 (2008)
13. Ajay, D.: Software Design Pattern for E-governance Solution Framework. In: Proceedings of the 4th National Conference; INDIACOM-2010 Computing for Nation Development, New Delhi, India, February 25 - 26 (2010)
14. Chang, C.-H., Chu, W.C., Hsueh, N.-L., Koong, C.-S.: A Case Study of Pattern-based Software Framework to Improve the Quality of Software Development. In: Proceedings of the 2009 ACM symposium on Applied Computing (2009); ISBN:978-1-60558-166-8
15. Parikh, A., Buddhdev, B.V.: E-Governance Solution Based on Observer Design Pattern. In: National Conference on architecture Future IT Systems, NCAFIS 2008 (2008)
16. Vassilis, M., Christos, D.: Bridging theory and practice in e-government: A set of guidelines for architectural design. Government Information Quarterly 27, 70–81 (2010)
17. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley Professional Computing Series. Addison-Wesley, Reading (2004)
18. Mittal, P.A., Kumar, M., et al.: A framework for eGovernment solutions. IBM Journal of Research and Development
19. Fanch, P.W.: Design Reuse through Frameworks and Pattern. IEEE Software 18(5), 71–76 (2001)
20. Dada, D.: The Failure of E-Government in Developing Countries: A Literature Review. EIJSDC 26(7), 1–10 (2006)

ICT Its Role in e-Governance and Rural Development

Deka Ganesh Chandra¹ and Dutta Borah Malaya²

¹ DGE&T, Ministry of Labour & Employment, New Delhi-1, India

² Dept of Computer Engineering, Delhi Technological University, Delhi -42, India
ganeshdeka2000@gmail.com, malayadutta@dce.ac.in

Abstract. Information and Communication Technologies (ICTs) play a key role in development & Economic growth of the Developing countries of the World. Political, Cultural, Socio-economic Developmental & Behavioral decisions today rests on the ability to access, gather, analyze and utilize Information and Knowledge. ICT is the conduits that transmit information and knowledge to individual to widen their choices for Economic and social empowerment. People around the Globe from few years from now will be carrying a handheld computer connected to the Web to get the information about the World at their fingertips. Government of India is having an ambitious objective of transforming the citizen-government interaction at all levels to by the electronic mode (e-Governance) by 2020. A successful ICT application in e-Governance giving one-stop solutions for rural community is the need of the hour. Implementation of IT Act 2000 is yet in a nebulous stage though the Act was enacted 10 years back.

In this paper, we have analyzed various issues relating to e-Governance, Administrative Reforms in Rural Development context. We have also included an introductory discussion about "Information and Communication Technologies" sector strategy approach paper 2011 which will define the road map for preparing the World Bank Groups new ICT sector strategy for the coming years for the development of the poor & developing countries of the Glob.

Keywords: e-Governance, ICT, NeGP, Common Services Centers (CSC), e-Panchayat.

1 Introduction

ICT is the conduits that transmit information and knowledge to individual to widen their choices for Economic and Social empowerment. Sociopolitical & Socio-economic condition of the people today rests on the ability to Access, Gather, Analyze and Utilize information and knowledge. Integrating Technology with Development gives effective and speedy solution for sustainable development. The core list contains four sets of indicators:

- ICT infrastructure and Access
- Access and use of ICT by households & individuals

- Use of ICT by businesses &
- ICT goods

Cell phones and other ICTs can provide a broad range of public and social services to the poor in remote and rural areas. The mobile phones are no more a luxury good but a vital utility for the underprivileged. In remote rural areas, availability of crop seeds, current market prices etc can be availed by farmers using mobile phones.

Increasingly land registration, education, health care, and voting are being conducted electronically using ICT. People around the Globe from few years from now will be carrying a handheld computer connected to the Web to get the information about the World at their fingertips.

Recent World Bank study shows that a 10% increase in mobile phone subscribers is associated with a 0.8% increase in economic growth while 10% increase in high-speed internet connections is related with a 1.3% increase in economic growth. ICT infrastructure development attracts foreign direct investment, generates fiscal revenues and creates employment opportunities. Local information technology service industries generate exports, improve a firm's productivity, and offer equalizing job opportunities, especially for youth and women.

These remarkable advancements in technology and understanding of how it affects growth highlight a strategic shift in the way ICT can influence development. *The distribution of the tool of wealth creation & knowledge are highly unequal amongst countries of the Globe.* At the current rate of technological advancement this disparities in access to ICT related development are large and likely to become larger in adoption amongst the countries around the Globe as more of the services in an economy come online. Those without access to this technological advancement will be marginalized.

The rest of this paper is organized into 8 Sections. **Section-2** is about e-government, **Section-3** is about the **Role of International Organization** while **Section-4** is about the various issues related to **ICT, Law and Administrative Reforms & Public Grievances**. **Section-5** is about Government of India Initiatives of e-Governance & **Section-6** is about **e-Panchayat**. **Section-7** is a **Case Study** about **ICT for Rural masses in North Easter India & Jammu & Kashmir**. In **Section-8** we have mentioned about our **Observations** and finally **Section -9** is about the **future scope for studies & research** in this area.

2 e-Government

e-Government is the use of Information Technology to exchange information and services with citizens, businesses and other arms of government.

The Legislature, Judiciary and Administration may apply e-governance in order to improve internal efficiency, the delivery of public services or processes of democratic governance. It also refers to the citizen to government interface including the feedback of policies.

e-governance has not only the ability to handle momentum and complexity but also to underpin the regulatory reform. Even though ICT is not substitute for good policy, it empowers the citizens to question the actions of regulators and brings systemic issues to the forefront.

The costs associated with Telecommunication infrastructure and human capital continues to hold back e- governance development. However, effective strategies and legal frameworks can recompense significantly, even in least developed countries.

2.1 e-Governance the Asian Scenario

In the United Nations e-Government Survey 2010, it was found that as a result of growing use of ICT by the public sector the citizens are much benefited from the e-service delivery, better access to information, more efficient government management and improved interactions with governments. According to **United Nations E-Government Survey 2010** the following is the ranking among the national portals in the region:

Table 1. World e-Government development ranking

Country	Year 2010	Year 2008
Maldives	92	95
Iran	102	108
Sri Lanka	111	101
India	119	113
Bangladesh	134	142
Pakistan	146	131
Bhutan	152	134
Nepal	153	150
Afghanistan	168	167

Most countries have published a tremendous amount of information online, many going beyond basic websites to provide national portals that serve as a major starting point for users to connect to government services in different Ministries [1].

3 Role of International Organization

3.1 World Bank

“You just have to be in any developing country, and you’ll see the power of ICT,” says World Bank Managing Director Ngozi Okonjo-Iweala. Over 3 billion mobile phones are now in the developing world and the World Bank’s target is to take it to the next level by supporting to use the mobile phones and the Internet as a platform for:

- Social Services
- Green growth and social accountability and
- Boosting up the well being and incomes of poor people.

All the way through the last decade the World Bank Group has supported most important reforms of telecom markets around the developing world, causative to attract huge amounts of private investment in mobile phone networks. Mobile

connectivity network has already reached more than 3 billion people in developing countries. Among others the World Bank Group has also financed & supported in the following area:

- Connectivity infrastructure
- Regulatory reform, including privatization of state-owned telecoms providers
- Liberalization efforts
- Human capacity development and
- ICT applications in Public Administration, Health & Education

“The ICT field is moving so fast that our strategy should be to leverage expertise outside the World Bank and make it available to our clients”, said World Bank Vice-President for Sustainable Development, **Inger Andersen**.

The International Finance Corporation’s (IFC) investment of US \$3.2 billion in private sector projects in developing countries has created 225 million new mobile subscribers and 57,000 IT jobs in 54 companies.

The Multilateral Investment Guarantee Agency (MIGA) has provided 38 investment guarantees in ICT sectors in developing countries, guaranteeing US \$1.3 billion for 21 projects, 12 of which are in Africa and contributed US \$6 billion foreign direct investment in the ICT sector.

“The information technology revolution is just beginning” said Philippe Dongier, Manager of the Bank's Information and Communication Technologies (ICT) sector.

The World Bank is preparing an Approach Paper proposing three strategic directions i.e. **Innovate, Connect & Transform** to shape the Bank Group’s work in ICTs in the coming years.

3.2 infoDev

infoDev supports global sharing of information on ICT for Development (ICT4D), helps to reduce duplication of efforts and investments as well as creates partnerships with public and private-sector organizations that focuses on **Innovate, Connect & Transform** of ICT as Tools of Development and Poverty Reduction.

4 ICT, Law and Administrative Reforms and Public Grievances in India

To make the use of ICT effective the parliament of India has passed THE INFORMATION TECHNOLOGY ACT, 2000.

Another act THE RIGHT TO INFORMATION ACT, 2005 also advocates the use of ICT (Internet) in the proper dissemination of information in public domain.

4.1 The Information Technology Act, 2000

The **CHAPTER III** of this act defines the ELECTRONIC GOVERNANCE. **Section-8** of this Chapter i.e. Chapter-III indicates the Publication of Rule, Regulation etc. in Electronic Gazette.

Any Rule, regulation, order, bye-law, notification or any other matter published in the Official Gazette should also be published in the Electronic Gazette.

CHAPTER X represents the THE CYBER REGULATIONS APPELLATE TRIBUNAL where **Section-48** indicates the **Establishment of Cyber Appellate Tribunal**.

1. The Central Government shall, by notification, establish one or more appellate tribunals to be known as the Cyber Regulations Appellate Tribunal.
2. The Central Government shall also specify, in the notification referred to in subsection
3. The matters and places in relation to which the Cyber Appellate Tribunal may exercise jurisdiction

The new Office and the Court Room of the Cyber Regulation Appellate Tribunal in New Delhi for handling the liability to check Cyber fraud, Cyber crime & Cyber terrorism was inaugurated on 27th July 2009, by the then Chief Justice of India Mr. Justice K.G. Balakrishnan.

4.2 The Right to Information Act, 2005

In this act the CHAPTER II (Sec-2 & Sec-4) represents the Right to information and obligations of public authorities

Section-2: It shall be a constant endeavour of every public authority to take steps in accordance with the requirements of clause (b) of sub-section (1) to provide as much information *suo motu* to the public at regular intervals through various means of communications, including internet, so that the public have minimum resort to the use of this Act to obtain information.

Section-4: All materials shall be disseminated taking into consideration the cost effectiveness, local language and the most effective method of communication in that local area and the information should be easily accessible, to the extent possible in electronic format with the Central Public Information Officer or State Public Information Officer, as the case may be, available free or at such cost of the medium or the print cost price as may be prescribed.

Explanation- For the purposes of Sub-Sections-4 "disseminated" means making known or communicated the information to the public through notice boards, newspapers, public announcements, media broadcasts, the internet or any other means, including inspection of offices of any public authority.

However, most of the Government organizations either not developed their Website or the few which have developed their Websites were not updated regularly as a result the requisite information regarding the activities of the organizations is not readily available to the cliental group online.

Further the WebPages of the various Government organizations under the same Ministry/ Departments were not uniform so far as the presentations of the contents in the WebPages of the were concerned.

If the websites of the Government Departments/ Ministries are updated regularly, it will not only enhance the transparency but will also be saving the valuable time otherwise spent entertaining the enquiries of the public as well as replying the questions asked under the Right to Information Act 2005.

4.3 The Information Technology (Amendment) Act, 2008, Section 69

This act authorize the Central Government/State Government/its authorized agency to intercept, monitor or decrypt any information generated, transmitted, received or stored in any computer resource if it is necessary or expedient so to do in the interest of the sovereignty or integrity of India, defense of India, security of the State, friendly relations with foreign States or public order or for preventing incitement to the commission of any cognizable offence or for investigation of any offence.

4.4 Administrative Reforms and Public Grievances

Perception of about government performance is that Government agencies have not delivered what was expected. The key determinant of organization performance is quality of management systems. To improve perception of government organization, it is necessary:

- Enough to deliver results
- Effectively manage interface with Citizens & Clients
- Enough to **communicate effectively** with citizens & clients

The **Department of Administrative Reforms & Public Grievances [3]** is the nodal agency to formulate policy guidelines for citizen-centric governance in the country. To redress the grievances of the citizens is the main objective of the department. DAR&PG formulates public complaint redress mechanism for efficient and well-timed redress/settlement of citizens' grievances. Public Grievances pertaining to identified issues in respect of 20 Central Government Organizations' are being handled by Directorate of Public Grievances (DPG), Cabinet Secretariat.

Department of Administrative Reforms & Public Grievances has put in place an online web enabled portal-**Centralized Public Grievance Redress and Monitoring System (CPGRAMS)**, which helps in streamlining of the registration, transmission, tracking and monitoring of grievances both at the citizen and departmental level in the website: pgportal.gov.in/Grievance.aspx.

Grievances relating to the following are not within jurisdiction of Department of Administrative Reforms & Public Grievances:

1. Policy matters
2. Commercial contracts
3. Decisions involving quasi-judicial procedure
4. Service Matters (excluding payment of Gratuity & GPF)
5. Matters which are pending before courts
6. The matters which are already disposed of at the Minister's level
7. Frivolous complaints

4.5 The Administrative Reforms Commission (ARC) 7(Seven) Step Model for Citizen Centricity

1. **Define** all services which are to be provided and identify the clients
2. **Lay down Standards** and norms for each service
3. **Capability Building** to meet the set standards
4. **Perform** to achieve the standards
5. **Monitor performance** against the set standards.
6. **Evaluate the impact** through an independent mechanism
7. **Continuous improvement** based on monitoring and evaluation of results

Some of the recommendation the government organization having the public interface can implement for citizen centricity:

- To drop the suggestions by the public Suggestion box to be placed in suitable locations of the organization/Department
- Periodical consultation with citizens & other stock holders for the short term & long-term strategies for the overall development of the organization.
- Provision for submission of Online **Feedback** in the Website of the organization.
- The website of the organization should be updated regularly so that it contains all the vital Information about the organization.
- All the Government Organizations/Department to publish citizens/Client model. Few of Government have done it so far. **Ministry of Tourism is the leader in this initiative.**

In the 2nd Administrative Reforms Commissions Reports the Commission has already made detailed recommendations on this issue in its 11th Report on e-Governance.

The Commission further recommended that “Each government organization should prepare a time-bound plan for providing of transactional information through their websites. To begin with, this could be done by updating the websites at regular intervals, while at the same time, re-engineering the back-end processes and putting them on computer networks. Ultimately, all the back-end processes should be computerized.”

The Commission also recommended that “A clear road map with a set of milestones should be outlined by Government of India with the ultimate objective of transforming the citizen-government interaction at all levels to the e-Governance made by 2020. This may be enshrined in a legal framework keeping in consideration the mammoth dimension of the task, the levels of required coordination between the Union and State Governments and the diverse field situations in which it should be implemented.”

5 e-Governance in India

The first step of e-Governance in India was the Computerization of Government Departments. Present e-Governance initiatives will be encapsulating the finer points of Governance for instance Citizen Centricity, Service Orientation & Transparency.

The most expected benefits of e-governance includes improved efficiency, convenience and better accessibility of public services.

Now a day's government can further strengthen the community confidence gained by transparency through free distribution of government Data based on open standards.

To speed up the e-governance implementation across the various arms of Government at National, State and Local levels utmost care has been taken by the Government of India to adopted common vision and strategy having the potential of huge costs savings in presenting a seamless view of Government to citizens by:

- Sharing of core and support infrastructure
- Enabling interoperability through standards

The main motto of e-governance is to "Make all Government services accessible to the common man in his locality, through common service delivery outlets, and ensure efficiency, transparency and reliability of such services at affordable costs to realize the basic needs of the common man".

Over the years various State Governments and Central Ministries have made sustained efforts to usher in an era of e-Government[2].Large number of initiative[5] have undertaken at multiple levels to improve the delivery of public services and simplify the process of accessing them.

5.1 The National e-Governance Plan (NeGP)

On 18th May 2006 Government of India approved NeGP which consists of 27 Mission Mode Projects (MMPs) of State Governments & Central Government and 8 Integrated Mission Mode Projects (MMPs). The theme of the NeGP is to "Make all Government services accessible to the common man in his locality, through common service delivery outlets, and ensure efficiency, transparency, and reliability of such services at affordable costs to realize the basic needs of the common man". Government approved the National e-Governance Plan (NeGP), comprising of the following:

- Vision
- Approach
- Strategy
- Key components
- Implementation methodology &
- Management Structure

Different accessible or continuing projects in the MMP category under the Central Ministries, States & State Departments would be properly improved and enhanced to align with the objectives of NeGP.

Table 2. Mission Mode Project(MMP) under the NeGP[2]

Sl No	Central MMPs	State MMPs	Integrated MMPs
1	Banking	Agriculture	CSC[1]
2	Central Excise & Customs	Commercial Taxes	e-Biz[2]
3	Income Tax (IT)	e-District	e-Courts[3]
4	Insurance	Employment Exchange	e-Procurement[4]
5	MCA21	Land Records	EDI for eTrade[5]
6	National Citizen Database	Municipalities	National e-governance Service Delivery Gateway[6]
7	Passport	Gram Panchayats	
8	Immigration, Visa and Foreigners Registration & Tracking	Police	
9	Pension	Road Transport	India Portal[7]
10	e-Office	Treasuries	

5.2 Common Services Centers (CSC)

The Integrated Mission Mode Project **CSC** was introduced as strategic foundation of the NeGP, particularly in rural areas to implement e-governance at the grassroots level.

The main objectives of these CSCs are to offer a plethora of web-enabled services to rural communities’ in the field of e-governance, e-commerce & e-Agriculture.

The main objective of the CSCs is to provide the basic e-governance like;

- Application forms
- Certificates and utility payments such as electricity, telephone and water bills.
- Under the Scheme, favorable atmosphere would be created for private sector and NGOs’ active participation for effective implementation of the CSC Scheme thereby becoming a partner of the government in the development of rural India.

An ideal CSC will be having floor are of 100–150 sq. ft, minimum 1 PC with UPS, 1 Printer, Digital/Web Camera, power connection/Genset/Inverter/Solar,OS and other application software, Wired/Wireless Broadband Connectivity, trained and incentivized manpower. Investment required is around US \$ 4000/CSC [4].

The proposed CSC scheme under the **Private Public Partnership (PPP) model** will be consisting of a 3-tier structure:

- Village Level Entrepreneur(VLE) will be at the bottom of the structure also called the CSC operator
- The 2nd layer above the VLE will be the Service Centre Agency (SCA) which will be responsible for a cluster of 500-1000 CSCs; and
- The topmost layer called State Designated Agency (SDA) identified by the State Government will be responsible for managing the implementation above the entire State.

6 e-Panchayat

NeGP identify e-panchayat as one of its **Mission Mode Project** with the objective of bringing ICT to all 2,50,000 Panchayati Raj Institutions(PRIs) at the Gram Panchayat(village),Block(intermediate) and Zilla Parishad (district) levels. For planning and implementing the Centrally Sponsored Schemes(CSSs) i.e. the National Rural Health Mission, Mid Day Meal Scheme, Indira Awas Yozana (IAY), Pradhan Mantri Gram Sadak Yozana(PMGSY) and National Rural Employment Guarantee Scheme a third of the central budget of Rs 3,50,000 crore is allocated to the panchayats annually[6].

For effective implementation and monitoring of the money being spent on development and upliftment of masses a complete Panchayat Enterprise Suite (PES) was developed to outline all function and services that panchayats were expected and mandated to deliver. The core applications meet the pre-requisites for:

- De-centralize data base
- PRI budgeting and accounting
- Implementation and monitoring of central and State sponsored schemes
- Citizens centric services
- Unique codes to panchayats
- Essential GIS based services
- Online self-learning medium for all 3 million elected representatives and around 1 million officials' functionaries of the panchayats.

12 applications, known as core applications to be developed centrally through National Informatics Centre (NIC) were identified which will address majority of the needs of the stake holders. Out of 12 applications 4 of them namely:

- PLANPlus
- PRIASoft
- National Panchayat Directory
- Panchayat profile (Partly); have been developed & deployed in all 2, 50,000 and the rest 8 is likely to be rolled out in all the panchayats by May 2011.

12 states have already computers and trained manpower at Gram panchayat level. For the rest States and UTs Ministry of Panchayati Raj has suggested a service procurement model wherein the service provider provides the hardware and trained operator for a period of 2 to 3 years during which the process can be internalized.

Presently Bharat Sanchar Nigam Limited (BSNL) has provided Broadband connectivity to 70,000 Zilla Parishads, Block Panchayats and Gram Panchayats Rest of the panchayats will be connected by broadband connectivity by **March 2012**.The 5000 Gram panchayats of Northeast will be connected by satellite connectivity (VSAT).

Ministry of Panchayati Raj will be launching training programme from April 2011 for training 2 representatives in each Panchayat. There will be 6 day module

educating representatives on basic ITC skills through identified Institutes and CDs loaded with Computer Based Training (CBT) materials in all vernacular languages for the first time users. These courses are designed to enable the panchayat officials to draft and upload documents via the core application platforms and browser through relevant sites for information on health, agriculture and the like. For training panchayat representatives and functionaries on CBT content the Department of Electronics and Accreditation of Computer Classes (DOEACC) has accredited 2,500 institutes nationwide. There will be 2 online examinations one for the Trainees & another for the Trainers. Those who will clear the examinations are likely to offer incentives from the state government concerned. The target is to train 25,000 in next fiscal year and follow it by 10, 00,000 in the year after.

7 Community Information Center (CIC), the ICT Experience in Rural Areas of North East and Jammu and Kashmir [7]

CIC Project was announced by the Prime Minister of India in January 2000 as a special Package for NE India & Jammu & Kashmir in order to bridge the digital divide and to speed up economic development as well as to overcome of the communication bottleneck. The implementation period for the Project slated for 2 years and the total time of the project was 5 years. The project envisages a community centre with Internet facility through VSAT at each administrative block.

Table 3. CIC infrastructure[Source: *cic.nic.in*]

Sl. No	Name of Equipment	Quantity
1	Server Machine	01
2	Client Systems	05
3	VSAT	01
4	Laser Printer, Dot Matrix Printer	01each
5	UPS (1KVA, 2 KVA)	01 each
6	LAN & Networking Equipments [Modem, LAN hub, TV, Webcam etc]	

The respective State Governments appointed 2 persons as CIC Operator for each CICs.To bring the uniformity of the presentation of the contents of the Websites the Websites were dynamically generated and managed using Community software solution framework eNRICH, jointly prepared by NIC and UNESCO.

The following table gives the details of the CIC in North Eastern India and the State of Jammu & Kashmir:

Table 4. List of the CIC in North Eastern India & the State of Jammu & Kashmir

Sl. No	Name of State	No. of CIC
1	Arunachal Pradesh	56
2	Assam	219
3	Jammu & Kashmir	134
4	Manipur	33
5	Meghalaya	32
6	Mizoram	26
7	Nagaland	52
8	Sikkim	40
9	Tripura	40
	Total	632

8 Conclusion

1. ICT is Cross Cutter, which can be used in any field for the Development of Mankind. ICT must reach the grass root level for bridging the Digital Divide and inclusive growth of Indian economy.
2. A successful ICT application in e-Governance giving one-stop solutions for rural community is the need of the hour. The best example of ICT application for the masses in India is India Railways Catering & Tourism Services online ticketing website *irctc.co.in*.
3. Implementation of IT Act 2000 is yet in a nebulous stage though the Act was enacted 10 years back. **E-mail & other electronic** form of record are still considered as secondary evidence in the court of Law in India. There is an urgent need of ICT Act to address all the Legal issues pertaining to ICT.
4. If the websites of the Government Departments/ Ministries were updated regularly it will not only enhance the transparency but will be saving the valuable time otherwise spent entertaining the enquiries of the public as well replying the questions asked under the **Right to Information Act 2005**. Electronic form of communications is rarely used for Official communications, due to existing bureaucratic system.
5. In some cases it will be difficult to find Village Level Entrepreneur (VLE) to operate a CSC in remote areas.

9 Future Ahead

1. There is great scope for study & research in the field of e-governance Administrative Reforms, Cyber Law & Implementation of Information Technology Acts.
2. The CSCs are likely to serve as Information Kiosks for easy & quick access for a plethora of Information to rural masses.

3. m-Government cell is being created under the Department of Information Technology's(DITs) e-Governance unit with a mix group of 1,500 officials, comprising members from the Academy, Industry stake holder, civil society(non – government organizations) and officials from DIT and Department of Telecommunication. The aim is to deliver all services presently available through computers through **Mobile Technology**.
4. As part of 100 day agenda, Ministry of Information Technology is planning to come up with Electronic Service to be passed in parliament as an electronic service delivery Act (ESDA) by the end of this year. This bill intends to make it mandatory for all government departments and ministries to deliver public services only in electronic mode from a cut date and that cut date will be fixed by the concerned department or the ministry depending upon their level of readiness.
5. Another emerging field of study in application of ICT in rural development context is cloud Computing.

References

1. United Nations E-Government Survey (2010)
2. Web portal of Department of Information Technology, Govt. of India, <http://mit.gov.in/content/national-e-governance-plan>
3. Portal for Public grievance, Dept. of Administrative Reforms and Public grievances, Govt. of India, <http://pgportal.gov.in/AboutUs.aspx>
4. Web portal of Department of Information Technology, Ministry of Communication and Information Technology, Govt.of India, <http://csc-india.org/Welcome/Home/tabid/617/language/en-GB/Default.aspx>
5. Web portal of infoDev, The World Bank, <http://infodev.org/en/Page.About.html>
6. Official Website of Govt of India, Ministry of Panchayati Raj
7. Deka, G.C.: Prospect of Rural Enterprise- An Approach to ICT as a Tool for Rural Employment and E-Governance. In: Proceedings of IETE 37th MTS, pp. 41–46. ICTIRD, Kolkata (2006)

Enhancing Sustainability of Software: A Case-Study with Monitoring Software for MGNREGS in India

C.K. Raju and Ashok Mishra

Indian Institute of Technology, Kharagpur, India
ckraju@agfe.iitkgp.ernet.in
amishra@agfe.iitkgp.ernet.in

Abstract. Software can be perceived either as a resource or as a property. Software is a resource if its ownership is in public domain, and a property if its ownership wrests with some entity. Quite often, the license under which a software is made available determines whether the software can be treated as a resource or as a property. Licenses like GNU General Public License enforce the resource characteristics upon the software. Proprietary licenses render the software under some form of private ownership. Usually sustainability of projects are determined by the resource/property characteristics of their constituents. Sustainability of software projects can be enhanced if software can be transformed to possess resource characteristics. Software deployed in ICT applications is often dependent on data standards and protocols. While ownership issues determine the characteristics of data standards, issues related to transparency determine resourcefulness of protocols. Fairness in implementation and accessibility to test or inspect are guaranteed before the public, only when software data standards and protocols are made available as resources. By adhering to open data standards and transparent protocols, the resource characteristics of software data standards and protocols can be enforced. Once the resource characteristics prevail over software, data standards and protocols, the sustainability of software would get enhanced.

1 Introduction

Sustainable development has been defined by a World Commission [2] constituted for the purpose by United Nations as one that

meets the needs of the present without compromising on the ability of the future to meet their needs

wherein aspects related to control and usage of resources were implicit.

The definition generated criticisms for the futility of solutions prescribed as well as admirations for its exhaustive analyses of the problems faced by the world [1]. Despite these differences, what got universal acceptance was the fact that

dwindling resources of earth called for more attention for their preservation or regeneration.

The work done here is an attempt to understand the variations with which the theme of sustainable development has evolved over the last decade and half, in the context of Information and Communication Technology (ICT) based projects. An attempt is made to study whether some of the core principles of the studies by World Commission on Sustainable Development could be applied to software in the realm of ICT applications. If it is possible to do so, then it should also be possible to derive a broad framework whereby sustainability of software in such ICT applications could be evaluated or enhanced if needed.

Towards this, a case study on an existing ICT project deployed in rural local governments in India, which has been declared as a success by the Planning Commission of India, has been taken up for analysis and evaluation. The ICT project pertains to an online solution developed by National Informatics Center (NIC) for monitoring activities of the Mahatma Gandhi National Rural Employment Guarantee Scheme (MGNREGS) of Government of India. If a framework could be developed, then it could set in motion a new approach to developing a sustainable software solution for monitoring MGNREGS.

2 Existing Studies on Sustainable ICT

A review done on recent studies involving sustainability of ICT projects reveal diverging conclusions on the concept of sustainability itself. A few typical cases analyzed are summarized in succeeding paragraphs.

Sustainability of ICT applications has been interpreted to be achieved if the applications were embedded within national projects that are public in nature [13]. According to this study, sustainability of such ICT projects materializes only when they are associated with social development of marginalized communities through their livelihoods. These projects were also supposed to be driven from the bottom to the top, i.e., from the marginally excluded digitally-poor societies to the digitally-rich societies.

In another study involving public services and citizens it was concluded that ICT projects needed to be integrated to existing citizen services and other public interfaces for increasing efficiency of delivery of services, introducing more transparency in administrative affairs and also for allowing more connectivity [11]. Sustainability of such forms of ICT projects had been interpreted as guaranteeing availability such digital services in future too.

Another interpretation had been one where sustainability would conclude with meeting challenges like providing access to infrastructure, overcoming limited formal education, handling training and capacity building efforts, overcoming financial or political constraints and integrating socio-cultural issues [16]. These conclusions were based on a study undertaken in a rural area and sustainability issues were closely linked to institutional, economic, political and technological sustainability.

In a study undertaken in Denmark, it had been concluded that bringing people close to network by providing them with broadband services through fiber

connectivity to residences, it would be able to achieve sustainability [17] to ICT initiatives. Internet penetration has also been identified as having close linkages with economic growth and thereby sustainable development [12]. Sustainability is thus linked with access to infrastructure. Sustainability has also been interpreted as a means to empower people to learn and evaluate development initiatives that use communication technologies [20], [9] or by using specific broadcasting utilities like community radio [10]. Increasing mobile communication facilities has also been pointed out as a means of sustainable growth for the economy [5] with ICT as an important factor. Regarding the resources pertaining to sustainability suggested here, the core argument relates resources to newer forms of digital or analogue media that could be created and making available such infrastructure for various purposes.

There are also arguments that portray sustainability as an objective attained on enhancing private-public partnerships [14], [7], [8] or doing so by improving mutual trust [19]. Newer forms of relationships that can be forged using ICT and strengthening existing relationships are hinted while mentioning sustainability of ICT projects in these studies.

Sustainability has also been projected as a measure of cost-effectiveness of ICT solutions [6] which at times can be achieved by using re-usability feature of open source projects [21]. In these cases, an indication of financial resources that are required for commissioning and maintaining ICT projects is implicit in the recommendations of the authors.

In some of these studies, the authors tend to link sustainability of ICT applications to economic growth which could be measured using GDP as a parameter. Sustainability of ICT applications, therefore, is indicative of sustaining GDP growth. A few authors seem to convey that transforming media into a more suitable form would create sustainability, some other authors seem to convey that amending the processes involved in the project would create sustainability and yet another set of authors seem to convey that altering the objectives or end results would create sustainability. A major assumption in the studies, though not explicitly stated, is that economic growth with ICT applications enhances the well-being of citizens. Factors like employment generation or redressal of income inequalities were taken for granted in these approaches.

OECD has attempted to analyse India's economic growth over the last two decades. A recent survey on four emerging economies [3] reports that India's growth in GDP had the lowest impact on employment generation when compared with countries like Brazil and South Africa. It also mentions that in India, income inequality figures deteriorated with increase in GDP, hinting that benefits of economic growth were skewed. Our inquiries, therefore, ought to ensure that these anomalies are not exemplified with the introduction of ICT enabled environment, rather it should attempt to redress them, if possible.

For calculating GDP, the consumption of services by government agencies is a factor. If software licenses are purchased by government agencies, GDP would improve. Proprietary software licenses are merely replicated by proprietary software firms and sold to public agencies, which doesn't involve employment creation

with increased spending. If Free Software were to be procured for assisting public services, it would mean that most of the expenditure that would be incurred would be spent on software maintenance or other services involving labour, as replicating Free Software does not involve significant spending. An ICT-enabled environment usually improves GDP as the frequency of transactions are higher in the enabled environment [4]. This has also been corroborated through a case study involving ICT investments in China [22]. However, when the software components are proprietary, the GDP will further continue to increase [4] neither guaranteeing employment generation nor improving income inequalities. A fresh approach on realising sustainability of software in ICT applications, therefore is imperative.

3 Sustainability of Software in ICT Projects

In the World Commission's report on sustainable development [2], it was found that the onus was on rearrangement of the resources involved. Natural resources were appropriated as private property and consumed for attaining development. This led to a thinking on issues like conservation or replenishment of resources.

Most of the cases that were taken up for analysis of sustainability in ICT projects, however, remained inconclusive on identifying resources that determine sustainability. Unanimity in the preference for resources too were missing in the discussions. Further, the diverging views implied that serious discussions on resources too were absent in the literature. In these discussions involving sustainable development of ICT based solutions, sufficient details too were not provided apriori, on dwindling resources that demanded attention. There were also no attempts to prioritize such resources.

The constituents in software projects could be intellectual or information-based ones. As resources, the constituents remain publicly accessible without excluding anyone. They transform into property when they lose their resource-worthiness. Moreover, it is the characteristics of such constituents that determine the demand for greater financial support. For instance, if an ICT solution included patents which have a royalty tag attached, then obviously their deployment would accompany demand for financial support. So is the case if information to be processed were to be encoded in proprietary formats. Its usage and management too would demand financial support. The control over deployment in multiple installations too would be dictated by the owners of intellectual property like patent holders or by the owners of the proprietary encoding formats like monopoly software developing establishments. Issues related to analysis, modification and re-distribution of such constituents too would be dependent on the nature of their ownership, which are under private control. The situation now deals with the environment of information processing as a property and not as a resource.

If one were to engage intellectual or information-based resources, then sustainable development of software ought to be on preserving the characteristics of the resources or promising their regeneration. Usually a property is created

by claiming exclusive ownership over the resource. Once a resource gets converted into property, there is a possibility of resource getting scarce. It has been brought out earlier that even infinitely reproducible resources like information-based ones that could be transmitted without much of a cost over a frictionless medium between any two points in the world, could be artificially controlled using legal mechanisms after appropriating rights over their ownership¹. Moreover, if, by the act of converting a resource into property, a situation of universal deprivation gets created, then the theme of sustainability becomes significant.

The quest for sustainability of software in contexts like these should be therefore, to suggest measures that preserve the resource-character, which can be implemented by disallowing exclusive ownership of any kind over the resource or by creating a conducive environment whereby such resources could be replenished or regenerated.

A case-study related to monitoring a rural public service in India is taken as a sample to relate and advance the arguments proposed herewith.

4 ICT through Planning Commission of India

Information Technology was projected as a sector that needed individual attention in the ninth five-year plan (1997-2002) formulated by Planning Commission of India [18]. E-Governance initiatives received a major impetus when the ninth five-year plan advocated earmarking of up to three percent of the budget on Information Technology to all ministries and government departments.

According to the plan document, the proposal would generate about one million additional jobs every year. It would also lead to increased productivity in various sectors, improved timeliness in implementation of projects, leading to the minimization of time and cost overruns, as well as creation of entirely new enterprises like Software Export, Internet-based Services, Electronic Commerce etc. The Commission expected that the impact of IT would be predominant in the social sectors like health, education, judiciary and rural development.

The ninth-five year plan period also saw setting up of a separate Ministry of Information Technology by Government of India. A comprehensive Information Technology (IT) Act 2000 was enacted in India, pursuant to the United Nations Model Law on Electronic Commerce 1996, where legal recognition was to be accorded to all commercial transactions through electronic media.

The tenth five-year plan (2002-2007) urged e-governance investments to be sustainable and suggested adoption of standards in electronic transactions. It is here that issues relating to sustainability with regard to projects in Information Technology get mentioned in plan documents initially. Subsequently, in the eleventh five-year plan (2007-2012), the Planning Commission declared that an online software implemented for monitoring National Rural Employment Guarantee Scheme (NREGS) was a success [18].

¹ Eben Moglen, Software Freedom Law Center had made this comment at the Annual Lecture at Scottish Society for Computers and Law in 2007. Eben had been the principal author of the initial version of GNU General Public License.

Table 1. Software Components in a Client Location

Software Component	Owned by
MS Windows XP SP-2	Microsoft Inc
MS IIS Web Server	Microsoft Inc
MS SQL Server 2000	Microsoft Inc
NREGASoft	National Informatics Center

Through such a declaration, it was thus implied that the monitoring software christened NREGASoft which was developed by NIC was sustainable. Since Planning Commission had earlier hinted on adoption of standards in electronic transactions for sustaining ICT applications, it also implied that NREGASoft adhered to such measures.

5 Financial Implication of NREGASoft

The ICT application NREGASoft which monitors the activities of NREGS is capable of operating in 'online mode' as well as in 'offline mode' [15]. The manual lists out details on both these different modes under which it can operate. The online mode is impractical in local self-government institutions where the software is deployed, owing to the large volume of data entry operations that is required of the project. In the offline mode of operation, which is the de-facto mode of operation, a local database server in a local body captures and processes data which is later synchronised with the central database server situated near to the headquarters of Ministry of Rural Development. It would be worthwhile to observe the cost of the entire project when fully functional and operated in the offline mode.

In February 2010, a Microsoft Windows Small Business Server which contained the components mentioned in Table 1 for successful operation of NREGASoft at a client location with 5 client licenses was priced² at 79,758.00 Rupees (US\$ 1899)³. The online monitoring software application NREGASoft was made available free of cost to the local bodies by NIC.

An approximation of the total cost of the project to the country is therefore the cost of licensed set of software for one client site multiplied by the total number of local bodies in India where NREGS is in operation. Though accurate figures are not available because of regular demarcation of local bodies by different election commissions, a rough estimate puts the total number of local bodies in India at over 240,000. The total cost for licenses alone in the ICT project developed by NIC which monitors NREGS activities in India would therefore

² The published prices at Microsoft's website <http://www.microsoft.com/sbs/en/us/pricing.aspx> keep revising, the prices of a premium version as on Jan 2011 is shown as US \$ 1899.

³ At a Currency Exchange Rate of US \$ 1 = 42 Indian Rupees.

come to 19,141,920,000 Rupees (US\$ 455.7 million). A budget allocation of such an amount, the beneficiary of which is a single software corporate, will now be deemed necessary, in order to monitor a rural employment guarantee scheme which aims to address the twin problem of poverty and scarcity of employment opportunities amongst rural population in India. It is pertinent to note that even after making such a payment to the software licensing company, there wouldn't be any ownership to any public agency or Government of India, over the software on which NREGASoft executes.

Knowledge of such figures might propel investigation for enhancing sustainability by probing alternate cost-effective solutions or by negotiating with the software firm that licenses such software. However, one needs to know whether characteristics of intellectual or information-based resources were responsible for reaching at this situation and how sustainability of software is related to the overall sustainable development of the ICT project. Therefore, a detailed scrutiny of other resources is required.

6 A Framework to Enhance Sustainability of Software

It is the nature of ownership that helps to distinguish between a property and a resource. At least two forms of ownership exist in the realm of information or intellectual domain, which are inclusive ownership and exclusive ownership. Exclusive ownership excludes others, while inclusive ownership doesn't. For instance, Free and Open Source projects which are licensed under Free and Open Public Licenses like GNU General Public License (GPL⁴) enforce inclusive ownership of software projects. With an inclusive ownership, there can be other owners too for the same resource. Inclusive ownership is as good as null ownership, as nobody has exclusive ownership which deprives others. Therefore if sustainability of a software project is to be enhanced, it is enough if the resource-worthiness of the software project is improved, which is by transforming services having exclusive ownership into services having inclusive ownership.

Pertinent issues to be raised in software projects therefore, should be on how resource-worthiness for such projects needs to be approached. A minimal framework would be to

1. perceive software component itself as a potential resource for the development project,
2. enhance resource-worthiness of sub-components that is used to build the software component and
3. ensure that any software components newly created should retain the inherent qualities mentioned herein.

Taking the first condition of the framework, the aim should be to perceive the software component itself as a resource. Here, it is possible to have the software

⁴ GNU stands for GNU's Not Unix a recursive acronym and GPL stands for GNU General Public License.

component either as a property or as a resource. In NREGASoft, for instance, the software environment in which the application resides is a property where a software corporate holds exclusive ownership (see Table II). NREGASoft can be altered to reside in a new software environment which is devoid of any sort of exclusive ownership. By preferring to have a variant of NREGASoft to work in Free and Open Source Software environment, i.e., software which are released under a GNU General Public License (GPL) or equivalent license, the resource-worthiness of the software component in ICT application can be significantly altered.

In the second condition of the framework, the aim should be to ensure that only components that are resources should go into the making of any software. Here, resources could be data standards which are also open standards, established protocols and any software components that do not have restrictive patents associated with them. In NREGASoft, public information is processed using protocols that are non-verifiable, data formats which are proprietary and software components that have patents associated with them. These characteristics are fundamental to an exclusively owned property.

In the last condition of the framework, any software product created by using resources of the ICT application, should also retain the characteristics of a software resource. By associating Free and Open Licenses to the software end-products, their resource worthiness could be ensured.

In resolving to apply some of the core principles of the World Commission's approach to sustainable development, it is therefore possible that resource-worthiness of software ICT projects could be significantly altered, and their sustainability enhanced. To validate the framework, an alternative solution needs to be developed which would demonstrate its operational efficacy.

7 An Approach That Enhances Sustainability of Software

Drawing necessary conclusions from the minimal framework that could be use to improve the resource-worthiness of ICT applications, certain definite directions could be rendered

1. to the nature of software environment that is to be deployed in similar ICT applications,
2. to the nature of data formats and protocols to be used or followed, and
3. to the nature of ownership with which the monitoring solution is to be developed and released for deployment.

Since an environment with Free Software components enhance the resource-worthiness of the environment, it is imperative that all software utilities that make the environment conform to Free and Open Licenses. In the alternative approach followed by the author, a GNU/Linux was used as the operating system. Apache and Nginx were both tried out as webserver software. Both these applications are available under Free/Open Licenses. For database server, MySQL released under GPL was used.

In a strict Free and Open Source environment, one can ensure fairness in usage of existing protocols as well as data standards for processing of data. This is not possible in a closed-source environment, where one has to necessarily rely on the claims made by the developer, with little or no way of knowing it conclusively. The database environment that NREGASoft uses at present is proprietary, owned by Microsoft Inc. In the new approach here, MySQL was used as the database server for deployment at client locations.

The last part of the mini framework that ensures resource-worthiness required any software developed from the software environment too to be released under Free and Open Source Licenses. Here, the scripts were developed after integrating Drupal, a content management software which is released under Free and Open Source License to the webserver. Since Drupal itself is a front-end to a database application, the final dump of contents is sufficient to recreate the original monitoring software. By attaching GPL conditions to the dump, the monitoring software and the contents could be ensured to remain as resource in future too.

The experiment conducted proves that resource-worthiness of software in ICT projects can be enhanced by bringing in changes to the characteristics of its sub-components. By deciding not to allow usage of software solutions that are released under proprietary licenses helps to improve the sustainability of software environment. Implementing software solutions that are released under Free/Open Licenses resource-worthiness of such solutions can be guaranteed in future too. Since sustainability of software in ICT solutions is integrated with the characteristics associated with resources, preventing exclusive ownership over resources is also essential, which is taken care by the legal aspects of Free/Open Licenses.

On the financial implications for the alternative approach, the cost requirement for licenses needed to be deployed in the 240,000 local bodies in India can be effectively nullified. The cost of licenses that NREGASoft carries for its implementation, can now be disbursed for generating additional employment opportunities in rural areas. Thus it can also be seen that financial cost of such ICT projects are heavily dependent on the characteristics of the software components, nature of formats or protocols used, and the extent of usage of patents or other restrictive practices.

It also validates the presupposition that if these characteristics are altered to remove contents that contain exclusive ownership, then the cost of projects too can be brought down, apart from enhancing sustainability of software in the project.

8 Conclusion

Sustainable development as a theme came into prominence in the last quarter of the twenty-first century. A study commissioned under the aegis of United Nations had concluded that key to sustainability lay on the means in which natural resources were appropriated. In the case of software in ICT applications too, the theme of sustainability holds great promise.

Software in ICT applications can be perceived as a resource which could include resources within itself and which are capable of generating new resources as end product. However, through extraneous interventions the resource-worthiness of software in ICT components could be altered significantly. If exclusive ownership is allowed, these resources get converted into private property. Sustainability of software in ICT applications can be enhanced if their resource-worthiness can be preserved or improved.

An alternative scalable experiment that is developed for monitoring NREGS activities validates the claim that some of the key findings of World Commission on sustainable development can be applied in the realm of software in ICT application. Adoption of Free and Open Source Software, adherence to open data standards and mandating use of transparent protocols would ensure that software in such ICT solutions improves its resource-worthiness and thereby, enhances its sustainability.

References

1. Adelina, M.M., Luciana, C.C.: Sustainable Resource Use and Sustainable Development: A Contradiction?! Center for Development Research. University of Bonn (November 2004)
2. Bruntland, G. (ed.): *Our Common Future: The World Commission on Environment and Development*. Oxford University Press, Oxford (1987)
3. Elena, A., Michael, F.: *Growth, Employment and Inequality in Brazil, China, India and South Africa: An Overview*. OECD Publishing, Paris (2011)
4. Ezell, S., Andes, S.: ICT R and D Policies. *An International Perspective* 14(4), 76–80 (2010)
5. Gerami, M.: Information and communication technologies in middle east countries. In: *International Conference on Networking and Information Technology (ICNIT 2010)*, pp. 338–342 (2010)
6. Grace, J., Kenny, C.: A short review of information and communication technologies and basic education in LDCs—what is useful, what is sustainable? *International Journal of Educational Development* 23(6), 627–636 (2003)
7. Hearn, G., Kimber, M., Lennie, J., Simpson, L.: A way forward: Sustainable ICTs and Regional sustainability. *The Journal of Community Informatics* 1(2) (2005)
8. Hosman, L., Fife, E.: Improving the prospects for sustainable ICT projects in the developing world. *International Journal of Media and Cultural Politics* 4(1), 3–8 (2008)
9. Hudson, H.: Overcoming the barriers of isolation: Strategies for small and isolated developing states. In: *Information Infrastructure Symposium, GIIS 2009, Global*, pp. 1–7 (2009)
10. Hussain, F., Tongia, R.: Community Radio for Development in South Asia: A Sustainability Study. In: *Proceedings of the International Conference on Information and Communication Technologies and Development, ICTD 2007*, pp. 1–13 (December 2007)
11. Jager, A., Reijswoud, V.: E-Governance in the Developing World in Action. *The Journal of Community Informatics* 4 (2008)
12. Kim, C.: A study of Internet Penetration Percents of Africa using digital divide models. In: *Proceedings of PICMET 2010 Technology Management for Global Economic Growth*, pp. 1–11 (2010)

13. Mallalieu, K., Roche, S.: Selecting Sustainable ICT Solutions for Pro-poor Intervention. In: *Digital Poverty: Latin American and Caribbean Perspectives*, p. 119. Practical Action Pub. (2007)
14. Mansell, R., Wehn, U.: *Knowledge of Societies - Information Technology for Sustainable Development*. Oxford University Press, New York (1998)
15. NIC, Government of India: *User Manual of NREGA*. Ministry of Rural Development, GoI (2007)
16. Pade, C., Brenda, M.: An exploration of the categories associated with ICT project sustainability in rural areas of developing countries: A case study of the Dwesa project. In: *SAICSIT 2006: Proceedings of the 2006 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries*, pp. 100–106. (2006)
17. Pedersen, J., Riaz, M.: Bringing Fiber To The Home to Rural Areas in Denmark. In: *Proceedings of the 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies, ISABEL 2009*, pp. 1–6 (November 2009)
18. Planning Commission, Government of India: *India's Five Year Plans: Complete Documents: First Five Year Plan (1951-56) to Tenth Five Year Plan 2002-2007*. Academic Foundation (2003)
19. Rajalekshmi, G.K.: E-governance services through telecenters: The role of human intermediary and issues of trust. *Inf. Technol. Int. Dev.* 4(1), 19–35 (2007)
20. Sundén, S., Wicander, G.: Bridging the Digital Divide ICT Solutions Supporting Economic and Social Development for the Unseen Majority. In: *HumanIT*, pp. 18–34 (2003)
21. Heike, W., Paterson, B.: Sustainable Software Development. In: *SAICSIT 2004: Proceedings of the 2004 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries*, pp. 274–278 (2004)
22. Yi, L., Zheng, L., Yan, Q., Li, Y.: The relationship between ICT investment and economic growth in China. In: *IEEE International Conference on Advanced Management Science (ICAMS)*, vol. 2, pp. 136–140 (2010)

Proficient Discovery of Service in Event Driven Service Oriented Architecture

P. Dharanyadevi¹, P. Dhavachelvan¹, S.K.V. Jayakumar¹, R. Baskaran²,
and V.S.K. Venkatachalapathy³

¹Department of Computer Science, Pondicherry University, Puducherry, 605014, India

²Department of Computer Science and Engineering, Anna University, Chennai, India

³Department of Mechanical Engineering, Sri Manakula Vinayagar Engineering College, India

{dharanyadevi, dhavachelvan, skvjey, baskaran.ramachandran,
vskvpathy}@gmail.com

Abstract. This paper proposes an architectural framework that able to take into account, in an asynchronous fashion. Service Oriented Architectures (SOAs) supports request reply pattern, it not suits for the stop- start (asynchronous) communication. Therefore, in this paper we integrate the SOA with the Event Driven Architecture (EDA) in web service discovery. The critical problem of service discovery in application-oriented web service technologies is how to match the user request with the web service advertisement, efficiently and accurately as per the user requirements. To meliorate the accuracy and efficiency in WS-discovery, the Proficient Matchmaking Technique (PMT) has been proposed in EDSOA. The performance evaluation shows that performance of the proposed technique is aloft than the existing technique.

Keywords: Web Service Discovery, Event Driven Architecture, Service Oriented Architecture, Performance.

1 Introduction

Web service discovery is the process of finding the appropriate services for a given task defined by the user [1], [10]. The web services are bootless when they do not discover the appropriate services. Web service discovery consists of three processes: matchmaking, selection and ranking, this paper deals with the matchmaking technique. Matchmaking is the key technique in web service discovery, it matches the user request with the web service advertisements [2] and its significant challenge is to discover the web services efficiently and accurately [8].

The existing SOA based WS Matchmaking process (WSM), the matchmaking is done based on two sequential processes: functional based matchmaking and non-functional based matchmaking [4]. The functional based matchmaking process matches the web service advertisements based on the functional requirements of the user. The non-functional based matchmaking process further filters the results of the functional process based on the non-functional requirements of the user. And then the appropriate services are again filtered by the selection and ranking process to

obtain the desired services. As in the existing WSM, the matchmaking is done in two sequentially process, it decreases the efficiency. The proposed PMT, which is the logical entity in charge of perform the retrieval of services according to both functional and non-functional requirements in parallel, with respect to ordinary web service matchmaking mechanisms considering such different requirements in a sequential order. To increase the efficiency in proposed the matchmaking is carried out in parallel process.

Event Driven Architecture (EDA) is an architectural style that prescribes that communication between components has to be performed on the basis of event notifications, where events are basically understood as changes in the state of something relevant for the system [5], [6]. The interaction between events and services is commonly referred as Event Driven Service Oriented Architecture (EDSOA).

The paper presented works on service discovery in EDSOA using proposed PMT. It explained the concept and algorithm. Within the context of web service discovery, the paper focuses on the "matchmaking" problem, which consists in finding the most desired service on the basis of the service consumer request. The paper proposes an architectural framework able to take into account, in an asynchronous fashion, non-functional service requirements related to context-awareness and quality of service issues. Experiments were carried out for evaluation purpose.

The rest of this paper is organized as follows: Section 2 describes the proposed PMT in EDSOA; Section 3 discusses the experimentation for the evaluation purpose; Section 4 concludes and discusses the further work.

2 Proposed- Proficient Matchmaking Technique in EDSOA

The automated program by PMT in EDSOA is resides on matchmaker engine that roams across accessible Universal Description, Discovery, and Integration (UDDI) [11] registries to discover the efficient and accurate service in pervasive environment. Fig 1, illustrate the proposed PMT in EDSOA.

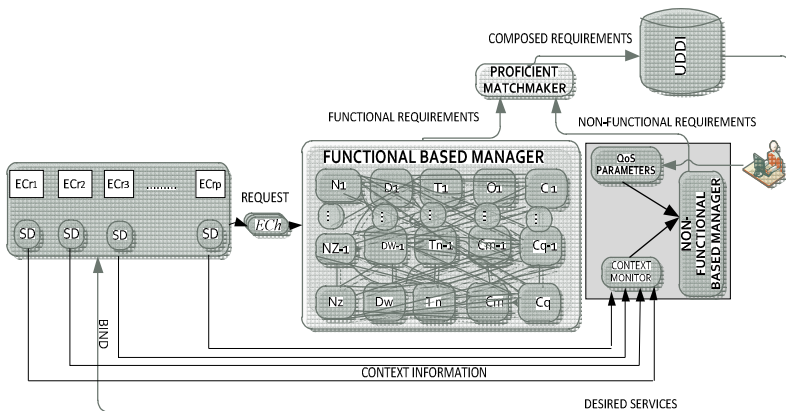


Fig. 1. Proposed Model- PMT in EDSOA

The proposed PMT in Event Driven Service Oriented Architecture (EDSOA) is discussed below,

Event Creator: As illustrated in the Fig 1, Event Creator (*ECr*) is the web users who initiate the technique by requesting the services.

Event Channel: As illustrated in the Fig 1, Event Channel (*ECh*) is the service request by the service consumer. As depicts in the Fig 1, the service request consists of the subset of Name (N), Description (D) [8], Tasks (T), Object (O) and Constraints (C). The name and description are used to identify and describe the property. The constraints may be date, time or location defined by the service consumers.

Event Processing: As illustrate in the Fig 1, the set of request (functional requirements) are send to the functional based manager and the functional based manager sends the functional requirements to the proficient matchmaker. The set of QoS parameters either defined by the developers or event creator. In this paper for implementation purpose we consider some of the QoS parameters such as, cost of service, capacity, integrity, performance, reliability, robustness, scalability and security [4]. *Cost of Service (CS)* is the cost per web service request or invocation. *Capacity (C)* is the limit of concurrent requests for guaranteed performance. *Integrity (I)* is the measure of the service's ability to prevent unauthorized access and preserve its data integrity. *Performance (P)* is categorized into throughput, response time, and latency. *Throughput (T)* is the number of requests served in a given time period. *Response Time* is defined as the delay from the request to getting a response from the service. *Latency (L)* is the time between client request and the start of its response. *Reliability (R)* is the guarantee that a service doesn't break within a given period of time in a way that the user (human or another service) notices the failure. *Robustness (RB)* is resilience to ill-formed input and incorrect invocation sequences. *Scalability (S)* defines whether the service capacity can increase as needed. *Security (ST)* includes the existence and type of authentication mechanisms the service offers confidentiality and data integrity of messages exchanged, non-repudiation of requests or messages, and resilience to denial of service attacks. ST includes the following concepts are authentication and encryption. *Authentication (A)* is the service either requires user authentication or accepts anonymous users. *Encryption (E)* is the type and strength of encryption technology used for storage and messaging; it is also a sub concept of supported standard.

As depict in the Fig 1, context information acquires and utilizes information about the context of a device to provide services that are appropriate to the particular people, place, time, event, etc. As depict in the Fig 1, the Sensor Device (SD) sense the context information's and the context monitor reads the context information's from the user such as, user identity, user preference, network characteristic, available band width, description of location, conditions of the physical environment,

timestamp, time span, order of events, date, day and year [7], [8], [9]. The non-functional requirements are obtained by composing QoS parameter and context information. The non-functional requirements are forwarded to the non functional based manager. The non functional based manager composes the non-functional requirements and forwarded to proficient matchmaker. The proficient matchmaker composes both the functional and non-functional requirements (consist of QoS and context information) in parallel and it matches with the web service advertisements either manually or automatically and retrieves the desired services. Then the desired services are given to the event creator.

Literally the time taken to transfer the functional requirements to proficient matchmaker is less than the time taken to transfer the non-functional requirements to proficient matchmaker; in this case functional requirements of the specific requestor should be idle until the non-functional requirements received. Existing web service architecture is SOA based, it is request reply pattern it won't suits for the stop-start (asynchronous) communication. The proposed architectural framework is in need of stop- start (asynchronous) fashion, so in the proposed architectural frame work the SOA has been integrated with the EDA.

2.1 Algorithm for Proficient Matchmaking Technique in EDSOA

The Proficient Matchmaking Technique in EDSOA matches the user request with the web advertisements in parallel and to support the stop-start (asynchronous) communication the SOA is integrated with EDA. As depicts in the Fig.2, the steps to be followed in the PMT are,

Step 1: The service consumer request for the service.

Step 2: Then the set of request are forwarded to functional based manager.

Step 3: The functional based manager forwards the functional requirements to the proficient matchmaker.

Step 4: The non-functional requirements consists of QoS attributes and context information. The QoS attributes are defined by the either event creator or developers.

Step 5: The QoS parameters are forwarded to the non-functional based manager.

Step 6: The Context Monitor (CM) sense the context information from the sensor in the event creator environment.

Step 7: Then context information's are sends to the non-functional based manager.

Step 8: The non-functional based manager compose the QoS parameters and context information and send to the proficient matchmaker.

Step 9: If the functional requirements of specific requestor are equal to the non-functional requirements of specific requestor, the proficient matchmaker composes the functional and non-functional requirements, if not functional requirements of specific requestor waits unit the non-functional requirements of specific requestor received.

Step 10: The composed (functional and non-functional) requirements matches with the web service advertisements registered in the UDDI and retrieve the desired services.

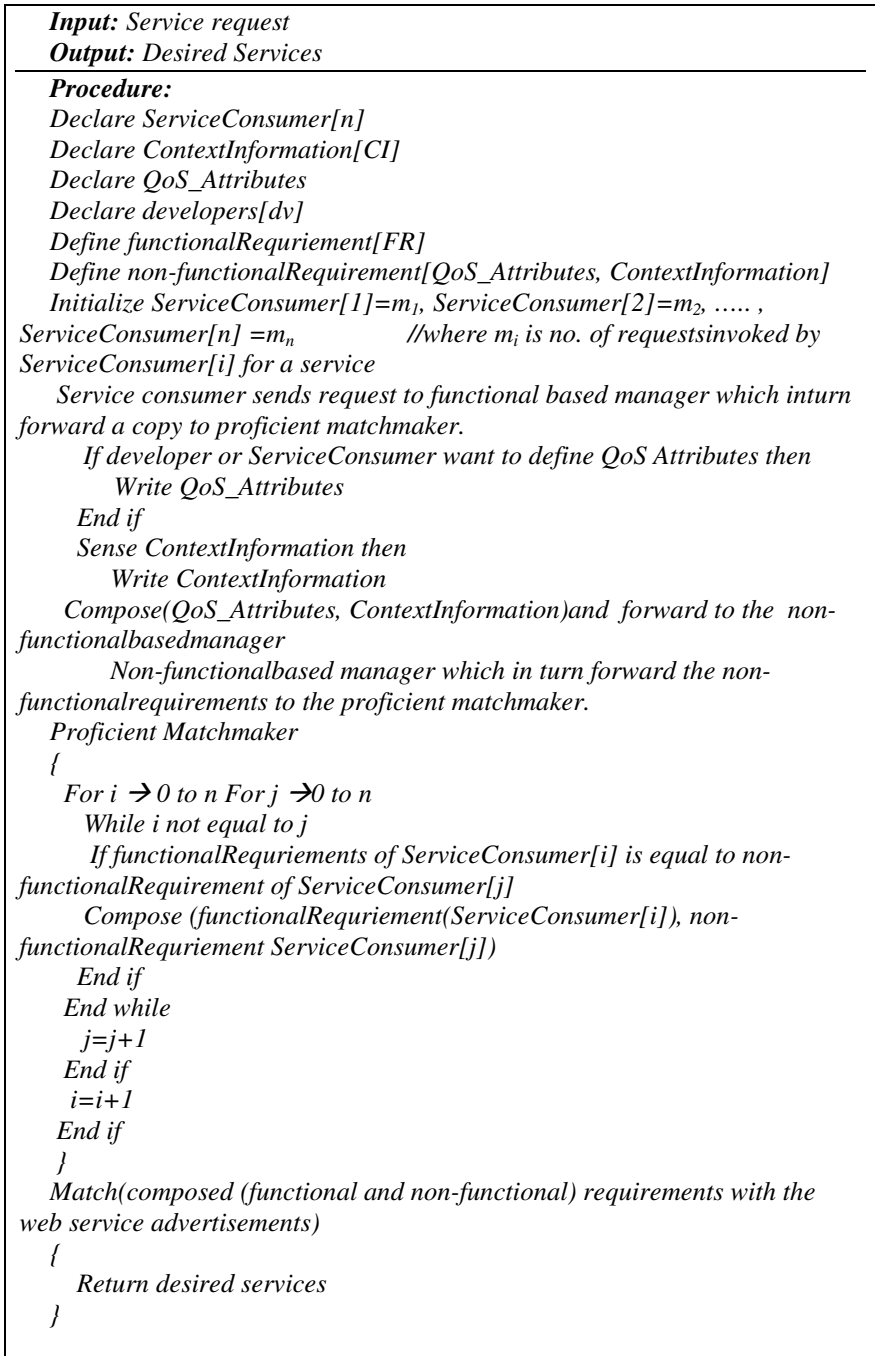


Fig. 2. Algorithm for Proficient Matchmaking Technique in EDSOA

3 Experimentation and Result Analysis

3.1 Experimentation Setup

The following techniques were implemented in Java (JDK 1.6.14) and Tomcat 6.2.29 web server, in order to compare the performance of the proposed technique with the existing technique. The list of web services such as “*CurrencyConvertor, Hotel, CarBrokerage, Airline, TravelAgent, Email, OnlineShopping, ELearning, EBanking and Hospital*” are stored in repository with the following details name of the service, WSDL URL and service description were stored in the Sqlserver 2008. We test the effect of each search to discover the services of the proposed technique and existing WSM process by each test scenarios by specifying the keyword. WSDL URL is used to utilize or integrate the service with user application. Description about the services is used to search the service.

User Interface: User can find the services through UI by giving the query as input, the given query will be sending to the searching component, and it will fetch the result as services. After finding the service, user can utilize the service in their application.

The existing WSM and PMT techniques were used to find out the exact services from the repository. After searching the list of services, the services can be filtered based on the exact matching and number of hits. In this implementation the system (laptop) is embedded with RFID tag, then the details of the system is recognized by the RFID readers, the context information’s such as date, time and location are fetched through readers and the information are send to the context information repository.

3.2 Result Analysis

The performance of the proposed PMT is analyzed through the metrics such as precision and f-measure. Precision is defined as the ratio between the total numbers of relevant services (Qi) that are retrieved to the total number of relevant and non-relevant services (NRi), retrieved from the registries [4]. When precision values are high, accuracy of the system is also high. The precision is calculated by the formulae,

$$Precision = Qi / (Qi + NRi) \quad (1)$$

F-measure that combines precision and recall is the harmonic mean of precision and recall. Recall is the ratio of the number of relevant services (Qi) retrieved and the total number of retrieved services (Ri) for the user requirements present in the registries [11]. When F-measure values are high, the accuracy and efficiency of the system are also high. The F-Measure metric can be calculated by the formulae,

$$F-Measure = (2.precision.recall) / (precision + recall) \quad (2)$$

Table 1, depicts the retrieved, relevant and non-relevant services obtained by the existing WSM and proposed PMT in EDSOA and calculated precision and f-measure values of existing and proposed techniques. Ri are the total number of retrieved services could consist of both the relevant and irrelevant services. Qi are the total

number of the relevant services are those services which are pick and chose by the user from the set of retrieved services as per the requirements, NR_i are the total number of non-relevant services are those services which not pick and chose by the user and i is the number of test scenario.

Table 1. Services obtained by the Existing WSM and Proposed PMT and its Evaluation by Precision and F-Measure

TS	WSM					PMT				
	Ri	Qi	NRi	Precision	F-Measure	Ri	Qi	NRi	Precision	F-Measure
1	26	12	14	0.46	0.75	08	5	4	0.62	0.90
2	9	6	3	0.66	0.92	03	2	1	0.67	0.93
3	19	12	7	0.63	0.90	05	4	1	0.8	0.98
4	18	11	7	0.61	0.89	04	3	1	0.75	0.96
5	16	9	7	0.5	0.85	05	4	1	0.8	0.98
6	19	9	10	0.47	0.76	06	4	2	0.67	0.93
7	9	6	3	0.66	0.92	03	2	1	0.66	0.93
8	17	8	9	0.47	0.76	07	4	3	0.57	0.86
9	16	7	9	0.43	0.73	07	4	3	0.57	0.86
10	22	9	13	0.4	0.70	09	6	3	0.67	0.93

Table 1 and Fig 3 depict the calculated precision values of existing WSM and proposed PMT. As illustrated in the Fig 3, the precision values of proposed PMT are higher than the precision value of existing WSM, when the precision values are high; the accuracy of the system is also high. So literally the accuracy is high in proposed PMT when compared to the existing WSM.

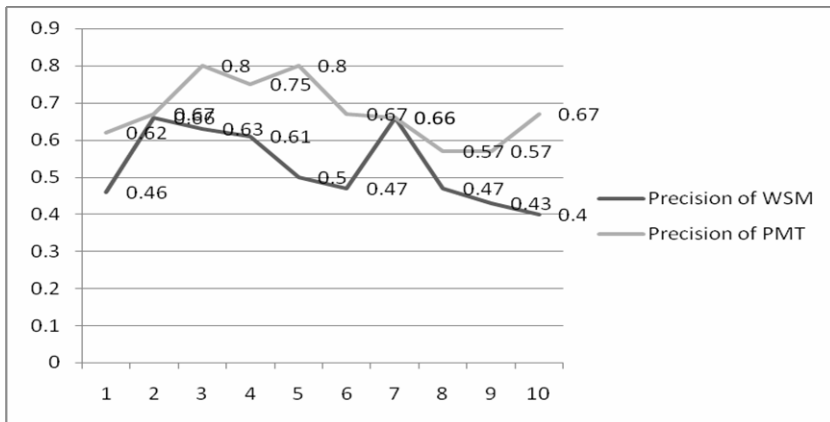


Fig. 3. Comparison of Existing WSM and Proposed PMT by Precision Metric

Table 1 and Fig 4 depict the calculated f-measure values of existing WSM and proposed PMT. As illustrated in the Fig 4, the f-measure values of proposed PMT are higher than the f-measure values of existing WSM, when the f-measure values are high, the accuracy and efficiency of the system are also high. So literally the accuracy and efficiency are high in proposed PMT when compared to the existing WSM.

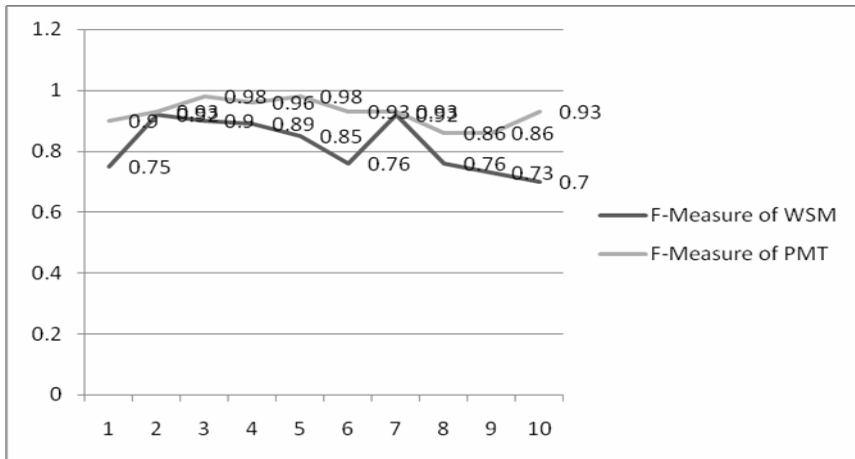


Fig. 4. Comparison of Existing WSM and Proposed PMT by F-Measure Metric

4 Conclusion

The proposed Proficient Matchmaking Technique (PMT), which is the logical entity in charge of perform the retrieval of services according to both functional and non-functional requirements in parallel, whereas the ordinary web service matchmaking mechanisms considering such different requirements in a sequential order. The paper proposes an architectural framework able to take into account, in an asynchronous fashion, non-functional service requirements related to context-awareness and quality of service issues. Experiments were carried out for evaluation purpose. According to the performance test, it is proved that, the proposed PMT has more accuracy and efficiency when compared to the existing WSM matchmaking process. A more complex and standardized semantics technology can help us express the proposed technique more powerfully in future.

Acknowledgments. This work is a part of the Research Project sponsored under the Fast track Scheme for young Scientists, India with the Reference No. as D.O.No.SR/FTP/ETA-112/2010 and the Departmental SAP Program. The authors would like to express their thanks for the support offered by the Sponsored Agency.

References

1. Ludwig, S.A., Rana, O.F., Padget, J., Naylor, W.: Matchmaking Framework for Mathematical Web Services. *Journal of Grid Computing*, 33–48 (2006)
2. Han, W., Shi, X., Chen, R.: Process-Context Aware Matchmaking for Web Service Composition. *Journal of Network and Computer Applications*, 559–576 (2008)
3. Peng, Y.-B., Zheng, Z.-J., Gao, J., Jiang, X.-Q., Ai, J.-Q.: Method Of Two Stages Semantic Service Discovery. In: *Proceedings of the Eighth International Conference on Machine Learning and Cybernetics*, Baoding, July 12-15 (2009)

4. Kritikos, K., Plexousakis, D.: Requirements for QoS-Based Web Service Description and Discovery. *IEEE Transactions on Services Computing* 2, 280–286 (2009)
5. Laliwala, Z., Chaudhary, S.: Event-driven Service-Oriented Architecture. In: *International Conference on Web Service* (2008)
6. Levina, O., Stantchev, V.: Member IEEE Realizing Event-Driven SOA. In: *Fourth International Conference on Internet and Web Applications and Services* (2009)
7. Dietze, S., Gugliotta, A., Domingue, J.: Towards Context-aware Semantic Web Service Discovery through Conceptual Situation Spaces. In: *ACM Conference SSSIA, Beijing, China* (2008)
8. Rong, W., Liu, K.: A Survey of Context Aware Web Service Discovery: From User's Perspective. In: *Fifth IEEE International Symposium on Service Oriented System Engineering* (2010)
9. Wang, H., Li, Z., Yang, B., Xia, H.: A Context-Aware Service Matchmaking method using Description Logic. In: *IEEE Asia-Pacific Services Computing Conference* (2007)
10. <http://xml.coverpages.org/uddi.html>
11. <http://www.w3.org/TR/wsa-reqs/>

Web User Session Clustering Using Modified K-Means Algorithm

G. Poornalatha¹ and Prakash S. Raghavendra²

Department of Information Technology,
National Institute of Technology Karnataka (NITK), Surathkal,
Mangalore, India

¹poornalathag@yahoo.com, ²srp@nitk.ac.in

Abstract. The proliferation of internet along with the attractiveness of the web in recent years has made web mining as the research area of great magnitude. Web mining essentially has many advantages which makes this technology attractive to researchers. The analysis of web user's navigational pattern within a web site can provide useful information for applications like, server performance enhancements, restructuring a web site, direct marketing in e-commerce etc. The navigation paths may be explored based on some similarity criteria, in order to get the useful inference about the usage of web. The objective of this paper is to propose an effective clustering technique to group users' sessions by modifying K-means algorithm and suggest a method to compute the distance between sessions based on similarity of their web access path, which takes care of the issue of the user sessions that are of variable length.

Keywords: web mining, clustering; K-means, Jaccard Index.

1 Introduction

Now the present generation is living in an information era. Moreover, the evolution of the internet along with the popularity of the web has made even an ordinary person to use the information available at his finger tips for various purposes. Web has been adopted as a critical communication and information medium by a majority of the population. Due to the rapid growth in the use of web the task of analyzing, understanding and producing useful information manually from a vast quantity of data available on the web is a very complicated and time consuming task. Thus, there is a requirement to develop techniques to get the valuable information, hidden in the web data, so as to improve the web performance.

This paper focuses on clustering web user sessions based on their navigation path which is of variable length. Clustering is a technique for grouping user sessions such that, within a single cluster the usage pattern is more similar while sessions in different groups are dissimilar. The knowledge discovered from the clustering may be used to analyze the pattern of usage of the web site by the user, to recommend for restructuring of web site, to pre-fetch or cache the pages and predict the next page

visited by the user to reduce the latency etc. As a result, realizing user's navigation patterns on a web site is an important activity for browser to pre-fetch as well as the web site designer to take decisions on redesigning the site.

A number of clustering approaches have been proposed in the literature. For example, Federico et al. [1] present a survey of the developments in the area of web usage mining, where the view points on various techniques like association rules, clustering, sequence patterns etc. are given. Yunjuan et al. [2] suggest that the focus of web usage mining should be shifted from single user session to group of user sessions and applied clustering for identifying such cluster of similar sessions. They introduce an effective clustering technique using belief function based on Dempster-shafer's theory. Chaofeng Li et al. [3] presented an algorithm for clustering of web session based on increase of similarities. Here number of clusters is defined according to the knowledge of application fields and uses ROCK to decide the initial point for each cluster.

Dariusz Krol et al. [4] investigated on the internet system user behavior using cluster analysis. Here sessions are represented as vectors where each dimension represents a web page and stores the value of user interest in each page of a session. The sessions are clustered using Hard C-Means algorithm. Yongjian Fu et al. [5] proposed a generalization based clustering method which employs the attribute-oriented induction method to reduce the large dimensionality of data. Prakash S Raghavendra et al. [6] modeled user behavior as a vector of the time spent at each URL. The cosine of the vector is taken as the similarity/distance measure, instead of euclidean distance and modified the standard k-means algorithm accordingly. Jin-HuaXu et al. [7] presented vector analysis and k-means based algorithm for mining user clusters.

In the web usage domain, there are two kinds of interesting clusters to be discovered: usage clusters and page clusters. In both applications, permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user/s query or past history of information [8]. George Pallis et al. [9] assessed the quality of user session clusters in order to make inferences regarding the users' navigation behavior.

The studies have shown that the most commonly used partitioning-based clustering algorithm, is the K-means algorithm, which is more suitable for large datasets. K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Euclidean distance is generally used as a metric. The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments.

In this paper, an effective method is proposed to compare variable length sessions and basic k-means algorithm is modified to get effective clusters, such that the initial centroid assignments will not have much impact on the clusters. Jaccard index is used to analyze the goodness of the clusters obtained, while [9] uses chi square test to validate the clusters obtained by using EM algorithm. The main contribution of this paper is to propose, improved way of comparing user sessions represented as vectors, that are of variable length inherently and employing Jaccard index for analyzing the

effectiveness of the clustering done on two standard set of web server logs. The results obtained by this proposed technique are encouraging.

The rest of the paper is organized as follows. The section 2 talks about the proposed method of clustering in detail. The section 3 discusses about the results, followed by conclusion in section 4.

2 Clustering

2.1 Modified K-Means

The basic K-means algorithm initially selects the cluster centroids randomly and finds the new cluster centroid based on the average value obtained within each cluster, in each iteration. In the modified K-means algorithm, the old cluster centroid is updated by the delta amount, where, delta is nothing but the average distance value of each cluster. i.e., instead of assigning a new point as a centroid, the existing centroid is moved by delta quantity in order to use the k-means for web session clustering, since web sessions are vectors and not data points.

The modified algorithm is as shown in Algorithm Modified K-means. The numerical example is presented as given in Table 1 to compare the basic and modified K-means algorithm for the data set $D=\{11,22,18,15,25,36,27,8,39,10\}$. The results reveal that, the modified K-means algorithm is better than the basic K-means algorithm in terms of number of iterations taken to converge and the quality of clusters formed irrespective of the initial centroids selected. Thus the empirical study shows that modified version of k-means is better than the basic K-means.

Table 1. Comparison of basic and modified K-means

No.	Initial centroids	Basic K-means Clusters	iterations	Modified K-means clusters	iterations
1	m1=8 m2=18 m3=36	c1=11,15,8,10 c2=22,18,25,27 c3=36,39	5	c1=11,18,15,8,10 c2=22,25,27 c3=36,39	3
2	m1=11 m2=22 m3=18	c1=11,15,8,10 c2=25,36,27,39 c3=22,18	4	c1=11,15,8,10 c2=36,39 c3=22,18,25,27	5
3	m1=27 m2=8 m3=10	c1=22,18,25,36,27,39 c2=8 c3=11,15,10	20	c1=36,39 c2=11,15,8,10 c3=22,18,25,27	6

Algorithm: Modified K-means

Input: a set of data $D= \{d_1, d_2, \dots, d_n\}$, the desired number of k clusters

Output: a set of clusters $C= \{c_1, c_2, \dots, c_k\}$ of D

Method:

Select any k data points $\{d_1, d_2, \dots, d_k\}$ from D and set $m_i=d_i$ to get $M= \{m_1, m_2, \dots, m_k\}$ where, $0 < i < k+1$
 newC=empty, newM=empty

Repeat

for each s_i , compute $D=\{d_1,d_2,\dots,d_k\}$ where, $d_i=|d_i-m_j|$, $0 < j < k+1$, $0 < i < n+1$

assign d_i to c_j where $d_j=\min(D)$, $0 < j < k+1$

for each c_j , $\text{delta}_j=\text{sum}(\text{distances of each } d_i \text{ in } c_j) / \text{number of sessions in } c_j$.

$\text{newM}=\{m_1+\text{delta}_1, m_2+\text{delta}_2,\dots,m_k+\text{delta}_k\}$

if ($C == \text{newC}$) or ($M == \text{newM}$) break;

copy C into newC , M to newM

until false

2.2 Modified K-Means for Web Session Clustering

In general, the web user sessions are not simple data points, but n -dimensional vectors. Suppose, a user visits pages p_1, p_2, p_7 of a web site in a sequence, then, the session is represented as a vector $s=\{P_1 P_2 P_7\}$. Before clustering web user sessions, the algorithm, Modified K-means is changed to suit the requirements as given in Algorithm Modified K-means for Web Session Clustering. To find the dissimilarity between any two sessions s_i and s_j , we propose an efficient function to compute variable length vector distance (VLVD) between any two sessions s_i and s_j as given in function VLVD.

Algorithm: Modified K-Means for Web Session Clustering

Input: a set of web user sessions $WS = \{s_1, s_2, \dots, s_n\}$, the desired number of k clusters

Output: a set of clusters $C = \{c_1, c_2, \dots, c_k\}$ of WS

Method:

Select any k sessions $\{s_1, s_2, \dots, s_k\}$ from WS and set $m_i=s_i$ to get $M = \{m_1, m_2, \dots, m_k\}$ where, $0 < i < k+1$

$\text{newC}=\text{empty}$, $\text{newM}=\text{empty}$

Repeat

for each s_i , compute $D=\{d_1,d_2,\dots,d_k\}$ where, $d_i= \text{VLVD}(s_i,m_j)$ and $0 < j < k+1$,

$0 < i < n+1$

assign s_i to c_j where $d_j:=\min(D)$ where $0 < j < k+1$

for each c_j , $\text{delta}_j:=\text{sum}(\text{distances of each } s_i \text{ in } c_j) / \text{number of sessions in } c_j$.

$\text{newM}=\{m_1+\text{delta}_1, m_2+\text{delta}_2,\dots,m_k+\text{delta}_k\}$

if ($C == \text{newC}$) or ($M == \text{newM}$) break;

copy C into newC , M to newM

until false

Function: VLVD

Input: two web user sessions s_i and s_j

Output: distance d between s_i and s_j

Method:

Set $l_1 = |s_i|$ where, $|s_i|$ is the length of the session s_i

Set $l_2 = |s_j|$ where, $|s_j|$ is the length of the session s_j

Set $C = s_i \cap s_j$

Set $\text{dist} = l_1 + l_2 - 2C$

Set $\text{len} = l_1 + l_2$

$d = \text{dist}/\text{len}$

return d

The majority of the algorithms discussed by the researchers represent each of the web session as a binary vector of length n , where n is the number of pages in a web site. Since, the issue of variable length of web user session vectors is not addressed efficiently by majority of the researchers; the function VLVD (s_i, s_j) tries to deal with the variable length session vectors to find the distance or dissimilarity between any two sessions. The VLVD function computes the number of pages that are different between any two sessions, similar to the hamming distance. To get the hamming distance, the two vectors that are taken into consideration should be of same length, but, the VLVD function overcomes this drawback.

The value of d lies in the range of 0 and 1. The value 1 indicates that the two sessions are completely different, where as 0 indicates that the sessions are completely similar. Consider an example data set with 5 sessions, to illustrate the VLVD function.

Example:

S1: P1 P2 P3 P4 P5

S2: P4 P5

S3: P1 P2 P5

S4: P6 P7

S5: P1 P2 P3 P4 P5

VLVD (S1, S2) = 0.42

VLVD (S1, S3) = 0.25

VLVD (S1, S4) = 1.0

VLVD (S1, S5) = 0.0

The example clearly shows that, the sessions S1 and S5 are similar whereas, S1 and S4 are entirely different. S3 is closer to S1 compare to S2. Thus it is possible to measure the distance between the sessions efficiently, though they are not equal length vectors.

3 Results and Discussions

To implement the modified k-means with VLVD function, two data sets are considered: The first set is NASA log taken from NASA Kennedy space center www server in Florida (<http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html>) which consists of approximately 10, 00,000 + entries. The log has the data collected from 00:00:00 July 1, 1995 through 23:59:59 July 31, 1995, a total of 31 days. The data is preprocessed and based on domain knowledge obtained after constructing distinct user requests, 30 categories of pages are formulated.

The second set is MSNBC data set taken from msnbc.com (<http://kdd.ics.uci.edu/databases/msnbc/msnbc.html>) that gives the page visits of users who visited msnbc.com on September 28, 1999. Visits are recorded at the level of URL category and are recorded in time order and therefore, preprocessing was not required for this data set. Table 2 summarizes the details of these two sets and description of page categories for these two data sets are given in Table 3 and 4 respectively.

Table 2. Dataset

Data Set	Time period	File size	Number of sessions considered	Number of page categories
NASA	1/7/1995 to 31/7/1995	117,532 KB	5000	30
MSNBC	28/9/1999	12,287 KB	5000	17

Table 3. Web page categories – NASA data set

P1	<i>/elv/</i>	P11	<i>/icon/</i>	P21	<i>/shuttle/countdown/</i>
P2	<i>/facilities/</i>	P12	<i>/images/</i>	P22	<i>/shuttle/movies/</i>
P3	<i>/shuttle/mission/</i>	P13	<i>/logistics/</i>	P23	<i>/software/</i>
P4	<i>/downs/</i>	P14	<i>/mdss/</i>	P24	<i>/statistics/</i>
P5	<i>/base-ops/</i>	P15	<i>/msfc/</i>	P25	<i>/history/apollo/</i>
P6	<i>/bio-med/</i>	P16	<i>/news/</i>	P26	<i>/history/gemini/</i>
P7	<i>/facts/</i>	P17	<i>/pao/</i>	P27	<i>/history/mercury/</i>
P8	<i>/finance/</i>	P18	<i>/payloads/</i>	P28	<i>/shuttle/</i>
P9	<i>/history/</i>	P19	<i>/persons/</i>	P29	<i>/shuttle/resources/</i>
P10	<i>/htbin/</i>	P20	<i>/procurement/</i>	P30	<i>/shuttle/technology/</i>

Table 4. Web page categories – MSNBC data set

P1	Front page	P7	Misc	P13	Msn-sports
P2	News	P8	Weather	P14	Sports
P3	Tech	P9	Msn-news	P15	Summary
P4	local	P10	Health	P16	Bbs
P5	Opinion	P11	Living	P17	Travel
P6	On-air	P12	Business		

3.1 Analysis of Clusters – NASA Data Set

Fig. 1 shows the frequency of access to various page categories in various clusters of NASA data. “/history/apollo/” and “/shuttle/missions/” categories are viewed more

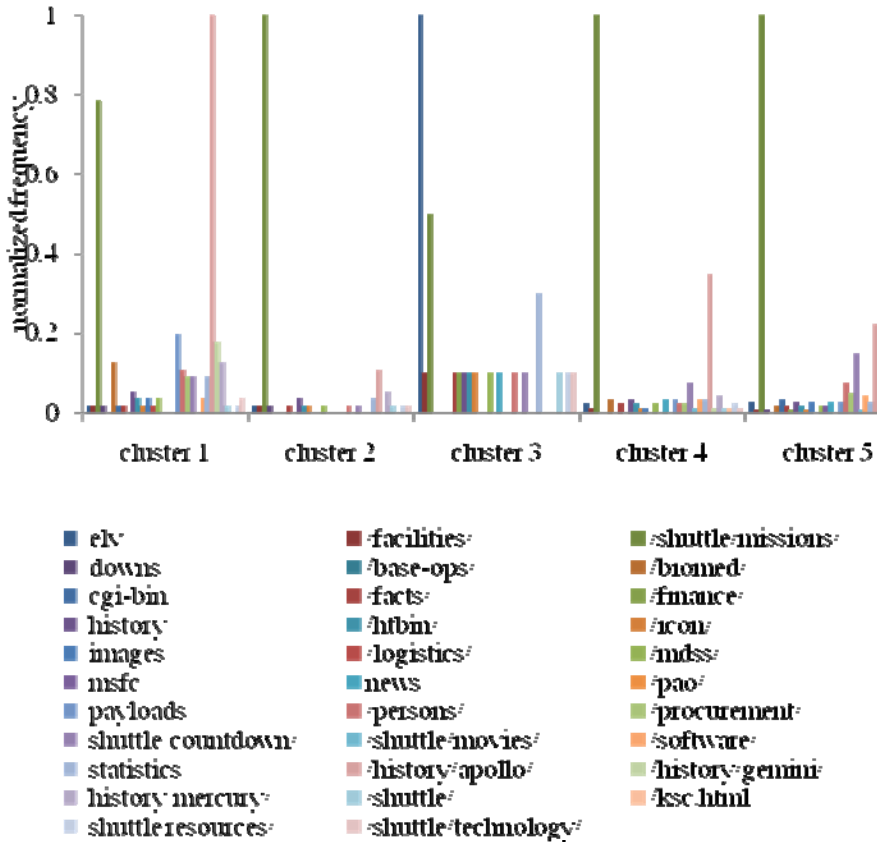


Fig. 1. Normalized frequency of web page categories (NASA dataset)

frequently in cluster 1 compare to other categories, while cluster 2 concentrates on “/shuttle/missions/” category most of the times. In cluster 3 the category “/elv/” is viewed majority of the times and 50% of frequency is to “/shuttle/missions/” category. The users in cluster 4 are more interested in “/shuttle/missions/” and “/history/apollo” categories. Similar to cluster 4, the frequency is more for categories “/shuttle/missions/” and “/history/apollo” in cluster 5, along with the category “/shuttle/countdown/”, where as cluster 4 users are not interested in “/shuttle/countdown/” because the frequency is zero for this category in cluster 4. It may look like the categories of cluster 1 and 4 are similar, but, the usage patterns of these two clusters are different. i.e., in cluster 1, “/history/apollo” is viewed more than “/shuttle/missions/” where as it is vice versa in cluster 4. In cluster 4 around 40% of frequency is to “/history/apollo/”. Overall, it is observed that, the most frequently visited category is “/shuttle/missions/” in this web site. Thus, the clusters formed show different patterns of usage in combination with “/shuttle/missions/” category.

3.2 Analysis of Clusters – MSNBC Data Set

Fig. 2 shows the frequency of access to various page categories in various clusters. More than 60% of times, request is to “misc”, “on-air” while 40% of times, for “weather” and “sports” categories in cluster 1. It shows that, users of this cluster show more interest in these categories. In cluster 2, the users visit “front page” followed by “news” and “local” categories majority of the times, indicating their interest in local information and news. The users in cluster 3 do not belong to any specific categories. They visit “front page” and just visit other categories, while cluster 4 clearly shows more than 50% of times the visit is to “misc” category. In contrast, users in cluster 5 are more interested in “opinion” and subsequently in “on air” and “summary” categories.

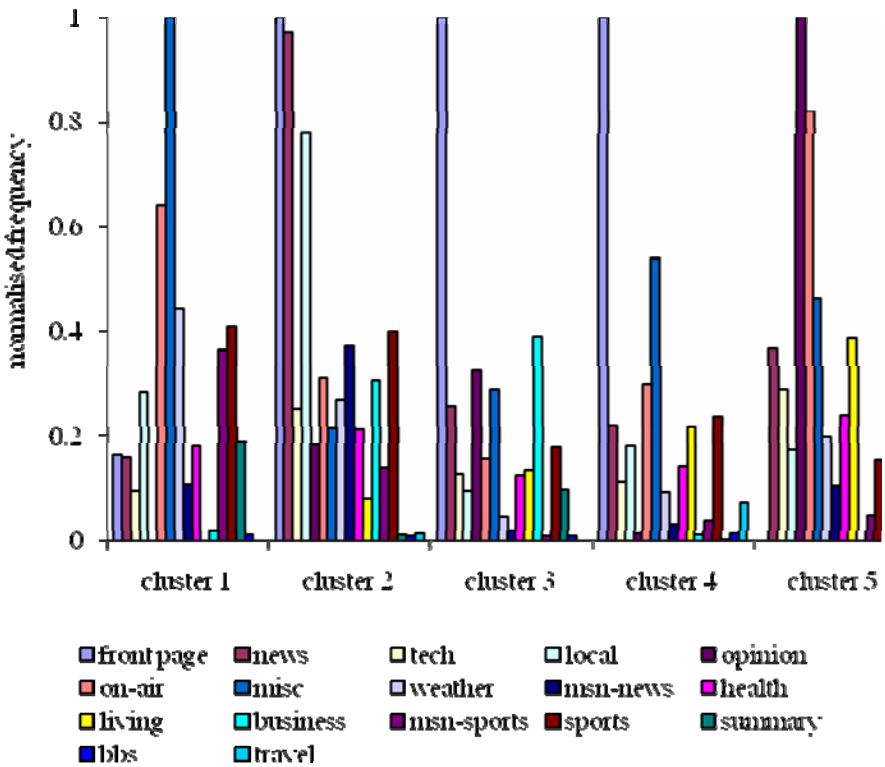


Fig. 2. Normalised frequency of web page categories (MSNBC dataset)

3.3 Analysis of the Clusters Formed by the Proposed Method

The graphs shown in Fig. 1 and 2, clearly indicates the patterns obtained by the proposed method for the two data sets. The Jaccard index, also known as the Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard index between two sample sets A and B is computed as:

$$\text{Jac}(A, B) = |A \cap B| / |A \cup B| \quad (1)$$

If $\text{Jac}(A, B)$ is equal to 1, it indicates that, the samples A and B are exactly similar. In our example, to compare the five clusters that were formed for the NASA data sets, (1) is used and the average value for each cluster is less than 0.3 as shown in Table 5. This indicates that, the clusters obtained are not exactly the same and hence the distance between the clusters is more across all the clusters. Thus, it could be inferred that the clustering done is reasonably good. Similar analysis could be done on the clusters of MSNBC data set provided we get the data regarding the actual pages of the site in each category along with the main page categories. Due to the unavailability of details regarding the pages, the Jaccard index is not applied to the clusters obtained for the MSNBC data set. However, the analysis done on the NASA data set proves the goodness of the proposed clustering method.

Table 5. Jaccard index for NASA data set

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
Average Jaccard index	0.25	0.23	0.12	0.28	0.28

4 Conclusion

With the explosive growth of the web-based applications, there is significant interest in analyzing the web usage data for the task of understanding the users' web page navigation and apply the outcome knowledge to better serve the needs of user. This paper presents a modified k-means algorithm and also the VLVD function to compute the distance between user sessions that takes care of the issue of the uneven lengths of sessions.

As a future work, it is planned to test the impact of this method to more number of user sessions and more number of clusters. Also, the clusters obtained by this proposed method, could be used to develop a recommender system as well as to design a web page prediction system that helps in reducing web page latency for the user. This would also help the web site administrator to reorganize the web site accordingly.

References

1. Facca, F.M., Lanzi, P.L.: Mining interesting knowledge from web logs: a survey. *Journal of Data and Knowledge Engineering* 53, 225–241 (2005)
2. Xie, Y., Phoha, V.V.: Web User clustering from Access Log Using Belief function. In: *Proceedings of the First International Conference On Knowledge Capture (K-CAP 2001)*, pp. 202–208. ACM Press, New York (2001)

3. Li, C.: Algorithm of Web Session Clustering Based on Increase of Similarities. In: Proceedings of International Conference on Information Management, Innovation Management and Industrial Engineering, pp. 316–319. IEEE, Los Alamitos (2008)
4. Krol, D., Scigajlo, M., Trawinski, B.: Investigation of Internet System User Behavior Using Cluster Analysis. In: Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, pp. 3408–3412. IEEE, Los Alamitos (2008)
5. Fu, Y., Sandhu, K., Shih, M.-Y.: Clustering of Web Users Based on Access Patterns. In: KDD workshop on Web Mining, San Diego, CA (1999)
6. Raghavendra, P.S., Chowdhury, S.R., Kameswari, S.V.: Comparative Study of Neural Networks and K-Means Classification in Web Usage Mining. In: Proceedings of 5th IEEE International Conference for Internet Technology and Secured Transaction (ICITST). IEEE, Los Alamitos (2010)
7. Xu, J.-H., Liu, H.: Web User Clustering Analysis based on KMeans Algorithm. In: Proceedings of 2010 International conference on Information, Networking and Automation (ICINA), pp. V26–V29. IEEE, Los Alamitos (2010)
8. Srivastava, J., Cooley, R., Deshpande, M.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In: ACM SIGKDD, vol. 1, pp. 12–23 (2000)
9. Pallis, G., Angelis, L., Vakali, A.: Validation and interpretation of Web users' sessions clusters. *Journal of Information Processing & Management* 43, 1348–1367 (2007)

FOL-Mine – A More Efficient Method for Mining Web Access Pattern

A. Rajimol¹ and G. Raju²

¹ School of computer Science, Mahatma Gandhi University, Kottayam, Kerala, India
rajikovoor@yahoo.com

² Department of IT, Kannur University, Kannur, Kerala, India
kurupgraju@rediffmail.com

Abstract. In this paper, we propose an efficient sequential access pattern mining algorithm, FOL-mine. The FOL-mine algorithm is based on the projected data base of each frequent event and eliminates the need for construction of pattern tree. First Occurrence List, the Basic data structure used in the algorithm, manages suffix building very efficiently. There is no need to rebuild the projection databases. Experimental analysis of the algorithms reveals significant performance gain over other access pattern mining algorithms.

Keywords: Web Access Pattern Mining, Pre-order Linked Web Access Pattern Mining, First-Occurrence Linked Pattern Tree mining, First-Occurrence Forest Mining, Conditional Sequence Mining, FOL-Mine.

1 Introduction

Web mining is the extraction of interesting and useful knowledge and implicit information from activity related to the WWW [2]. Web usage mining, also known as Web log mining, discover interesting and frequent user access patterns from the web browsing details stored in server web logs, proxy server logs or browser logs [1]. Web log mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, network traffic flow analysis and so on [2].

Sequential pattern mining, an important data mining tool used for pattern retrieval, can be used to capture frequent navigational paths among user trails. Web Access Pattern is a sequential pattern in a large set of pieces of web logs. Sequential pattern mining that discover frequent pattern in a Web Access Sequence Data Base (WASD) was first introduced by Agarwal and Srikant in [3]. A web log is a sequence of pairs: user-id and access information. For the purpose of study of sequential pattern mining, preprocessing [6] is applied to the original log file and WASD is generated.

There are mainly two heuristics in sequential pattern mining - Apriori based and Pattern Growth based methods. Apriori algorithms face the problem of tedious support counting and generation of explosive number of candidates when mining

large sequence databases having numerous and/or long patterns. Pattern Growth methods grow frequent patterns by mining increasingly smaller projection databases, and thus, are faster than apriori-based algorithms [14].

In this paper, we propose an efficient, pattern growth, web access pattern mining algorithm, FOL-Mine (First Occurrence List Mine). The proposed FOL-Mine algorithm uses a highly efficient linked structure and algorithms, to hold the web access sequences and then to mine the access patterns. Concept of first occurrence of symbols in FOF-Mine [12] is employed in our method for improving performance. Occurrence of an event in a sequence is a first-Occurrence if it appears for the first time in the sequence. Efficiency of the process of finding out the first occurrences of a symbol is crucial in improving the performance of access pattern mining. In the proposed FOL-Mine, a linked list with header node, stores the first occurrences of each event in a very compact and efficient way.

All WAP-Tree based methods involve two data base scan, first to find out the frequent items and second to create tree using frequent sub-sequences. In our method only one data base scan is required. Earlier WAP-Tree based methods either use complicated linkages of tree nodes or construct intermediate trees for projection data bases for finding out the first occurrences. In the proposed method this is managed very efficiently by a simple linked structure.

The rest of the paper is organized as follows. Section 2 provides a study of related works, section 3 describes our new algorithm and section 4 presents the evaluation of performance of the algorithm. Conclusion is given in section 5.

2 Related Work

Srikant and Agarwal introduced three algorithms, namely, Apriori, Apriori All, and Apriori Some, using the idea of association rule mining presented in [3]. GSP Mine (Generalized Sequential Pattern Mining), an algorithm to mine sequential patterns was introduced by the same authors in [5].

Pei. et al. proposed Web Access Pattern Tree(WAP Tree) in [6], which is a compressed data structure to store web access sequences. This data structure facilitates the development of an efficient pattern mining algorithm, WAP-Mine. This algorithm avoids the problem of generating exponential candidate patterns which is the major drawback of all apriori methods. WAP-Mine is based on suffix heuristic i.e. it generates access patterns by building suffix sequences [6].

Pre-order Linked WAP Tree (PLWAP Tree) algorithm, proposed by Lu and Ezeife, in [7] eliminates the intermediate reconstruction of WAP-Trees during mining. In PLWAP-Mine, the nodes of the same events are linked in pre-order fashion. Pre-order linkage is used to identify the sub-trees under consideration during mining. Binary position codes attached to each node identify the ancestor/descendants relationships between nodes of the tree. This algorithm is based on prefix building. PLWAP Tree method outperforms both GSP and WAP-Mine when the number of frequent patterns increases and the minimum support threshold is low [7]. Performance of PLWAP degrades when the length of sequence is more than 20 because of the increase in the size of position codes as the depth of the tree increases. Though the PLWAP-tree

algorithm saves memory, it has to go through all its occurrences, including those that are not part of the projection database, to find the first-occurrences of a symbol [13].

CS-Mine (Conditional Sequence mine) algorithm presented in [8] also is based on WAP-tree, but it uses WAP-tree only for generating initial conditional sequence base. The conditional sequence base of an event ei based on prefix sequence *Sprefix*, is the set of all long suffix sequences of ei in sequences of a certain dataset. Mining is done directly on the conditional sequence base of each frequent event and eliminated the need for costly reconstruction of intermediate conditional WAP trees. This algorithm outperforms WAP mine significantly, especially when the support threshold become smaller and the size of the web access sequence gets larger [8]. But the major drawback of the method is the recursive generation of the sub conditional sequence base and the test to see whether the single sequence forms part of the final sequential access patterns which induce unnecessary delay in the mining process.

B. Y. Zhou et al., the authors of CS-Mine [8] propose CSB-mine algorithm in [9] as an improvement of CS-mine by eliminating the building of WAP Tree at all. Initially in the Preprocessing step, the initial conditional sequence base is constructed from the Web Access Sequence Database (WASD). Then conditional frequent events, events that satisfy the support threshold are identified. Frequent events are used to construct the Header Table. For each conditional frequent event ei , a linked-list structure ei -queue with the first item labeled ei in sequences of *CSB* (Sc) is created. The head pointer of each event queue is recorded in the Header Table. And finally all non-frequent events are deleted. All other procedures remain the same as that of CS-Mine without any improvement.

First Occurrence Linked WAP Tree Mining (FLWAP-Mine) introduced in [10] uses FLWAP-Tree, a WAP-Tree in which only the first occurrences of each symbol are linked instead of linking all occurrences as in previous methods. It uses first occurrence of events to build projection trees instead of following the link of whole WAP-Tree as in PLWAP. FLWAP mining out performs the PLWAP mining consistently and significantly [12]. The major drawback of FLWAP-tree algorithm is the rebuilding of every projection database, and thus using a lot of memory. Also, the linking of tree nodes increases the complexity of tree construction.

In [11], P. Tang et al. use the frame work given in [10] and propose a modified approach for finding out the first occurrences of items and thus to improve pattern building. They used forest of first-occurrence sub-trees as the basic data structure for representing projection database. FOF-Mine uses an extended aggregate tree with the root node representing the empty symbol ϵ . The count of the root node is the total number of sequences in the database. A recursive mining function, FOF-Mine, is used to mine patterns from the aggregate tree. Given a symbol a , each sub tree rooted at a first-occurrence of it, is the first-occurrence sub tree of a . A list of pointers to the first-occurrences of a in the aggregate tree is the forest of first-occurrence (FOF) sub trees of a symbol. The sum of the counts of the root nodes gives the support of the symbol a in the projection database D_a . The sub-trees rooted at the children of the nodes in FOF represent the projection database D_a . As all the nodes are already present in the tree there is no need to reconstruct intermediate trees, which is time and space consuming.

In FOF-Mine, for each frequent event a , the projection database D_a of the current database with respect to the symbol a , is identified. If the support of the symbol a ,

satisfies the required support threshold then a is appended to the previous pattern and the new pattern is added to the set of pattern. Now, the new database is recursively mined to generate all pattern starting with the current event. FOF algorithm outperforms both PLWAP and FLWAP [11].

3 Mining Access Pattern with FOL-Mine

3.1 Motivations

The compact WAP-Tree structure and the corresponding WAP-Mine algorithm [6] performed better than all other earlier apriori methods. Many modifications of the WAP-Tree were introduced [7, 10, 11] to improve the efficiency of mining. All these WAP-Tree based algorithms require two database scans; one for finding out the frequent items and the other for creating the aggregate tree using frequent subsequences. Moreover, in spite of its compactness, the size of node and complexity of building up the WAP-Tree are disadvantageous [9].

CSB-Mine algorithm [9] is a web access pattern mining method that does not use WAP-tree. It uses a database structure called conditional sequence base to store the frequent access sequences. Using this database, intermediate projection databases are generated. The performance analysis shows that CSB-Mine outperforms WAP-mine algorithm especially when support becomes smaller. Scalability also is more for CSB-Mine algorithm when the input database size becomes larger [9]. So, in FOL-Mine we have adopted the idea of not using WAP-Tree.

CSB-Mine has got certain drawbacks. Sub conditional sequence base are generated in each recursive call of CSB-Mine. Also, to find whether a symbol is to be appended as a part of pattern, it uses a procedure that test whether all sequences in the sub conditional sequence base can be combined into a single sequence. All these processes are space and time consuming.

All WAP-Tree based mining algorithms except WAP-Mine employ prefix building to generate patterns recursively. All these methods use various techniques to locate the First Occurrences of each symbol so as to build patterns. Efficiency of the process of finding out the first occurrences of each symbol becomes crucial in the mining process. In FOF-Mine [11] the authors give the idea of collecting the first occurrences together through the concept of Forest of First Occurrences. This concept is used in the proposed algorithm to improve the efficiency. This avoids the reconstruction of projection databases and thus saves space and time considerably.

In the proposed FOL-Mine, a linked list with header node, stores the first occurrences of each event in a very compact and efficient way. Header node stores the count of first occurrences which gives the total support. The proposed algorithm avoids the unnecessary generations of intermediate projected databases and tedious support counting.

FOL-Mine does not use any of the functions used either in CSB-Mine [9] or in FOF-Mine [11].

3.2 Theoretical Back Ground

Let E be the set of symbols, each symbol representing a web page. A non-empty web access sequence S is a finite sequence of symbols from E , $S = s_1s_2\dots s_m$ such that $s_i \in E$ for all $1 \leq i \leq m$ and s_i and s_j are not necessarily different for $i \neq j$. The length of the sequence $S = s_1s_2\dots s_m$ is m [6]. The empty sequence ε is a special web access sequence of length 0 and $\varepsilon.s = s.\varepsilon = s$ for any sequence s where ‘.’ is the concatenation operator [11]. A web access database D is a multi-set of web access sequences including the possible empty sequence. A web access sequence $S' = s'_1s'_2\dots s'_n$ is a subsequence of sequence $S = s_1s_2\dots s_m$, if and only if $n \leq m$ and there exist i_1, i_2, \dots, i_n such that $1 < i_1 < i_2 < \dots < i_n \leq m$ and $s'_j = s_{i_j}$ for all $1 \leq j \leq n$. The empty sequence ε is a subsequence of any sequence. A web access sequences S in D is said to support pattern p if p is a subsequence of S . The support of pattern p in D , denoted as $\text{SupD}(p)$, is the number of web access sequences in D that support p . Given a threshold η in interval $[0,1]$, a pattern p is frequent with respect to ξ and D if $\text{SupD}(p) \geq \xi * |D|$, where $|D|$ is the number of web sequences in D . $\xi |D|$ is called the absolute threshold and denoted as η . The web access pattern mining problem is to find all frequent web access patterns in D with respect to ξ [6].

Given a web access sequence S and a symbol a from E such that a in s , the a -prefix of s is the prefix of s from the first symbol (the leftmost symbol) to the first occurrence of a inclusive. The a -projection of s is what is left after the a -prefix is deleted. If a occurs only once as the last symbol in S , the a -prefix is S and the a -projection is the empty sequence ε [10].

Given the database D and a symbol a in E , the a -projection database of D , denoted as Da , is the multi-set of a -projections of the web access sequences in D that support a . The same sequence may repeat in a projection database. Given a symbol a from E and database D , the support of pattern p (including empty pattern) in the a -projection database Da of D is equal to the support of pattern $a.p$ in the original database D .

Given a symbol a from E and database D , the support of a in D is same as the number of sequences in a -projection database. The set of frequent pattern is empty if $|D|$ is less than the absolute threshold. Otherwise it is the union of all sequences that are prefixed by a for each $a \in E$ having support $\geq \eta$.

3.3 The Proposed Algorithm

The structure of the database used for storing the web access sequences is a linked list with each node is of the form

```
struct node {symbol item; next *node; }
```

The start address of each linked list is stored in an array of pointers of the type `ptr *node`. The main algorithm of FOL-Mine is given in Fig. 1.

Input :

1. WAS, the Web Access Database.
2. Support = Support threshold.

Fig. 1. Algorithm: main

Output:

1. F = the set of sequential access patterns

Method:

1. $m=0$; $L= \text{null}$;

2. While eof (WAS)

i. Read a Sequence from disk file and construct linked list registering the start address in Head list[m].

ii. Update the set of items and their support

iii. Increment m .

End while.

3. $\eta = \text{support} * m$; // absolute support

4. Generate the set of frequent event Σ

5. FOL-Mine (ε , L , η)

6. Return.

Fig. 1. (continued)

Step-1 initializes m , the number of access sequences in WAS and the initial FOL-list. Step-2 retrieves elements of a sequence from the disk file and forms the linked list. The start address of the linked list is registered in an array of pointers, Head list. List of elements and their support are updated each time a sequence is read. Step 3 calculates the absolute support, η . Events that have support greater than or equal to the absolute support, η are identified as set of frequent event Σ in step 4. Step 5 of algorithm generates the access patterns by calling the recursive mining functions FOL-Mine given in Fig. 2.

FOL-Mine uses a linked list of type FOL-list to store the first occurrences of an event. Each node of FOL-list is pointer of the first occurrence of an item in a web access sequence. Head node stores the total number of occurrences and start address of the FOL-list. Structure of Header node and FOL-list are as described below.

```
struct htype { int supp; symbol item; folnode next; }
struct folnode { wasdlist *occr; folnode *next };
```

Algorithm FOL-Mine (pattern q , FOL L , η)

1. for each $a \in \Sigma$ do

i. $L_a \leftarrow \text{construct FOL}(L, a)$

ii. if support of $a > \eta$

Fig. 2. Algorithm FOL-Mine

```

    F ← F U { q . a }
    F' ← F U FOL-Mine ( q . a , La, η)
  end if
iii. delete La
end for
3. return

```

Fig. 2.(continued)

First step of the algorithm, builds the First Occurrence List of the symbol a , that stores the first occurrences of symbol a in each sequence using the algorithm construct FOL (L, a), given in Fig.3.

Algorithm: construct-FOL (L, a)

```

1. Count-a=0
2. La = create an header node with label a
3. if L is not empty // FOL-list is made using
   FOL-list in the previous call//
   for each item in L
     find the next occurrence of item a
     If (exists)
       create a node and attach to La
       increment Count-a.
   end for.
else
  for each web access sequence
    find the first occurrence of item a.
    if (exists) create a node and attach to La
    increment Count-a.
  end for.
end if.
4. update header node with Count-a
5. return FOL(La, a)

```

Fig. 3. Algorithm: Construct FOL (L, a)

4 Performance Evaluation

Frequent sequential pattern mining algorithms fall into two categories; apriori based and pattern growth. Many studies reveals that pattern growth methods out performs apriori ones. As FOL-Mine is a pattern growth method, comparison with other pattern growth based algorithms is enough to prove the efficiency.

In CSB-Mine intermediate sub conditional sequence bases are generated in each recursive call. This takes up a significant amount of time and space. But, in FOL-Mine once the sequences are represented in a liked format, rest of the mining is done using it. No intermediate reconstruction is required.

CSB-Mine includes two more procedure to carry out the mining. During each call it has to build an event queue. Algorithm Single Sequence Test is used to verify whether all sequences in sub conditional sequence bases can be combined into a single sequence. If so, the mining of that sub conditional sequence bases will be terminated. This single sequence will be used to form a part of the final sequential access patterns. Otherwise, sub conditional sequence base is constructed and mining is done recursively. This also introduces significant amount of time in the actual CSB-Mine. From the discussion it evident that FOL-Mine will perform better than CSB-Mine.

Now, to prove FOL-Mine to be a better performer, the only we have to do is to compare it with the WAP-Tree methods. In [7] authors proved PLWAP-Mine performs better than the WAP-Mine method in [6]. The authors of [10] claim the better performance of FLWAP-Mine over PLWAP-Mine [7], in their performance evaluation. Experimental results in [11], claim that the execution time and memory requirement is less for FOF-Mine than FLWAP [10]. So, to verify the efficiency of the new FOL-Mine, it is enough to compare it with FOF-Mine.

Time and Memory are the two major concerns in sequential pattern mining. So, these two factors are considered for performance evaluation of the proposed algorithm.

4.1 Execution Time Comparison

All the tests were performed on a 1.79Ghz AMD Sempron(tm) machine with 448 MB RAM and running Microsoft Windows XP Professional version 2002. Both FOF-Mine and FOL-Mine are implemented in Microsoft visual C++ 6.0.

Data sets used for the experiments are T25i10D10K and T10i4D100k. These Synthetic datasets are generated using the publicly available synthetic data generation program of the IBM Quest data mining project at <http://www.almaden.ibm.com/cs/quest/>, which has been used in most sequential pattern mining studies[5,6,10,11,12]. T25i10D10K is a 948 KB data base with 10000 sequences and T10i4D100k is of 3.83 MB with 1 lakh sequences.

The execution time of the two algorithms was thoroughly tested. Codes are added in the original implementation to measure the total execution time. Fig. 4(i) and 4(ii) shows the comparison of execution time in both data sets, T25i10D10K and T10i4D100k, respectively. Experiment was repeated by varying the support threshold. FOL-Mine algorithm is shown to out perform FOF-Mine especially when the support value becomes smaller.

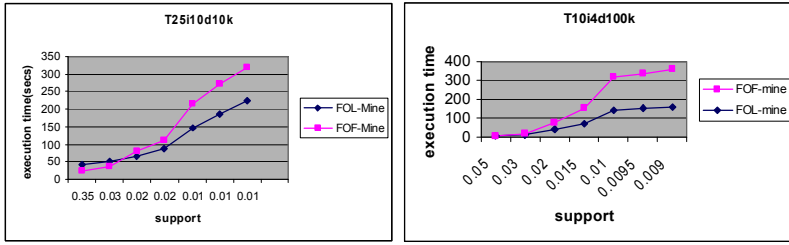


Fig. 4. Execution Time trend for different minimum support for (i) T25i10D10K and (ii) T10i4D100K

4.2 Scale-Up Experiments

These experiments test how the execution time changes when the size of the data base changed while keeping the support threshold same. Experiment is done using the Dataset T10i4D100K. Data size is changed from 20k to 100k. Fig. 10 shows the result of the scale-up experiments. From the result it is very clear that FOL-Mine have good scalability, furthermore, FOF-Mine shows a better scalability than FOF-Mine.

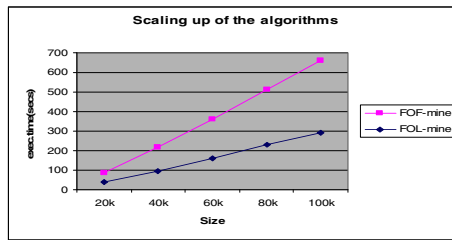


Fig. 5. Execution Time trend for different Data Size. Fixed minimum Support 0.005 and Data Base T10i4D100K

4.3 Memory Considerations

Generally, tree structure is very compact, but the Aggregate tree used in WAP-Tree methods, node structure requires more space. In the linked-database used in FOL-Mine, each node holds only two different fields, item and a pointer. But each node of aggregate tree in FOF-Mine store four different information: item, count, pointer to right-sibling and a pointer to left- child.



Fig. 6. (i) node in FOL-Mine (ii) node of tree in FOF-Mine

So, only when the number of node in the tree is less than half of the number of nodes in the linked-database of FOL-Mine, tree structure has got any advantage. This happens only when the access sequences in WASD are almost similar.

5 Conclusion

We have presented a pattern growth web access pattern mining method FOL-Mine, using FOL-list structure to facilitate efficient mining. We have evaluated the method using synthetic data. The results show significant improvement in the execution time. The results of speed up experiments reveal a linear increase as the increase in the size of data base.

References

1. Kosala, R., Blockeel, H.: Web Mining Research: A Survey. *ACM SIGKDD Explorations* 2, 1–15 (2000)
2. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *ACM SIGKDD Explorations* 1, 12–23 (2000)
3. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: 20th International Conference on Very Large Databases, Santiago, Chile, pp. 487–499 (1994)
4. Agrawal, R., Srikant, R.: Mining Sequential Patterns. In: 11th International Conference on Data Engineering, Taipei, Taiwan, pp. 3–14 (1995)
5. Srikant, R., Agrawal, R.: Mining Sequential Patterns: Generalizations and Performance improvements. In: 5th International Conference on Extending Database Technology (EDBT), Avignon, France, pp. 3–17 (1996)
6. Pei, J., Han, J., Mortazavi-asl, B., Zhu, H.: Mining Access Patterns Efficiently from Web Logs. In: Terano, T., Chen, A.L.P. (eds.) *PAKDD 2000*. LNCS, vol. 1805, pp. 396–407. Springer, Heidelberg (2000)
7. Lu, Y., Ezeife, C.I.: Position Coded Pre-order Linked WAP-Tree for Web Log Sequential Pattern Mining. In: Whang, K.-Y., Jeon, J., Shim, K., Srivastava, J. (eds.) *PAKDD 2003*. LNCS (LNAI), vol. 2637, pp. 337–349. Springer, Heidelberg (2003)
8. Zhou, B., Hui, S.C., Fong, A.: CS-Mine: An Efficient WAP-Tree Mining for Web Access Patterns. In: Yu, J.X., Lin, X., Lu, H., Zhang, Y. (eds.) *APWeb 2004*. LNCS, vol. 3007, pp. 523–532. Springer, Heidelberg (2004)
9. Zhou, B., Hui, S.C., Fong, A.C.M.: Efficient Sequential Access Pattern Mining for Web Recommendations. *Int.J. Knowledge based and Intelligent Engineering Systems* (2006)
10. Tang, P., Turkia, M.P., Gallivan, K.A.: Mining web access patterns with first-occurrence linked WAP-trees. In: 16th International Conference on Software Engineering and Data Engineering (SEDE 2007), Las Vegas, USA, pp. 247–252 (2007)
11. Pearson, E.A., Tang, P.: Mining Frequent Sequential Patterns with First-Occurrence Forests. In: 46th ACM Southeastern Conference (ACMSE), Auburn, Alabama, pp. 34–39 (2008)
12. Lu, Y., Ezeife, C.I.: PLWAP sequential Mining: open source code. In: First International Workshop on Open Source Data Mining: Frequent Patterns Mining Implementation, Chicago, Illinois, pp. 26–35 (2005)
13. Rajimol, A., Raju, G.: Web Acces Pattern Mining – A Survey. In: 2nd International Conference on Data Engineering and Management (ICDEM 2010), LNCS, vol. 6144, Springer, Heidelberg (2010)
14. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*. Morgan Kaufman, San Francisco (2007)

Semantic Association Mining on Spatial Patterns in Medical Images

S. Saritha and G. SanthoshKumar

Department of Computer Science, Cochin University of Science & Technology, Kochi
sarithas.sarithas@gmail.com, sancochin@gmail.com

Abstract. The advancement in the field of image mining owes the credibility to the discovery of significant image patterns from the archive. An important pattern of an image is the spatial displacement of objects in an image, and the term is coined as spatial relationship feature. In this paper, we incorporate the concept of spatial relationship into medical images. The spatial relationship of structures in the medical image is expressed in terms of fuzzy set theory, thus making the scenario closer to the semantics of the images. The fuzzy spatial relationships existing between the structures in the medical images are mined to identify relevant spatial patterns, so that characteristic knowledge generated is highly authentic for the medical domain under consideration. From the spatial patterns generated spatial association rules are deduced that can steer as an aid to medical diagnosis or rather new diagnosis rules.

Keywords: Association rule mining, image features, image mining, semantic mining, spatial pattern, spatial relationship.

1 Introduction

The incredible increase in the amount of image collection has led to the evolvement of image mining as a major field of data mining. Extracting knowledge from the image data is difficult due to the fact that the images have amorphous arrangement of data. Image mining deals with the extraction of implicit knowledge, image data relationship or other patterns not explicitly stored in the image databases. It is an interdisciplinary endeavor that essentially draws upon proficiency in computer vision, image processing, image retrieval, data mining, machine learning, database and artificial intelligence [4]. The glory of image mining lies in the fact that it explores image meanings, as per human insight and detects relevant patterns in images or relations between numerous images in the archive. The various types of image archive consist of satellite images, medical images and photographs. Among these, the medical images are of numerous types like CT images, ECT images, MR images and so on. The clinical significance of each medical image varies with respect to the functional importance of the organs of the human body. MR and CT images of the brain hold important position in medical images, as the brain is the center of the body.

Concepts gained from medical images are very much powerful as they have copious interpretations associated with them. This is inherent on the features of the medical images. Features constitute the visual properties of an image as well as its semantics. The competence of image mining is deeply rooted in the extracted features of the images. The features range from the syntactic features like color, shape and texture in the pixel level, spatial relationship features in the object level and semantic features described using ontology, depending on the domain under consideration.

The different paradigms of image mining are image classification, image clustering and image association mining. Image classification and clustering [5] are the supervised and unsupervised classification of images into appropriate groups respectively. Association mining of images involves extraction of features from images as relationships between the objects/scenes in the images, identifying the interesting/frequent pattern existing in the image set and generating knowledge. Knowledge can be of various forms like characteristic rules, structures, clusters or association rules.

Medical images contain different types of anatomical structures. The structures have its own features, attributes and descriptions. Also there exists some spatial relationship between these structures, which can be explicit as well as implicit. Associations mining of spatial relationship in medical images results in obtaining significant spatial patterns of the anatomical structures in images and also yield spatial association rules which can validate existing rules in the medical knowledge or to provide new knowledge.

Spatial association mining was first introduced in architectural images by Hsu & Lee through Viewpoint Miner [7]. Also a different approach to perform the same task on simple images is given by [2], [3]. An initial breakthrough in medical image mining was achieved in 2002 by Megalooikonomou et al. by generating association rules relating the structure and function of the brain. Recently there has been efficient association rule based methods to support medical image diagnosis of mammograms [8]. Association rule based mining of medical images is used to validate the diagnosis in [6], [9]. However, there has been no attempt to perform a spatial association mining in medical images to the best of our knowledge up till date. Implicit spatial information in medical images is highly critical for interpretation of the image, and association rules mined from medical images will aid us in concluding to an interesting spatial relationship existing between the structures existing in images.

The remaining of the paper is organized as follows. Section 2 introduces the architecture of the proposed system and also describes the different steps involved in the process. Section 3 outlines the implementation details and interprets the significance of the association rules mined. In Section 4 the paper is concluded.

2 System Description

The proposed system here is based on the relationship existing between the objects in an image. This type of feature which describes each object in the scene/image with

respect to each other is termed the spatial relationship feature. Spatial relationship features are devised based on how humans differentiate one spatial situation from another or identify the spatial situations as same. Spatial relationships can be broadly classified as of two kinds, topological and directional. Topological relations are of different types formulated for different needs as seen in [1]. Directional relations are highly used in geographic information systems, remotely sensed imagery, medical imaging for the purpose of improved recognition, visualization and interpretation of the systems.

Medical imagery contains images of human anatomical structures. It is evident that there exists a defined spatial relationship between the anatomical structures for normal situations. Of course, the spatial relationships deviate when there is an anomaly. The medical advisors or the expert systems read the spatial deviation exhibited by anatomical structures for diagnostic purposes. The readability of spatial relationships between certain objects is highly prominent, where as the readability of spatial relationships between some other objects is ambiguous and if missed out will result in erroneous diagnosis. Therefore it is highly necessary to form spatial association rules of the form *antecedent* \rightarrow *consequent*, where (i) *antecedent* contains spatial patterns of anatomical structures and *consequent* contains the anomaly description or diagnosis result or (ii) *antecedent* contains spatial patterns between prominent structures and *consequent* contains spatial patterns between ambiguous structures. This can be achieved by mining the relevant spatial relationship pattern existing in the medical image archive.

In the proposed system, the spatial relationships existing between the anatomical structures in already classified medical images are defined by means of fuzzy set theory [1]. Fuzzy set approach provides a measure for imprecise spatial representation, and also the computational representation associated with it is relatively simple. The fuzzy spatial relationship features of the structures are then mined to identify the significant pattern and thus generate association rules which also exhibit fuzziness. So the association rules extracted after mining relates the spatial displacement of anatomical structures to either the categorized image or to the spatial displacement of another set of anatomical structures.

The medical imagery archive consists of CT scans, MRI scans, PET scans, 3D renderings and so on. The medical images considered here are MRI scans of the brain. The fuzzy spatial relationships between the anatomical structures in the brain images are mined to discover the frequent spatial pattern and thereby generating new spatial associations as mentioned earlier. The overall design of the proposed scheme can be depicted using the block diagram in Figure 1.

The proposed methodology can be implemented in four modules as follows

1. Anatomical Structure Identification
2. Fuzzy Spatial Relationship Detection
3. Frequent Spatial Pattern Discovery
4. Spatial Association Mining

Each module is described in detail below.

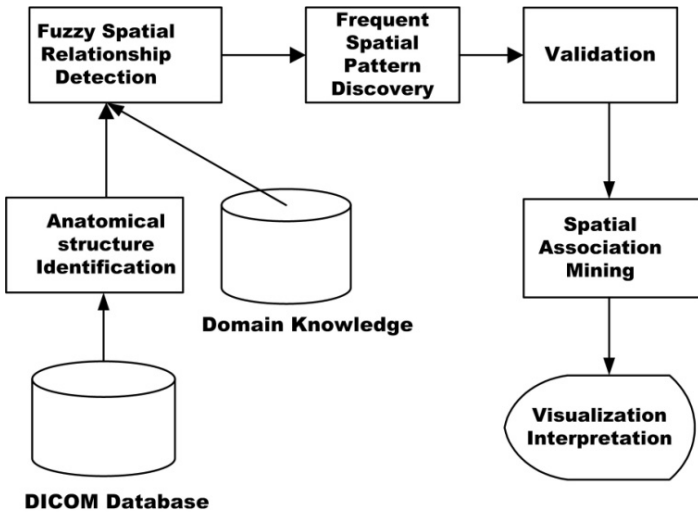


Fig. 1. Design of the proposed scheme

2.1 Anatomical Structure Identification

MRI scan consists of sagittal renderings, coronal and axial slices. The structures identified in each category of slices are different. The axial slices of the MRI scan are considered here for experiments. A sample of the MRI axial scan is shown in Figure 2. A few anatomical structures like Caudate Nucleus (CN), Thalamus (TH), Lateral Ventricle (LV) and Putamen (PU) are marked in the figure. The right hemisphere of the brain is normal, whereas the left hemisphere shows an anomaly. It is evident from the figure that the spatial relationship existing between the anatomical structures in normal and pathological situations is different. Spatial patterns that deviate from the normal situation point towards an anomaly.

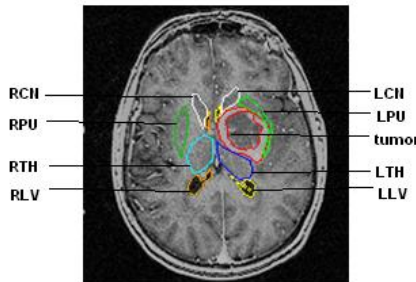


Fig. 2. Axial slice of a MRI brain image

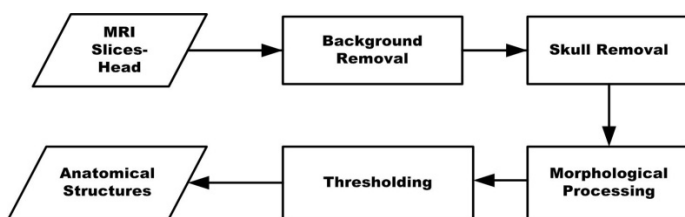


Fig. 3. Anatomical Structure Identification Process

The MRI slice has to be segmented to identify the anatomical structures. The step involved in segmenting the images is as in Figure 3.

We can identify different anatomical structures in the MRI images. Some of the anatomical structures identified after the segmentation process is RPU, RLV, RTH, RCN, LPU, LLV, LTH, LCN etc. The following Table 1 summarizes the MRI images after anatomical structure identification. Col I are the MRI image ID's. Col II are the structure ID's. Col III describes the position of each structure identified in terms of spatial displacement and Col IV is the description regarding the image under consideration.

Table 1. MRI image expression with anatomical structures identified

Col I	Col II	Col III		Col IV
<i>Image Id</i>	<i>Structure Id</i>	<i>Disp.(X)</i>	<i>Disp.(Y)</i>	<i>Image Attribute</i>
I1	RTH	X_{RTH_1}	Y_{RTH_1}	Normal(N)
I1	RCN	X_{RCN_1}	Y_{RCN_1}	Normal(N)
I1	RPU	X_{RPU_1}	Y_{RPU_1}	Normal(N)
....
I2	LLV	X_{LLV_2}	Y_{LLV_2}	DiseaseI(D1)
I2	LPU	X_{LPU_2}	Y_{LPU_2}	DiseaseI(D1)
.....
I3	RTH	X_{RTH_3}	Y_{RTH_3}	DiseaseII(D2)

2.2 Fuzzy Spatial Relationship Detection

Once the anatomical structures are identified, the spatial relationship between the structures is detected. The spatial relationships existing between the structures in the normal and abnormal situations are different as evident from Figure 2. Also it is obvious that the spatial relationships existing between the structures vary with respect to different abnormalities of the brain. The spatial relationships between structures are here modeled in terms of fuzzy set theory. Fuzzy spatial relationships are computed in four primitive directions, left, right, above or below, and the corresponding mathematical representation is given by [1]. There is a degree of membership

associated in the four directions for every structure in the image. There are a number of fuzzy model evaluation methods in existence, which is used extensively depending upon the application under consideration. Here we are considering the centroid method as it has linear complexity.

If ‘k’ anatomical structures are identified from a medical image, we are considering only k (k-1)/2 spatial patterns existing between them. It is to be noted that the rest of the spatial relationships are redundant from these.

The following Table 2 gives the image expression in terms of fuzzy spatial relationship existing between samples of four structures identified in the medical image. For example image I1 has four structures identified as LCN, LLV, LPU and LTH. The fuzzy spatial relationship existing between these structures are given in Col III of the table. F_{12} gives the spatial relationship existing between structures 1 and 2 with 1 as reference. Table 3 details an example of Col III for the first image in terms of fuzzy set theory. Col IV gives the description of the image. Table 3 gives the membership degree of each structure with respect to another in terms of the primitive directions.

Table 2. Fuzzy spatial relationship in image data for four structures

<i>Col I</i>	<i>Col II</i>				<i>Col III</i>						<i>Col IV</i>
<i>Img Id</i>	<i>Str1</i>	<i>Str2</i>	<i>Str3</i>	<i>Str4</i>	F_{12}	F_{13}	F_{14}	F_{23}	F_{24}	F_{34}	<i>Img Attr.</i>
I1	LCN	LLV	LPU	LTH	μ_{12}	μ_{13}	μ_{14}	μ_{23}	μ_{24}	μ_{34}	N
I2	LCN	LLV	LPU	LTH	μ_{12}	μ_{13}	μ_{14}	μ_{23}	μ_{24}	μ_{34}	D1
I3	RCN	RLV	RPU	RTH	μ_{12}	μ_{13}	μ_{14}	μ_{23}	μ_{24}	μ_{34}	D2
...

Table 3. Membership degree in four primitive directions above, below, left and right

<i>Col I</i>	<i>Col II</i>	<i>Col III</i>			
<i>Image Id</i>	<i>Fuzzy Spatial Pointer</i>	μ_{right}	μ_{left}	μ_{above}	μ_{below}
I1	μ_{12}	0.01	0	0	0.99
I1	μ_{13}	0.27	0	0	0.73
I1	μ_{14}	0	0.02	0	0.98

2.3 Frequent Spatial Pattern Discovery

A spatial pattern consists of two or more anatomical structures and the spatial relationships among the structures. As an example, a spatial pattern for four structures can be of the form $(S_1, S_2, S_3, S_4, S_{r12}, S_{r13}, S_{r14}, S_{r23}, S_{r24}, S_{r34})$, where S_i indicates an anatomical structure and S_{rij} denotes spatial relationship between the structures S_i and S_j . The length of a spatial pattern is the number of spatial relationships existing between structures in the pattern. Even though it appears simple to discover frequent patterns using the traditional Apriori algorithm from the image expression of Table 2, there exist some significant challenges as outlined below.

a. *Quantitative Analysis of the data* - Two k -patterns is joinable if the first $(k-1)$ items and the corresponding spatial relations between them are identical in both k -patterns. For fuzzy spatial relationships, the patterns are joinable if the weighted fuzzy membership values are above a threshold value, in any of the four directions, as per the guidance of domain knowledge.

b. *Medically meaningful patterns* - Brain is composed of left and right hemispheres, and meaningful relations are easily recognizable inside the hemispheres separately, rather than together. So a group constraint function (group $(S_i) =L$ or group $(S_i) =R$) is assigned to each structure, with the aid of priori knowledge.

c. *Frequent Pattern Length* - The number of structures and corresponding spatial relations in a frequent pattern should be restricted in size, so that the patterns are easy to interpret and thus help in generation of simple association rules in the next phase. Let S_{max} be the maximum number of structures appearing in a pattern.

The concepts of the Apriori algorithm is borrowed and modified as follows by taking into consideration the above challenges.

Algorithm 1: To find frequent spatial patterns in medical images

Input: (i) Image expression data in Tabular format (Refer Table 2 and 3) (ii)

Support count min_sup

Output: Frequent spatial pattern $-L$

1. $C_2 \leftarrow \{ \text{candidate 2 length pattern set} \}$
2. $L_2 \leftarrow \{ \text{frequent 2 length pattern set} \}$ from min_sup
3. $k=2$; // number of structures present in the pattern
4. while $(k < S_{max}) \{$
5. $L_{k+1} = \Phi$
6. For each pattern $\alpha = (P_1, P_2, P_3, \dots, P_k, P_{r12}, P_{r13}, \dots, P_{rk(k-1)/2})$ in L_k
7. For each pattern $\beta = (Q_1, Q_2, Q_3, \dots, Q_k, Q_{r12}, Q_{r13}, \dots, Q_{rk(k-1)/2})$ in L_k
8. Joining α to β , to form C_{k+1}
 - 8.1 The weighted threshold spatial relation in α is in alignment with that in β in the desired direction (as per domain knowledge)
 - 8.2 $group\{(P_1, P_2, P_3, \dots, P_k)\} = group\{(Q_1, Q_2, Q_3, \dots, Q_k)\}$
9. Check the support for each candidate, say c , in C_{k+1} to be above min_sup to be added to form L_{k+1}
10. $k=k+1$;}
11. $L = \cup L_i$

2.4 Spatial Association Mining

A spatial association rule in medical images is a rule that associates spatial relations among anatomical structures to themselves or to the image characteristics. Some examples of such rules are (i) (P_1, P_2, P_{r12}) and $(P_1, P_3, P_{r13}) \rightarrow$ Disease 1 (ii) $(P_1, P_2, P_3, P_{r12}, P_{r13}, P_{r23}) \rightarrow$ Disease 2 and (iii) $(P_1, P_2, P_{r12}) \rightarrow (P_3, P_4, P_{r34})$ and Normal, with necessary confidence. To strengthen association rule generation phase, in addition to the confidence, we are introducing a constraint function, such that $\text{cons}(S_i) = a$, if the structure S_i should appear in the antecedent of the rule and $\text{cons}(S_i) = c$, if it should appear in the consequent. An association rule of the form $\alpha \rightarrow \beta$ is valid if and only if (i) it has minimum confidence and (ii) $\text{cons}(S_\alpha) = a$ and $\text{cons}(S_\beta) = c$.

3 Implementation and Results

The objective of the experiment was to associate with MRI slices of brain tumor patients to the spatial deviation of the anatomical structures, for the purpose of validating actual diagnosis rules. Here the intention is to confirm the validity of established diagnosis. Also, we are attempting to find new rules. The dataset was MRI images which for simplicity was categorized as Normal (N) or Abnormal (A) as per diagnosis records and was obtained from [10],[11]. The MRI images considered here are DICOM images. The axial slices of MRI images are considered here for experimental purposes. Different anatomical structures can be identified in the MRI scan; of course, the readability of the structures varies in regard to the slices. Here four different anatomical structures in both hemispheres' are identified in the 'anatomical structure identification' process. They are LCN, LLV, LPU, LTH in the left hemisphere and RCN, RLV, RPU, RTH in the right hemisphere.

The experiments were run with 60% confidence and $S_{\max}=4$. With variations in support value, interesting rules of different length pattern were obtained, that validated the actual diagnosis and domain knowledge. Some examples of rules that validate the actual diagnosis and domain knowledge are given below.

Rule 1: LCN is not totally above LLV \rightarrow Abnormal (conf: 100%)

Interpretation of Rule 1: With the aid of domain knowledge, the anatomical structure LCN should be located above w.r.t the structure LLV, i.e., explaining in terms of fuzzy spatial relationship ontology, the membership degree of LCN w.r.t LLV in the 'Above' direction is 0.9 for all 'Normal' situations. Analyzing our dataset, when the membership degree of LCN w.r.t LLV in the 'Above' direction varied from 0.9, it always resulted in 'Abnormal' situations. And hence the rule can be expressed in imprecise terms as shown above.

Rule 2: LTH is not totally below LCN \rightarrow Abnormal (conf: 100%)

Interpretation of Rule 2: With the aid of domain knowledge, the anatomical structure LTH should be located below w.r.t the structure LCN, i.e., explaining in terms of fuzzy spatial relationship ontology, the membership degree of LTH w.r.t LCN in the 'Below' direction is 0.98 for all 'Normal' situations. Analyzing our dataset, when the membership degree of LTH w.r.t LCN in the 'Below' direction

varied from 0.98, it always resulted in ‘Abnormal’ situations. And hence the rule can be expressed in imprecise terms as shown above.

Rule 3: LCN is not totally above LLV \rightarrow LCN is not totally above LTH (conf: 60%)

Interpretation of Rule 3: This is a rule correlating the spatial relation between three structures LCN, LLV and LTH. The rule was attained at a confidence of 60%. When LCN is not totally above LLV, then it is also not totally above LTH.

The rules derived had necessary confidence and was verified in concurrence with human analysis. With the aid of group () and cons () functions, we could eliminate practically insignificant rules. The imprecise terms given in the rule is given by the degree in which the value pertains to the fuzzy set describing the term. Although the medical database store quantitative attributes containing precise values, as human beings we employ rules relating imprecise terms rather than precise terms. As our intention was to find association rules with improved semantics, we are relating imprecise terms with clear semantic content from a database containing precise values.

4 Conclusion

In this paper, the concept of fuzzy spatial relationship and association rule mining is extended to medical images. As per the design of the system, the medical image is segmented to identify the anatomical structures as well as the spatial location. Once the structures are identified, their spatial relationship is estimated in terms of fuzzy set theory, which brings the situation more close to real scenarios. These spatial relationships are then mined to obtain relevant spatial patterns between the structures. From this, simple and interesting association rules relating the spatial relationships with image attributes are derived. Future work includes obtaining new diagnosis rule for MRI images. The entire system can also be viewed from the perspective of an application. It can be used to compute the success of a tumor removal surgery in brain. The spatial displacement of the anatomical structures ‘before’ and ‘after’ the surgery, along with domain knowledge guidance, will give us an implication of how far the surgery was a success.

References

1. Bloch, I., Hudelot, C., Jamal, A.: Fuzzy spatial relation ontology for image interpretation. *J. Fuzzy Sets and Systems* 159, 1929–1951 (2008)
2. Lee, A.J.T., Hong, R.W., Ko, W.M., Tsao, W.K., Lin, H.H.: Mining spatial association rules in image databases. *Information Sciences* 177(7), 1593–1608 (2007)
3. Lee, A.J.T., Liu, Y.H., Tsai, H.M., Lin, H.H., Wu, H.W.: Mining frequent patterns in image databases with 9D-SPA representation. *Systems and Software* 82, 603–618 (2009)
4. Burl, M.C.: Mining for image content. In: *Systemics, Cybernetics, and Informatics /Information Systems: Analysis and Synthesis* (1999)
5. Jain, A.K., Murt, M.N., Flynn, P.J.: Data clustering: a review. *ACM Computing Survey* 31(3) (1999)
6. Marcela, X., Caetano, T.: An Association Rule-Based Method to Support Medical Image Diagnosis with Efficiency. *IEEE Transactions on Multimedia* 10(2) (2008)

7. Hsu, W., Dai, J., Lee, M.: Mining viewpoint patterns in image databases. In: 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC (2003)
8. Megalooikonomou, V., Barnathan, M., Zhang, J., Kontos, D., Bakic, P., Maidment, A.: Analyzing Tree-Like Structures in Biomedical Images Based on Texture and Branching: An Application To Breast Imaging. In: International Workshop on Digital Mammography (IWDM), Tucson, AZ (2008)
9. Pan, H., Han, Q., Yin, G.: A ROI-Based Mining Method with Medical Domain Knowledge Guidance. In: IEEE International Conference on Internet Computing in Science and Engineering (2008)
10. Brain Web Project, <http://mouldy.bic.mni.mcgill.ca/brainweb/>
11. Harvard Medical School, <http://www.med.harvard.edu/AANLIB/>

FCHC: A Social Semantic Focused Crawler

Anjali Thukral¹, Varun Mendiratta¹, Abhishek Behl¹, Hema Banati²,
and Punam Bedi¹

¹Department of Computer Science, University of Delhi, Delhi, India
athukral@cs.du.ac.in, varunlibra3@yahoo.com,
abhishek_8988@yahoo.co.in, punambedi@ieee.org

²Department of Computer Science, Dyal Singh College,
University of Delhi, Delhi, India
hema.banati@gmail.com

Abstract. The World Wide Web is a huge collection of web pages where every second, new piece of information is added. Searching and retrieving relevant web resources is a protracted task and finding relevant resources w.r.t. some topic, without any explicit or implicit feedback adds more intricacy to the process. Focused crawling in such scenarios provides a better alternate to generic crawling especially when topic specific or personalized information is required. This paper presents a crawling approach FCHC that uses human cognition for focused crawl on a social bookmarking site initiated with the seeds retrieved from a search engine. It utilizes social bookmark tags as implicit feedback to compute eResource relevance and Vector Space Model to rank the retrieved eResources. A well established metric called harvest ratio is used to compare the results of the proposed approach with the semantic focused crawler and the classic focused crawler. The analysis of the results shows a better performance of social semantic focused crawlers over the semantic and classic focused crawlers.

Keywords: Focused Crawling, Social Bookmarking, Web Resource Retrieval, Vector Space Model, Resource Relevance, Resource Ranking.

1 Introduction

The search engines in general and web crawlers in particular are facing challenges of ever increasing volume of the WWW. Every day thousands of web pages are being added to the web. With the passage of time, it is becoming difficult to crawl and update the complete web in short time. In such circumstances a goal-directed crawling or Focused crawling is a promising alternate solution to a generic crawler. A focused crawler, the term coined by Chakrabarti et al. [1] is a topic-driven web crawler which selectively retrieve web pages that are relevant to a pre-defined set of topics. A focused crawler yields latest resources (web pages) relevant to the need of individuals while utilizing minimum storage space, time and network bandwidth [2]. Applications of the focused crawler include business intelligence (to keep track of publically available information about their potential competitors) [3], generating web based

recommendations, retrieving domain/topic relevant eLearning web resources [4], scientific paper repositories and many more.

The focused web crawlers are designed to retrieve web pages based on the guidelines that identify relevant pages or/ and priority criterions to sequence the web pages to be crawled and add them to the local database. This database may then serve different application needs. Focused crawlers are grouped into two broad categories namely, the classic focused crawler and the learning focused crawlers. Both have their own variations depending on various algorithms [3] applied on them. However the main difference between the two is that the former one follows the predefined and fixed guidelines or criterions for crawling whereas the latter learns or adapts the crawling guidelines based on the dynamically updating training set. Semantic crawler and social semantic crawler are such two variations of the classic focused crawler.

This paper presents the design and implementation of three variations of the classic focused crawler. These variations include classic focused crawler based on the link structure of the web by applying best sequence of web pages, semantic crawler to semantically analyze relevance of the web page and prioritize the next crawl, and the proposed social semantic crawler FCHC (Focused Crawling using Human Cognition) that crawls a social bookmarking site using a systematic pattern to retrieve relevant resources by analyzing semantically relevant tags to the topic. A social bookmarking site is an online social network collaborative bookmarking system that facilitates to locate many relevant resources, which a normal crawler in a search engine is usually unable to find. Social networks have been analyzed for a couple of decades to find useful information related to web pages [1].

The rest of the paper is organized as follows: Section 2 present the basic concepts of focused crawler and Social Bookmarking Sites (SBS). Section 3 describes the design of different types of focused crawlers including the proposed social semantic focused crawler. Section 4 presents the evaluation of the focused crawlers explained in section 3. Section 5 concludes the paper.

2 Background

2.1 Web Crawler

Web crawlers, also known as bots, spider etc. are the software programs that are designed to assemble URLs and other web page attributes locally by exploring the link structure of the web. Focused crawlers are the specialized form of the web crawlers that selectively seek out web pages that are relevant to a topic. The baseline idea of the focused crawler is to maximize the retrieval percentage of relevant web pages while keeping the total number of fetched pages at the minimum [1]. Therefore for a focused crawler it is always crucial to search for relevant group/ cluster of pages on the web. There exist many algorithms that predict the probability of getting relevant clusters through different link paths [3]. In this regard, Social Network Sites, in general and Social Bookmarking Sites (SBS) in particular provide the promising solution. Here, a user could find a number of relevant pages along with the collection of tags that define or at least classify the web resource related to a topic. There are over one thousand social bookmarking sites where web users tag millions of web

resources for their later reference. The focused crawlers can be broadly classified into two categories - Classic focused crawlers and Learning crawlers (Fig. 1).

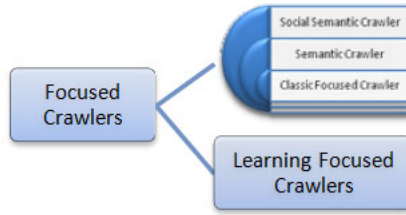


Fig. 1. Focused Web Crawlers Taxonomy

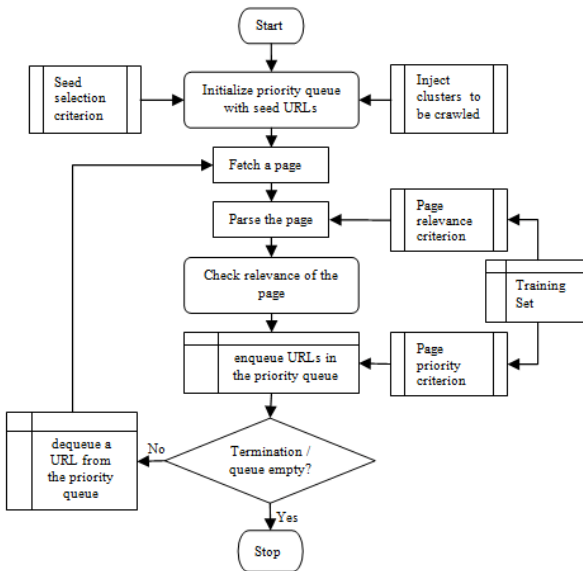


Fig. 2. Baseline flow of a Classic Focused Crawler

Classic focused crawler: The baseline of a focused crawler, shown in Fig. 2 is same for all its variants. It initiates with a set of pre-selected seed URLs. The crawler first fetches a page and then parses it to extract all the links and to check topic relevance of the page, based on the page relevance criterion. The page relevance criterion is similar to the classifier [1], which evaluates the relevance of hypertext document. All the parsed URLs are enqueued in the priority queue if the page is considered relevant according to the page relevance criterion. The priority of the page is further governed by the page priority criterion, which resembles the distiller [1]. The distiller identifies hypertext nodes that have good access points to many relevant pages within few links. Different priority queues may be used here depending on the application’s

requirements. If a page is considered non relevant then the links on that page are not put on the queue i.e. it is not crawled further. This is where it differs from a generic crawler used by search engines. After checking termination condition, which could be the number of URLs to be crawled, time limit or the empty queue, it decides to either stop or dequeue a top priority URL from the queue and repeat the crawl process. There exist many variations to this basic focused crawler by varying the four important selection criterions. i) The seed selection criterion, ii) Inject clusters to be crawled, iii) Page relevance criterion and iv) Page priority criterion.

One such prominent variation of the classic focused crawler is Semantic focused crawler that uses semantic knowledge (hyponyms, hypernyms, synonyms, antonyms etc. of the topic) to define page relevance criterion and variation to this is Social semantic focused crawler (the proposed crawler) which also specify an area on the web to be crawled. In our case the area of the web is Social Network Site. They are explained in subsequent sections.

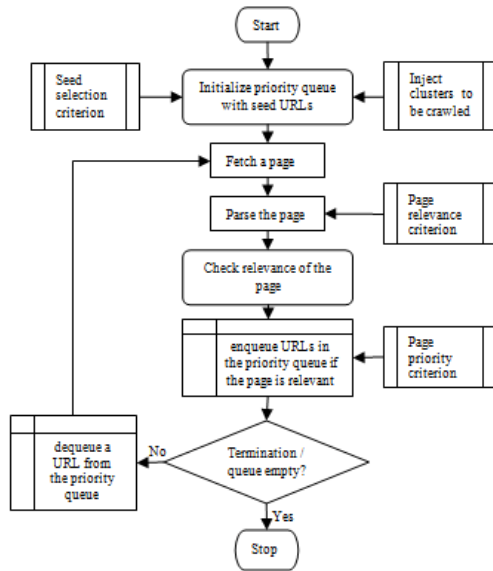


Fig. 3. Learning Crawler

Learning crawler: The other type of the focused crawler is Learning Crawler. Learning crawlers (Fig. 3) apply a training set to govern one or more criterions of the classic focused crawler. The training set consists of a set of example pages related to the topic. The crawler identifies relevant and non relevant pages by making use of the training set. Learning crawlers use various methods based on Bayesian classifier, context graphs [5] and Hidden Markov Model [2] to estimate the link distance between a crawled page and the relevant pages. This data is used to train the crawlers by setting the criterion for page relevance and fetching priority.

2.2 Social Bookmarking Site

A Social Bookmarking site is a specialized type of Social Networking site where web users tag the web resources of their interest for later use. These sites facilitate their users to tag, organize and share the web resources. The users usually use various keywords as tags to bookmark the web resources that define their content. People having same interests may thus find and share resources bookmarked by other users of the community. Due to the free access and complete freedom to tag bookmarks, there exists lot of noise along with the useful information [6]. But searching relevant documents in a defined pattern similar to the search made by humans in SBS, leads to filter the noise and retrieve useful resources. The experimental studies to recommend ranked resources using SBS in [7], [8] have shown promising results. The effectiveness of a web search engine can also be improved by using tag based search algorithms [9] and re-rank the search results [10], [11].

The required information in the SBS can be viewed as a set of bookmarks, B , where each bookmark, $b \in B$, is a set of triples: $\{r, u, t_1\}, \{r, u, t_2\} \dots \{r, u, t_k\}$; resource (of any format), user and tags. Thus a single triple $\langle r_i, u_j, t_k \rangle$ shares a relation signifying resource r_i being tagged by user u_j with tag t_k . For each set of resource and user, there are a number of tags. It may be noted further that each bookmark b represents a single resource.

3 Focused Crawler Design

The design of three crawlers, classic FC (Focused Crawler), Semantic FC and the proposed social semantic FC, FCHC are presented in this section. The classic FC crawls on the anchor text relevance, Semantic FC crawls on the content semantic similarity of the web pages and the FCHC crawls on the semantic relevance of bookmarked tags.

3.1 Design of the Classic Focused Crawler

The classic focused crawler uses the links on the web page to crawl which initiates with the select URLs. During parsing of the page, only relevant links (URLs) are put in the queue. The links are considered relevant if the parsed anchor text contain semantically relevant terms of the topic. The semantically relevant terms are extracted from the domain ontology corresponding to the topic [4]. The priority criterion is FCFS and the crawl is run for a number of URLs. The topic relevance of each retrieved document is then computed using semantic relevance for evaluation.

3.2 Design of the Semantic Focused Crawler

The terms semantically similar to the topic are extracted from the ontology and are compared with the terms contained by the page to determine semantic similarity between the topic and the page. Instead of using hypernyms, hyponyms, synonyms or antonyms from WordNet which is a taxonomic hierarchy of natural language terms [2], the crawler uses conceptually related terms from the manually created domain ontology. This is because the technical terms usually do not have synonyms,

hyponyms etc. rather they have sub concepts, super concepts and sibling concepts, which is more significant when extracted from the domain ontology constructed by experts. The semantic relevance is computed by summing the semantic distance of all semantically related words present on the web page [13]. If the page is semantically relevant, it is put in the queue for further crawl. The priority is set depending on the computed semantic relevance.

3.3 Design of the Proposed Social Semantic Focused Crawler

FCHC (Focused Crawling using Human Cognition), the proposed crawler is the social web and semantic knowledge to crawl relevant resources and therefore it can be considered as one of the variation of social semantic focused crawler. At the baseline the technique is a variant of the classic focused crawler that uses domain ontology to expand the search topic semantically and starting with the seed URLs searches relevant resources in a SBS. The searching pattern used to search relevant resources in SBS makes this crawler different from others. The framework of the FCHC Crawler is illustrated in Fig. 4. Unlike other focused crawlers that parse every web page to calculate or predict relevance of the web page or path to reach relevant resources, the proposed crawler make use of the tags and the bookmarks tagged by similar users. The crawler using a pattern based on human cognition explores the pages of SBS users who have bookmarked web resources with the semantically relevant tags. During the crawl, the crawler checks the resource relevance by the presence of semantically relevant terms in the tags. However, the quantitative relevance of web resources is computed later for ranking purpose. This reduces the crawling time of the crawler.

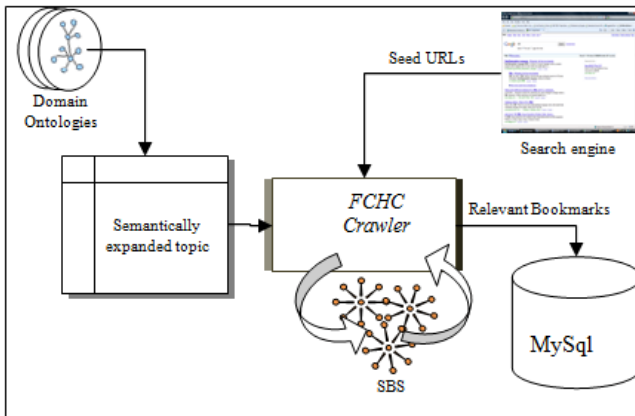


Fig. 4. Framework of the FCHC Crawler

The important selection criterions for FCHC from the design of classic focused crawler’s (Fig. 2) perspective are detailed below.

Seed selection criterion: Seed URLs are simply picked from the topic search made on the search engine. Top ten URLs are considered as seed to initiate crawl on the web.

Inject SBS to be crawled: A URL of the SBS is provided to the crawler to search for relevant resources. The crawler searches relevant pages from the SBS using the seed URLs.

Parse the page: These are the SBS pages, parsed for extracting bookmarks that are tagged with relevant terms and users who have tagged relevant resources.

Page relevant criterion: A subjective relevance is used to check the relevance of page during the crawl. Later, after the complete crawl a quantitative social semantic relevance of each resource is computed.

Page priority criterion: A systematic search pattern motivated by human cognition is used as the priority criterion for the crawler. This is illustrated in Fig. 5. Based on the design of SBS (that we are using for experimentation) the pages are crawled using two types of pattern: Breadth first pattern (BFP) and Depth first pattern (DFP). In a BFP the crawler enqueue all users of all seed URLs that exist on the pages retrieved through the URLs. Then from the queue one by one the tags page of each user is parsed to reach the relevant resources. All these relevant resources are then again put in the queue to be parsed like the seed URLs.

In general, a crawler hardly uses DFP but in particular case of SBS, DFP also shows promising results. This pattern first reaches the relevant resources of the first user of first seed URL and enqueue them. It then iterates the process for all users of the same seed URL. It similarly works upon all other seed URLs. The difference between BFP and DFP is that the latter completes retrieval of relevant tagged resources of the user one by one, whereas the former first queue up all related users and then moves to their tagged resource pages one by one. The other termination criterion that we are using in FCHC social semantic crawler is to crawl up to level 1 or level 2 (Fig. 5) for DFP. It actually indicates the depth of the crawl.

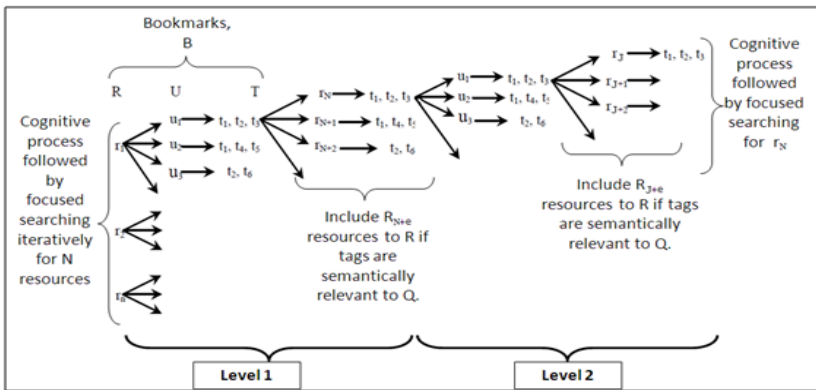


Fig. 5. The crawling pattern followed by FCHC-DFP

Termination criterion: If the number of URLs to be crawled is specified then the crawler uses it as a termination criterion otherwise when the priority queue becomes empty, the crawler stops.

The crawler uses Social Semantic Similarity [4] based on VSM to measure topic relevance of the retrieved URLs. It uses tagged resources and their popularity count to compute the similarity between topic and the web resources. The bookmarking site delicious.com¹ is used in this paper for crawling and retrieving social semantic relevant web resources. The SBS is a collection of bookmarks where each bookmark is a set of triplet $\langle r_i, u_j, t_k \rangle$, where r_i is a resource tagged by user u_j with a tag t_k .

Following formula is used to compute the similarity (social semantic) between the topic and each resource.

$$\text{Cosine } \theta_{Q_0, r_i} = \frac{\sum_{q_t=t_k} (Wt_{Q_0, q_t} \cdot Wt_{t_k, r_i})}{\sqrt{\sum_{q_t} Wt_{Q_0, q_t}^2} \sqrt{\sum_{t_k} Wt_{t_k, r_i}^2}}$$

Wt_{t_k, r_i} is the tag weight of the resource r_i computed with the semantic distance between the tag and the topic, and the popularity count of the tag in SBS. Wt_{Q_0, q_t} is computed using semantic distance between the expanded term q_t and the topic Q_0 .

4 Evaluation

The experimental study is conducted on Intel core 2 Duo processor, 2.4 GHz, 2 GB RAM, 32-bit OS. The classic focused crawler, semantic focused crawler and social semantic focused crawlers (BFP, DFP up to level 1 and DFP up to level 2) are implemented in java and MySQL. The resources are crawled on the topic ‘DML’ (Data Manipulation Language), a database concept. The comparisons are made over the retrieved resources by the crawlers, and their computed semantic relevance to the topic. Afore mentioned crawlers use different relevance criteria during the crawl, but they are evaluated by applying similar relevance computation method.

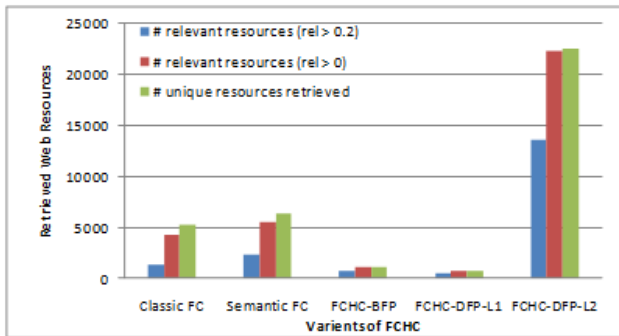


Fig. 6. Number of resources retrieved through the variations of FCHC crawler on the same topic and same set of seed URLs

¹ <http://delicious.com>

The FCHC crawler with BFP approach retrieved 1108 resources in 1hr 49 min; DFP-L1 retrieved 774 resources in 1 hr and DFP-L2 retrieved 22,552 resources in 20 hrs 9 min. Though DFP-L2 took longest time to retrieve resources, it retrieved maximum (13584) number of relevant resources when compared to BFP (688) and DFP-L1 (513). Fig. 6 shows the difference between the relevant resources retrieved and the total number of retrieved resources by classic FC, semantic FC and FCHC crawler while using BFP, DFP-L1 and DFP-L2 approaches.

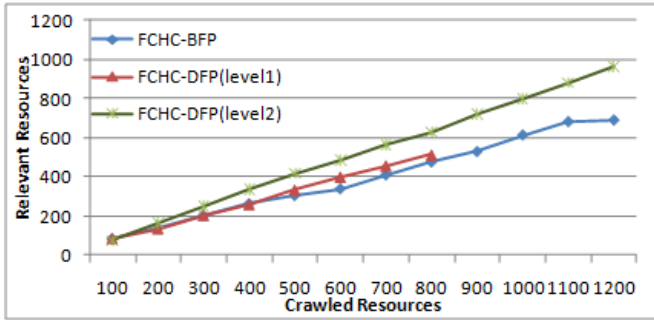


Fig. 7. Performance comparison among three types of Social Semantic Focused Crawlers (FCHC)

The FCHC-DFP-L2 crawler shows the best performance (Fig. 7), when compared with its other variants. The structure of the SBS is such that the crawler is able to retrieve better results when the resources are traversed in Depth first pattern and it is best when retrieval is deeper.

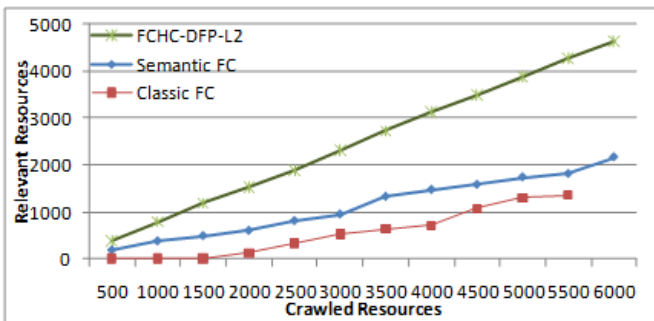


Fig. 8. Performance comparison among three types of focused crawlers

Fig. 8 shows the comparison among the three types of crawlers: Classic FC, Semantic FC and FCHC-DFP-L2. Though the semantic focused crawler retrieved many relevant resources from the web, it was observed that it also retrieved similar pages written in different languages, which increases the computed relevance of a web page. Classic FC retrieved least relevant resources in comparison to others.

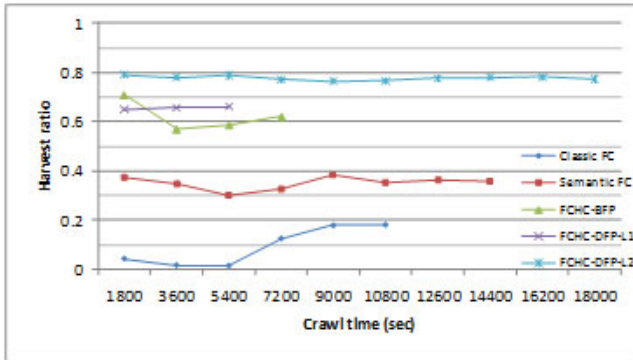


Fig. 9. Harvest rate of all crawlers

Fig. 9 shows the harvest rate of the retrieved resources by all five focused crawlers. Harvest ratio (*hr*) [1], [2], [13], [14] is the fraction of web pages crawled that satisfy the crawling target (relevant pages) $\#r$ among the crawled pages $\#p$. Thus $hr = \#r/\#p$, $hr \in [0,1]$ and harvest rate is the harvest ratio per unit of time. FCHC-DFP-L2 shows the best harvest rate among all the crawlers.

5 Conclusion

The design, implementation and evaluation of various types of focused crawlers have been presented in this paper. These include classic focused crawler, semantic focused crawler and three variations (BFP, DFP-L1 and DFP-L2) of the proposed social semantic focused crawler FCHC.

The classic focused crawler and the semantic focused crawler uses the link structure of the web. However, the classic focused crawler parse the content of retrieved web pages by analyzing the anchor text whereas the semantic focused crawler analyzes semantically relevant terms to compute the relevance of web pages and prioritize the sequence of web pages to be crawled.

The social semantic focused crawler FCHC is based on the similar design as the classic focused crawler with the differences in the ways of selecting the region of the web as SBS, parsing the SBS page instead of the web resource, relevance criterion on semantic terms and priority criterion based on human cognitive search pattern.

The crawlers were evaluated on the semantic relevance of the crawled web pages using the well-known performance metric, harvest rate. FCHC-DFP-L2 showed exceptionally good performance an average harvest rate of 0.78 as compared to FCHC-DFP-L1 (0.67), FCHC-BFP (0.62), Semantic FC (0.35) and Classic FC (0.09). Although BFP and DFP-L1 crawled less number of resources as compared to semantic FC and classic FC, they show a better harvest ratio. Overall, all three variants of social semantic focused crawler showed better performance when compared to semantic and classic focused crawlers. It is apparent from the evaluation that crawling social bookmarking site for retrieving relevant resources yield better results when compared to the web crawling. However, there may be relevant web

pages that are not bookmarked by any user in SBS. Such pages are not retrieved through the FCHC, which will form the part of future work.

References

1. Chakrabarti, S., Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks* 31, 1623–1640 (1999)
2. Batsakisa, S., Petrakisa, G.M.E., Milios, E.: Improving the performance of focused web crawlers. *Data & Knowledge Engineering* 68, 1001–1013 (2009)
3. Pant, G., Srinivasan, P. and Menczer, F.: *Crawling the web. s.l. : Springer-Verlag, 2004, Web Dynamics: Adapting to Change in Content, Size, Topology and Use(2004).*
4. Bedi, P., Banati, H., Thukral, A.: Social semantic retrieval and ranking of eResources. In: *ACEEE 2010: Second International Conference on Advances in Recent Technologies in Communication and Computing, Kerela, India, pp. 343–347 (2010)*
5. Diligenti, M., et al.: Focused crawling using context graphs. In: *26th International Conference on Very Large Databases, Cairo, Egypt, pp. 527–534 (2000)*
6. Greg, P., Chowdhury, A., Torgeson, C.: A picture of search. Hong Kong Spink. In: *1st International Conference on Scalable Information Systems (2006)*
7. Bischoff, K., et al.: Can all tags be used for search? In: *17th ACM Conference on Information and Knowledge Management, pp. 193–202. ACM, New York (2008)*
8. Firan, C.S., Nejdil, W., Paiu, R.: The benefit of using tag-based profiles. In: *LA-WEB (2007)*
9. Agrahri, A.K., Anand, T.M.D., Riedl, J.: Can people collaborate to improve the relevance of search results? Switzerland : Lausanne, 2008. In: *2nd ACM International Conference on Recommender Systems, October 23-25, pp. 283–286 (2008)*
10. Bao, S., et al.: Optimizing web search using social annotation. In: *16th International Conference on World Wide Web. ACM, Banff (2007)*
11. Chen, S.-Y., Yi, Z.: Improve web search ranking with social tagging. In: *1st International Workshop on Mining Social media and 13th Conference of the Spanish Association for Artificial Intelligence, Sevilla, Spain (2009)*
12. Ehrig, M., Maedche, A.: Ontology-focused crawling of web documents. In: *Symposium on Applied Computing. ACM, Melbourn (2003)*
13. Menczer, F., et al.: Evaluating topic-driven web crawlers. In: *24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 241–249. ACM, New York (2001)*
14. Zheng, H.-T., Kang, B.-Y., Kim, H.-G.: Learnable Focused Crawling Based on Ontology. In: Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G. (eds.) *AIRS 2008. LNCS, vol. 4993, pp. 264–275. Springer, Heidelberg (2008)*

A Dynamic Seller Selection Model for an Agent Mediated e-Market

Vibha Gaur and Neeraj Kumar Sharma

Department of Computer Science,
University of Delhi, Delhi, India
{3.vibha,neerajraj100}@gmail.com

Abstract. Due to the virtuality of online trade, e-market has an information asymmetrical system, and confirmation of sellers' behavior and goods' attributes mostly depend on the assessment of buyers. So, in a virtual environment, how the buyers utilize all existing information and choose the right sellers to guarantee the utility of a deal is of vital importance. This paper proposes an agent mediated e-market methodology that aims to help a buyer agent in selecting a seller that offers a good with the highest expected value. Due to the need to assess multiple attributes of a good in the virtual environment, this problem belongs to the class of fuzzy multi-attribute decision making. This problem is fuzzy in nature due to the lack of precision in assessing the relative importance of different attributes and the performance rating of goods from different sellers with respect to each attribute. The proposed methodology addresses these issues by first computing the expected value of a good being offered by different sellers using fuzzy set theory and then selecting a seller that offers the good with the highest expected value. The proposed methodology is relatively dynamic as it is sensitive to the changing experience of buyer agents in the e-market. This ensures that after sufficiently large number of transactions by the same buyer for a particular good, instead of incurring the overhead of computing the subjective weights, buyer can utilize the attribute weight information from the previous transactions.

Keywords: Linguistic Terms; Fuzzy Attribute Weights; Fuzzy Sets; e-market.

1 Introduction

The world today has embraced the Internet business as a part of its everyday life. Agent mediated e-commerce provides a platform where buyers and sellers exchange goods and services in e-market by employing software agents. Due to the information asymmetric environment of the on-line trade, the capability of buyers to use existing information for choosing right sellers is of utmost importance in the success of e-commerce [10].

The environment of e-market is dynamic by nature as it undergoes continuous changes with different agents joining and leaving the e-market at will [9]. The power of an agent mediated e-commerce can be utilized to the optimum if different process

models inherent to the e-transactions like deciding about pricing of goods, computing and distributing reputation of participants and selection of sellers are also dynamic. A dynamic model must be sensitive to the changing e-market environment and must adapt to changing experience of buyer and seller agents with each transaction. Dynamic e-commerce models would provide virtually instantaneous knowledge about the changing e-market environment and would utilise Internets' capacity for continuous interactivity.

Various seller selection methodologies have been proposed in the literature that include linear weighting models [19], categorical model [19], weighted point model, total cost of ownership, multi-attribute utility theory, principal component analysis [20], analytical network process [11], AHP [13, 14], fuzzy AHP [18] and linear programming [16], among others. AHP has been used for various ranking problems [13, 14] but with a shortcoming that number of comparison increases quickly with the increase in the number of criteria. Further, as very often buyers' express their views in linguistic terms due to unquantifiable information, incomplete information, unobtainable information, and partial ignorance [15], so fuzzy AHP [17, 18] is used in literature for expressing relative importance of attributes of a good. The seller selection models described in the literature [11, 13, 14, 16, 18, 19, 20] are relatively static as they are not sensitive to the changing experience of buyer agents in the e-market. Hence, this paper proposes a dynamic seller selection methodology that is sensitive to the changing experience of buyers in the e-market.

In an e-market environment having buyer and seller agents, the problem of selecting an appropriate seller for buying a particular good is a challenging one and belongs to the class of multiple attribute decision making as the process entails the evaluation of different suppliers based on multiple attributes. Decision making may be characterised as a process of choosing or selecting 'sufficiently good' alternative from a set of available alternatives, to attain a goal [1].

The major factors that affect buyers' selection of a seller for buying a specific good include the perception of a seller and evaluation of the good to be purchased. In the agent mediated e-commerce, the perception of the seller is generally represented by the reputation of a seller in the e-market where reputation is expressed as "a quantity derived from the underlying social network which is globally visible to all members of the network" [7]. This paper deals with the problem of selecting a seller by evaluating a good being offered by various sellers.

In the proposed methodology, a buyer agent first computes the weights of different attributes of the good to be purchased based on relative importance of its different attributes and then computes the expected value of the good offered by each seller based on its attributes. Finally, it chooses a seller with highest expected value of the good i.e. good with highest expected utility for the buyer. Expected requirement of a buyer from a good constitutes buyers' estimation of goods' attributes and is subjective and fuzzy in nature. It is subjective as relative priority of attributes of a good would vary with each good and with each buyer. It is fuzzy as generally buyers' expectations related to a particular attribute are specified in fuzzy terms like "low", "high" or "very high". Similarly, a buyer has to map linguistic assessment of goods being offered by different sellers based on their attributes to the fuzzy scale. Hence this paper uses fuzzy set theory to allow a buyer agent to compute attribute weights of a good and to select a seller that offers the good with highest expected value.

The seller selection methodology proposed in this paper is relatively dynamic as in computing the weights of different attributes of a good from the buyers’ perspective, it gives importance to the increasing experience of a buyer, as with each successive purchase of the same good by a particular buyer, the importance of empirical component of the weight increases and that of subjective weight component decreases. This ensures that after purchasing a good repeatedly in a large number of transactions, as the buyer gains sufficient experience, instead of computing the subjective weight component, it can utilise the attribute weight information from the previous transactions.

The rest of this paper is organized as follows. Concepts of fuzzy set theory are introduced in section 2. Section 3 presents a dynamic seller selection methodology. Section 4 comprises of a case study and section 5 concludes the paper.

2 Fuzzy Set Theory

Fuzzy set theory mathematically represents uncertainty and vagueness of human thinking to provide formalized tools to deal with ambiguity embedded in a large number of problems. Basic concepts of fuzzy set theory and fuzzy membership functions used in this paper are described next in this section.

2.1 Overview

In classical set theory, elements have either complete membership or complete non-membership in a given set. With fuzzy set theory, intermediate degrees of membership are allowed. The degree of membership of each elements in the set is defined as the membership function of the fuzzy set in the range [0,1].

A fuzzy number is completely defined by its membership function in the range [0,1]. Let \tilde{A} be a fuzzy number whose membership function $f_{\tilde{A}}(x)$ is defined by (1).

$$\begin{aligned}
 f_{\tilde{A}}(x) &= \frac{x-a}{b-a}, & a \leq x \leq b \\
 &= 1, & b \leq x \leq c \\
 &= \frac{d-x}{d-c}, & c \leq x \leq d \\
 &= 0, & \text{otherwise}
 \end{aligned}
 \tag{1}$$

The trapezoidal fuzzy number \tilde{A} is illustrated in Fig. 1.

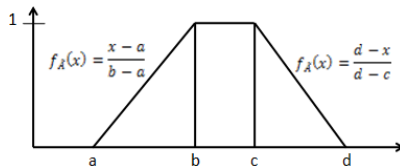


Fig. 1. Trapezoidal Fuzzy Number

The two most popular membership function examples for fuzzy numbers comprise of triangular and trapezoidal fuzzy numbers. If the membership function $f_{\tilde{A}}(x)$ is piecewise linear, then \tilde{A} is referred to as trapezoidal fuzzy number and is usually denoted by $\tilde{A} = [a, b, c, d]$. In particular when $b = c$, the trapezoidal fuzzy number is reduced to a triangular fuzzy number denoted by $\tilde{A} = [a, b, c]$. Hence triangular fuzzy numbers are special cases of trapezoidal fuzzy numbers.

Fuzzy logic is very useful in an environment that requires human perceptions as inputs where ambiguity and vagueness exists. The major constituents of seller selection problem like the assessment of relative importance of goods' attributes and the assessment of sellers' offers are often subjective and can best be expressed in linguistic terms. Therefore, the use of fuzzy set theory in designing the proposed technique for selecting a seller appears to be a logical choice. In the next section the membership functions used in the evaluation process are presented.

2.2 Fuzzy Membership Functions

The problem of computing the estimated value of a good constitutes two components: estimating weights of the goods' attributes and estimating the performance of a good being offered by different sellers for each attribute.

The buyer agent specifies the relative importance of each attribute of the good in linguistic terms like "Equally Important", "Moderately Important", "Highly Important" and "Very Highly Important" to facilitate pairwise comparison. The fuzzy scale for linguistic weights of these attributes is illustrated below in Fig. 2.

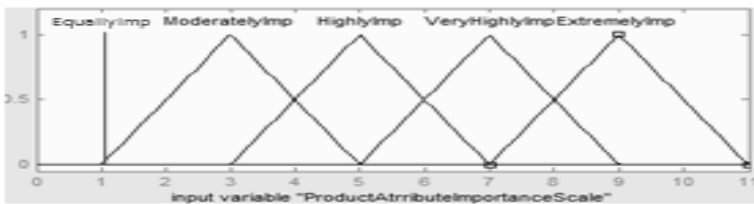


Fig. 2. Fuzzy Number Scale for Relative Importance of Attributes

Mapping of linguistic variables representing attribute importance to trapezoidal fuzzy numbers according to the scale given in Fig. 2 is shown in the Table 1 below.

Table 1. Mapping of Linguistic Variables to Trapezoidal Fuzzy Values

Linguistic Term	Trapezoidal Fuzzy Value
Equally Important (E)	(1, 1, 1, 1)
Moderately Important (M)	(1, 3, 3, 5)
Highly Important (H)	(3, 5, 5, 7)
Very Highly Important (VH)	(5, 7, 7, 9)
Extremely Important (EI)	(7, 9, 9, 11)

The value of attributes of a good offered by various sellers with respect to different attributes is expressed in linguistic terms such as “poor”, “average”, “good”, “very good” and “excellent”. The corresponding fuzzy scale of these linguistic membership functions for different attributes in the range 0-13 is given below in Fig. 3.



Fig. 3. Fuzzy Scale for Linguistic Performance of Sellers

Mapping of a goods’ linguistic performance to fuzzy values using scale given in Fig. 3 is shown in the Table 2 below.

Table 2. Mapping of Linguistic Performance Variables to Trapezoidal Fuzzy Values

Linguistic Term	Trapezoidal Fuzzy Value
Poor (P)	(0, 0, 1, 3)
Average (A)	(1, 3, 4, 6)
High (H)	(4, 6, 7, 9)
Very High (VH)	(7, 9, 10, 12)
Excellent (EX)	(10, 12, 13, 13).

Fuzzy Arithmetic. The methodology proposed in this paper uses three important algebraic operations on fuzzy numbers: inverse, addition and multiplication. The result of addition of two trapezoidal fuzzy numbers is a trapezoidal fuzzy number. Therefore, if $\tilde{A} = (a_1, a_2, a_3, a_4)$ and $\tilde{B} = (b_1, b_2, b_3, b_4)$ are two positive trapezoidal fuzzy numbers then, the fuzzy addition of \tilde{A} and \tilde{B} is given below in (2).

$$\tilde{A} + \tilde{B} = (a_1 + b_1, a_2 + b_2, a_3 + b_3, a_4 + b_4) \tag{2}$$

The inverse of a fuzzy number \tilde{A} represented as \tilde{A}^{-1} is shown below in (3).

$$\tilde{A}^{-1} = \left(\frac{1}{a_4}, \frac{1}{a_3}, \frac{1}{a_2}, \frac{1}{a_1} \right) \tag{3}$$

Unlike addition and subtraction, product of two trapezoidal fuzzy numbers may not result into a trapezoidal number [3, 4]. From the computational point of view, though, it is recommendable to use trapezoidal fuzzy numbers. Therefore, this paper uses an approximation of the product of two trapezoidal fuzzy numbers to a new trapezoidal fuzzy number [3]. The product of two trapezoidal fuzzy numbers, A and B given

above is approximated by the trapezoidal fuzzy number $\tilde{C} = \tilde{A} \times \tilde{B} = (c_1, c_2, c_3, c_4)$ as proposed in [3, 4] where,

$$\begin{aligned}
 c_1 &= \frac{3}{2}(a_2 - a_1)(b_2 - b_1) + 2[(a_2 - a_1)b_1 + (b_2 - b_1)a_1] + 3a_1b_1 - 2a_2b_2, \\
 c_2 &= a_2b_2, \\
 c_3 &= a_3b_3, \\
 c_4 &= \frac{3}{2}(a_4 - a_3)(b_4 - b_3) - 2[(a_4 - a_3)b_4 + (b_4 - b_3)a_4] + 3a_4b_4 - 2a_3b_3.
 \end{aligned}
 \tag{4}$$

For defuzzifying, this paper uses centre of area (COA) or centroid method. For a fuzzy number $\tilde{A} = (a_1, a_2, a_3, a_4)$, its COA is computed as: $(a_1 + a_2 + a_3 + a_4)/4$. The proposed dynamic seller selection methodology is described in the next section.

3 Dynamic Seller Selection Methodology

The seller selection methodology proposed in this section is based on the e-market model having a set of buyers and sellers [9]. In this e-market model, sellers are divided into four categories, namely, reputed, non-reputed, dis-reputed and new sellers. Buyers allocate reputation rating to sellers in the range [0,1]. At any given time a buyer preferably selects a seller from the list of reputed sellers but in no case, it selects a dis-reputed seller. Before purchasing a good, a buyer computes expected value of the offered good by each seller and places an order to the seller that is offering the good with the highest expected value. Computing the expected value of a good offered by different sellers involves seller selection and is of extreme importance as it leads to greater satisfaction of buyer agents and also encourages sellers to offer high quality goods.

The proposed seller selection methodology based on the concept of fuzzy set theory is divided into three phases. In Phase I, the buyers’ pairwise preferences for different attributes of a good are elicited in linguistic terms. These preferences are mapped to fuzzy values by means of trapezoidal fuzzy numbers and their subjective weights are computed by combining extent analysis method [6] with fuzzy AHP technique. Then, the overall weight of each attribute of a good is computed by combining the subjective weight component with the empirical weight component of the attribute. In Phase II, utilities of the good being offered by different sellers are computed by means of a method that approximates the product of two trapezoidal fuzzy numbers to a new trapezoidal fuzzy number. Finally, in Phase III, crisp alternatives are ranked after defuzzifying the trapezoidal fuzzy numbers representing their overall utilities. A systematic representation of this methodology divided into Phase I, Phase II and Phase III is given below.

Phase I:

1. Obtain the buyers’ assessment of pairwise comparison of different attributes of a good in linguistic terms like “E”, “M”, “H”, “VH” or “EI” as illustrated in Fig. 2.
2. Using the appropriate fuzzy membership function, map the linguistic terms to fuzzy values as represented in Table 1.

3. Compute subjective fuzzy weights from the buyer’s perspective by combining extent analysis method [6] with fuzzy AHP. Let $\widetilde{\text{FPM}}$ (Fuzzy Pairwise Matrix) represents the fuzzy reciprocal $n \times n$ matrix representing all pairwise comparisons \tilde{a}_{ij} for all $i, j \in \{1, 2, \dots, n\}$.

$$\widetilde{\text{FPM}} = \begin{bmatrix} (1,1,1,1) & \tilde{a}_{12} & \dots & \tilde{a}_{1n} \\ \tilde{a}_{21} & (1,1,1,1) & \dots & \tilde{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{a}_{n1} & \tilde{a}_{n2} & \dots & (1,1,1,1) \end{bmatrix} \tag{5}$$

Where $\tilde{a}_{ij} = \tilde{a}_{ji}^{-1}$ and all \tilde{a}_{ij} are trapezoidal fuzzy numbers. Now fuzzy subjective weights are computed based on the method of extent analysis [6]. The subjective weight computation of attribute a_i denoted as $\widetilde{\text{sw}}_{a_i}$ is given in (6).

$$\widetilde{\text{sw}}_{a_i} = \sum_{j=1}^n \tilde{a}_{ij} \times \left[\sum_{t=1}^n \sum_{j=1}^n \tilde{a}_{ij} \right]^{-1} \tag{6}$$

Similarly, compute $\widetilde{\text{sw}}_{a_i}$ for $i = 1, 2, \dots, n$, i.e. for all attributes of a good represented by $\widetilde{\text{SW}}$ in (7) below.

$$\widetilde{\text{SW}} = \begin{bmatrix} \widetilde{\text{sw}}_{a_1} \\ \vdots \\ \widetilde{\text{sw}}_{a_n} \end{bmatrix} \tag{7}$$

4. Compute the empirical weight component $\widetilde{\text{ew}}_{a_i}$, i.e. the average of fuzzy weight of each attribute, for $i = 1, 2, \dots, n$, in last n transactions by the same buyer for the same good represented by $\widetilde{\text{EW}}$ in (8).

$$\widetilde{\text{EW}} = \begin{bmatrix} \widetilde{\text{ew}}_{a_1} \\ \vdots \\ \widetilde{\text{ew}}_{a_n} \end{bmatrix} \tag{8}$$

5. Obtain the overall fuzzy attribute weight $\widetilde{\text{w}}_{a_i}$ of a good by using (9) given below.

$$\widetilde{\text{w}}_{a_i} = \alpha * \widetilde{\text{ew}}_{a_i} + (1 - \alpha) * \widetilde{\text{sw}}_{a_i} \tag{9}$$

Similarly, compute $\widetilde{\text{w}}_{a_i}$ for $i = 1, 2, \dots, n$, represented by $\widetilde{\text{W}}$ as shown below.

$$\widetilde{\text{W}} = \begin{bmatrix} \widetilde{\text{w}}_{a_1} \\ \vdots \\ \widetilde{\text{w}}_{a_n} \end{bmatrix} \tag{10}$$

Where, in (9), the value of α is zero in the case of a buyer purchasing a good for the first time. With each subsequent purchase of the same good by the same buyer, the value of α increases by a small fraction, say 0.01. This ensures that initially when a buyer has no experience of buying a good, the overall weight of a goods’ attributes depends only on subjective weight component of each attribute of the good i.e., $\widetilde{\text{sw}}_{a_i}$. As buyer gains experience of buying a particular good with each subsequent purchase, the importance of its empirical weight component i.e. $\widetilde{\text{ew}}_{a_i}$ increases and the importance of $\widetilde{\text{sw}}_{a_i}$ decreases

proportionately. After sufficiently large number of purchases of the same good, say after 100 transactions for an α increment rate of 0.01, the importance of \widetilde{sw}_{a_i} becomes negligible. This means that after participating in sufficiently large number of transactions by the same buyer for a particular good, it is not necessary for a buyer to incur the overhead of computing the subjective weights of the goods' attributes and instead utilise the previous transactions weight information about the goods' attributes.

6. Obtain the list of reputed sellers who have offered the good to the buyer.
7. Solicit the buyers' perception of each seller's offer for the good in linguistic terms "P", "A", "G", "VG" or "EX" based on fuzzy membership scale of Fig. 3.
8. Using Table 2, map these linguistic terms into fuzzy performance ratings of the offers by different sellers for that good. Let pr_{ij} represents the performance ratings of a seller i for attribute j using trapezoidal fuzzy numbers. Fuzzy performance of each seller i , for $i = 1, 2, \dots, m$ and for each attribute j , for $j = 1, 2, \dots, n$ is represented by fuzzy attribute performance matrix \widetilde{PR} shown in (11).

$$\widetilde{PR} = \begin{bmatrix} \widetilde{pr}_{11} & \widetilde{pr}_{12} & \dots & \widetilde{pr}_{1n} \\ \widetilde{pr}_{21} & \widetilde{pr}_{22} & \dots & \widetilde{pr}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{pr}_{m1} & \widetilde{pr}_{m2} & \dots & \widetilde{pr}_{mn} \end{bmatrix} \tag{11}$$

According to (11), if seller 1's goods' performance with respect to attribute 2 is "VG" then its fuzzy performance according to Table 2 is $\widetilde{pr}_{12} = (7, 9, 10, 12)$.

Phase II:

9. Compute the fuzzy value of the seller i 's good \widetilde{fvs}_i as: $\widetilde{fvs}_i = \sum_{j=1}^n \widetilde{pr}_{ij} \widetilde{w}_{a_j}$. The fuzzy value matrix of each seller i 's good, for $i = 1, 2, \dots, m$ represented by \widetilde{FVS} is shown in (12) below.

$$\widetilde{FVS} = \begin{bmatrix} \widetilde{fvs}_1 \\ \widetilde{fvs}_2 \\ \vdots \\ \widetilde{fvs}_m \end{bmatrix} = \begin{bmatrix} \widetilde{pr}_{11} & \widetilde{pr}_{12} & \dots & \widetilde{pr}_{1n} \\ \widetilde{pr}_{21} & \widetilde{pr}_{22} & \dots & \widetilde{pr}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{pr}_{m1} & \widetilde{pr}_{m2} & \dots & \widetilde{pr}_{mn} \end{bmatrix} \times \begin{bmatrix} \widetilde{w}_{a_1} \\ \widetilde{w}_{a_2} \\ \vdots \\ \widetilde{w}_{a_n} \end{bmatrix} \tag{12}$$

Phase III:

10. Perform defuzzification on the resultant fuzzy matrix \widetilde{FVS} to obtain crisp value matrix CVS using Centre of Area (COA) approach.
11. Select the seller with the highest expected value of the good.

This methodology is relatively dynamic as in computing the weights of different attributes of a good from the buyers' perspective, it takes into cognizance the buyers' experience of buying a good, as with each successive purchase of the same good by the same buyer, the importance of empirical weight component from previous transactions increases and that of subjective weight component decreases.

As the importance of attributes of a good and the expected performance of the offers from various sellers are fuzzy in nature, this methodology keeps both of these

terms fuzzy except at the end where a crisp score is obtained to rate the sellers. Performing all computations in fuzzy terms rather than unnecessary conversion to crisp values prevents any loss of information due to unnecessary conversion.

4 Case Study

To illustrate the application of proposed reputation framework, a case study was conducted by simulating an electronic marketplace having four buyers and six sellers, i.e. $B = \{b_i \text{ where } i = 1..4\}$ and $S = \{s_i \text{ where } i = 1...6\}$, where B is the set of buyers and S is the set of sellers in the marketplace for a good g . An overview of the proposed model simulated in MATLAB is illustrated in Fig. 4 below.

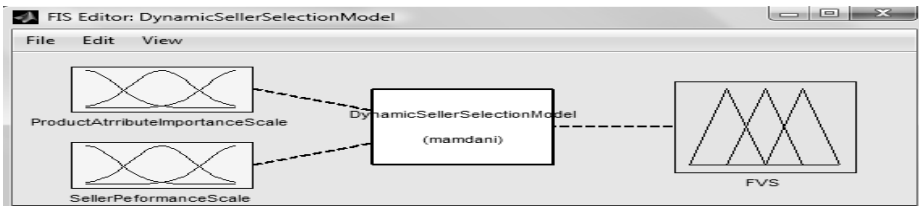


Fig. 4. An overview of Dynamic Seller Selection Model

A scenario was considered in e-market where buyer b_1 wanted to buy good g . The buyer b_1 specified the pairwise importance of different attributes of good g i.e. Price (P), Quality (Q), Delivery Period (DP) and Service Offered (SO) in linguistic terms. Their equivalent fuzzy values based on the fuzzy scale of Fig. 2 are illustrated in (13).

$$\begin{matrix} & & P & Q & DP & SO \\ \widetilde{FPM} = & P & (1,1,1) & (\frac{1}{5}, \frac{1}{3}, \frac{1}{3}, \frac{1}{1}) & (\frac{1}{7}, \frac{1}{5}, \frac{1}{5}, \frac{1}{3}) & (\frac{1}{5}, \frac{1}{3}, \frac{1}{3}, \frac{1}{1}) \\ & Q & (1, 3, 3, 5) & (1,1,1,1) & (3, 5, 5, 7) & (\frac{1}{5}, \frac{1}{3}, \frac{1}{3}, \frac{1}{1}) \\ & DP & (3, 5, 5, 7) & (\frac{1}{7}, \frac{1}{5}, \frac{1}{5}, \frac{1}{3}) & (1,1,1,1) & (\frac{1}{5}, \frac{1}{3}, \frac{1}{3}, \frac{1}{1}) \\ & SO & (1,3, 3, 5) & (1,3, 3, 5) & (1, 3, 3, 5) & (1,1,1,1) \end{matrix} \tag{13}$$

Using (6), the subjective weight of attribute price was computed as shown in (14).

$$\begin{aligned} \widetilde{sw}_P &= (1.543, 1.866, 1.866, 3.333) \times (\frac{1}{42.666}, \frac{1}{27.732}, \frac{1}{27.732}, \frac{1}{15.086}) \\ &= (0.0341, 0.0673, 0.0673, 0.1988) \end{aligned} \tag{14}$$

Similarly, subjective weights of other attributes were computed and are shown as SW in Fig. 5. Further, the average of the weights in the previous transactions were $\widetilde{ew}_P = (0.0405, 0.1265, 0.1265, 0.2435)$, $\widetilde{ew}_Q = (0.11, 0.44, 0.44, 0.87)$, $\widetilde{ew}_{DP} = (0.074, 0.196, 0.196, 0.443)$ and, $\widetilde{ew}_{SO} = (0.0725, 0.327, 0.327, 0.718)$ represented as EW in Fig. 5. Using (9) and (10), the resulting overall weights of different attributes of the good g as computed in MATLAB with $\alpha = 0.27$ are illustrated in Fig. 5 below.

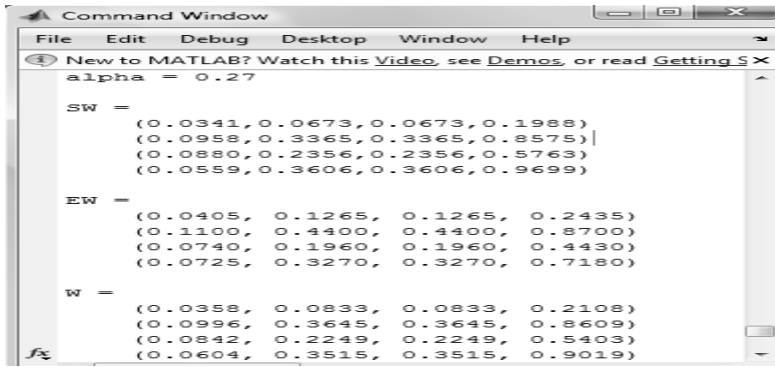


Fig. 5. Overall weight computation of attributes of good g by buyer b

Sellers s_1, s_3, s_4, s_6 , responded to sell good g to buyer b_1 . Now, buyer b_1 computed the expected value of the product being offered by the four sellers as explained below.

The buyers’ assessment of each seller’s product offer for the attributes Price (P), Quality (Q), Delivery Period (DP) and Service Offered (SO) in linguistic terms is shown in the Table 3.

Table 3. Buyers’ assessment of offers by different sellers in linguistic terms

↓Sellers / Attributes→	P	Q	DP	SO
s_1	VH	VH	H	A
s_3	H	H	VH	A
s_4	H	EX	H	H
s_6	EX	VH	H	A

By mapping linguistic terms of Table 3 to fuzzy values, fuzzy performance matrix \widetilde{PR} representing performance of various sellers’ offer for good g is shown in (15).

$$\widetilde{PR} = \begin{matrix} & & \begin{matrix} P & Q & DP & SO \end{matrix} \\ \begin{matrix} s_1 \\ s_3 \\ s_4 \\ s_6 \end{matrix} & \left[\begin{array}{cccc} (7, 9, 10, 12) & (7, 9, 10, 12) & (4, 6, 7, 9) & (1, 3, 4, 6) \\ (4, 6, 7, 9) & (4, 6, 7, 9) & (7, 9, 10, 12) & (1, 3, 4, 6) \\ (4, 6, 7, 9) & (10, 12, 13, 13) & (4, 6, 7, 9) & (4, 6, 7, 9) \\ (10, 12, 13, 13) & (7, 9, 10, 12) & (4, 6, 7, 9) & (1, 3, 4, 6) \end{array} \right] & (15) \end{matrix}$$

Using (12), \widetilde{FVS} was computed as, $\widetilde{FVS} = \widetilde{PR} \times \widetilde{W}$ and after defuzzifying the resultant crisp expected value (CVS) of the good g for each seller s_i for $i = 1, 3, 4, 6$, as computed using MATLAB is illustrated in Fig. 6.

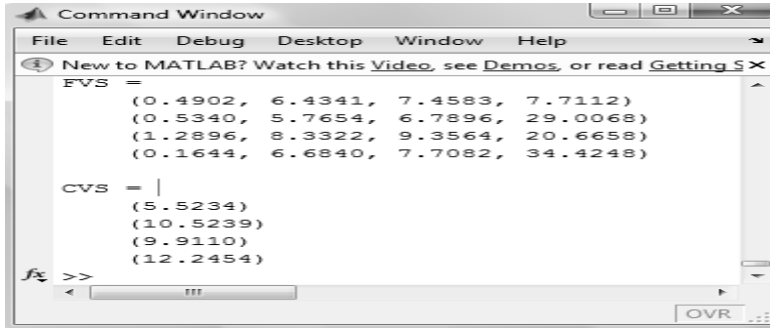


Fig. 6. Fuzzy (FVS) and Crisp (CVS) values of offers by different sellers

Finally, seller s_6 with the highest expected value of the good g equalling 12.2454 is selected by buyer b for purchase.

5 Conclusion

This paper proposed a dynamic seller selection methodology for buyer agents to enable a buyer to make an informed decision of buying a good by selecting a seller that offers a good with the highest expected value. The paper introduced a model that captures buyers' opinion in linguistic terms and map these terms to fuzzy values as it made managing imprecise and linguistic information feasible. The proposed model is relatively dynamic as it is sensitive to the changing experience of buyers in the e-market. Due to its dynamic nature, with each successive transaction of the same good by a particular buyer, the importance of empirical weight component increases and, that of subjective weight component decreases and becomes negligible after sufficiently large number of transactions. This ensures that after gaining sufficient experience of buying a good repeatedly, buyer can avoid the overhead of computing subjective weights of goods' attributes and instead utilise the weight information of different attributes from the previous transactions. Further, as the assessment of importance of different attribute of a good and the expected performance of the offers from different sellers are fuzzy in nature, this methodology performs all computations in fuzzy terms rather than unnecessary converting them to crisp value preventing any loss of information due to unnecessary conversion.

References

1. Ribeiro, R.A.: Fuzzy multiple attribute decision making: A review and new preference elicitation techniques. *Fuzzy Sets and Systems*, 155–181 (1996)
2. Abbasbandy, S., Amirfakhrian, M.: The Nearest Trapezoidal Form of a Generalized Left Right Fuzzy Number. *Int. J. Approx. Reas.* 43, 166–178 (2006)
3. Chou, C.-C.: The representation of multiplication operation on fuzzy numbers and application to solving fuzzy multiple criteria decision making problems. In: Yang, Q., Webb, G. (eds.) *PRICAI 2006. LNCS (LNAI)*, vol. 4099, pp. 161–169. Springer, Heidelberg (2006)

4. Yeh, C.-H., Chang, Y.-H.: Modelling subjective evaluation for fuzzy group multi-criteria decision making. *European Journal of Operational Research*, 464–473 (2009)
5. Mateos, A., Jiménez, A.: A Trapezoidal Fuzzy Numbers-Based Approach for Aggregating Group Preferences and Ranking Decision Alternatives in MCDM. In: Ehr Gott, M., Fonseca, C.M., Gandibleux, X., Hao, J.-K., Sevaux, M. (eds.) *EMO 2009. LNCS*, vol. 5467, pp. 365–379. Springer, Heidelberg (2009)
6. Chang, D.-Y.: Application of extent analysis method on fuzzy AHP. *European Journal of Operations research* 95, 649–655 (2006)
7. Dellarocas, C.: Reputation Mechanism Design in Online Trading Environments with Pure moral Hazard. *Information Systems Research* 16(2), 209–230 (2005)
8. Mulralidharan, C., Anantharaman, N., Deshmukh, S.G.: A Multi-criteria Group Decision Making Model for Supplier Rating. *The Journal of Supply Chain Management: A Global Review of purchasing and Supply*, 22–35 (2002)
9. Gaur, V., Sharma, N.K., Bedi, P.: Evaluating Reputation Systems for Agent Mediated e-Commerce. In: *ACEEE International Conference on Advances in Computer Science*, ACS 2010, Kerala, India (December 2010)
10. Wu, S., Wang, S., Liu, X.: A Seller Selection Model Study Based on Online Credit analysis in C2C e-Business environment. *IEEE*, Los Alamitos (2010)
11. Elachezhian, C., Vijaya Ramnath, B., Kesavan, R.: Vendor Evaluation using Multi-Criteria decision making Technique. *International Journal of Computer Applications* (0975-8887) 5(9) (August 2010)
12. Deng, S., Wortzel, L.H.: Importer Purchase Behaviour: guidelines of Asian Exporters. *Journal of business Research* 32, 41–47 (1995)
13. Murlidharan, C., Anantharaman, N., Deshmukh, S.G.: A Multi-Criteria Group Decision Making Model for Supplier Rating. *The Journal of Supply Chain Management*, 22–35 (2002)
14. Mohanty, R.P., Deshmukh, S.G.: Use of AHP for Evaluating Sources of Supply. *International Journal of Physical Distribution and Logistics Management*, 45–47 (1993)
15. Fenton, N., Wang, W.: Risk and confidence analysis of fuzzy multicriteria decision making. *Knowledge Based Systems* 19, 430–437 (2006)
16. Chris, E.I., Dunu, E., Gebremikael, F.: An Analysis of Strategic supplier Selection and Evaluation in a generic pharmaceutical firm supply chain. In: *ASBBS Annual Conference*, Las Vegas, vol. 17(1), pp. 77–91 (2010)
17. Mohanty, B.K., Bhasker, B.: Product classification in the Internet Business. *Decision Support Systems* 38, 611–619 (2005)
18. Kahraman, C., Cebeciand, U., Ulukan, Z.: Multi-criteria supplier selection using fuzzy AHP. *Logistics Information Management* 16, 382–394 (2003)
19. Timmerman, E.: An approach to vendor performance evaluation. *Journal of Purchasing and Materials management* 26(4), 2–8
20. Petroni, A.: Vendor Selection Using Principal Component Analysis. *The Journal of Supply Chain Management*, 63–69 (2000)

A Modified Ontology Based Personalized Search Engine Using Bond Energy Algorithm

Bhaskara Rao Boddu¹ and Valli Kumari Vatsavayi²

¹ Department of I T, G I T, GITAM University, Visakhapatnam, Andhra Pradesh-45, India
bhaskararaoboddu@gmail.com

² Department of CS&SE, College of Engg., Andhra University, Andhra Pradesh-03, India
vallikumari@gmail.com

Abstract. Search engines play an important function in increasing the speed of access to web information. As the volume of the content on the web is dynamically increasing, the general purpose web search engines are becoming inadequate. Since users with different backgrounds give queries in different contexts expecting different responses. The large number of irrelevant results returned by a search engine usually disappoints the user. The personalization of search engine overcomes this problem by ranking the results of web documents based on the inherent relations and closeness between the query concept and web document. In this paper we personalize the search engine using the Fuzzy Concept Network (FCN) and the Bond Energy Algorithm (BEA). The BEA calculates the closeness called affinity. Our main idea is to employ the concept network built based on the user's profile and ranks based on the Bond Energy Algorithm for searching and quickly retrieving of web pages. The advantage is that the most relevant search results can be retrieved. We examined our approach with data sets and prove our claims.

Keywords: Personalized Search Engine, Fuzzy Concept Network, Profile, Ranks, Affinity, Bond Energy Algorithm, and Ontology.

1 Introduction

Until recently, search engines were limited in their capability of searching the relevant pages as the web was built using languages and syntax rules which do not consider the semantics of the content. With the advent of semantic web technologies, personalized search engines [1] today are becoming intuitive, and have the capability to pull up results (web pages) based on semantics [2] of keywords and phrases. The advantage of semantic web is that, structures can be built and calculate the bonds [3] for the meaningful content of the web pages. Software agents that can roam from page to page can traverse these structures for information with well defined meaning. The semantic web [4] provides a common framework that allows data to be shared and reused across applications, enterprise and community boundaries. The imprecise and uncertain information comes from three major aspects in an information retrieval system environment including the representation of users' queries, the representations

of documents and the relevance relationship between both [5]. At present most of the business information retrieval systems implement the Boolean logic model [6]. The information retrieval systems based on the Boolean logic model are unable to represent uncertain information. Therefore if there is uncertain information, the query processing [7] of these systems is not handled properly. In order to represent and process the imprecise and uncertain information in information retrieval systems, the fuzzy set theory has been applied to information retrieval systems [8, 9]. Lot of fuzzy information retrieval methods based on the fuzzy set theory have been proposed which provide a sound mathematical framework to deal with the uncertainty of document representation, query specification, the document retrieval process and improves the disadvantages of the Boolean logic model. In this context, fuzzy concept networks are used in personalizing search engines [1]. The personalization process extracts users' preference and builds a profile for every individual user, and then it provides immediate response by sorting and re-ranking the results with respect to individual user's interest.

This paper proposes a system that searches web documents based on usage and fuzzy concept network. We can expect more quality results, because it searches based on the calculation of degree of relevance matrix and utilizes the fuzzy concept network to get enriched document description matrix [10, 11]. Fuzzy concept network calculates the relevance among concepts using fuzzy logic. The construction of fuzzy concept network is based on user profile and Bond Energy Algorithm [3]. It is more intelligent and more flexible than the existing methods since it can automatically construct and evaluate the knowledge base. It evaluates the degree that a document satisfies the users' queries by considering both context-independent relationships and context-dependent relationships between concepts. These concepts are involved in the documents and users' queries when the search contexts are considered. Search engine selects appropriate web pages for user by processing fuzzy document retrieval using fuzzy concept network as user knowledge. Fuzzy concept network and fuzzy document retrieval system can be used for effective personalized method. The rest of this paper is organized as follows. In section 2, the related work is introduced. Section 3 presents our work related to personal web search engine using fuzzy concept network with BEA. Section 4 presents personalization search results. Conclusions are discussed in section 5.

2 Related Works

In [12], Lucarella et al. proposed a knowledge-based document retrieval technique based on the fuzzy set theory [13], where the knowledge base is represented by a fuzzy concept network. Through the inference based on the links in a fuzzy concept network, the implicit relationships between concepts can be derived. However, since the fuzzy inference process must be performed every time when the users submit a query, the method presented in [9] is not efficient enough. The reason is that strengths between two concept nodes or between a concept node and a document node in fuzzy concept network are specified by experts. This assumption may be impractical when the application domain contains a large amount of concepts and documents.

In [6], Chen et al. used relevance matrices to model fuzzy concept networks. By calculating the transitive closure of a relevance matrix, the implicit relevance degrees between concepts could be obtained based on the context-dependent relationships. These methods assume that relationships between concepts are unchangeable in all cases. In fact, the relationship between two concepts may vary according to different contexts. For a more cooperative information retrieval system, the possible context-dependent relationships between concepts should also be discussed, and the proper relationship between concepts is adopted when the user specifies search contexts.

Chang and Chen [14] introduced a flexible query expression by combining range query and point query to solve the problem of Chen and Wang’s work. To support the conceptual query, they adopt the concept network to specify the relationships among the concepts. It fails to generalize the work for all queries. In all previous works the fuzzy concept matrix is constructed from the user profile that contains some of the relevance between concepts. In this paper, we use a new technique that creates an automatic concept network, so that the degree of relevance between concepts is computed automatically.

3 Our Work

In this section we generate an enriched fuzzy concept network [10, 11] using word net ontology. We then produce the concept of affinity [3] to rearrange the documents retrieved as to match the user’s interests as close as possible.

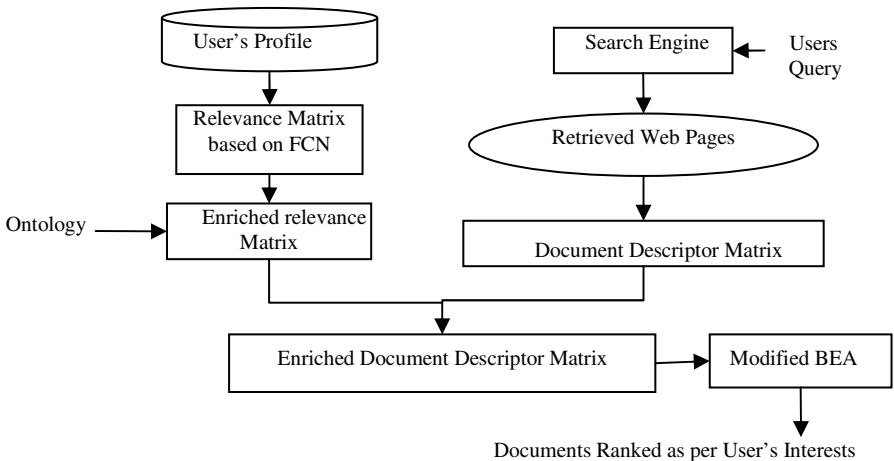


Fig. 1. The Work Flow of Modified Personalized Search Engine

WordNet [15] is a large lexical database in English that uses semantics and is used to get the Synsets (the set of Synonyms of the words), Hypernyms (the set of words that are more generalized than the current word), and Hyponyms (the set of words that are more specialized than the current word). The work flow of Modified Personalized

Search Engine (MPSE) is shown in fig. 1. A fuzzy concept network consists of both the concepts and documents connected by weighted edges each edge represents the degree of relevance between the two nodes represented by $\alpha \in [0,1]$. Each node represents a concept $c_i \in C$ or document $d_j \in D$. Each edge connects either two concepts or a concept c_i and a document d_j . If $c_i \xrightarrow{\alpha} c_j$ then it indicates that the degree of relevance from concept c_i to concept c_j is α . If $c_i \xrightarrow{\alpha} d_j$ then it indicates that the degree of relevance of document d_j with respect to concept c_i is α , where $\alpha \in [0,1]$. Expression $c_i \xrightarrow{\alpha} c_j$ is represented with $f(c_i, c_j) = \alpha$ and expression $c_i \xrightarrow{\alpha} d_j$ is represented with $g(c_i, d_j) = \alpha$ where f and g are mapping functions. In a concept network a document has a different relevance value with respect to each concept. Assume that the concept network shown in Figure 2 consists of four documents D_1, D_2, D_3, D_4 and seven concepts $C_1, C_2, C_3, \dots, C_7$. The relevant value of document D_2 with respect to concept C_4 can be calculated as follows: $\text{Min}(1.0, 0.9) = 0.9$ a route via C_1 , in the second route it is 0.2 and in the third route i.e., via C_2 it is $\text{Min}(0.4, 0.5) = 0.4$.

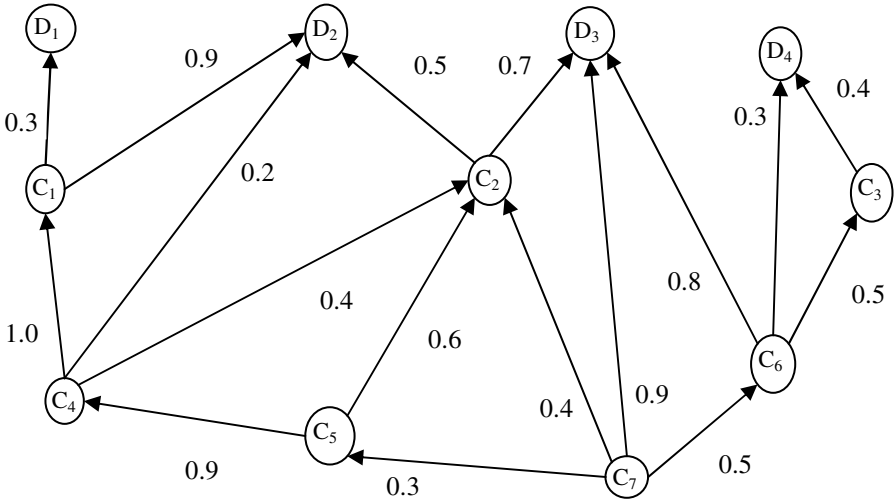


Fig. 2. Fuzzy Concept Network

Then, based on [12], we can see that the relevant value of the document D_2 with respect to the concept C_4 is: $\text{Max}(0.9, 0.2, 0.4) = 0.9$. Firstly, the user's profile along with the visited web pages is collected. After preprocessing the profile, a fuzzy

concept network is generated according to the predefined concepts vector. The degree of relevance for two concepts c_i and c_j is obtained from the formula:

$$f(c_i, c_j) = \frac{f_{c_i} + f_{c_j} - |f_{c_i} - f_{c_j}|}{N} \tag{1}$$

Where f_{c_i} is the frequency with which the concept c_i has appeared in the user’s profile and the f_{c_j} is the frequency with which the concept c_j has appeared in the user’s profile and N is the total number of concepts used in by the user along with the synsets, hyponyms, hypernyms and stem of each and every concept. The concepts from user profile are collected into a word vector called Enriched Profile(C) i.e.

$$\begin{aligned} w_1 &= \{c / \forall q \in Q \exists c, d \in q, q = c + d \text{ or } q = d + c\} \\ w_2 &= \{d \in S / \exists c \in q, \sigma(c, d) = 1 \text{ or } \sigma(d, c) = 1\} \\ w_3 &= \{d / \exists c \in q, d \leq c\} \\ w_4 &= \{d / \exists c \in q, d \geq c\} \\ C &= \bigcup_{i=1}^4 W_i \end{aligned} \tag{2}$$

Now $f(c_i, c_j)$ values are calculated from the enriched profile. These values are tabulated in a fuzzy concept matrix, R. The matrix R is called Degree of relevance between the concepts in C. Each value in the matrix represents the similarity between the corresponding concepts, where $1 \leq i \leq n$ and $1 \leq j \leq n$ and n is the total number of concepts.

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & \dots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \dots & \dots & r_{nn} \end{bmatrix} \tag{3}$$

Where r_{ij} is the degree of relevance between concept c_i and concept c_j and $r_{ij} = f(c_i, c_j)$ and $c_i, c_j \in C$ where C is the set of concepts obtained from user’s query. Similarly a document descriptor matrix V is tabulated such that each element represents the degree of relevance between the corresponding concept and document.

$$V = \begin{bmatrix} v_{11} & v_{12} & \dots & \dots & v_{1n} \\ v_{21} & v_{22} & \dots & \dots & v_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{n1} & v_{n2} & \dots & \dots & v_{nn} \end{bmatrix} \tag{4}$$

Where v_{ij} represents the degree of relevance between concept c_i and document d_j , $1 \leq i \leq n$ and $1 \leq j \leq m$, m is the number of documents retrieved. The value of v_{ij} is calculated as $v_{ij} = g(c_i, d_j)$, $\sigma: D \times C \rightarrow [0, 1]$. The degree of relevance, g is given by the formula

$$g(c_i, d_j) = \frac{\text{fre}_{d_j}(c_i)}{\sum \forall d_j \text{fre}_{d_j}(c_i)} \tag{5}$$

Where $\text{fre}_{d_j}(c_i)$ is the term frequency of concept c_i in document d_j , $\sum \forall d_j \text{fre}_{d_j}(c_i)$ is the term frequency of concept c_i in all the documents. $R^2 = R \Theta R$ is the multiplication of the fuzzy concept matrix [10] which results in a matrix where every element is evaluated as $v_{ij}^n(r_{ij} \wedge r_{ji})$, $1 \leq i, j \leq n$. \vee and \wedge represent the max and min operation respectively. Then, there exists an integer $\rho \leq n-1$, such that $R^\rho = R^{\rho-1} = R^{\rho-2} = R^{n-1}$. Let $R^* = R^\rho$, then R^* is called the transitive closure of the concept matrix. Missed information of fuzzy concept network can be inferred from the transitive closure of itself.

$$R^2 = R \Theta R = \begin{bmatrix} \bigvee_{i=1}^n (r_{li} \wedge r_{il}) & \bigvee_{i=1}^n (r_{li} \wedge r_{i1}) & \dots & \dots & \bigvee_{i=1}^n (r_{li} \wedge r_{in}) \\ \bigvee_{i=1}^n (r_{2i} \wedge r_{i1}) & \bigvee_{i=1}^n (r_{2i} \wedge r_{i2}) & \dots & \dots & \bigvee_{i=1}^n (r_{2i} \wedge r_{in}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \bigvee_{i=1}^n (r_{ni} \wedge r_{i1}) & \bigvee_{i=1}^n (r_{ni} \wedge r_{i2}) & \dots & \dots & \bigvee_{i=1}^n (r_{ni} \wedge r_{in}) \end{bmatrix} \tag{6}$$

The relevance degree of each document, with respect to a specific concept, can be improved by computing the multiplication of the document descriptor matrix V and the transitive closure of concept matrix R as follows: $V^* = R^* \Theta V$ Where V^* is called the Enriched Document Descriptor matrix

$$R^* \Theta V = V^* = \begin{bmatrix} \bigvee_{i=1}^n (r_{1i} \wedge v_{i1}) & \bigvee_{i=1}^n (r_{1i} \wedge v_{i2}) & \dots & \dots & \bigvee_{i=1}^n (r_{1i} \wedge v_{im}) \\ \bigvee_{i=1}^n (r_{2i} \wedge v_{i1}) & \bigvee_{i=1}^n (r_{2i} \wedge v_{i2}) & \dots & \dots & \bigvee_{i=1}^n (r_{2i} \wedge v_{im}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \bigvee_{i=1}^n (r_{ni} \wedge v_{i1}) & \bigvee_{i=1}^n (r_{ni} \wedge v_{i2}) & \dots & \dots & \bigvee_{i=1}^n (r_{ni} \wedge v_{im}) \end{bmatrix} \tag{7}$$

Bond Energy Algorithm

Now, using the enriched document descriptor matrix, the Initial Affinity Vector (I) is generated. The initial affinity vector consists of affinity values of each and every retrieved document with the concepts obtained from user’s query. The affinity value is calculated with respect to the user’s interests collected from the user profile. Affinity [3] measures the closeness of the document with every concept in user’s query altogether. Generation of Initial affinity Vector is done in two steps:

1. Initialization: The document descriptor matrix is taken as input. The size of the input matrix is N x M. where N indicates number of concepts and M indicates number of concepts.
2. Affinity Calculation: For every document, affinity is calculated. These affinity values constitute the Initial affinity Vector. The initial retrieval of documents is rearranged so as to match the user’s interests. The affinity value is calculated as follows:

$$aff_{d_j} = \sum_{\bigvee_{c_i}} w_{c_i} v_{ij} \tag{8}$$

Where w_{c_i} Represents the weight of the concept c_i as per the user’s previous searches and is given as :

$$w_{c_i} = \frac{f_{c_i}}{N} \tag{9}$$

The affinity value of a document with the concepts obtained from user's query is calculated with respect to the user's interests collected from the user profile. Affinity measures the closeness of the document with every concept in user's query altogether. The affinity vector is obtained where each element represents the affinity values in the order of the documents retrieved.

$$I = (aff_{d_1}, aff_{d_2}, aff_{d_3}, \dots \dots, aff_{d_m}) \quad (10)$$

Here m represents number of documents. Now, the affinity vector is sorted using a sorting algorithm in such a way that the document with the highest affinity value comes to the first position and the documents with the next highest affinities will follow in order. The sorted affinity vector is represented as F.

Algorithm 1. Affinity Calculation

Input: Initial Affinity Vector (I).

Output: Generating Ranking Vector (F)

Method: Affinity is Calculating as follows

1. **Begin**
2. **For** each d_k in Affinity Vector I
3. Find a Document d_j that satisfies the condition

$$aff_{d_j} \geq aff_{d_k} \quad \forall k = \{1, 2, 3 \dots, j-1, j+1 \dots m\}$$

4. Insert d_j into final ranking vector (F) at position i.
 5. $I = I + 1$
 6. Remove d_j from Initial Affinity Vector (I).
 7. **end for**
 8. **end Begin**
-

If a threshold value is specified, then all the documents having their affinity value greater than the threshold value are presented to the user as per their ordering in the Final ranking Vector F.

4 Experimental Analysis

We have tested this with a predefined user profile, {matter, solids, sea}. Then the refined enriched profile of the above user profile is {(matter, 3), (solids, 2), (sea, 1), (liquid, 2), (gas, 1), (substance, 3), (medium, 1), (issue, 1), (shape, 2), (water, 1)}. The degree of relevance between the concepts (R) and document descriptor matrix (V) are calculated below.

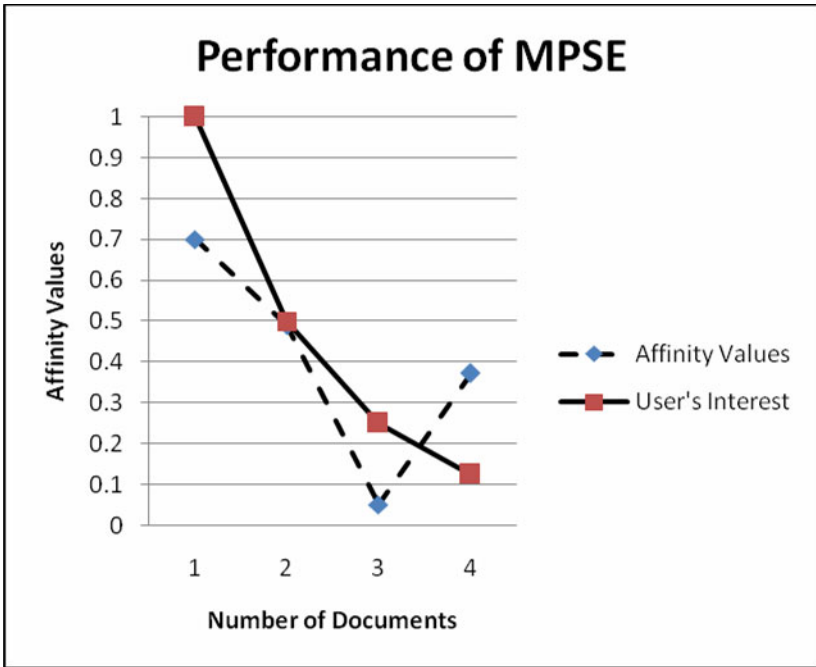
$$R = \begin{matrix} & C_1 & C_2 & C_3 & C_4 & C_5 & C_6 & C_7 & C_8 & C_9 & C_{10} \\ C_1 & 0.6 & 0.4 & 0.2 & 0.4 & 0.2 & 0.6 & 0.2 & 0.2 & 0.4 & 0.2 \\ C_2 & 0.4 & 0.4 & 0.2 & 0.4 & 0.2 & 0.4 & 0.2 & 0.2 & 0.4 & 0.2 \\ C_3 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ C_4 & 0.4 & 0.4 & 0.2 & 0.4 & 0.2 & 0.4 & 0.2 & 0.2 & 0.4 & 0.2 \\ C_5 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ C_6 & 0.6 & 0.4 & 0.2 & 0.4 & 0.2 & 0.2 & 0.2 & 0.2 & 0.4 & 0.2 \\ C_7 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ C_8 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ C_9 & 0.4 & 0.4 & 0.2 & 0.4 & 0.2 & 0.4 & 0.2 & 0.2 & 0.4 & 0.2 \\ C_{10} & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{matrix}$$

$$R^* = \begin{matrix} & C_1 & C_2 & C_3 & C_4 & C_5 & C_6 & C_7 & C_8 & C_9 & C_{10} \\ C_1 & 0.6 & 0.4 & 0.2 & 0.4 & 0.2 & 0.6 & 0.2 & 0.2 & 0.4 & 0.2 \\ C_2 & 0.4 & 0.4 & 0.2 & 0.4 & 0.2 & 0.4 & 0.2 & 0.2 & 0.4 & 0.2 \\ C_3 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ C_4 & 0.4 & 0.4 & 0.2 & 0.4 & 0.2 & 0.4 & 0.2 & 0.2 & 0.4 & 0.2 \\ C_5 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ C_6 & 0.6 & 0.4 & 0.2 & 0.4 & 0.2 & 0.6 & 0.2 & 0.2 & 0.4 & 0.2 \\ C_7 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ C_8 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ C_9 & 0.4 & 0.4 & 0.2 & 0.4 & 0.2 & 0.4 & 0.2 & 0.2 & 0.4 & 0.2 \\ C_{10} & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{matrix}$$

$$V = \begin{matrix} & D_1 & D_2 & D_3 & D_4 \\ C_1 & 0.23 & 0.71 & 0.03 & 0.03 \\ C_2 & 0 & 0.63 & 0.35 & 0.02 \\ C_3 & 0 & 1 & 0 & 0 \\ C_4 & 0 & 0.97 & 0 & 0.03 \\ C_5 & 0 & 0.86 & 0.14 & 0 \\ C_6 & 0 & 1 & 0 & 0 \\ C_7 & 0 & 1 & 0 & 0 \\ C_8 & 1 & 0 & 0 & 0 \\ C_9 & 0 & 1 & 0 & 0 \\ C_{10} & 0 & 0.66 & 0.33 & 0 \end{matrix}$$

$$V^* = \begin{matrix} & D_1 & D_2 & D_3 & D_4 \\ C_1 & 0.23 & 0.6 & 0.35 & 0.03 \\ C_2 & 0.23 & 0.4 & 0.35 & 0.03 \\ C_3 & 0.2 & 0.2 & 0.14 & 0.03 \\ C_4 & 0.23 & 0.4 & 0.35 & 0.03 \\ C_5 & 0.2 & 0.2 & 0.14 & 0.03 \\ C_6 & 0.23 & 0.6 & 0.35 & 0.03 \\ C_7 & 0.2 & 0.2 & 0.14 & 0.03 \\ C_8 & 1.2 & 0.2 & 0.14 & 0.03 \\ C_9 & 0.23 & 0.4 & 0.35 & 0.03 \\ C_{10} & 0.2 & 0.2 & 0.14 & 0.03 \end{matrix}$$

$I = (0.376, 0.7, 0.49, 0.0510)$ and Final Vector $(V) = (D_2, D_3, D_1, D_4)$. Ranks as per user's Interest is $(U) = (D_2, D_3, D_4, D_1)$ and Ranks as per Google is (D_1, D_2, D_3, D_4) . The vectors V and U are very close to each other i.e., we have succeed to retrieve the documents that are needed by the user.



5 Conclusions

To personalize search engine results, the proposed system used the idea of enriched fuzzy concept networks. In order to accomplish this, we employed the concepts of ontology on the User’s profile. The ranking results using the enriched fuzzy concept networks and also the modified Bond Energy Algorithm are more relevant with user’s interests than the obtained ranking using common fuzzy concept networks. Using multi-agent systems one can optimize performance of the system. Further improvement is to minimize the search result by using the concept of similarity comparison between documents which in-turn increases the relevance of the search results as per the user’s interest.

References

1. Pretschner, A., Gauch, S.: Ontology Based Personalized Search. *Web Intelligence and Agent Systems* 1(3-4) (December 2003)
2. Bhaskara Rao, B., Valli Kumari, V., Raju, K.: Semantic Similarity Computation: Ant colony Optimization Algorithm Using Ontology. In: *ICINC 2010*, pp. V2–199. IEEE, Los Alamitos (2011) 978-1-4244-8271-9/10
3. Tamer Ozsu, M., Valduriez, P.: *Principles of Distributed Database Systems*, 131–150
4. Matthews, B.: Semantic web technologies. *JISC Technology and Standards Watch*, CCLRC Rutherford Appleton Laboratory
5. Chen, S.M., Horng, Y.J., Lee, C.H.: Fuzzy Information Retrieval based on Multi-relationship Fuzzy Concept Networks. *Fuzzy Sets and Systems* 140, 183–205 (2003)

6. Chen, S.M., Wang, J.Y.: Document Retrieval Using Knowledge-based Fuzzy Information Retrieval Techniques. *IEEE Trans. Syst. Man Cybern.* 25, 793–803 (1995)
7. Widiantoro, D.H., Yen, J.: Using Fuzzy Ontology for Query Refinement in a Personalized Abstract Search Engine. *IEEE, Los Alamitos* (2001); ISBN 0-7803-3/01
8. Chen, S.M., Horng, Y.J., Lee, C.H.: Document Retrieval Using Fuzzy Valued Concept Networks. *IEEE Trans. Syst. Man Cybern.* 31, 111–118 (2001)
9. Kim, K.J., Cho, S.B.: A Personalized Web Search Engine Using Fuzzy Concept Network with Link Structure. In: *Proc. IFSA World Cong. NAFIPS Conf.*, pp. 81–86 (2001)
10. Akhlaghian, F., Arzanian, B., Moradi, P.: A Personalized Search Engine Using Ontology Based Fuzzy Concept Networks. In: *International Conference on Data Storage and Data Engineering, DSDE 2010*, pp. 137–141. *IEEE CS Digital Library, Los Alamitos* (2010)
11. Akhlaghian, F., Arzanian, B., Moradi, P.: A Multi-agent Based Personalized Meta-search Engine Using Automatic Fuzzy Concept Networks. In: *Proceedings 3rd International Conference on Knowledge Discovery and Data Mining, DSDE*, pp. 208–211 (2010) 978-0-7695-3923-2/10
12. Lucarella, D., Morara, R.: FIRST: Fuzzy information retrieval system. *Journal of Information Science* 17(1), 81–91 (1991)
13. Zadeh, L.A.: Fuzzy sets. *Information and Control* 8, 338–353 (1965)
14. Chang, C.-S., Chen, A.L.P.: Supporting Conceptual and Neighborhood Queries on the World Wide Web. *IEEE Trans. Man Cybern* 28, 300–308
15. WordNet. A lexical database for English. Princeton University, Princeton, <http://www.wordnet.princeton.edu>

A Client Perceived Performance Evaluation of Web Servers

Ash Mohammad Abbas and Ravindra Kumar

Department of Computer Engineering,
Zakir Husain College of Engineering and Technology,
Aligarh Muslim University, Aligarh -202002, India
am.abbas.ce@amu.ac.in, ravi_amu2002@rediffmail.com

Abstract. In this paper, we evaluate the performance of Web Servers as perceived by a client. We evaluate the performance of Web Servers in case when (i) there is no data flowing between a Web Server and a client, and (ii) there is data flowing from the Web Server to the client. The performance parameters that we focused are: round trip latencies, access rate, and the connection throughput.

Keywords: Client perceived performance, throughput, access rate, latencies.

1 Introduction

It is increasingly important to measure the performance of a Web Server. However, dynamic traffic conditions, and different types of configurations of edge routers, core routers and access networks, make the evaluation of the performance of Web Servers a challenging task.

A lot of research is directed towards the measurement of performance of Web Servers. In [1], there is a description of how to generate the load for measuring the performance of multi-tier software systems. In [2], a testbed for analyzing the performance of pseudo-serving for congestion control in temporarily busy servers is presented. In [3], a tool is presented for measuring the performance of Web Servers. User perceived performance of Web Servers is described in [5]. A description of measuring the performance of Apache Web Servers is presented in [6].

The performance of a Web Server measured by different clients that are remote to a Web Server may be different depending upon their requirements and depending upon the configuration of the client, edge router, etc. In other words, the performance of a Web Server measured by a client is the performance of the server as viewed by the client, and we call it to be the client perceived performance.

In this paper, we are mainly concerned with the measurement of client perceived performance of Web Servers. Specifically, we are interested in how a Web Server behaves in terms of the round trip latencies, connection throughput, and the access rate for the contents that are hosted on the Web Server.

The rest of the paper is organized as follows. In Section 2, we present the measurement of round trip latencies. In Section 3, we describe the measurement of connection throughput. In Section 4, we describe the measurement of access rate. In Section 5, we conclude the paper.

2 Measurement of Latencies

Measurement of latencies or delays is an important step in analyzing the performance of any Web Server. In this section, we present methods for measuring the client perceived latencies of different Web Servers.

To measure the performance of a Web Server, one needs to measure the delays or latencies involved from a client to the Web Server. We call these latencies or delays as client perceived latencies as these latencies will be different for different clients, and will depend upon many factors that constitute the client side including the bandwidth of the link through which the client is connected to the Internet, the edge router, etc. Therefore, the delay measured by a client is the delay that is perceived by the client or the delay that is visible to the client.

In what follows, we present methods to measure the client perceived latencies.

2.1 Methodology for Latency Measurement

The latencies can be measured using a tool such as Ping. The standard Ping tool uses ICMP as the underlying protocol. We developed a perl script so that in Ping one may specify and use either of ICMP, TCP or UDP as the underlying protocol. The client generates a Ping request by specifying the protocol and the either the URL or IP address of the Web Server. The client sends the Ping request to the Web Server, sets the timeout, and starts the timer. The client waits for a Ping response from the Web Server until the timeout. If the Ping response from the Web Server is received before the timeout, the client stops the timer and computes the difference between the time of sending Ping request and time of receiving Ping response. The client then generates the statistics for the Web Server. Otherwise, the client simply generates a message such as "Request Timeout". The Web Server receives the Ping request, determines the protocol specified in the Ping request. It then generates a Ping response and sends towards the client using the protocol specified in the Ping request.

The modeling of latencies depends on the transport protocol used. For example, consider TCP as the transport layer protocol. Let RTT be the round trip time, O be the size of the ping request/response message, P be the number of times TCP idles at the server, S be the maximum segment size (MSS). Let R be the bandwidth of the link connecting the client and the server. Then, an expression of latencies can be written as follows (for a description of how to derive it, refer to [9]).

$$\delta = 2RTT + \frac{O}{R} + P \left[RTT + \frac{S}{R} \right] - (2^P - 1) \frac{S}{R}. \quad (1)$$

This expression is written here only for the sake of understanding of how one can model the latencies. The actual modeling of the latencies incurred is little more involved. For example, one has to consider all links between the client and the remote server and their bandwidths might be different. For the purpose of simplifying the analysis, one can use the minimum of the bandwidths of the links from the client and the server as the value of R in (1).

2.2 Performance Results for Latencies

We measured latencies from the point when a client sends a Ping request to a Web Server till it receives the Ping response from the Web Server.

Figure 1 shows a comparison of the latencies for different Web Servers using Ping with different transport layer protocol such as TCP, UDP, and ICMP. From Figure 1, we observe that the latencies in case of the Ping with UDP are the smallest, and the latencies in case of the Ping with TCP are the largest. The reason is that TCP provides a connection-oriented transport service using a procedure of handshaking while UDP provides a connectionless transport service. The process of handshaking consumes one *Round Trip Time (RTT)*. Therefore, delays or latencies in case of UDP ping are much less.

The latencies in case of Ping with ICMP are in larger than those of UDP Ping, and are smaller than those of TCP Ping. Note that the Ping that is often used is based on ICMP. In case of ICMP, packets are handled differently as compared to TCP or UDP. ICMP packets get lower priority. If the network is congested, routers may be configured to ignore ICMP packets entirely. For this reason ICMP is not accurate in measuring the latency of remote machine. But it is a good first approximation.

3 Measurement of Access Rate

In this section, we discuss how a client can measure the rate of access of the contents that are lying on the Web Server. Using access rate, one can evaluate how fast one can access the contents from a Web Server. The access rate is defined to be the rate at which the client receives the content that are lying on the Web Server. The rate at which a client retrieves a file from the Web Server depends upon how fast is the client and its access network. Therefore, the access rate for the same file stored on a Web Server may be different for different clients. This is the reason that we call the access rate as client perceived access rate.

In what follows, we describe how one can measure the client perceived access rate.

3.1 Methodology for Access Rate Measurement

To measure the access rate client generates HTTP request specifying the URL/IP Address of the Web Server. At the time, when the HTTP request is sent to the Web Server, the client starts the timer. When the HTTP request reaches the Web Server, it

receives the request at port 80. It then examines the request whether it contains a GET method or a HEAD method. If the HTTP request contains the GET method, the Web Server generates an HTTP response, appends the desired file to it, and sends the response to the client. Otherwise, if the HTTP request contains a HEAD method, the Web Server generates an HTTP response without appending the file, and sends the HTTP response to the client.

When the client receives the HTTP response from the Web Server, it stops the timer. It then computes the time duration between the instant when it sent the HTTP request to the Web Server and when it received the HTTP response from the Web Server. The access rate is the file size divided by the time duration. Let F_s be the file size, and Δ be the time difference between the time instants when a request is fired until the response is received. The access rate, r_a , is given by

$$r_a = \frac{F_s}{\Delta}. \quad (2)$$

Note that Δ is the overall time starting from the firing of HTTP request by the client until the client receives the HTTP response from the web server. As the underlying transport protocol for HTTP is TCP, Δ includes the time taken incurred in handshaking, the time incurred in the slow start phase of TCP, queuing delays at the intermediate routers, propagation delays, etc.

To compute the access rate, the size of the file contained in HTTP response is required. The size of the Web page is contained in the content length field of the header of the HTTP response.

In what follows, we present results and discussion.

3.2 Performance Results for Access Rate

We considered web sites of some of the Indian Universities and some of the Universities of United States with the viewpoint that the web sites of universities are comparatively static. By the word “static”, we mean that the contents do not change frequently. Also, we considered web sites of several news channels within India and abroad that can be considered to be relatively dynamic sites. By the word “dynamic”, we mean that the contents of the web sites often change frequently.

Figure 2 through Figure 5 show the access rate for different Web Servers. The Web page sizes of the Web Servers are in an increasing order. As a result, one may correlate the access rate with the size of the Web pages lying on Web Servers. Each point represents an average of 10 runs. Note that as the size of the Web page increases, there is no specific trend observed for the access rate.

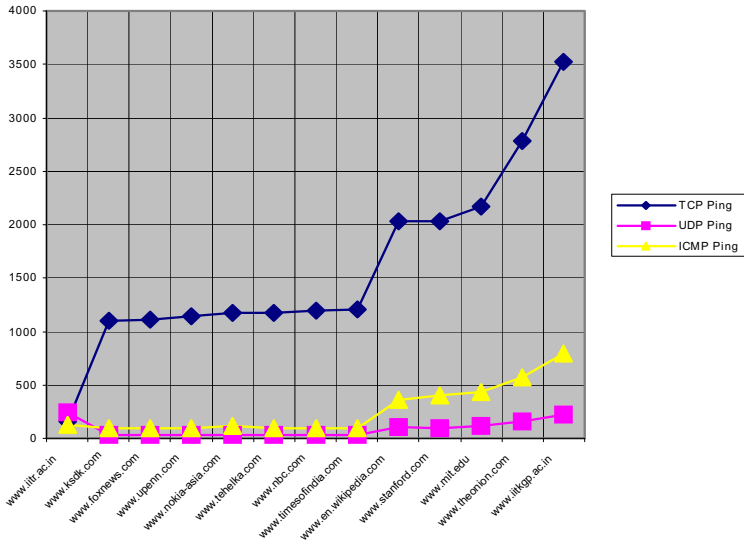


Fig. 1. A comparison of latencies for different Web Servers using Ping with different underlying protocols

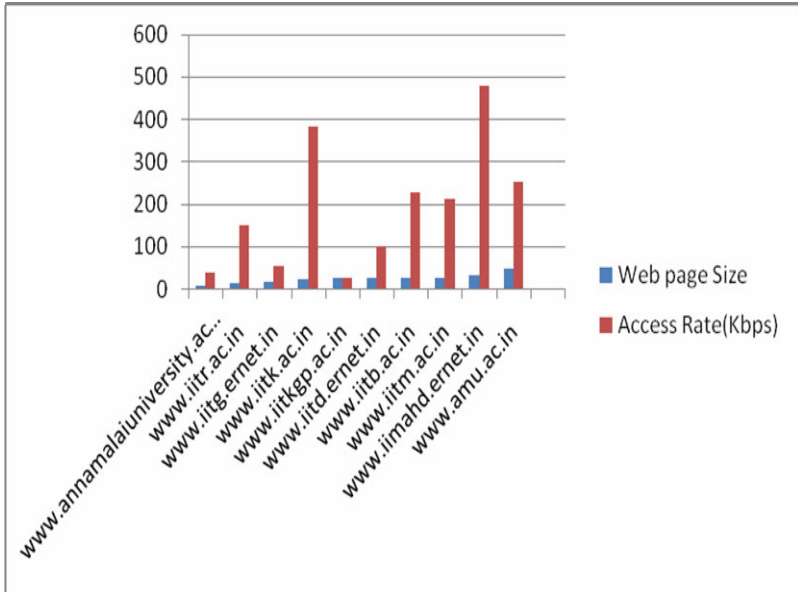


Fig. 2. Access rate versus Web page size for Indian Universities

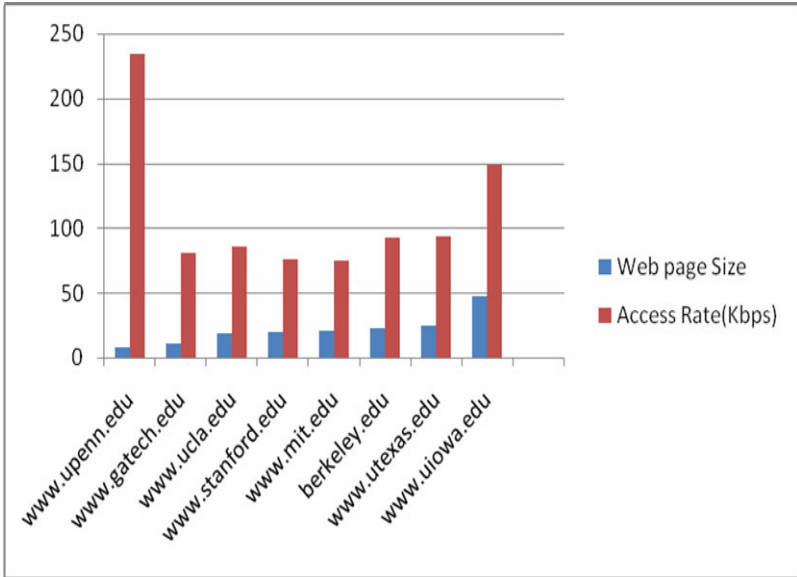


Fig. 3. Access rate versus Web page size for US Universities

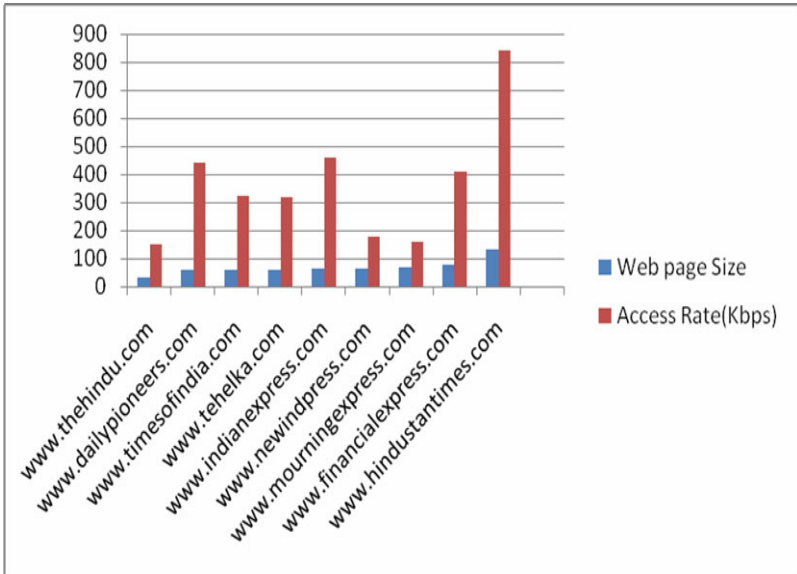


Fig. 4. Access rate versus Web page size for Indian news papers

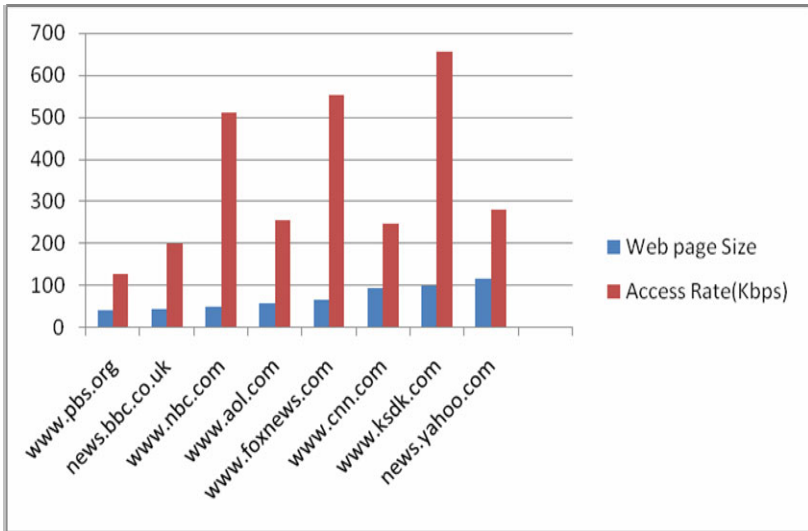


Fig. 5. Access rate versus Web page size for US news channels

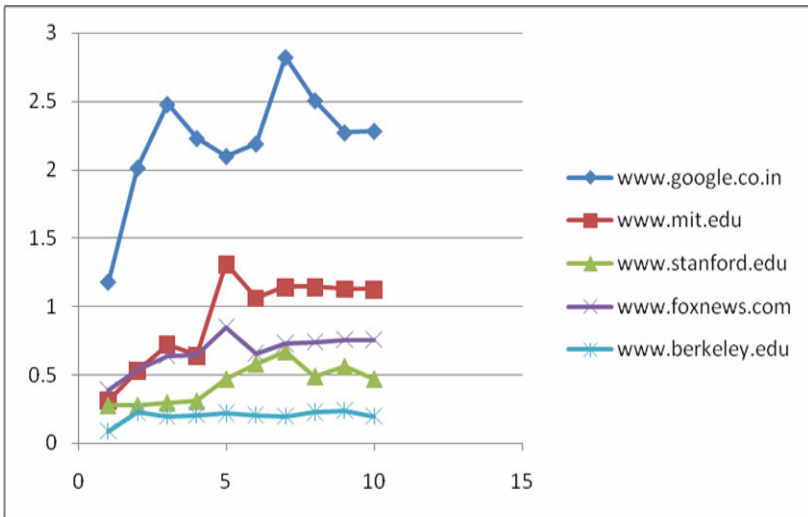


Fig. 6. Comparison of connection throughput versus number of parallel connections for different Web Servers

We observe that the web page sizes actually measured are fairly close to those provided by the Google. All Web Servers are listed in an increasing order of their Web page size. When the size of the Web page increases, there is no specific trend for the round trip latencies and the access rate. This is due to the reason that the round trip latencies and access rates depend not only on the size of the Web page but also on

some other factors such as the speed of the access network, edge router and congestion in the network. As a result, there is no specific trend observed for round trip latencies as well as access rates of Web Servers.

In what follows, we describe how one can measure the performance of a Web Server when a client tries to overload the Web Server.

4 Measurement of Connection Throughput

In this section, we discuss the effect on the performance of a Web Server when the load on the Web Server increases. The client can overload a server by firing requests for multiple connections. We call the number of successful connections per unit time as the connection throughput. In other words, connection throughput is defined to be the number of connections per unit time that are accepted by a Web Server and corresponding responses are sent to the client.

4.1 Methodology for Connection Throughput Measurement

To measure the connection throughput, we send multiple HTTP requests to initiate parallel connections to a Web Server. Note that the underlying transport layer protocol for HTTP is TCP. At the instant of time, when multiple requests to establish multiple connections are sent by the client, a timer is started. The Web Server to which the requests are destined may either accept all requests or drops some of them depending upon the load on the Web Server. If the Web Server accepts a request, it forms a response and sends the response towards the client. The response contains a status code and a status phrase “200 OK” that designates that the request to make a connection with the client is accepted by the Web Server. In case, some of the requests out of those sent by the client are not accepted by the Web Server, then the Web Server sends a response that contains a status code and a status phrase such as “304 Bad Request”. The client measures the time from the instance when it sent requests to make connections with the Web Server till it receives all responses from the Server.

We would like to mention that since the client is sending requests to initiate TCP connection with the Web Server and TCP provides a reliable transport service, therefore, there is no question of a response being dropped between the client and the Web Server. When the client receives the last response from the Web Server, it computes the difference of times of sending requests towards the Web Server and receiving replies from the Web Server. The connection throughput is computed to be the number of successful connections divided by the time. Let the number of successful connections accepted by the server be C_s , and let T be the time taken from sending requests by the client towards the server and receiving replies from the server at the client. Then, connection throughput is computed by

$$\eta = \frac{C_s}{T}. \quad (3)$$

As mentioned earlier, the time T may depend on a number of factors such as the transport layer protocol, time incurred in handshaking, slow start phase (in case of TCP is used as the transport layer protocol), queuing at the routers, and propagation delays, etc.

4.2 Performance Results for Connection Throughput

Figure 6 shows a comparison among the connection throughput as a function of number of parallel connections for different Web Servers. We observe that as the number of parallel connections is increased, the connection throughput first increases and then becomes almost constant. This is due to the fact that, increasing the number of parallel connections beyond a certain limit increases the load on the Web Server and a overloaded Web Server may not respond quickly as compared to the scenario when the number of parallel connections are only a few.

We observe that the connection throughput for the Google is the largest. The reason is that the clients that we used for performance evaluation are geographically close to Google Web Servers as compared to other Web Servers. Therefore, propagation delays for Google are far less as compared to other Web Servers. As a result, the Google Web Server receives HTTP requests sent by the client earlier than the requests of other Web Servers. Similarly, the client receives HTTP responses sent by the Google Web Server earlier than the responses sent by other Web Servers.

In the next section, we conclude the paper.

5 Conclusion

In this paper, we evaluated client perceived performance of Web servers. The performance parameters that we focused on are: round trip latencies, access rate, and the connection throughput. Our contributions are as follows.

- We measured the round trip latencies of different Web Servers using Ping with TCP, UDP and ICMP as underlying protocols. We observed that the performance of Ping with UDP as an underlying protocol comes out to be the best of all three protocols.
- We measured access latencies from the point when a client sends a request for a file to the Web Server till the client receives the desired file.
- We analyzed the performance of different Web Servers in terms of connection throughput. We observed that the connection throughput initially increases with increasing the number of parallel connections and beyond a certain limit it becomes almost constant as the Web server becomes overloaded.

In future, one may simulate the scenarios at the client side as well as at the server side in terms of link bandwidth, access network, characteristics of edge and core routers, etc., and one may compare the results obtained through simulations with those obtained experimentally.

References

1. Shirodkar, S., Apte, V.: AutoPerf: An Automated Load Generator and Performance Measurement Tool for Multi-tier Software Systems. In: Proceedings of World Wide Web Conference (WWW), pp. 1291–1292. ACM Press, New York (2007)
2. Abbas, A.M., Naqvi, S.H.R.: JServRMRS: A Testbed for Performance Analysis of Pseudo-Serving for Congestion Control in Temporarily Busy Web Servers. In: Proceedings of IEEE Asia-Pacific Conference on Communications (APCC), pp. 161–165. IEEE Press, Piscataway (2003)
3. Mosberger, D., Jin, T.: httpperf: A Tool for Measuring Web Server. In: Proceedings of IEEE Workshop on Internet Server Performance (WISP), pp. 59–67. IEEE Press, Piscataway (1998)
4. Iyengar, A., MacNair, E., Nguyen, T.: An Analysis of Web Server Performance. In: Proceedings of IEEE Global Telecommunication Conference (GLOBECOM), pp. 1943–1947. IEEE Press, Piscataway (1997)
5. Dalal, A.C., Jordan, S.: Improving the User Perceived Performance at a World Wide Web Server. In: Proceedings of IEEE Global Telecommunication Conference (GLOBECOM), pp. 2465–2469. IEEE Press, Piscataway (2001)
6. Hu, Y., Nanda, A., Yang, Q.: Measurement, Analysis and Performance Improvement of the Apache Web Server. In: Proceedings of IEEE International Performance, Computing and Communications Conference (IPCCC), pp. 261–267. IEEE Press, Piscataway (1999)
7. Rhee, Y.J., Choi, C.W., Kim, T.W., Kim, T.Y.: Client-side Mechanism for Improving Busy Web Server Performance. In: Proceedings of International Conference on Information Technology and Information Networks (ICII), pp. 95–99. IEEE Press, Piscataway (2001)
8. Krishnamurthy, B., Wills, C.E.: Improving Web Server Performance by Client Characterization Driven Server Adaptation. In: Proceedings of World Wide Web Conference (WWW), pp. 305–316. ACM Press, New York (2002)
9. Kurose, J.F., Ross, S.M.: Computer Networking: A Top-Down Approach Featuring the Internet, 5th edn. Pearson Education, Boston (2010)
10. Abbas, A.M., Sharma, V.S., Jain, A.: A Domain Based Prioritized Model for Web Servers. In: Proceedings of 9th IEEE International Conference on Information Technology (ICIT), pp. 265–268. IEEE Press, Piscataway (2006)
11. Kumar, R.: Client Perceived Performance of Web Servers. M.Tech. Dissertation, Department of Computer Engineering. Aligarh Muslim University, Aligarh (2008)

Enhanced Quality of Experience through IVR Mashup to Access Same Service Multiple Operator Services

Imran Ahmed and Sunil Kumar Kopparapu

TCS Innovation Labs Mumbai, Tata Consultancy Services Limited, Yantra Park,
Thane(W) 400 601, India

{ahmed.imran,sunilkumar.kopparapu}@tcs.com

<http://www.tcs.com>

Abstract. On one hand a telephone voice call has become an easy and a convenient means to inquire, seek information or book services and on other hand any frequently required service by the masses, like Taxi usually has multiple service operators. While each of these service operators provide the same type of service to the end user, their access points are different and require the end user to choose a priori which service operator to call. The Quality of Experience (QoE) faces major shortcomings in such a Same Service Multiple Operator (SSMO) scenario in terms of (a) long hold times (b) no way the user can connect to the the service operator who can provide him/her, the service he/she needs, fastest. Unlike a web portal or a web mashup, there is no way for the user to comparatively and simultaneously check the offerings from SSMO and choose the best among them. In this paper, we propose an interface which enables an enhanced QoE in a SSMO scenario while using the telephone channel which is achieved through a IVR Mashup [1].

Keywords: Mashup, Hold time, Telequeues, Call Center.

1 Introduction

Services like Radio-Cabs, Travel Portals, Movie-show Bookings, local Yellow Pages search etc are becoming popular and users still prefer to be serviced through a telephone call. Statistics shows that 80% of a firm's customer interaction is through call centers [2]. At the same time any popularly used service, say Taxi-on-call, usually has multiple service operators. Though these different service operators provide the same type of service to the end user, the end user has to call a different service number to make use of the service of that particular service operator. This essentially means even if the end user does not have a preferred service operator, to be serviced, they have to choose a priori a particular service operator. Consider a simple scenario of two different service operators say SP A and SP B who provide the same service. At a given time t assume that the service operator SP A is able to provide the required service while SP B is not. Let us look at the scenario where a user who wants to avail

of the service (has no preference for SP A or SP B) calls up SP B. Since SP B can not provide the required service, the user ends the call with SP B and next calls SP A and avails of the service. In sum total the QoE of the user is poor even though SP A was able to provide the service. The QoE is poor because of wasted telephone calls (calls SP B and then SP A) and time (time taken to find SP B is not available and then time taken to actually ask for services of SP A). If there was a way of knowing a priori that the service was available with SP A and not with SP B then the user would have called SP A and completed the transaction with a better QoE. This paper proposes a system that enables richer QoE in terms of both time and naturalness.

2 Typical Scenario

When a user calls any service over the telephone, the call is answered by a live agent or by an automated IVR (Interactive Voice Response) system. Invariably, in either situation, the caller is put on hold by playing some music. After a certain hold time, the caller is able to interact with the service operator (can talk to the agent/IVR) and is able to make the service request. In many cases the caller request may not be served in a single call to a service operator; for reasons like operator could not provide with the requested service or the caller wants to get information from another similar but different service operators and then make a decision. The caller has to dial the next service operator where the user may be put to hold again and the cycle repeats and this is very inconvenient if the number of the service operators is large. Figure 1 illustrates a timeline describing a typical scenario where a user needs a service and has the option of obtaining this service from multiple operators. As shown in Figure 1, when the user calls a service operator, the user is put on hold, until the service operator (in most cases a human agent) is able to service the user request. The user then dials the next service operator only after the previous call is complete; and the wait cycle repeats. Thus the user gains access to the service only after being put on wait by each of the operators and additionally there is an interaction time (speaking to the human agent of the service operators) until finally served by a particular service operator. During the wait times usually some music or advertisement promotions are played, which are as it is today irrelevant and most of the times frustrating to the caller; even in the most optimistic scenario when the user is required to call only a single service operator. This unproductive wait time often degrades the overall QoE and hampers the simplicity and ease of using the telephone services.

In such a scenario, using a telephone call, there is no way the user can connect to the operator which can service the user fastest or first. Unlike a web portal or a web mashup, there is no way for the user to check the offerings from different service operators and choose the best among them using a telephone call. As seen in the example scenario the QoE is poor and the user in a worst case scenario may be left with no information even though there are several operators for the service. To summarize, the telephone call service channel faces following shortcomings:

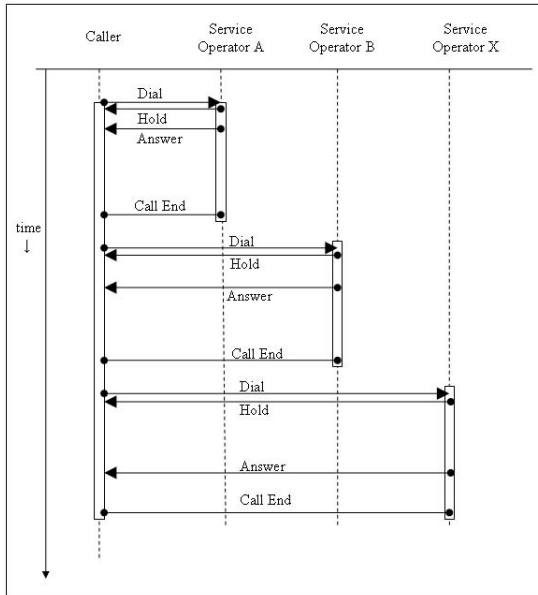


Fig. 1. Timeline describing a typical scenario where a user has to call several same service providing operators to book a service

- i. Long hold times and no effective use of the hold time. Usually some music or repetitive advertisements are played which are extraneous to the caller.
- ii. Inability to connect to the least wait time service operator.
- iii. Sequential connection, one after the other, to service operators or sources of information. There is no way to get information from various service operators in the same call.
- iv. The user has to speak the same request for service information every time the user talks to the new service operators, a redundant repeat.

Thus even though speaking over a telephone to get a service done is a preferable channel over other channels like website and messaging, it is comparatively less efficient and in the current form does not address the poor QoE. Therefore, there is a need for a system that addresses the shortcomings mentioned above and enables a mashup like interface over the telephone, providing both ease to the access of services and information and also enhances caller experience. Recently there have been systems which have addressed the problem of call hold time. Virtual Queuing [3] allows callers to hangup on long hold times; while keeping a virtual placeholder for the caller, this placeholder is used to call back the users when their turn arrives. Deep Dialing [4] system skips hold time and also skips users having to enter choices from a telephone keypad to reach a particular type of service in an IVR menus and sub menus. It allows a caller to directly reach a particular node in a IVR or directly get connected to the live agent behind an automated IVR. However, [3] and [4] do not provide for a mashup interface,

and do not address either the ease to access services from multiple operators, or the ability to connect to the service operator with minimum wait time. The proposed IVR Mashup system addresses these shortcomings.

3 The Mashup Service

Figure 2 illustrates the block diagram of the IVR Mashup platform. As depicted in Figure 2 the IVR Mashup is an IVR interface which interacts with the caller on one end and with multiple service operators on the other end. In a typical scenario a user who wishes to use the IVR Mashup service registers on the web with minimal details. The information about the user like his interests, preferences, etc is gathered during this registration and stored in an organized database. When the user wants to make use of the system the user calls the IVR Mashup service. The mashup service answers the user's call and then automatically places simultaneous calls to multiple service operators through the Outbound Dialer system. The mashup platform uses the Call Connection Status Analyzer to monitor the outbound call status. In the meantime it determines for how long and what advertisement (for a set of advertisements) must be played to the user, using the basic details available in the user profile/DB at the time of registration. This helps play targeted or personalized infotainment messages and advertisements to the user. Once connected to a service operator the mashup service handles the call as per the type of mashup service selected by the user. The preference of type of mashup service is marked in the user profile or chosen by the caller after dialing to the mashup service.

Figure 3 illustrates the timeline describing a typical transaction of mashup service. The steps involved in the transaction can be elaborated as follows:

- i. The transaction begins when a registered caller dials the Mashup service.
- ii. The Mashup service then simultaneously calls different service operators, and waits for them to answer. On the other end it refers to the caller's profile and plays relevant infotainment messages and advertisements.
- iii. When a service operator (operator with minimum wait time) answer's the call from the mashup service, the mashup system connects the caller to this service operator. The caller can now directly speak to the service operator agent or IVR.

Additionally, the Mashup service can call and connect only to that service operator which is preferred by the user. The preferred service operator is known from the user's profile. In this scenario the user has to wait for the time that s/he would have to wait even when s/he did not use the mashup service, however since the mashup service plays relevant infotainment messages and advertisements to the user till it gets connected the QoE is enhanced.

The IVR mashup service can also serve like a web mashup and provide service from all the service operators in the same call. The mashup service talks to the service operators after the wait period using advanced Automatic Speech Recognition and Text to Speech software. The mashup service is capable of

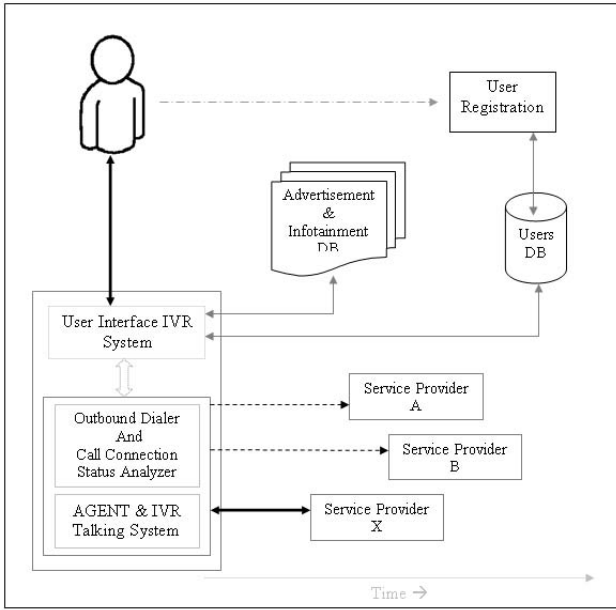


Fig. 2. Block diagram of the proposed IVR Mashup platform

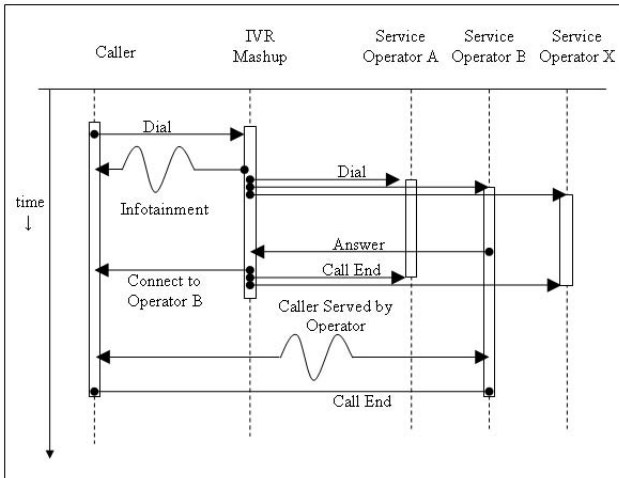


Fig. 3. Timeline describing a typical transaction completed by the mashup service by connecting the caller to the immediate available service operator

talking simultaneously to more than one service operator. The mashup service then provides the answers to the user’s request as and when it completes a transaction with a service operator or provides an update of information together at the end. It is also possible for the mashup service to complete the call with the service operators and later call the user and furnish the service details.

4 Mashup Service Analysis

4.1 Minimum Wait Advantage

The Mashup service discussed has a clear advantage of connecting the user to the service operator which can answer him first. As shown in Figure 4, for the purpose of analysis, let us consider a caller has the option to make use of any of the three available service providers, say A,B,C. Let $p(A)$, $p(B)$, $p(C)$ be the probabilities that the caller calls service operators A, B, C respectively. Let WA , WB and WC be the average waiting time faced by any caller calling the service operators A, B and C respectively. In the absence of the mashup service the caller has to wait (hold) for a time

$$t = p(A)WA + p(B)WB + p(C)WC \tag{1}$$

which is average probable and could be more. While use of the mashup connect service, the caller gets connected to the first available service operator. Hence the wait time is

$$t = \min\{p(A)WA, p(B)WB, p(C)WC\} \tag{2}$$

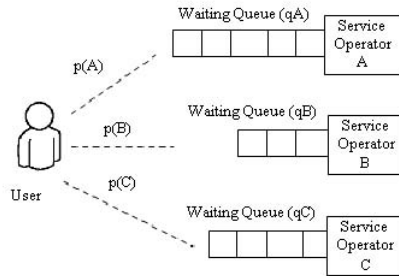


Fig. 4. Analysis of Minimum Wait Advantage

To understand this better, consider the Erlang C queuing model as described in 5. Using this model, the probability that a caller has to wait in the queue when s/he calls a service operator with N agents is given as:

$$C(N, R) = 1 - \frac{\sum_{m=0}^{N-1} (R^m / m!)}{\sum_{m=0}^{N-1} (R^m / m!) + (R^N / N!)(1 / (1 - R/N))} \tag{3}$$

$$R = \frac{\lambda}{\mu} \tag{4}$$

where λ is the call arrival rate and μ is the service rate. And the probability that the caller has to wait for time t is given as:

$$P(> t) = C(N, R) * e^{-(N\mu-\lambda)t} \tag{5}$$

Consider a service operator SP1 having 100 agents ($N = 100$), call service time of 10 minutes ($\mu = 1/10$) and average call arrival rate of 9.9 calls per minute. Consider another service operator SP2 having same number of agents and call service time, but average call arrival rate that is 90% of that of SP1. Using (1) to (5) we can plot the probability of waiting for time t and compare these for any caller trying to call service operators SP1 or SP2 directly and callers using mashup service. This is illustrated in the graph in Figure 5.

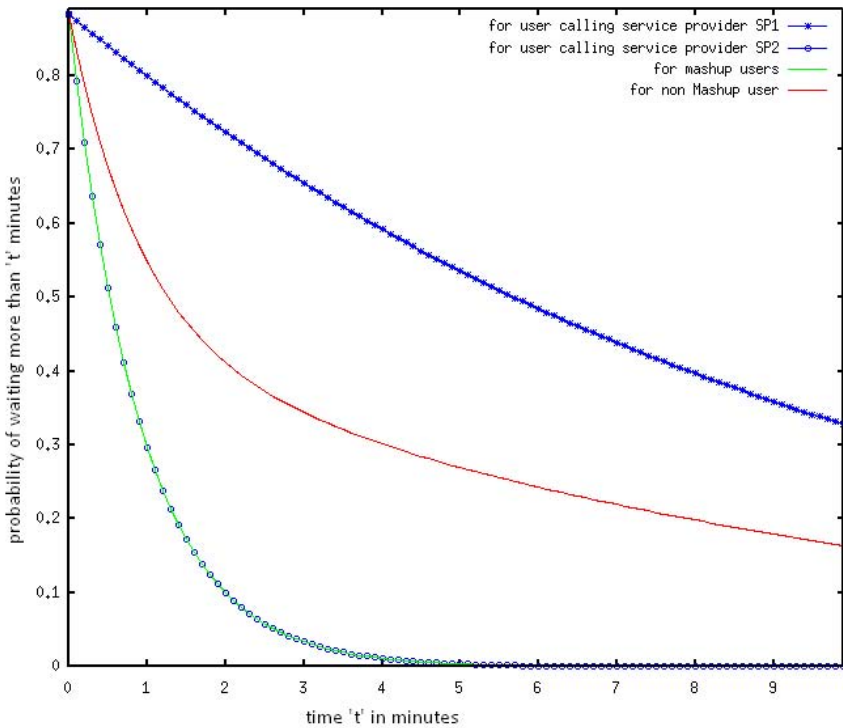


Fig. 5. Graph of probability of waiting greater than time t

As shown in Figure 5, for a caller particularly calling SP2 the probability of waiting greater than time t is lesser than that for a caller particularly calling SP1. However, for a non mashup user, who can arbitrarily call SP1 or SP2, the average probability of waiting is greater than that of SP2 (the service operator having minimum probability of waiting in this case). It must be noted that in a practical situation the service operator with minimum wait time is not known to a caller

(non mashup user) and also it may not be the same service operator which has minimum wait time throughout. For a mashup user, since the mashup service ensures to connect the user to the service operator with least wait time, the probability of waiting is always minimum and equal to that of service operator having minimum probability of waiting (SP2 in this case). Therefore, the graph of the probability of waiting for the mashup user overlaps that of SP2.

Similarly, using the Erlang C queuing model in [5] the average wait time for any caller in the queue is given as:

$$Tq = \frac{C(N, R)}{(N\mu - \lambda)} \quad (6)$$

The average waiting time can be plotted as a function of the call arrival rate (λ). Figure 6 shows a graph that compares the average wait time for callers trying to call service operators SP1 or SP2 directly and callers using mashup service, at different arrival rates. (Average call arrival rate for SP2 is 90% of that of SP1.)

As shown in Figure 6, for a caller particularly calling SP2 the average waiting time is lesser than that for a caller particularly calling SP1. However, for a non mashup user, who can arbitrarily call SP1 or SP2, the average waiting time is greater than that of SP2 (least waiting time in this case). For a mashup user, the average waiting time is minimum and equal to least waiting time (given by SP2 in this case). Therefore, the graph of waiting time for the mashup user overlaps that provided by SP2.

Thus it can be observed from Figures 5 and 6 that the probability of waiting greater time t , as well as the average waiting time for a mashup user is minimum when compared to a non mashup user. In the worst case scenario where the waiting times for all service operators is same, the IVR mashup service gives the advantage of enhanced QoE as discussed in the next section.

4.2 Enhanced Quality of Experience

The Telephone Mashup service provides an overall enhanced QoE. The factors enhancing QoE are:

- i. Minimum wait time to connect to a service operator.
- ii. The wait is not boring because the caller gets to listen to something that is appealing to his/her personal interests.
- iii. It allows user to speak personalized and natural voice requests.
- iv. User can get information from multiple service operators in a single call.

4.3 Effect on Call Center Queues

The primary function of the IVR Mashup service is to call a service operator and wait in the queue of the call center (of the service operator) on behalf of the caller. Different scenarios of the Mashup service have a different effect on the call center queues. For cases where the IVR Mashup calls the service operator preferred by

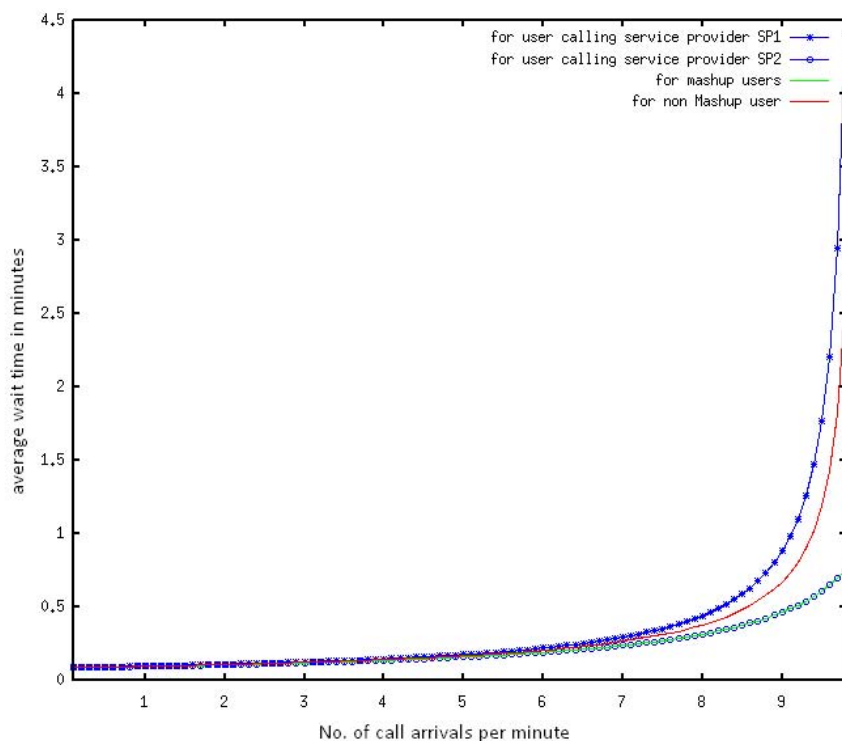


Fig. 6. Average waiting time for different call arrival rates

the user, it is same as the caller calling and waiting in the queue, and hence the queue remains unaffected. In other case where the mashup service connects to the fastest answering service operator, the mashup service calls multiple service operators and stays in the queue of the respective call centers. However, once the call is connected to a service operator (the operator with minimum wait time) the mashup service disconnects calls to remaining service operators, thus clearing places in their queues. Hence, even though the queues appear to be increasing, it actually is reducing at the same time. When the Mashup service calls multiple service operators and stays in the queue of the respective call centers, and talks to each of the service operators, it increases the queue length, but the waiting is still done by the Mashup service thus increasing the QoE.

5 Conclusion

Even though a telephone voice call is highly popular, for booking services and getting information about services, it faces major shortcomings like long hold times and inability to get information from multiple service operators in the same call. There is an express need to develop a platform which can integrate services offered by different service operators, through the telephone channel, to

enhance QoE. The proposed IVR Mashup solution is a step towards that. The advantage of such a system is that a caller not only spends less time to enable a transaction but also gets a better calling experience.

References

1. Sunil, K., Imran, A., Pande, A.: System and method to enable access same service multiple providers on telephone. Patent 2422/MUM/2010 (2010)
2. Sun, B.: Technology Innovation and Implications for Customer Relationship Management. *Marketing Science* 25(6), 594–597 (2006)
3. http://en.wikipedia.org/wiki/Virtual_queue
4. <http://fonolo.com/overview/deepdialing>
5. Gans, N., Koole, G., Mandelbaum, A.: Telephone Call Centers: Tutorial, Review, and Research Prospects *Manufacturing and Service Operations Management*, vol. 5(2), pp. 79–141 (2003)

Information Content Based Semantic Similarity Approaches for Multiple Biomedical Ontologies

K. Saruladha, G. Aghila, and A. Bhuvaneshwary

Pondicherry Engineering College, Puducherry, India
bhuvana787@gmail.com, charusanthaprasad@yahoo.com,
aghilaa@yahoo.com

Abstract. Semantic similarity mechanism is mandatory in information retrieval, information integration, ontology mapping and psycholinguistics. The objective of this work is to develop a computational approach which assesses semantic similarity among concepts from different and independent ontologies without constructing a priori a shared ontology. The proposed approach is based on Tversky similarity model and is mapped to information theoretic domain. This paper also explores the possibility of adapting the existing single ontology information content based approaches and propose methods for assessing semantic similarity among concepts from different multiple ontologies. The proposed approaches are corpus independent and they correlate well with the human judgements. The proposed approaches have been experimented with two biomedical ontologies: SNOMED-CT (Systemized nomenclature of medical clinical terms) and Mesh (Medical subject headings) and the results are reported. The proposed four approaches outperform the path length based computational method as it achieves the highest correlation.

Keywords: Semantic similarity, multiple Ontologies, information Content, MeSH, SNOMED-CT.

1 Introduction

Evaluating semantic relatedness using network representations is a problem with a long history in artificial intelligence and psychology [1]. There is a need for consistent computational method to assess semantic similarity between words and concepts. For this the intended meaning of word or concept should be well understood. This is difficult as a single word has many meanings (polysemy) and a single word has an equivalent word (synonymy). These characteristics of polysemy and synonymy words have posed challenges in determining intended meaning of a word or concept. The intended meaning of the word could be disambiguated by resolving the semantic relations between the words and concepts.

To determine computationally the semantic similarity the information could be organized as a taxonomy (a static knowledge source) which is a structured way to characterize the words/concepts and their relationships. Several researchers have proposed various methods to quantify the semantic similarity between words/concepts organized in taxonomy or in ontology.

Resnik [8] has argued in his work that uniform distances between links could not better quantify the informativeness and suggested that the theory posed by Ross, could be used to quantify informativeness of concepts. This quantitative characterization was adopted by Resnik and he proposed a semantic similarity method based on information theory or information content. Lin [9] measure of semantic similarity has been extended from Resnik. It accounts the shared information and specific information of each concept in the considered corpus. Jiang and Cornath [10] proposed a combined approach combining edge and node based methods. The semantic space constructed by the taxonomy (ontology) and the length separating concepts is quantified with the computational measure derived from the corpus statistics. As the statistical approach is complimented by the taxonomic structure the semantic similarity between words would reasonably be assessed. All of Information Content based approaches considers hierarchical relations which maps well with the cognitive notion of classifications.

The remainder of this paper is organized as follows. Section 2 discusses the motivation of the similarity measure. Section 3 provides some background information regarding SNOMED-CT and MeSH. Section 4 discusses the existing similarity measures for single and cross ontologies. Section 5 presents the proposed Tversky based IC approach and adapting single ontology IC based approaches for multiple ontologies. Section 6 discusses the new datasets used to analyse and compare similarity metrics by correlating them to human judgements. Section 7 compares the results of the proposed similarity measure with the existing measure. Finally, Section 8 concludes the paper.

2 Motivation

Recent researchers in information retrieval and data integration accentuate the development of ontologies and semantic similarity mechanisms for comparing concepts that can be retrieved or merged across heterogeneous information sources (ontologies). These approaches could be used in ontology mapping and alignment environments.

3 Knowledge Source for Biomedical Domain

Well defined ontologies for biomedical domain are MeSH [14], SNOMED-CT [15]. SNOMED-CT [15]. It is a comprehensive clinical terminology with coverage of diseases, clinical findings. Fig. 1 represents the SNOMED-CT and MeSH hierarchy level comprising of concepts, terms and relationships to represent clinical information. The current version contains more than 13,136,022 concepts, 975,000 synonyms and 1,450,000 relationships organized into 19 hierarchies/sub trees/categories. MeSH [14], stands for Medical Subject Headings is one of the main source

vocabularies used in UMLS with the primary purpose of supporting indexing, cataloguing, and retrieval of medical literature articles stored in NLM MEDLINE database, and includes about 16 high-level categories) concept structure in Mesh. Currently there are 50956 concepts.

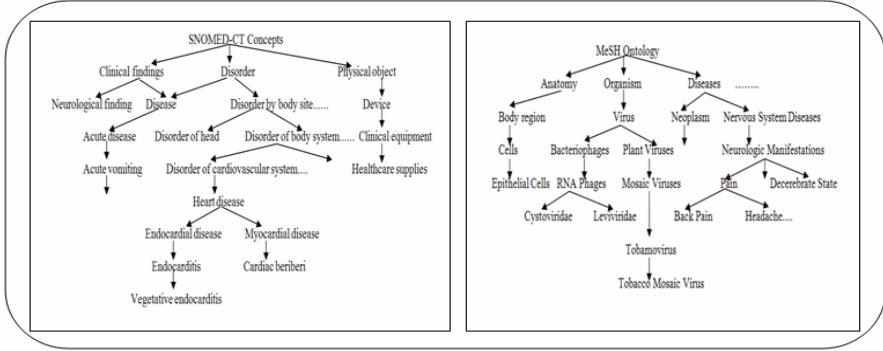


Fig. 1. SNOMED-CT & MeSH hierarchy

4 Related Work

Many techniques have been proposed for evaluating semantic similarity between concepts. They are classified into two categories: 1) Semantic similarity for single ontology, 2) Semantic similarity for cross ontologies. Measuring similarity between concepts is mainly classified into four different approaches. They are Ontology based approach, Information Content based approach, Hybrid based approach and Feature based approach. The semantic similarity method for comparing concepts belonging to multiple biomedical ontologies is limited.

Ontology based approach [3], [4], [6], [7] require consistent and rich ontologies to assess semantic similarity between two concepts. It is mainly classified under two categories. Path length approaches computes similarity by counting the number of nodes/edges between two concepts in terms of the shortest path in the taxonomy. Depth relative approaches takes into account the depth of the taxonomy by calculating the depth from the root to the target concept. Al-Mubaid & Nguyen Method [4] proposed a new ontology-structure-based technique for measuring semantic similarity in single ontology and across multiple ontologies in the biomedical domain. The semantic similarity between cross ontological terms is measured by considering one ontology as primary and another as secondary ontology. The secondary ontology is connected to the primary ontology by joining the common nodes called as bridge node of two ontologies.

Information theoretic approaches [8], [9], [10] usually employ the notion of Information Content (IC), which can be considered as a measure quantifying the amount of information a concept expresses. Seco [12] has proposed an Information Content (IC) calculation which is corpus independent (single ontology). Table 1 compares the various IC based approaches.

Table 1. Comparison Table for Information Content based Approach

Method	Relations considered	TEST data used	Corpus statistics (used)	Corpus considered	Correlation
Resnik [8]	IS-A	M&C	Yes	Brown Corpus	0.793
Lin [9]	IS-A	M&C	Yes	SemCor Corpus	0.823
Jiang & Conrath [10]	IS-A	M&C	Yes	SemCor Corpus	0.859

Feature based approach similarity takes into account the features that are common to two concepts and also the differentiating features specific to each. The similarity of a concept C_1 to a concept C_2 is a function of the features common to C_1 and C_2 , those in C_1 but not in C_2 and those in C_2 but not in C_1 . According to Rodriguez & Egenhofer [13], a concept is considered as an entity class. They have proposed a new method to determine the semantic similarity among different ontologies is based on Tversky [5] measure by taking into account the similarity between entity classes by means of a matching process over synonym sets S_w , semantic neighborhoods S_n , and distinguishing features S_u that are classified into parts, functions, and attributes is given by equation 1.

$$S(a, b) = \frac{|A \cap B|}{|A \cap B| + \alpha(a, b)|A / B + (1 - \alpha(a, b))|B / A|} \quad (1)$$

5 Proposed Similarity Method

Ontologies are central to the semantic web because they allow applications to agree on the terms that they use when communicating. Shared ontologies and ontology extension allow a certain degree of interoperability. When two ontologies overlap, they can be linked together in order to enable exchange of their underlying knowledge. Using Rodriguez & Egenhofer [13] method has achieved less correlation coefficient of 0.71 and path length approach achieved less correlation of 0.665. Pirró [11], [12], [17] has mentioned in his work that the IC based approaches could be extended to compute semantic similarity between concepts belonging to different ontologies if the problem of finding the Most Specific Common Abstraction (MSCA) concept is addressed. MSCA is defined as a concept that subsumes the concept C_1 and C_2 . Finding the common abstraction concept in a single ontology is easy. But when C_1 belongs to ontology O_1 and C_2 belongs to ontology O_2 finding the most specific common abstraction concept is difficult.

Our Contribution. Two computational approaches have been proposed for finding semantic similarity among concepts belonging to different ontologies.

1. Tversky based Information Content approach (TBIC),
2. Adapting the single ontology IC based approaches for assessment of semantic similarity of concepts belonging to multiple ontologies.

5.1 Proposed Tversky Based IC Approach

The Proposed similarity method considers a corpus independent information content based similarity computation to assess asymmetric similarity between biomedical concepts belonging to multiple ontologies based on Tversky [5] psychological model. In the proposed approach the two considered ontologies are connected to a virtual root which paves a way for finding the most specific common abstraction concept. It also eliminates the complexity of constructing apriori a shared ontology. The proposed Tversky based Information content (TBIC) approach compute semantic similarity among cross ontologies by taking into consideration the informativeness of the shared concept and the information of the unique concepts scaled by the position of the concept in the ontology. For the semantic similarity computation, the ontologies considered could be shallow or deep. The deep ontology is designated as primary and shallow ontology as secondary. The proposed Tversky based Information Content (TBIC) measure is defined as

$$Sim_{TBIC}(C_1, C_2) = \frac{IC(MSCA(C))}{IC(MSCA(C)) + \alpha(C_1, C_2) \cdot (IC(C_1)) + (1 - \alpha(C_1, C_2)) \cdot (IC(C_2))} \cdot (2)$$

$$\alpha(C_1, C_2) = \begin{cases} \frac{Depth(C_1^{O_1})}{depth(C_1^{O_1}) + depth(C_2^{O_2})} & \text{if } depth(C_1) \leq depth(C_2) \\ 1 - \frac{(Depth(C_1^{O_1}))}{Depth(C_1^{O_1}) + Depth(C_2^{O_2})} & \text{if } depth(C_1) \geq depth(C_2) \end{cases} \quad (3)$$

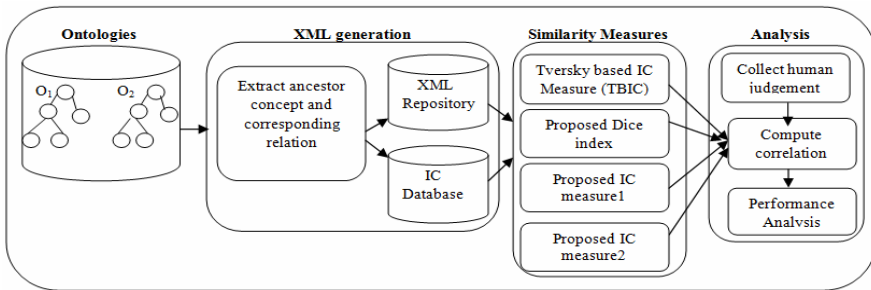


Fig. 2. Framework for semantic similarity assessment for Multiple Ontologies

where α is a function given in the equation (3) determines the relative importance of the non common characteristics and is defined using depth (α) of the considered ontology. The value of α was experimentally determined and found to be in the range of 0 to 0.9. The term IC (MSCA) quantifies the common characteristics of the compared concepts and is equivalent to the $|A \cap B|$ in Tversky’s model. The IC (MSCA) is computed based on two ontologies as C_1 belongs to O_1 and C_2 belongs to

O₂. In a single ontology common concept is present in the same ontology. But finding MSCA of concepts belonging to different ontologies is possible only when both the ontologies are connected through a virtual root (VR). There may be more than one common abstraction concepts present for the compared concepts. In such a situation the concept at the deepest level will be considered as the MSCA. The IC (MSCA) is computed using the following formula

$$IC(MSCA(c)) = 1 - \frac{\log(\max(\text{hypo}(O1(C1), O2(C2))) + 1)}{\log(\max\text{ con})} \tag{4}$$

where function (max(hypo(O1(C1), O2(C2)))) returns the taxonomy which is having maximum hyponymy of the concept. max con returns maximum concepts that occurs in the considered taxonomy. On computing IC (MSCA) the common characteristics of compared concepts was quantified. The unique characteristic of the compared concepts are to be determined by using the Seco [12] formula and is given by equation 5.

$$IC(C) = 1 - \frac{\log(\text{hypo}(C1, C2) + 1)}{\log(\max\text{ con})} \tag{5}$$

Variant of Tversky Similarity Index (Dice Coefficient). It is a variation of Tversky similarity model. For sets A and B, the Dice’s coefficient may be defined as twice the shared information over the combined set.

$$S(A, B) = \frac{|A \cap B|}{\frac{1}{2}(|A| + |B|)} \tag{6}$$

The shared information is quantified as (|A|∩|B|) and the information over the combined set as (|A|+|B|). When α=β=0.5, the Tversky measure reduces to Dice coefficient. Dice coefficient without any changes could be used for measuring similarity among concepts among single ontology. The value α=0.5 could be achieved if the concepts belonging to different ontologies were at the same level. If one concept is at the abstract level and another is at the deepest level (near leaf nodes) then α value could be 0.9. As this work was focused for adapting Dice coefficient for computing semantic similarity among concepts belonging to multiple ontologies, it was found that α=β=0.5 could not be true for all the compared concepts. Hence the Dice coefficient for quantifying semantic similarity between concepts belonging to multiple ontologies is refined as

$$S(A, B) = \frac{|A \cap B|}{\alpha(|A|) + \beta(|B|)} \tag{7}$$

The shared information between concepts is quantified as done in Information Content based approach (IC (MSCA)) for more than one ontologies and is given by

equation 4. The information over the combined set is quantified as the arithmetic addition of information content of individual concepts. Therefore the proposed refined Dice's coefficient for measuring semantic similarity between concepts belonging to multiple ontologies is given by

$$Sim_{Dice}(C_1, C_2) = \frac{IC(MSCA(C_1, C_2))}{\alpha(|IC(C_1)|) + \beta(|IC(C_2)|)} . \quad (8)$$

5.2 Adaptation of the Single Ontology IC Based Approaches for Multiple Ontology Semantic Similarity

The single ontology IC based approaches (Resnik, Lin) have been adapted for measuring semantic similarity among concepts belonging to multiple ontologies. The Information Content based computations used by IC based approaches are corpus independent. The IC computation done using Seco [12] formula is used in Refined Resnik and Lin methods.

Proposed IC Measure1. For measuring semantic similarity between multiple ontological concepts the existing Resnik measure has been adapted to proposed IC measure1 (Refined Resnik). The existing Resnik [8] measure for single ontology is given by

$$Sim_{res}(C1, C2) = \max_{C \in S(C1, C2)} [IC(C)] . \quad (9)$$

The proposed IC measure1 (refined Resnik measure) for multiple ontological concepts C_1 belongs to O_1 and C_2 belongs to O_2 is given by

$$Sim_{res}(C1, C2) = \max_{C \in S(O1(C1), O2(C2))} [IC(C)] . \quad (10)$$

Proposed IC Measure2. The existing Lin [9] measure for single ontology is given by

$$Sim_{Lin}(C1, C2) = 2 * \frac{IC(MSCA(C1, C2))}{IC(C1) + IC(C2)} . \quad (11)$$

The proposed IC measure 2 (refined Lin measure) for multiple ontological concepts C_1 belongs to O_1 and C_2 belongs to O_2 is given by

$$Sim_{Lin}(C1, C2) = 2 * \frac{IC(MSCA(O1(C1), O2(C2)))}{IC(C1) + IC(C2)} . \quad (12)$$

6 Experimental Setup

The quality of the proposed computational method could be assessed best if it is compared against human judgements. The biomedical concepts used in datasets were

extracted from MeSH and SNOMED-CT. Hence two commonly used biomedical datasets of concepts was distributed to 100 human subjects who are experts in the field of biomedicine. They were asked to assess the similarity of concept pairs of the two datasets on a scale 0(semantically unrelated) to 1(highly synonyms). For each concept of the dataset an XML file was generated which expresses the relation they have with other concepts and the depth of the concept in the considered ontology. A total of 132 XML files were generated for the concept pairs of the datasets. Figure 2 shows the framework of the semantic similarity assessment for multiple ontologies. The datasets used for the experiments is biomedical dataset containing 36 biomedical term pairs [4].

6.1 Evaluation Methodology

There are two types of evaluation methodology for finding semantic similarity. 1) Intrinsic Evaluation, 2) Extrinsic Evaluation. Usually accepted evaluation methodology uses intrinsic evaluation which is done by correlating the computational value against the human judgements. Pearson Correlation was used to assess the strength of the relation between human judgements of similarity and computational values obtained through proposed Information Content based method. By calculating the p value for the Pearson Correlation Coefficient the significance of the relation can be evaluated.

7 Results and Discussion

The correlation coefficient computed for the proposed semantic similarity methods has been tabulated in table 2. It shows that the proposed method based on Tversky achieves the highest correlation of 0.920. This could be because the MSCA calculated based on MeSH and SNOMED-CT better expresses the amount of information shared by two concepts. The results are said to be proven as the lower p value generates the significance of the results. Figure 3 & 4 shows the semantic similarity ratings graph for proposed Tversky based measure (TBIC), proposed dice index measure and proposed IC based measures 1 & 2 which have been compared with existing path length measure and human judgment. From the analysis of the graph, we could infer that similarity ratings obtained by the existing path length measure gives more deviation against human judgment. For example considers pairs such as pain-ache, malnutrition-nutritional deficiency, measles-rubeola which gives higher similarity values since each pairs are semantically similar concepts. In the case of path length measure the similarity value obtained for these pairs is 0 because shortest path length could not be found for semantically similar concepts. Hence it results in 0. In the proposed four measures contribute maximum similarity value against human judgment.

Table 2. Comparison Table

Approaches	Dataset	Correlation Coefficient
Proposed Tversky based IC measure	MeSH & SNOMED-CT (Dataset1)	0.920
Proposed Dice index Measure	MeSH & SNOMED-CT (Dataset1)	0.859
Proposed IC Measure 1	MeSH&SNOMED-CT (Dataset1)	0.911
Proposed IC Measure 2	MeSH & SNOMED-CT (Dataset1)	0.810
Path Length (Existing Method)	Dataset1, Dataset2& Dataset3	0.665

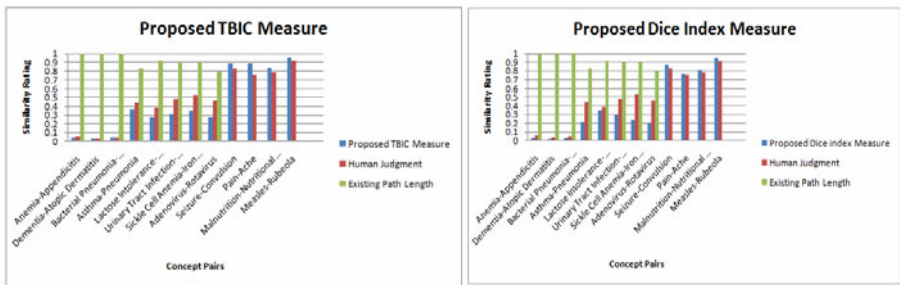


Fig. 3. Proposed Tversky and Dice based similarity assessment Ratings for Dataset1

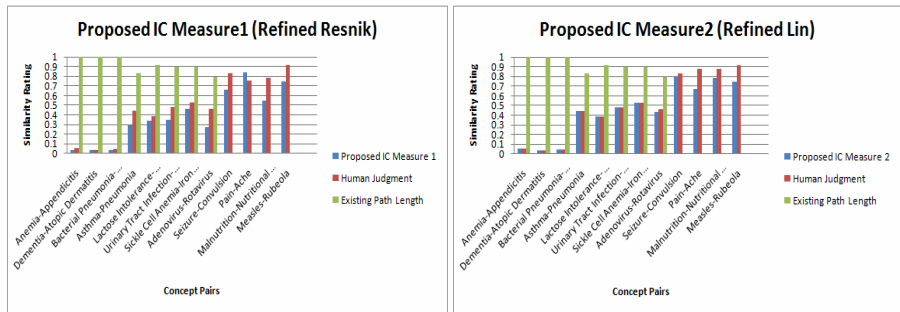


Fig. 4. Semantic similarity ratings using proposed IC based measures 1 & 2 for Dataset1

8 Conclusion

This paper presented computational approaches for measuring semantic similarity among concepts belonging to independent different ontologies. The Tversky similarity model was translated to information theoretic domain and the existing single ontology semantic similarity were adapted to measure semantic similarity

across cross ontological concepts. The results show the proposed Tversky based semantic similarity outperforms the path length and adapted IC based approaches.

References

1. Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. *J. Biomedical Informatics*. 40, 288–299 (2007)
2. Saruladha, K., Aghila, G., Bhuvaneshwary, A.: Computation of Semantic Similarity among Cross Ontological Concepts for Biomedical Domain. *J. Comp.* 2, 111–118 (2010)
3. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of semantic distance. *Comp. Ling.* 32, 13–47 (2006)
4. Nguyen, H.A., Al-Mubaid, H.: “Measuring Semantic Similarity between Biomedical Concepts within Multiple Ontologies. *IEEE Trans. on Systems, Man, and Cybernetics* 39, 339–398 (2009)
5. Tversky, A.: Features of similarity. *Psychological Review* 84, 327–352 (1977)
6. Rada, M.H., Bicknell, M., Blettner, E.: Development and application of a metric on semantic nets. *IEEE Trans. on Systems, Man, and Cybernetics* 19, 17–30 (1989)
7. Leacock, C., Chodorow, M.: Combining local context and Word-Net Similarity for Word Sense Identification: Christiane Fellbaum (1998)
8. Resnik, P.: Using Information content to evaluate semantic similarity in taxonomy. In: 14th International Joint Conference on Artificial Intelligence, pp. 448–453 (1995)
9. Lin, D.: An information-theoretic definition of similarity. In: 15th International Conference on Machine Learning, pp. 296–304 (1998)
10. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: International Conference on Research in Computational Linguistics (1997)
11. Pirró, G., Seco, N.: A new semantic similarity metric combining features and intrinsic information content. *J. Data & Knowledge Engg.*, 1271–1288 (2009)
12. Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in WordNet. In: 16th European Conference on Artificial Intelligence, pp. 1089–1090 (2004)
13. Rodriguez, M., Egenhofer, M.: Determining semantic similarity among entity classes from different ontologies. *IEEE Trans. on Knowledge and Data Engg.* 15, 442–456 (2003)
14. MeSH Browser (2010), <http://www.nlm.nih.gov/mesh/MBrowser.html>
15. SNOMED-CT (2010), <http://www.snomed.org/index.html>
16. Scharffe, F., Fensel, D.: Correspondence Patterns for Ontology Alignment. In: 16th International Conference on knowledge Engineering (2008)
17. Pirro, G., Euzenat, J.: A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness (2010)

Taking Project Tiger to the Classroom: A Virtual Lab Case Study

Harilal Parasuram¹, Bipin Nair¹, Krishnashree Achuthan², and Shyam Diwakar¹

¹ Amrita School of Biotechnology, Amrita Vishwa Vidyapeetham (Amrita University),
Amritapuri, Clappana P.O., 690525, Kollam, Kerala, India
shyam@amrita.edu

² Amrita School of Engineering, Amrita Vishwa Vidyapeetham (Amrita University),
Amritapuri, Clappana P.O., 690525, Kollam, Kerala, India

Abstract. Understanding how population dynamics change over time is critical to many practical problems as pest control, endangered species protection etc. Teaching population ecology is not easy since data is usually collected over a very long period. This paper discusses a specific tiger population case study relating to growth rate predictions using an online virtual lab. Studying tiger populations and introduction of such data in classrooms help in creating awareness and support new pedagogies to estimate animal population dynamics. We have used online virtual labs which are ready-made tools to perform simple experiments and analysis. An important and usually complex case of population analysis as in tiger populations in India is studied in this paper. Although some major parameters like food, transient movement, and ecosystem details have been ignored, predicted data for tiger population follows closely to actual data for previous years and even predicts the growth rate with a small standard deviation of 10%. Our results with tiger populations come close to the actual census values. We propose the use of simple mathematical models to make assessment of transient animal populations such as tigers, and sharks. Also use of such ready-made pro-academic online tools encourages new studies and an enhanced pedagogy to population ecology for mathematicians, biotechnologists, wildlife institute personnel among many other cross-disciplinary scientists.

Keywords: Virtual Labs; Tiger population in India; Population Ecology; e-learning.

1 Introduction

Half of the tiger population in the world is in India. Due to reduction in their population in large numbers, from 1969 onwards the ‘tiger’ was declared as an endangered species (by CITES). India launched a project called “Project Tiger” with the goal of saving the tiger and its habitat. Studying and understanding the dynamics of tiger population in our country is very essential to predict (build/tune model) the tiger population in future. These predictions could help to a great extent in protecting

from mass disappearance of endangered species. Estimating tiger population behaviors are also useful for demographic analysis and studies. We have employed a new approach of using online virtual lab tools to make studies on tiger population.

Typical tiger population studies need to cover spatial and temporal scales since the effects of anthropogenic pressures that come into conflict with ecological needs [15]. In tiger population studies, one major problem is the relative imprecision of single-year abundance estimates. Many tiger reserves are sensitive in terms of landscape matrix hostile to tigers leading to transient tiger populations [example see 15]. Using virtual lab online tools for tiger population studies is the major role of this paper. Although several critical parameters such as those related to food, disease and inter-competition were ignored, it was possible to connect the data to a feasible and surprisingly simple model based on exponential growth.

In this paper, we suggest on the applicative use of population ecology simulators as classroom models to complete the learning experience for a population ecology laboratory course. Taking such projects from response labs to classroom will help generating awareness and activity-oriented research. The paper reports the analysis, interpretation and some preliminary predictions in variations of tiger population in India.

1.1 Virtual Population Ecology Lab

Over past two decades, the areas of online teaching and web-based learning initiatives have grown tremendously especially in the areas of biology [1]. A Virtual Lab (VL) is an innovative, computer-based educational program designed to learn, understand the basic principle of experimental and theoretical science and their applications. Since practical experience is an important component in learning science subjects, the continuous practice in an experiment would make one an expert in that subject. The opportunity to deal with real system of education is not reaching to all students of all institutions across the country. These constraints could be resolved by using learning materials such as information technology (more specific e-learning platform) which includes accessibility of these labs through web sites, virtual field trips, computer simulations and virtual laboratories [2].

Early studies have found that hands-on learning, particularly computer-based (online) instruction (a VL environment), is effective in enhancing science achievement [3] and [4]. A VL is a virtual reality environment that simulates the real world for the purpose of discovery learning. It allows one to simulate/emulate real experiment and operation on the basis of its underlying basic principle [5]. Even though a VL cannot be a replacement or equal to traditional laboratory or wet laboratory, it is worth to consider the many benefits that it offers.

Large numbers of institutions are working strong to support for creating an active, engaged learning environment to enhance student learning [6] and [7]. Literature studies indicate that study materials could engage and motivate students when they were user friendly, interactive, and problem oriented [8]. The use of virtual labs will give new instructional practice that will help to create the engaged and active learning experience that is supported by the literature [8]. This method may also extend learning for those students who cannot reach the study material or experimental setup ("If you can't come to lab, lab will come to you") and our recent study indicated that

virtual Labs played a significant role for enhanced facilitation of biotechnology education in developing nations (for analysis, implementation and case-studies see [9]).

Each experiment in virtual population ecology lab discusses modeling in detail, with examples (classical models in population dynamics) and simulators with assignments. These experiments focused on simple models that incorporate variability and density dependence, matrix models that incorporate age and stage structure, variability and density dependence, and meta-population models with spatial structure, variability and density dependence.

Our goal was to test and try if such simple tools suffice to make reasonable prediction for transient populations of wildlife. Since the amount of data depends on habitat and sustainability, our models justify the variation for a short time-span. The main goal of this paper is to indicate that certain simple models suffice for short duration studies on transient tiger populations.

2 Methods

Main objectives of population with continuous or discrete growth model experiment is as follows 1) To study the growth pattern of a population if there are no factors to limit its growth, 2) To understand the various parameters of a population such as per capita rate of increase, per capita rate of birth and per capita rate of death, 3) To understand how these parameters affect the rate of growth of a population.

2.1 Tiger Population Study

Indian tigers (*Panther Tigris*) are the largest among all the living wild cats belongs to the Class: Mammalia, order: Carnivora, Family: Felidae. Habitat and Distribution: Tiger is found practically throughout the country, from the Himalayas to Cape Comorin, except in Punjab, Kutch and the deserts of Rajasthan. Main predator for tiger is mankind. They have been trapped or killed by humans for skins, tiger claws (are used for necklets) and body parts of tiger used as traditional medicine. Due to large number reduction in their population, in 1969 the tiger was declared as an endangered species. Buying or selling tiger parts has been banned by the Convention on International Trade in Endangered Species (CITES) in 1975[14]. As an international effort, India launched a project called "Project Tiger" with the help of the World Wildlife Fund (WWF) in 1973 with the goal of saving the tiger and its habitat in India. By 1970, this project could establish 17 tiger reserves [12]. Other nations such as Indonesia, Thailand and Nepal also focus on creating and enhancing their own tiger reserves.

2.2 Data Collection

Statistical data for this study was collected from Project Tiger which includes the tiger population from 1972 – 2002 of various tiger reserves [13]. Total number of tigers correspond to a year are the total sum of tigers in reserves in India (see Table. 1). And the second data set was the crime reported for the numbers of tigers that have been killed in past few years were from WPSI's Wildlife Crime Database [14]. Growth rate

has been calculated by using the formula, Growth rate = Total number of individuals at time (t+1) – (t)/ Total number of individuals at time (t) where ‘t’ represent the time in years.

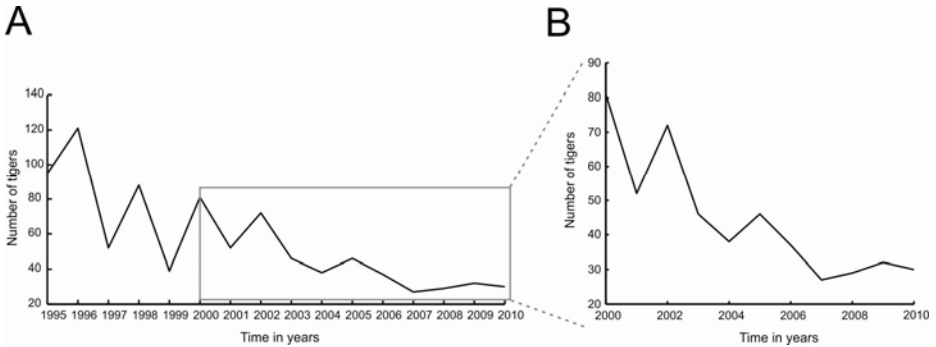


Fig. 1. Decline in tiger population. This data was taken from [14]

Calculated average decline rate was equal to -0.03. The nature of population decline (see Fig. 1) indicated that within a small time-range, tiger populations are less influenced by mortality rate. Hence we chose to ignore parameters such as food, climatic affects assuming linearity in considered region of interest.

2.3 Linearity in Geographic Range and Its Effect on Population

Tigers in India are usually spread on narrow geographic ranges where the species are supposedly sedentary. Comparisons at the generic level should provide results that are least encumbered by phylogenetic variation in morphology and life-history traits. Most tiger data are estimated with closed-intervals. The decline in tigers has been due to the principal threat to its sustainability mainly due to habitat destruction [20], decline in prey base caused by over hunting [15, 18], poaching [17, 18, 19] and poor tiger-human conflict management [16]. Usually while collecting data, the trap-rates (number of pictures per day) are very low, suggesting changes are low among transient populations. With a case of linearity in geographic range, we assumed in this geographic range variations of populations are more or less exponential with slight changes due to other properties such as food, prey densities etc.

2.4 Parameters and Variations

Population ecology virtual lab population growth simulator uses population with continuous growth and discrete growth models. A population with no defined reproductive season exhibits continuous growth or in other words births occur at all times of the year. Human population is a good example for this type of growth. The rate of continuous growth of a population is given by eq (1)

$$\frac{dN}{dt} = rN \tag{1}$$

Where N represent the number of individuals in the population at each time step. dN/dt is the change in number of individuals over time and r is the population growth rate. This differential equation predicts the total number of individuals in the population at time t , the slope at each point is equal to r times the number of individuals of the population. This differential equation (1) can be used to predict future continuous growth of species. Since continuous population growth follows exponential function, equation can be written as:

$$N_t = N_0 e^{rt} \quad (2)$$

Where, N_t is the total number of individuals in the population after t years, N_0 is the initial population size.

The individuals in a population which breed only in seasons define populations having discrete growth. An example is the population dynamics of ground squirrels, which breed only in spring. In these classifications of species the growth occurs in discrete time period.

Population size after t years in discrete growth is given by:

$$N_{t+1} = RN_t \quad (3)$$

$$R = 1 + (B - D) \quad (4)$$

$$N_t = (R)^t N_0 \quad (5)$$

Where N_{t+1} is the total number of individuals in a population after one year from now, N_t is the current population size, B is the birth rate, D is the death rate. Equation (5) can be used to calculate the rate of growth of population for n years. Population ecology virtual lab population growth simulator uses this equation to predict tiger population in India for next few years based on the statistical data collected from [3,14].

2.5 Simulator and Data Analysis

Virtualizing a population ecology experiment included a set of tasks. First step was to select an experiment followed by selecting a mathematical model which was described in the experiment from the set of experiments. Then the mathematical model was validated via implementations in platforms like MATLAB (Mathworks, USA) for simulating/validating and understanding the dynamics of the model. An implementation in Action Script (Adobe, USA) (with increased efficiency) was done for the web-based usage and also to decrease the reliability on unnecessary software. Action script was used for programming, since the mathematical models are not too detailed and it could be run at client-end. Documentations for an experiment are also made available. Updated/edited/modified version was uploaded as final version to be accessed through online from [22].

Population ecology simulator consists of three regions, the simulator's viewable window, the growth model menu and the simulation control menu. We have used the

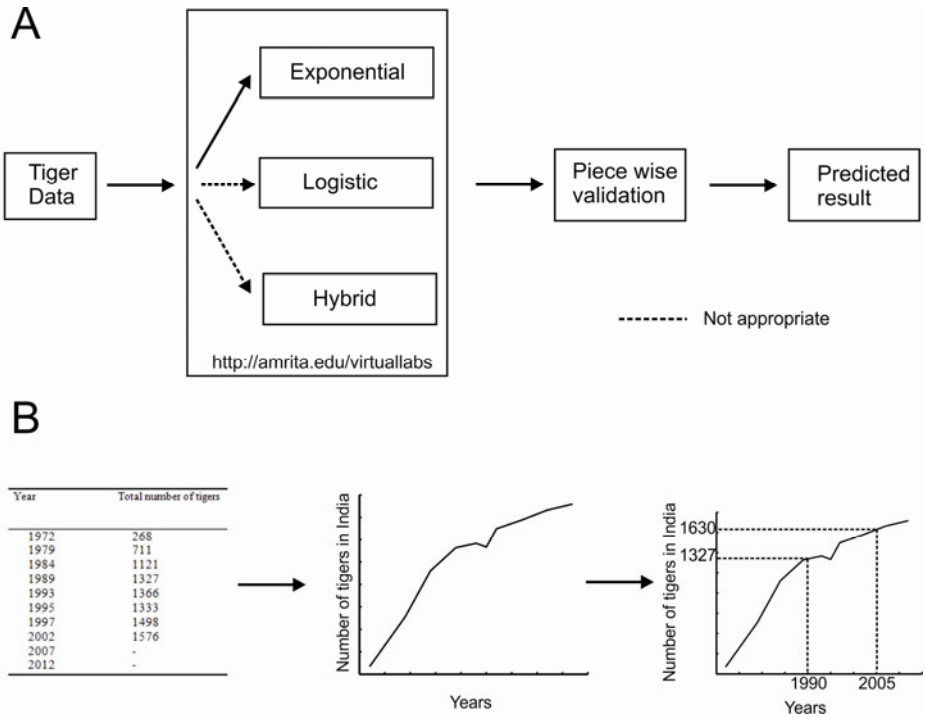


Fig. 2. Schematic representation of tiger population prediction using virtual population ecology simulator. Note that exponential model was chosen based on decline in populations. Predictions with other models such as Lotka-Volterra and Logistic growth were inappropriate or had errors.

continuous growth model in our simulations for study of tiger populations. In growth model menu, user can enter model parameters for both continuous and discrete model [12].

An adaptive growth rate model (see Fig. 2) for different periods as shown in Table 1 was used in estimating tiger populations. Simulator output is shown in Fig. 3.

2.6 Assumptions with Tiger Populations and Growth Model

The following assumptions were used in our study. The base assumption is that estimates are done for a short span of time.

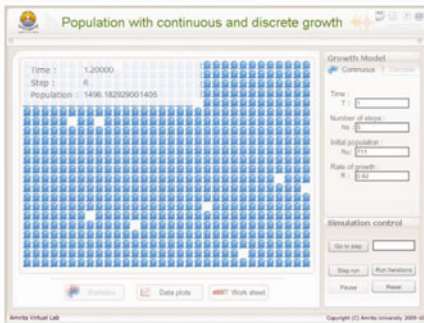
- i. Tiger population grows/declines exponentially within a short duration (10 years).
- ii. The rate of change of tiger population is proportional to its size.
- iii. During the process, the environment does not change in favor of one species and the genetic adaptation is sufficiently slow.

Table 1. Shows the statistical data for tiger population from 1972 – 2002 and extended (prediction column) the curve to 2012 using continuous growth model simulator (data collected from [13])

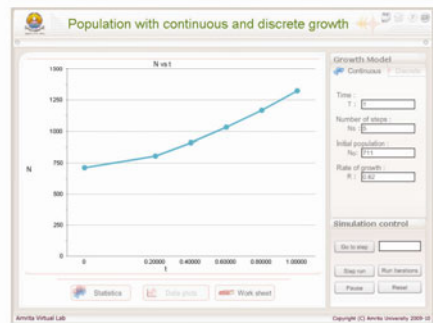
Year	Total number of tigers	Total number of tigers predicted by the model	Growth rate
1972	268	-	-
1979	711	-	0.6230
1984	1121	-	0.3657
1989	1327	-	0.1552
1993	1366	-	0.0285
1995	1333	-	-0.0247
1997	1498	-	0.1236
2002	1576	1586	0.0409
2007	-	1664	0.0468
2012	-	1718	0.0314

Calculated average growth rate was equal to 0.1545.

A



B



C

Population with continuous model				
Trial No	Time	Initial population	Rate of growth	Population
1	1	1333	0.12	1502.953
2	1	1502	0.014	1523.175

Fig. 3. Adaptations of exponential growth patterns. A. The simulator shows population with continuous growth model which has been used to study piece-wise properties of tiger populations in India. B. The simulated piece-wise continuous growth of tiger populations in India C. The model implemented using an online spreadsheet as seen in the user-end of the experiments. Simulator has three main tabs, 1) Statistics button will show the growth of population while the simulation is running (see Fig. 3A), 2) Data plots button, will show population size vs time (see Fig. 3B), 3) Worksheet button is an implementation of the model in excel (see Fig. 3C).

Although the assumptions make it difficult to actually call the simulation ‘realistic’, the validations showed a realistic trend and hence for this model, a simple exponential growth simulator was used.

2.7 Effect of Food, Climatic Conditions and Other Estimates

In most population studies, the tiger and associated prey populations within the sampled area were reasonably well protected [15]. Our model was also based that during the short span of time (estimations for a decade), tendencies such as permitted transience, temporary emigration, and variation in probabilities of initial capture and recapture among individual animals were almost stationary. This suggested that tiger populations in India remained consistently predictable in terms of exponential behavior.

In previous works [17], abundance (number of tigers in the sampled area) was always modeled as time dependent, resulting in one parameter for each primary sampling occasion. Survival was expressed on an annual scale. Annual survival was not modeled in our case. Also, the most parts of the study become inaccurate during the monsoon, which leads to a concern for a specific protection strategy for the specific problems.

3 Results

Taking real data to class rooms have been very difficult with population ecology due to its high unpredictability and model-related unreliability. However, using a simple growth rate simulator and using patterns from a short period, the model shows promise.

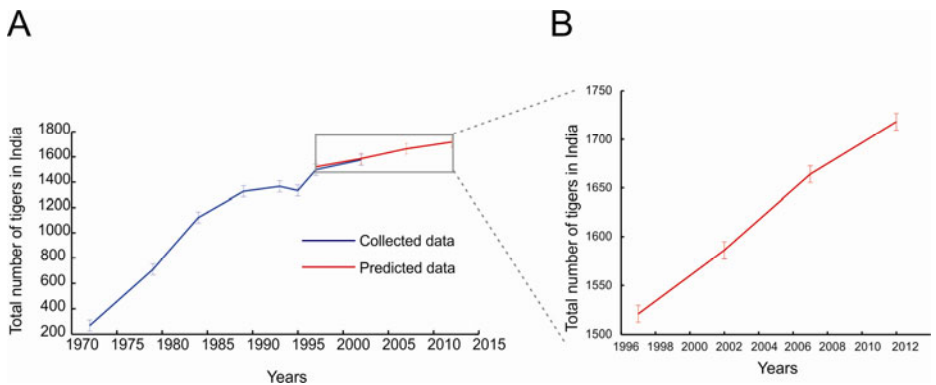


Fig. 4. Growth of tigers with predictions. A. The plot shows the nature / pattern of statistical data for tiger population in India from 1972 – 2002 (blue line) and an extended prediction (red line) of the curve to 2012 using continuous growth model simulator. The model assumes a 10% standard deviation shown by the error bar. B. The plot shows the enlarged curve of predicted piece-wise continuous growth of tiger populations in India.

The simulation showed predictability in the growth of tiger population in India for the year 2007 was 1664. Which is around the upper range to the statistical report released from National Tiger Conservation Authority (ranging between 1,165 and 1,657) in 2008. The difference in number of tigers from predicted to this statistical report may be because of some environmental factors, number of tigers that have been killed in past few years (see Fig. 5A) and the census by National Tiger Conservation Authority was only partially included West Bengal. And the recent release of the 'tiger census 2011' at the International Tiger Conference hosted by Environment and Forests, stated that 'The number of tigers in India's wild has gone up by 20%', the total population of tigers in India now reached 1706 from 1411. The prediction of total population of Indian tiger from our model by 2012 was 1718. The value predicted from our results approximately equal to count from the tiger census 2011.

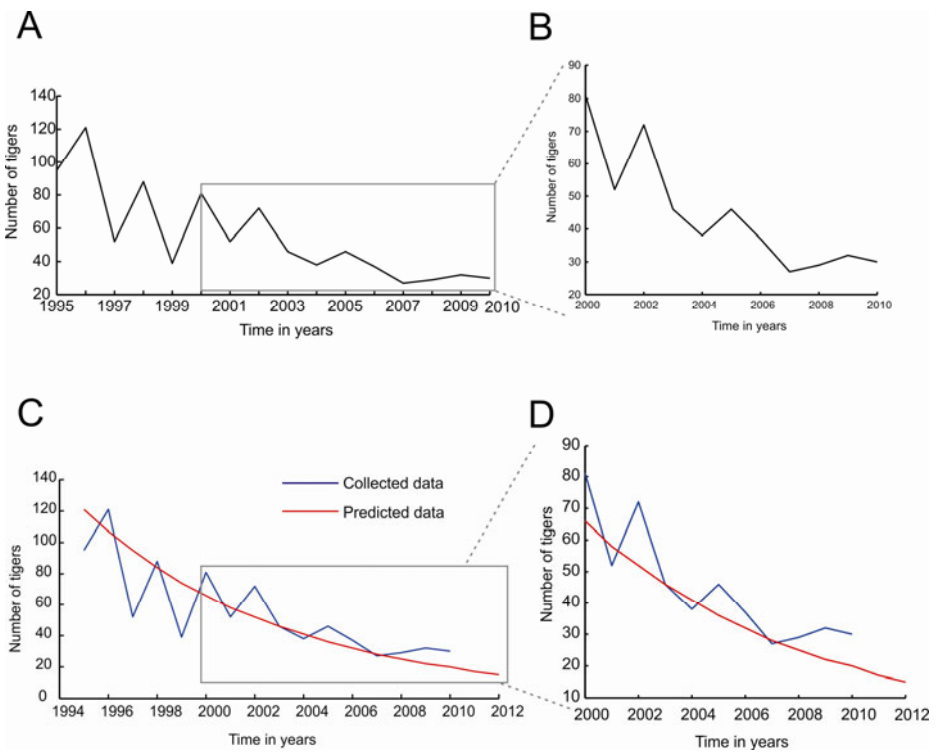


Fig. 5. Decline in tiger population. A. Real data indicative of the decrease in tigers showing the number killed or dead in past few years. Time (in years) is x-axis vs. number of tigers in y-axis (validations for data from [14]). B. The plot shows the zoomed lag part of curve from A, indicate the sudden decrease in number of tigers killed in past few years. Predictions of decline are shown in red in C. Note zoomed portion (D) shows that decline is not easily trended although predictions indicate decline is low as shown by the trend in data (A).

These above validation of results highlight the importance, attention and significant impact on predicting tiger population using online growth model from Virtual Population Ecology Lab.

The model showed a close match with data [14] although in predicting decline of tiger population (see Fig. 5A, C), model followed the trend without too much inaccuracy. Precision in predicting decline with exponential models was unreliable (see Fig. 5D).

The results indicates that project data from wildlife and geographic spread can be studied using tools such as virtual population ecology labs (see [22],[23]). The advantages were that students showed increased awareness in the use and understanding of population data and the use of virtual tools enhanced teaching by 23% as observed in some other studies (see Neuroscience case study in [9]).

4 Conclusion

Our model was parameterized specifically to deal with tiger transience (see [21]), and data selection results provided strong support for these models. The predicted data (Table.1, third column) for tiger population in India showed standard deviation of 10% from real data. With some assumptions, it was possible to use simple models like exponential growth models for studying tiger population. For a very short duration (such as in the data shown in Table 1), basic growth show a slowly saturating exponential and hence data matched the predictions (see Fig. 4A). We also show that using simple dynamics certain behaviors could be generalized in tiger population.

Online population ecology experiments developed on the basis of mathematical equations could help students to get a deeper understanding on model dynamics by exploring the parameter space provided by the model. Also it is always feasible for the user to supply the real data as input and observe the corresponding dynamics. The possibility to study such experiments has value. Biotechnology studies often include data collection and such models allow building simple hypothesis based on the dynamics. This new e-learning environment engaged and motivated the students to practice and explore the parametric space provided for the population ecology experiments.

Newer studies for analyzing fish populations and deer populations are being developed as part of the ongoing process. Such data will be made available as a virtual lab for continued use and study. We also noticed that the undergraduate and postgraduate students show an increased attention to details when we trained them on virtual labs instead of plainly explaining the theory. There was a 23% (metric not shown) improvement in interest to critically analyze population models among students who were introduced to population ecology studies directly virtual labs.

Acknowledgments. This work derives direction and ideas from the Chancellor of Amrita University, Sri Mata Amritanandamayi Devi. The authors would like to acknowledge Prof. Raghu Raman, Head, Center for Research in Advanced Technologies in Education (CREATE @Amrita) and Prof. Prema Nedungadi, Joint Director, CREATE @ Amrita and the entire CREATE team for the development of software simulations and animations of Amrita Virtual Labs. This work was

supported by the Sakshat project of National Mission on Education through ICT, Department of Higher Education, Ministry of Human Resource Department, Government of India.

References

1. Gilman, S.L.: Do Online Labs Work? An Assessment of an Online Lab on Cell Division. *American Biology Teacher* 68(9), 131–134 (2006)
2. Gumaraes, E., Maffei, A., Preire, J., Russo, B., Cardoso, E., Bergeman, M., Magalhaes, M.: REAL: A virtual laboratory for mobile robot experiments. *IEEE Trans. Educ.*, 37–42 (2003)
3. Bijlani, K., Manoj, P., Rangan, V.: VIEW: A Framework for Interactive eLearning in a Virtual World. In: *Proceedings of the Workshop on E-Learning for Business Needs*, BIS, Austria, pp. 177–187 (2008)
4. Mowryn, M.M., Shegog, R., Murray, G.N.: An Innovative Approach to Impacting Student Academic Achievement and Attitudes: Pilot Study of the HEADS UP Virtual Molecular Biology Lab, *Advances in Teaching and Learning Day*, 4–20 (2007)
5. Mahdavi, A., Metzger, G.: Towards a virtual laboratory for building performance and control. In: *Trapp, R. (ed.) Cybernetics and Systems 2002*, pp. 281–286. Austrian Society for Cybernetic Studies, Vienna (2002)
6. Lim, C.P.: Engaging Learners in Online Learning Environments. *TechTrends: Linking Research & Practice to Improve Learning* 48(4), 1623 (2004)
7. Quitadamo, I.J., Brown, A.: Effective teaching styles and instructional design for online learning environments. In: *National Educational Computing Conference*, IL. ERIC Chicago (2001)
8. McDonald, J.: Is “as good as facetoface” as good as it gets? *Journal of Asynchronous Learning* 2(2), 1223 (2002)
9. Diwakar, S., Achuthan, K., Nedungadi, P., Nair, B.: Enhanced Facilitation of Biotechnology Education in Developing Nations via Virtual Labs: Analysis, Implementation and Case-studies. *International Journal of Computer Theory and Engineering* 3(1), 1–8 (2011)
10. Endangered Species,
http://www.theinsite.org/earth/earth_es_tiger.html
11. Tigers, <http://www.edu.pe.ca/southernkings/tiger.htm>
12. Larry, R., Bertola, G.: *Introduction to Population Ecology*. Blackwell Pub., Cambridge (2006)
13. Tiger Project India,
<http://projecttiger.nic.in/populationinstate.asp>
14. WPSI's Tiger Poaching Statistics,
<http://www.wpsi-india.org/statistics/index.php>
15. Karanth, K., Nichols, U., Kumar, J.D., Hines, N.S., Assesing, J.E.: tiger population dynamics using photographic capture-recapture sampling. *Ecology* 87(11), 2925–2937 (2006)
16. Ajan, J., Sharma, S.K.: Camera trapping the Indochinese Tiger, *Panthera tigris corbetti*, in a secondary forest in peninsular Malaysia. *The Raffles Bulletin of Zoology* 51(2), 421–427 (2003)
17. Plowden, C., Bowles, D.: The illegal market in tiger parts in northern Sumatra, Indonesia. *Oryx* 31, 59–66 (1997)

18. Rabinowitz, A.: Chasing the dragon's tail: the struggle to save Thailand's wild cats. Doubleday, New York, 280 (1991)
19. Shaharuddin, W., Potential, N.: poaching threat to large mammals in Belum and Taman Negara. *The Journal of Wildlife and Parks* 16, 135–139 (1999)
20. Seidensticker, J.: Large carnivore and the consequences of habitat insularisation: ecology and conservation of tigers in Indonesia and Bangladesh. In: Miller, S.D., Everett, D.D. (eds.) *Cats of the world biology, conservation and management*, pp. 1–41. National Wildlife Federation, Washington DC
21. Pradel, R., Hines, J.: Capture–recapture survival models taking account of “transients.”. *Biometrics* 53, 60–72 (1997)
22. <http://amrita.edu/virtuallabs>
23. <http://sakshat.amrita.ac.in/VirtualLab/?sub=BIOTECH&brch=POE&ln=1&sim=Population-with-continuous-discrete-growth&cnt=theory>

Green Communications through Network Redesign

Sami J. Habib¹, Paulvanna N. Marimuthu¹, and Naser Zaeri²

¹ Kuwait University

Computer Engineering Department, P. O. Box 5969, Safat 13060 Kuwait

sami.habib@ku.edu.kw

² Arab Open University

Faculty of Computer Studies, P. O. Box 3322, Safat 13033 Kuwait

Abstract. We have extended our redesign network methodology with the capability to explore the redesign space for an enterprise network (EN) to meet the green communication constraints. These constraints attempt to save energy/power those results in efficient usage of existing network resources. The power consumption in EN is gaining importance due to its contribution towards making EN environmental friendly. In this paper, we have viewed EN as a flexible infrastructure, where each node in EN can be moved to join a cluster with high association with its nodes rather than remaining in its original cluster. We have proposed a number of redesign operations at a node-level to change the existing traffic flow by reducing the traffic at the backbone. We have modeled the enterprise network redesign problem as an optimization problem and utilized Simulated Annealing to search for enhanced association between nodes and clusters. We have estimated the power saving through the power spectral density (PSD) of the utilized polar non-return to zero (NRZ) encoding signal. The simulation results demonstrate an overall drop in the backbone traffic, which estimates by 10% power saving.

Keywords: green communications, redesign, simulated annealing, optimization, polar non-return to zero, power spectral density.

1 Introduction

A green enterprise network involves a minimal environmental impact in its entire lifespan from design, operation, and disposal. One of the ideas to introduce the green communication in an enterprise network (EN) is to reduce the traffic at the backbone through redesign operations. By relocating nodes, we were able to reduce the backbone traffic, thereby reducing the power consumption at the backbone. Here, we have defined the EN as a multitier network offering collection of services to the nodes; moreover, it is viewed as a group of subnets (clusters) comprising of heterogeneous nodes. Each cluster is defined with certain internal and external traffic initially. By redesigning the existing clusters, the traffic at the backbone is reduced and the performance of the network is increased.

In this paper, we have extended our redesign network methodology [11] to investigate the impact of our redesign process on power saving and environmental

responsive. We have computed the changes at the backbone after the proposed redesign operations of a clustered EN with prior traffic matrix pattern. We have proposed few redesign operations on node-level namely, move and swap operations to relocate the nodes into a better cluster, thereby reducing the traffic at the backbone. This reduction in the backbone traffic shows reduction in power consumption at the backbone links. We have employed Simulated Annealing as a search algorithm to explore the best clusters to relocate all nodes. By giving an opportunity to move the nodes to proper clusters, we are decreasing the external traffic at the backbone. We have utilized the polar non-return to zero (Polar NRZ) encoding scheme to transmit the generated data; moreover, we have estimated the power consumption at the backbone. Our simulation results demonstrate an overall decrease in the backbone traffic in *move* and *swap* redesign operations causes 10% and 3% savings in power consumption respectively. Overall, the redesign operations decrease the power consumption at the backbone.

2 Related Work

Increase in number of nodes in an enterprise network increases the energy consumption of the network. Many of the researchers discussed about the energy consumption of the network. Fisher et al. [1] considered the links in the core network as bundles of multiple cables and line cards that can be shut down independently. They evaluated their heuristics based on linear optimization technique to reduce the energy consumption.

Barroso and Holzle [2] suggested energy proportionality should be a primary design goal and they also discussed about achieving energy-proportionality at various levels through significant improvements in the energy usage profile of every system component, particularly the memory and disk subsystems. Nguyen et al. [3] developed an analytical model and applied the model to evaluate performance of the existing energy efficient Ethernet and wireless technologies through the number and the average throughput of clients. They also examined the combination of both techniques.

Yang et al. [4] studied the performance impact of traffic patterns on energy consumption through the analysis of usage time of links via simulation. They performed the simulation using OPNET Modeler. Chiaraviglio et al. [5] discussed the energy savings issues at the backbone network by turning off the spare devices. They claimed that they were able to save 23% energy per year. Banerjee et al. [6] provided analytical tools for the integrated evaluation of job management with respect to dynamic cooling behavior and they also discussed that by designing cooling-aware job management can reduce the data center energy consumption.

Gupta and Singh [7] designed a dynamic Ethernet link shutdown (DELS) algorithm that utilizes current technology leading to significant benefits in energy savings with little impact on packet loss or delay. The authors utilize buffer occupancy, the behavior of previous packet arrival times and a configurable maximum bounded delay in their algorithm to make sleeping decisions. The general network topology redesign problem has been discussed by various researchers in different dimensions. Park and Delis [8] discussed the problem of node realignment

and they proposed an online re-clustering framework to adaptively redistribute the node locations. They implemented it using the knowledge of nodes' data access patterns in a three-tier architecture that features logical node clustering.

Xu et al. [9] described the logical topology reconfiguration of wavelength division multiplexing (WDM) optical networks to optimize the network resource utilization corresponding to the changing traffic conditions. They employed a two-stage Simulated Annealing algorithm for multi-objective optimization.

In this paper, we have extended our redesign network methodology [11] to investigate the impact of the methodology on power savings at the backbone for now.

3 Network Model

We view an enterprise network as a multitier comprising of two types of clusters. The first cluster type is a set of subnets containing all nodes, and the second cluster type is a subnet (backbone) connecting all the nodes' clusters, as illustrated in Figure 1. We have considered a medium enterprise network comprised of 100 nodes, with prior known traffic pattern, distributed uniformly in an area of 200m x 200m. The nodes are grouped into 10 clusters. The traffic matrix describes the volume of traffic between each pair of nodes, and it describes the ideal network condition at the initial design time. The enterprise network, with known local and outgoing traffic as given in Table 1 for each cluster reflects the practical environment for simulating various redesign operations, leading to change the existing topology to reduce power consumption at the backbone.

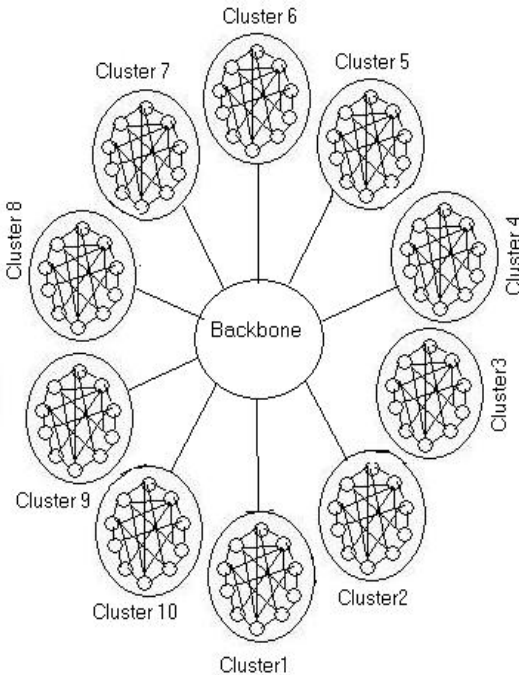


Fig. 1. Enterprise network model

Table 1. Initial traffic per cluster

Cluster	Local Traffic (Kbits/sec)	Outgoing Traffic (Kbits/sec)
1	799	7321
2	1208	7099
3	919	7410
4	1220	7764
5	1606	8045
6	794	7347
7	934	7914
8	1417	7738
9	1690	8794
10	1867	9735

4 Enterprise Network Redesign Problem Formulation

4.1 Network Redesign

We have formulated the topology redesign problem as a single optimization problem, where the main objective is to minimize the extra-traffic as stated in Equation (1). The term $\Psi_{i,j}$ represents the traffic flow from node i to node j belonging to different clusters (C_k and C_l). Constraint (2) indicates that the total number of clusters within the network should be bound between 2 clusters and the ratio of N over 2 to the nearest integer, where N represents the total number of nodes in the network. This constraint is formulated with the assumption that the minimum number of nodes in the network will be greater than or equal to 4. Constraint (3) ensures that a node is bound to a single cluster, as the selection of nodes is made random during each redesign operations. Constraint (4) guarantees that the summation of all nodes in all the clusters should be equal to N . Thus, there is no overlooked node and no added new node. Constraint (5) is added for wired nodes and it makes certain that the total cable length within a cluster is greater than zero and less than or equal to a given threshold values, D_{TH} . The function $COM()$ estimates the center-of-mass of a given cluster; moreover, the function $Dist()$ calculates the Euclidian distance from a node to its center-of-mass, whereas (x_i, y_i) represents the spatial distribution of each node in the network. In our redesign operations, we have fixed a distance threshold value of 300m to be the Euclidian distance from the center of mass of each cluster to their nodes, which is the optimum range of distribution of oppnet nodes. Constraint (6) ensures that the total number of bound nodes to a cluster should be balanced.

$$Min \sum_{i \in C_k \text{ and } j \in C_l} \Psi_{i,j} \quad (1)$$

$i \neq j \text{ and } k \neq l$

subject to

$$2 \leq \sum_{k=1} C_k \leq \left\lfloor \frac{N}{2} \right\rfloor \quad (2)$$

$$\sum_{k=1}^{\left\lfloor \frac{N}{2} \right\rfloor} \alpha_{i,k} = 1 \text{ for } i = 1, 2, \dots, N \quad (3)$$

$$\sum_{k=1}^{\left\lfloor \frac{N}{2} \right\rfloor} |C_k| = N \quad (4)$$

$$0 < \sum_{i \in C_k} Dist((x_i, y_i), COM(C_k)) \leq D_{TH} \text{ for } k = 1, 2, \dots, \left\lfloor \frac{N}{2} \right\rfloor \quad (5)$$

$$2 \leq \sum_{i \in C_k} \alpha_{i,k} \leq \left\lfloor \frac{N}{2} \right\rfloor \text{ for } k = 1, 2, \dots, \left\lfloor \frac{N}{2} \right\rfloor \quad (6)$$

5 Power Estimation Utilizing Polar NRZ Scheme

The average power is an important concept in communication systems. If a traffic volume is resistive (unity power factor), and then the average power is given by Equation (7).

$$P = \frac{\langle v^2(t) \rangle}{R} = \langle i^2(t) \rangle R \quad (7)$$

R is the value of the resistive traffic volume, $v(t)$ denote the voltage across a set of circuit terminals, $i(t)$ denote the current into the terminal, and $\langle [\cdot] \rangle$ is the time average operator. The concept of normalized power is often used by communication research community. In this paper, the notion R is assumed to be 1Ω . Consequently, the average normalized power is given by Equation (8).

$$P = \langle w^2(t) \rangle = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} w^2(t) dt \quad (8)$$

$w(t)$ represents a real voltage or current waveform. In dealing with communication systems, it is very useful to relate the normalized power of a waveform to its frequency domain description by the use of power spectral density (PSD), which describes how the power content of signals is affected by the devices in communication systems. The PSD for a deterministic power waveform is defined by Equation (9).

$$P_w(f) \triangleq \lim_{T \rightarrow \infty} \left(\frac{|W_T(f)|^2}{T} \right) \quad (9)$$

$w_T(t)$ is the truncated version of the waveform, $W_T(f)$ is the Fourier transform of $w_T(t)$ and $P_w(f)$ has units of watts per hertz. Then the normalized average power is given by Equation (10). The area under the PSD function is the normalized power.

$$P = \langle w^2(t) \rangle = \int_{-\infty}^{\infty} P_w(f) df \quad (10)$$

Binary 1's and 0's in digital communication may be represented in various serial-bit signaling formats called line codes. The waveforms for the line code may be classified according to the rule that is used to assign voltage levels to represent the binary data. The PSD for a line code can be evaluated using the stochastic approach.

We have selected polar NRZ encoding to represent our traffic volume. The PSD of a polar NRZ signal [12] with unit amplitude and rectangular pulse is given by Equation (11).

$$P_{polar-NRZ}(f) = T_b \left(\frac{\sin(\pi \times f \times T_b)}{\pi \times f \times T_b} \right)^2 \quad (11)$$

The bit rate (BR) = $1/T_b$. This is a sinc function, which implements that for a certain frequency as T_b is increased (the bit rate is decreased) the power is decreased.

6 An Overview of Network Redesign Operations

We have proposed various redesign operations namely node *move* and *swap* operations to change the localities of existing nodes. Let C be the number of clusters available initially and N be the total number of nodes in the network. Relocation of nodes by move operation has been implemented by the random selection of two clusters (C_i and C_j), followed by the random movement of a node n_i in cluster C_i into a second cluster C_j . We restrict the selection of C_i further by excluding it from the main list of C clusters. The swap operation interchanges two randomly selected nodes n_i and n_j in two randomly selected clusters (C_i and C_j).

The *move* and *swap* operations act as two neighbourhood functions in Simulated Annealing (SA) and the simulation results are generated after repeated trials of each neighborhood operations within SA.

7 Results and Discussion

We have started with an EN network of 100 nodes distributed evenly in 10 clusters as illustrated in Figure 1, having backbone traffic of around 100,000 Kbits/sec. We have employed redesign operations on nodes such as *move* and *swap* within Simulated Annealing [10] to redesign the given clusters to reduce the traffic at the backbone. In the simulation study using SA, we have selected the initial high temperature to be 5000°C , α , the cooling rate to be 5% and the β , the factor increasing the number of iterations to be 5%. The maximum time is selected as 100,000 units.

We have adopted polar NRZ signaling scheme to encode the transmitted data. Using the concept of digital transmission as described in section 5 and using Equation (11), we have estimated the power consumed by the traffic flow at the backbone in our enterprise network before and after the redesign operations. Figure 2 depicts the average power consumed during various traffic volumes by its PSD for two scenarios (a) and (b). From Figure 2, it is understood that the traffic volume in scenario (b) is less than its counterpart in scenario (a). It is observed that as the traffic volume is decreased, the average power consumption of the signal is also decreased.

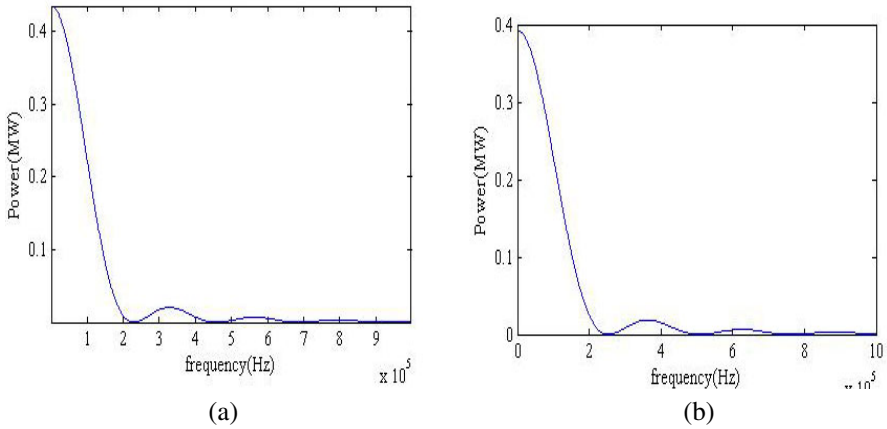


Fig. 2. PSD computed for a various traffic volume utilizing polar NRZ encoding

From our experimental results, it is observed that the move operation on nodes within SA results reduction in traffic after optimization, thereby showing a power saving of 10%. From Figure 3, it is also observed that the power reduction is seems to be prominent during the initial stages of optimization and the difference decreases towards the end. This concludes that the redesign operations reaches a saturation point, further optimization does not improve the topology.

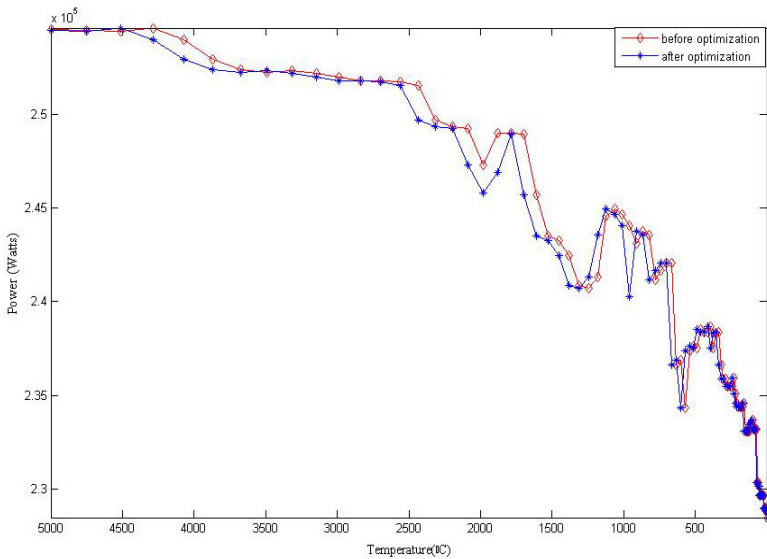


Fig. 3. Power savings during move node operation

Figure 4 explains the power savings during swap node operation. It is also observed that the topology redesign operations along with the selection of optimal annealing conditions cause reduction in the total extra-traffic, thereby saving 3% power in the swap neighborhood operations. It is also observed that the random swapping operations on nodes shares the gain between the clusters, causing oscillation in the traffic flow, thereby causing ups and downs in the power consumption curve.

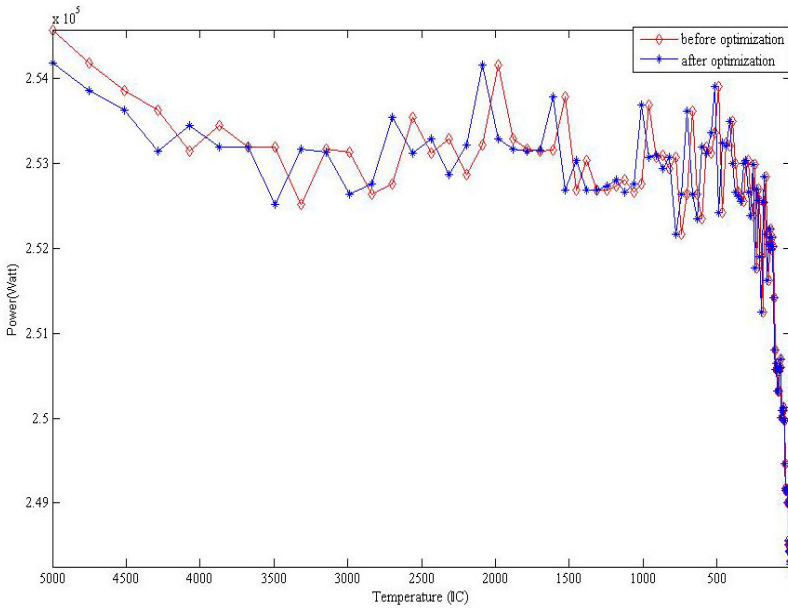


Fig. 4. Power savings during swap node operation

8 Conclusion and Future Work

We have extended our network redesign methodology to meet the green communication constraints by relocating nodes among the enterprise's clusters and reducing the traffic volume at the backbone. Thus, the reduction of traffic volume has a proportional effect on the power saving. We have utilized Simulated Annealing with the proposed redesign operations to search for nodes' relocations subject to consume less power. With polar NRZ encoding scheme to transmit the data, our redesign operations within SA show a maximum power gain of 10% at the backbone.

We are continuing our research by considering other power estimation schemes and also applying other optimization algorithms to compare our redesign process.

References

- [1] Fisher, W., Suchara, M., Rexford, J.: Greening Backbone Networks: Reducing Energy Consumption by Shutting off Cables in Bundled Links. In: International Workshop on Research Challenges in Security of Privacy for Mobile and Wireless Networks (WSPWN 2006), Miami, Florida, Florida (March 2006)
- [2] Barroso, L., Holzle, U.: The Case for Energy Proportional Computing. *IEEE Journal of Computer* 40(12), 134–141 (2007)
- [3] Le Nguyen, P., Morohashi, T., Imaizumi, H., Morikawa, H.: A Performance Evaluation of Energy Efficient Schemes for Green Office Networks. In: IEEE Green Technologies Conference, Grapevine, TX, April 15-16, pp. 1–9 (2010)
- [4] Yang, W.H., Kang, D.-K., Tong, F., Kim, Y.-C.: Performance Analysis of Energy Savings according to Traffic Patterns in Ethernet with Rate Adaptation. In: IEEE 12th International Conference on Computer Modeling and Simulation, Cambridge, UK, March 24-26, pp. 619–624 (2010)
- [5] Chiaraviglio, L., Mellia, M., Neri, F.: Energy-aware Backbone Networks: a Case Study. In: First International Workshop on Green Communications, Dresden, Germany, June 14-18 (2009)
- [6] Banerjee, A., Mukherjee, T., Varsamopoulos, G., Gupta, S.K.S.: Cooling-aware and Thermal-aware Workload Placement for Green HPC Data Centers. In: IEEE Green Computing Conference, Chicago, USA, August 15-18, pp. 1–12 (2010)
- [7] Gupta, M., Singh, S.: Dynamic Ethernet Link Shutdown for Power Conservation on Ethernet Links. In: IEEE International Conference on Communications, ICC 2007, Glasgow, Scotland, June 24-28 (2007)
- [8] Park, J., Delis, A.: On-line Realignment of Clients in Networked Databases. In: 21st IEEE International Conference on Distributed Computing Systems, Phoenix, Arizona, USA, pp. 421–428 (April 2001)
- [9] Xu, S., Sezaki, K., Tanaka, Y.: A two-stage simulated annealing Logical Topology Reconfiguration in IP over WDM Networks. *IEICE Transactions on Communications* E88-B(6), 2483–2494 (2005)
- [10] Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. *Journal of Science* 220(4598), 671–680 (1983)
- [11] Habib, S., Marimuthu, P.N., Taha, M.: Networks Consolidation through Soft Computing. In: Rauch, J., Raś, Z.W., Berka, P., Elomaa, T. (eds.) ISMIS 2009. LNCS, vol. 5722, pp. 542–551. Springer, Heidelberg (2009)
- [12] Couch II, L.W.: Digital and Analog Communication Systems. Pearson International Edition, USA (2007)

Unsupervised Modified Adaptive Floating Search Feature Selection

D. Devakumari¹ and K. Thangavel²

¹ Assistant Professor, Department of Computer Science, L.R.G. Government Arts College for Women, Tirupur

² Professor & Head, Department of Computer Science, Periyar University, Salem

Abstract. In feature selection, a search problem of finding a subset of features from a given set of measurements has been of interest for a long time. An unsupervised criterion, based on SVD-entropy (Singular Value Decomposition), selects a feature according to its contribution to the entropy (CE) calculated on a leave-one-out basis. Based on this criterion, this paper proposes a Modified Adaptive Floating Search feature selection method (MAFS) with flexible backtracking capabilities. Experimental results show that the proposed method performs better in selecting an optimal set of the relevant features. Features thus selected are evaluated using K-Means clustering algorithm. The clusters are validated by comparing the clustering results with the known classification. It is found that the clusters formed with selected features are as good as clusters formed with all features.

Keywords: Data Mining, Unsupervised Feature Selection, Contribution Entropy (CE), Adaptive Floating Search (AFS), Modified Adaptive Floating Search (MAFS), Clustering.

1 Introduction

Feature selection involves selecting a particular set of features of the original problem. Feature filtering is a process of selecting features without referring back to the data classification or any other target function. Hence we find filtering as a more suitable process that may be applied in an unsupervised manner [1].

Existing methods of unsupervised feature filtering include ranking of features according to range or variance [2, 4], selection according to highest rank of the first principal component [5, 6] and other statistical criteria. An intuitive, efficient and deterministic principle, depending on authentic properties of the data, which serves as a reliable criterion for feature ranking is based on SVD-entropy, selecting a feature according to its contribution to the entropy (CE) calculated on a leave-one-out basis [7]. It has been demonstrated that this principle can be turned into efficient and successful feature selection methods like simple ranking according to CE values (SR), sequential forward selection by accumulating features according to which set produces highest entropy (SFS1); sequential forward selection by accumulating

features through the choice of the best CE out of the remaining ones (SFS2); sequential backward elimination (SBE) of features with the lowest CE.

Here, we present another sequential search method, the Modified Adaptive Floating Search Feature Selection (MAFS), which attempts to compensate for the weaknesses of SFS and SBE with some backtracking capabilities. This method is an extension of the LRS algorithms which repeatedly adds L features and removes R features. Rather than fixing the values of L and R, the floating search method allows those values to be determined from the data.

The organization of the paper is as follows: Section 2 presents the background of the proposed work. The proposed work is discussed in section 3. The experimental results are provided in section 4. Analysis and Discussion of the results are provided in section 5. This paper concludes in section 6.

2 Background

Alter et al., [14] have defined a SVD based entropy E of the dataset as follows:

$$E = - \frac{1}{\log(n)} \sum_{j=1}^n V_j \log(V_j) \tag{1}$$

The contribution of the i^{th} feature to the entropy (CE_i) by a leave-one-out comparison according to

$$CE_i = E(A_{[n \times m]}) - E(A_{[n \times (m-1)]}) \tag{2}$$

where the i^{th} feature was removed in $A_{[n \times (m-1)]}$. Let us define the average of all CE to be c. We distinguish then, between three groups of features:

- (i) $CE_i > c$, features with high contribution
- (ii) $CE_i = c$, features with average contribution
- (iii) $CE_i < c$, features with low (usually negative) contribution

Entropy maximization has been implemented in three different ways: Simple ranking, Sequential Forward Selection and Sequential Backward Elimination.

3 Proposed Work

The Adaptive Floating Search Feature Selection (AFS) method attempts to compensate for the weaknesses of SFS and SBE with some backtracking capabilities. This method is an extension of the LRS algorithms which repeatedly adds L features and removes R features. Rather than fixing the values of L and R, the floating search method allows those values to be determined from the data.

The main limitation of LRS is the lack of a theory to help predict the optimal values of L and R. The AFS method is an extension of LRS since there is no need to specify any parameters such as L and R. The number of forward (adding) or backward (removing) steps is determined dynamically during the method's run so as to maximize the criterion function.

Two inequalities had already been proposed for AFS method, based on which we can determine the number of features to be added and removed respectively.

The first inequality is as follows:

$$\text{Max CE} - \text{Average CE} \leq \text{CE}(m) \leq \text{Max CE} \quad (3)$$

Each feature m which satisfies inequality (3) will be selected in one step.

The second inequality is as follows:

$$\text{Min CE} \leq \text{CE}(m) \leq \text{Min CE} + \text{Average CE} \quad (4)$$

If any feature m which satisfies inequality (4) has been included in the selected list, it is removed. Both set of selected and removed features are eliminated from the full set of features. CE values are calculated for the remaining features and the process is repeated till optimal features are selected.

In both these inequalities, the Average CE is calculated taking into account positive, negative and zero CE values. If in a dataset, features with negative and zero CE are more, then the average CE value will be very less. Therefore the number of features selected or eliminated during an iteration will be reduced. Hence it would be better if only positive CE values are considered while calculating the average CE. This led to the modified version of AFS namely MAFS.

The above mentioned inequalities (3) and (4) are modified as follows to be used in MAFS method:

$$\text{Max CE} - \text{Positive Average CE} \leq \text{CE}(m) \leq \text{Max CE} \quad (5)$$

$$\text{Min CE} \leq \text{CE}(m) \leq \text{Min CE} + \text{Positive Average CE} \quad (6)$$

The pseudo code of MAFS method is given in below:

1. Start with $M_{\text{AFS}} = \text{Empty set}$ and $M'_{\text{PLUS}} = M'_{\text{MINUS}} = M$.
2. Calculate the CE score of each element in M .
3. Select each element m from M'_{PLUS} which satisfies inequality (5).
4. Remove it from M'_{PLUS} and insert into M_{AFS} .
5. Select each element m from M'_{MINUS} which satisfies inequality (6).
6. Remove it from M'_{MINUS} and also from M_{AFS} (if it is already included in M_{AFS}).
7. While size of M'_{PLUS} and $M'_{\text{MINUS}} > 0$
 - a) For each element m in M'_{PLUS} and M'_{MINUS} (that is not included in M_{AFS}) recalculate its CE score.
 - b) Go to step 3
8. End.

Fig. 1. Pseudo code for Modified Adaptive Floating Search (MAFS) method

4 Experimental Results

The MAFS1 method of feature selection is experimented with different data sets from the UCI machine learning repository [www.archive.ics.uci.edu]. Features are selected using Adaptive Floating Search method (AFS1) and MAFS1. The results are tabulated in Table 1:

Table 1. Features selected by the AFS and MAFS1 methods using contribution entropy for different datasets

Data set	No of Instances	Total No of Features	No of Features Selected		No of Iterations	
			AFS	MAFS	AFS	MAFS
Lung Cancer	32	56	23	20	8	3
Cardiac Tomography	187	44	9	10	6	3
Audiology	226	67	43	48	10	6
Dermatology	366	33	19	18	5	4
Radar data from Ionosphere	351	34	20	16	2	1
Sonar signals	208	60	18	28	13	9

The MAFS feature selection method for two of the above data sets is explained below:

Ionosphere Dataset: Out of the total 34 features, 16 features have positive CE and the remaining features have negative CE. The AFS method selects all the 16 features and eliminates only 10 features during first iteration. For the remaining features CE is recalculated. During the second iteration, 4 features are selected and 4 are eliminated. Hence total features selected is 20 and number of iterations are 2.

MAFS method selects all the 16 features and eliminates all the remaining 18 features during the first iteration itself.

Sonar Dataset: Out of the total 60 features, 11 features have 0 CE values, 12 features have negative CE values and the remaining features have positive CE values. The features with higher CE values are given in table 2:

During the first iteration, the AFS method selects features {35, 36}. When CE values are recalculated for the remaining features during the second iteration, the CE value of feature 17 decreases. Hence, during successive iterations, features {37, 38, 39} are selected by AFS method, but feature 17 is eliminated during the 9th iteration. Proceeding in this way, the AFS method selects 18 features in 13 iterations.

Since the MAFS method considers only positive CE values for average calculation, the average CE value is more when compared with AFS method. Therefore, the MAFS method selects features {17, 35, 36, 37, 38, 39} during the first iteration itself. Hence MAFS method selects 28 features in 9 iterations.

Table 2. Features with high CE values from Sonar dataset

Sonar Dataset	
Feature Number	CE Value
36	0.0025
35	0.0023
37	0.0022
38	0.0016
39	0.0016
17	0.0016

The optimality of the features selected by the proposed method is evaluated using K-Means clustering algorithm. The aim of K-Means algorithm is to minimize the objective function or performance function. Objective function measures the total sum of distances i.e., the within-cluster sums of point-to-centroid distances. Results show that the MAFS1 method considerably reduces the objective function when compared to the existing methods as illustrated in Table 3.

Table 3. Values of Objective function computed by K-Means clustering algorithm for features selected by AFS and MAFS methods.

Data set	Objective Function Values		
	All features	AFS	MAFS
Lung Cancer	395.762	209.242	186.714
Cardiac Tomography	563966	171989	190732
Audiology	580.247	179.837	196.701
Dermatology	3682.29	2427.82	2264.53
Radar data from Ionosphere	2419.36	1579.24	1210.51
Sonar signals	235.031	91.791	139.542

5 Analysis and Discussion

Next the quality of the clusters constructed with all features and with selected features are assessed using some common indices like Rand Index and Jaccard Coefficient. Since AFS1 is an unsupervised feature selection method, we have omitted the class labels that were originally present in the datasets. Now, for the purpose of cluster validation, we consider the classification details. Clustering results are compared with actual classification to check the reliability of the clusters. The results are tabulated as follows:

Table 4. Rand Index and Jaccard Coefficient values for features selected by AFS and MAFS methods

Datasets	Rand Index			Jaccard Score		
	All Features	AFS Method	MAFS Method	All Features	AFS Method	MAFS Method
Lung Cancer	0.6598	0.6708	0.6351	0.3422	0.3460	0.3262
Cardiac Tomography	0.5938	0.5892	0.5892	0.5801	0.5750	0.5750
Audiology	0.7913	0.7803	0.7608	0.0954	0.0915	0.1082
Dermatology	0.8484	0.9363	0.9089	0.5455	0.6407	0.6295
Radar data from Ionosphere	0.5901	0.5141	0.5161	0.4358	0.3792	0.4345
Sonar signals	0.5056	0.5118	0.5134	0.3401	0.3616	0.3578

From Table 4, we find that, for all the datasets, both Rand Index and Jaccard Coefficient values are more or less same when calculated with all features, with features selected by AFS method and with features selected by the proposed MAFS method. The following figures present the graphical representations of cluster validations using Rand Index and Jaccard Coefficient respectively, for chosen datasets.

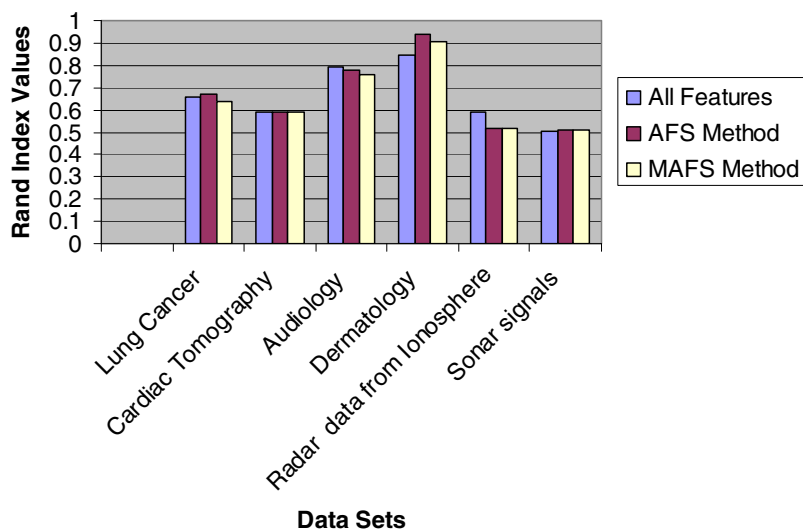


Fig. 2. Graphical representation of Rand Index values for clusters formed with all features and features selected with AFS and MAFS methods

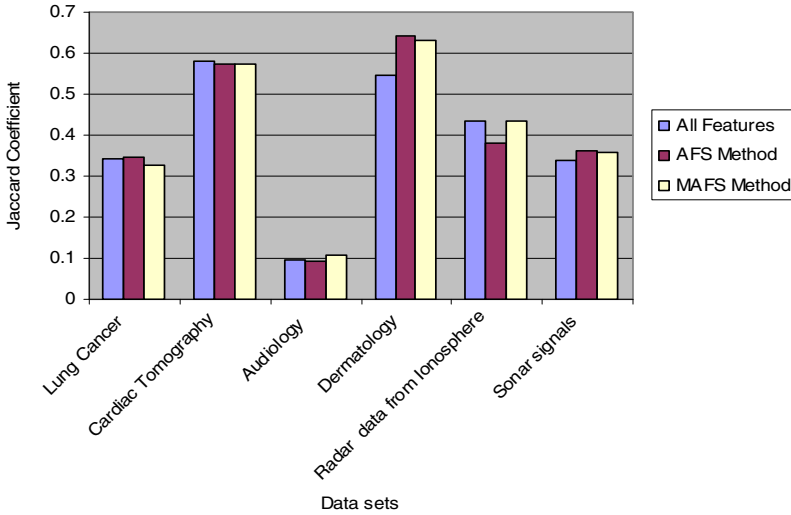


Fig. 3. Graphical representation of Jaccard Coefficient values for clusters formed with all features and features selected with AFS and MAFS methods

This illustrates that, an optimal set of features selected using AFS1 method is sufficient to maintain the quality of clusters equivalent to those clusters formed with all the features. Hence, we conclude that the proposed method effectively reduces the number of features while at the same time preserving the classification accuracy.

6 Conclusion

A novel principle for unsupervised feature filtering is based on maximization of SVD-entropy. The features are ranked according to their CE values. Based on this principle, four feature selection methods have already been implemented. This paper proposes the Modified Adaptive Floating Search feature selection (MAFS) method in which features are added and deleted according to their CE values. This method attempts to compensate for the weaknesses of forward selection and backward elimination methods by allowing flexible backtracking capabilities. The proposed method is experimented and proved to select better features as compared with the other existing methods.

Further more, the selected features were evaluated by K-Means clustering algorithm and it was found to reduce the objective function. The clusters were validated by comparing the clustering results with the known classification. Common indices like Rand Index and Jaccard Coefficient were used for this purpose. It was found that the clusters formed with selected features were as good as clusters formed with all features.

References

1. Handl, J., Knowles, J.: Feature Subset Selection in Unsupervised Learning via Multiobjective Optimization. *International Journal of Computational Intelligence Research* 2(3), 217–238 (2006)
2. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 3, 1157–1182 (2003)
3. Liu, H., Li, J., Wong, L.: A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. In: Lathrop, R., Miyano, K.N.S., Takagi, T., Kanehisa, M. (eds.) *13th International Conference on Genome Informatics*, pp. 51–60. Universal Academy Press, Tokyo (2002)
4. Herrero, J., Diaz-Uriarte, R., Dopazo, J.: Gene expression data preprocessing. *Bioinformatics* 19, 655–656 (2003)
5. Ding, C.H.Q.: Unsupervised Feature Selection Via Two-way Ordering in Gene Expression Analysis. *Bioinformatics* 19, 1259–1266 (2003)
6. Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., Brown, P.: Gene Shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* (2000)
7. Varshavsky, R., Gottlieb, A., Linial, M., Horn, D.: Novel Unsupervised Feature Filtering of Biological Data. *Bioinformatics* 283, 1–5 (2005)
8. Liu, H., Yu, L.: Towards Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering* 17, 491–502 (2005)
9. Dy, J.G., Brodley, C.E.: Feature selection for unsupervised learning. *Journal of Machine Learning Research* 889, 845 (2004)
10. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine intelligence* 312, 301 (2002)
11. Sondberg-Madsen, N., Thomsen, C., Pena, J.M.: Unsupervised feature subset selection. In: *Proceedings of the Workshop on Probabilistic Graphical Models for Classification*, vol. 82, p. 71 (2003)
12. Guo, D., Gahegan, M., Peuquet, D., MacEachren, A.: Breaking down dimensionality: An effective feature selection method for high-dimensional clustering. In: *Proceedings of the Third SIAM International Conference on Data Mining*, vol. 42, p. 29 (2003)
13. Dash, M., Liu, H.: Handling large unsupervised data via dimensionality reduction. In: *Proceedings of the ACM SIGMOD Workshop on Research Numbers in Data Mining and Knowledge Discovery* (1999)
14. Alter, O., Brown, P.O., Botstein, D.: Singular value decomposition for genome-wide expression data processing and modeling. *PNAS* 97, 10101–10106 (2000)
15. Somol, P., Pudil, P., Novovicova, J., Paclik, P.: Adaptive Floating Search Methods in Feature Selection. *Pattern Recognition Letters* 20, 1157–1163 (1999)

Fast and Efficient Mining of Web Access Sequences Using Prefix Based Minimized Trees

M. Thilagu¹ and R. Nadarajan²

¹ VLB Janakiammal College of Engineering and Technology, Coimbatore, India
mthilagu@gmail.com

² PSG College of Technology, Coimbatore, India
nadarajan_psg@yahoo.co.in

Abstract. Web access sequence mining discovers hidden information or knowledge from weblogs containing web usage patterns. The discovered knowledge is useful in many ways for web designers or decision makers to improve the website organization. Several algorithms have been proposed to mine web access sequence patterns and in general they generate candidate sequences and test them during the mining process. This paper describes a fast and efficient algorithm to discover web access sequences by constructing a data structure called prefix based minimized WAS-tree with maximal potential sequence patterns. The tree is recursively constructed and mined to find all the patterns in the database, satisfying the given min-sup. To prove that our algorithm is fast and efficient when compared to an existing algorithm, we have done experimental studies on a real dataset.

Keywords: Web Usage Mining, Web Access Sequence Tree, Maximal Potential Sequence Pattern, Data Mining.

1 Introduction

With the tremendous growth of the amount of information available online, the World Wide Web has rapidly become the main source of information to the majority of users in many domains. The huge data available in the World Wide Web can be mined mainly in three different dimensions, which are web content mining, web structure mining and web usage mining. This paper is related to the web usage mining which can be defined as the application of data mining techniques to web log data in order to discover user access patterns[12]. Web usage mining has various applications such as link prediction, site reorganization and web personalization. Web usage patterns of users can be found at server side, proxy side or client side as a web log file. Web access sequence mining is based on the web log file. A web log file registers the access information of users, including IP address, access time, request URL, referrer, user-agent and so on [2]. We can find out all the access sequential patterns from web access log. But the raw data in the web access log can't be mined directly. Hence, preprocessing is needed on those raw data and patterns are discovered by applying a sequence mining algorithm and finally patterns are analyzed and interpreted [10][11].

Web access sequence mining or web usage mining discovers interesting navigation patterns of web users. Most of the web sequential pattern mining methods are based on session level. Each sequence is the click-stream in a user session. Chen *et al.* [1] developed two efficient algorithms for determining large reference sequences for web traversal paths. The first one, called full-scan (FS) algorithm solved discrepancy between traversal patterns and association rules. The second one, called selective-scan (SS) algorithm is able to avoid database scans in some passes so as to reduce the disk I/O cost involved. IncWTP and WssWTP algorithms [3] are designed for incremental and interactive mining of web traversal patterns. Pei *et al.* proposed an algorithm web access pattern mine (WAP-mine) by using a tree structure called WAP-tree [4]. WAP-mine algorithm adopts a sequential pattern growth mining approach to avoid the level-wise candidate generation-and-test for the existing algorithms. Another approach was proposed to perform efficient mining of web access sequences with position coded pre-ordered linked WAP-tree [5].

The existing algorithms for web access sequence mining discussed above are either apriori-based, pattern-growth or tree-based approaches [6][7][8][9]. Apriori-based and pattern-growth methods suffer from candidate generation and multiple scans of the database, whereas tree-based algorithms has the advantage of mining the patterns without candidate generation. However they suffer from the maintenance of lot of links and paths if the database has scattered items. Hence it is needed to develop an algorithm to reduce the number of scans with minimum number of candidate sequences without much complexity process before mining the patterns especially for the scattered database. This paper introduces a web sequence mining algorithm that constructs prefix based minimized trees with maximal potential patterns. Hence the number of generate-and-test strategies done by existing approaches is reduced a lot. The rest of the paper is organized as follows: In Section 2, we give out the problem statement and Section 3 explains the web sequence mining algorithm with an example. Section 4 presents the experimental results and we summarize our work in Section 5.

2 Problem Statement

Let $E = \{e_1, e_2, \dots, e_n\}$ be a set of events. A web access sequence database, WASD, is a set of sequences in which each sequence denoted as a tuple $\langle tid, S \rangle$, contains a unique tid and a sequence S with set of ordered events. A sequence $a = \langle a_1, a_2, \dots, a_n \rangle$ is a subsequence of another sequence $b = \langle b_1, b_2, \dots, b_m \rangle$, $m \geq n$, if there exists an integer i , $1 \leq i \leq m-n+1$ and $a_1 \subseteq b_i, a_2 \subseteq b_{i+1}, \dots, a_n \subseteq b_{i+n-1}$. In web access sequence mining, gap between events is not allowed.

Let web access sequence database WASD is represented as a set $\{S_1, S_2, \dots, S_m\}$ where each S_i ($1 \leq i \leq m$) is a web access sequence. A sequence S is said to be a web access pattern of WAS, if $\text{sup}(S) \geq \xi$ (min-sup). A constraint in the problem is that a sequence is a set of ordered events without repetition. That is sequences with forward reference of pages are only allowed in this problem.

The problem of mining access pattern is: Given Web access sequence database WASD and a support threshold ξ , mine the complete set of ξ -pattern of WASD.

3 Web Access Sequence Mining-A Fast and Efficient Method

In this section, the proposed algorithm for web access sequence pattern mining algorithm is explained briefly. Our algorithm initially finds all frequent 1-itemset and frequent 2-itemset like previous approaches. Then, it applies divide and conquer method to discover patterns as follows. That is, it divides the search space of the database based on each frequent 2-itemset α and mines the patterns prefixed with α . In the divided search space prefixed with α , the algorithm finds α -based frequent 2-itemset again and a prefix pair table is created with these itemsets and their support count. Patterns in the α -based frequent 2-itemset are recursively merged to generate maximal potential sequence patterns prefixed with α . A constraint used at the time of merging patterns p_i and p_j is that items in p_j should have items in p_i as prefixes satisfying the min-sup and the length of the merged pattern is truncated if its length is greater than the maximum sequence length of the database. This is achieved by using the prefix pair table created above. Then, the generated maximal potential patterns are sorted based on their length and a prefix based minimized tree is constructed with those maximal potential sequence patterns. Finally, the database is scanned once to update the support count of the patterns in the tree and then patterns satisfying the min-sup are identified as frequent patterns. The above process is repeated for all other prefix based patterns generated as mentioned earlier.

Since the proposed algorithm generates prefix based maximal potential sequence patterns as candidate sequences, database scanning is required only for those sequences in order to search for frequent patterns and becomes faster. To make the algorithm as an efficient, we create a minimized WAS-tree with maximal potential sequence patterns to reduce the database scans further. The size of the tree is minimized because the tree is constructed with prefix based maximal potential patterns generated. Here, a header table associated with the tree is created with two fields namely node and link. The node field represents a web page and the corresponding link field represents the link information of that web page in the tree. In this algorithm, prefix based trees are recursively constructed and mined to find all the patterns in the database.

Example: Consider the web access sequence database as given in Table 1. The algorithm first generates frequent 1-itemset as {a,b,c,d,e,f} and frequent 2-itemset prefixed with 'a' as {ab,af} with min-sup=2.

Table 1. Web Access Sequence Database

Seq-id	Sequence
1	a b e f c d
2	a f d b e c
3	f a b e c d
4	g a f d b e
5	g a b e f c

From the divided search space prefixed with 'ab', the frequent 2-itemset discovered are {be,ef,fc,cd} in the database. Then, these patterns are recursively

merged to generate maximal potential patterns prefixed with 'ab', by satisfying the above mentioned constraint at the time of merging. In this case, only one pattern is generated as a maximal potential sequence pattern that is abefc and a prefix based tree is constructed with a single path having the pattern as abefc. To speed up the tree traversal, a header table is created with nodes and their adjacent links. Each node is labeled with an item and its corresponding support count. Once the tree is constructed, the database is scanned again to update the support count of abefc. Here, the database sequences are projected with prefix 'ab' and the pattern abefc is searched for. If the projected sequence has the suffix pattern as a subset, their support count is incremented. The above step is repeated for all the sequences in which the suffix pattern exists. Finally, the tree is traversed to mine all the patterns satisfying the given min-sup. To show that WAS-tree constructed by our algorithm is a minimized tree, we initially construct a WAS-tree prefixed with 'ab' without generating maximal potential sequence patterns and it is compared with our minimized tree. In this case, the tree constructed using database sequences prefixed with 'ab' has two paths and it is shown in Fig 1. Fig 2 shows a WAS-tree with maximal potential patterns prefixed with 'ab' generated by our algorithm. Because of maximal potential patterns, the tree size is here minimized by reducing a path.

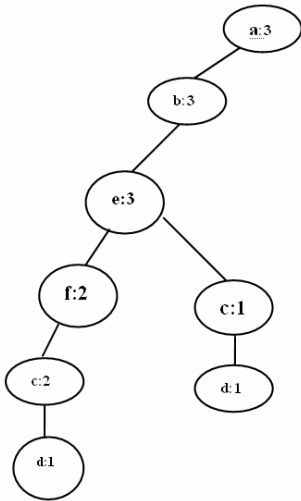


Fig. 1. Prefix Based WAS-Tree

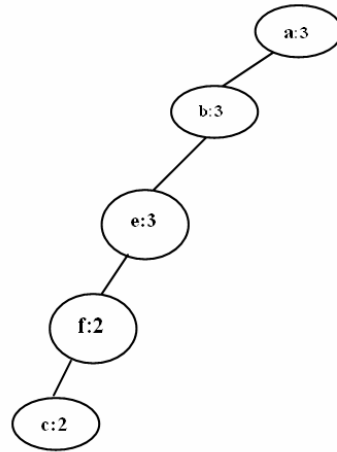


Fig. 2. Prefix Based Minimized WAS-Tree

Similarly, prefix based minimized trees are constructed for all the remaining frequent 2-itemset as a prefix. The maximum number of database scans required by our algorithm depends on the number of frequent 2-itemset generated in the database after generating frequent 1-itemset and 2-itemset. Thus, the proposed algorithm merits are stated as follows: 1) No level-by-level candidate generation 2) Minimized tree construction. To conclude, the algorithm becomes faster because of maximal potential sequence patterns and efficient because of prefix based minimized trees.

WAS Mining Algorithm

Input : Web Access Sequence Database WASD, min-sup

Output : Frequent Sequence Patterns

Step 1: Read the database WASD twice and find all the frequent 1-itemset and frequent 2-itemset

Step 2: Repeat for each frequent 2-itemset α

2.1. Divide the search space and find α -based frequent 2-itemset αFP_s

2.2. Generate maximal potential sequence patterns MPP_s with prefix α by merging patterns in αFP_s recursively

2.3 Arrange the maximal potential patterns MPP_s in order

2.4 Construct a prefix based WAS-tree with patterns in MPP_s

2.5 Scan the database once and update the count of the patterns in WAS-tree

2.6 Return patterns satisfying the min-sup as frequent patterns

Tree Construction Algorithm

Input : Maximal Potential Sequence Patterns MPP_s

Output : WAS-tree

Method:

1. Create a root node T for WAS-tree.
2. For (each sequence S_i in MPP_s)
 Call CreateTree(T, S_i)
3. Return WAS-tree.

Function CreateTree(tree root node T, sequence p-S)

1. Create node N with N.node-name=p
2. Create edge T-N with T-N.count = 1
3. Append edge T-N to $HT_{T,N}.Link$
4. If (S is non-empty) { Call CreateTree(N,S) }

Tree Updation Algorithm

Input : WAS-tree, Web Access Sequence Database WASD

Output : Frequent Access Patterns

Method:

1. For (each sequence S_i in WASD)
 Call UpdateTree(T, S_i)
2. Return Frequent Patterns

Function UpdateTree(tree root node T, sequence p-S)

1. If (T has a child N and N.node-name = p)
2. { T-N.count = T-N.count+ 1 }
3. If (S is non-empty) { Call UpdateTree(N,S) }

4 Experimental Results and Performance Evaluation

To evaluate the performance of the proposed web sequence mining algorithm, it is compared with an efficient algorithm PrefixSpan and both are implemented with C# language running under .NET. All experiments are performed on 2.16GHz Dell Computer with 1GB of RAM. To illustrate the performance comparisons we used freely available web access logs on internet. In general, PrefixSpan algorithm requires 'n' number of scans, where n is the length of the pattern. Our algorithm requires only four scans for a prefix based pattern of any length. That is, three scans for a maximal potential patterns generation plus one scan for mining all prefix based maximal potential patterns are needed as per the algorithm steps discussed earlier. The performance of our proposed work has been improved and performs better when compared to an existing approach because of generation of maximal potential patterns and minimizing the size of the tree.

The dataset used for performance evaluation is the CTI Dataset containing the preprocessed and filtered sessionized data for the main DePaul CTI Web server (<http://www.cs.depaul.edu>). Here, the file used for our experiment is the `cti.tra` file which contains the filtered sessionized data in transaction format. Each line in this file corresponds to the sequence of pages visited during one session. While the order of occurrence of pageview in each session represents the order in which these pageviews were visited, the transactions do not contain repeated visited to the same pageview in the same session. Thus, only the first access to a pageview is recorded as part of the transaction. To make the mining process easier, some preprocessing is performed on `cti.tra` file to translate pageviews into page ids using `cti`. Extensive experiments have been conducted to test the effectiveness of the algorithm by varying the min-sup and datasizes. Table 2 exhibits the performance of the proposed algorithm by varying the min-sup from 1% to 5% in CTI dataset. The performance of the algorithms is evaluated based on their time complexity and it is shown as a comparative study on the execution time of both algorithms for the min-sup=0.01 by varying the data size as given in Table 3. Fig 3 shows the performance analysis of these two algorithms on CTI dataset and a graph is drawn by using the execution time details of both the algorithms given in Table 3. That is, the analysis is performed by varying the data sizes with the minimum support count as 0.01. From the performance analysis report, it is found that the proposed algorithm performs better than PrefixSpan.

Table 2. Performance Analysis on CTI-dataset with varying min-sup (1% - 5%)

Min-Support in %	Proposed Algorithm
1	37
2	33
3	33
4	33
5	34

Table 3. A Comparative Study on Running Time in Seconds for CTI Dataset

No of Sessions	DataSize in KB	Proposed Algorithm	PrefixSpan
4000	81K	22	137
6000	122K	27	386
8000	156K	31	630
10000	202K	34	1038
12500	256K	37	1244

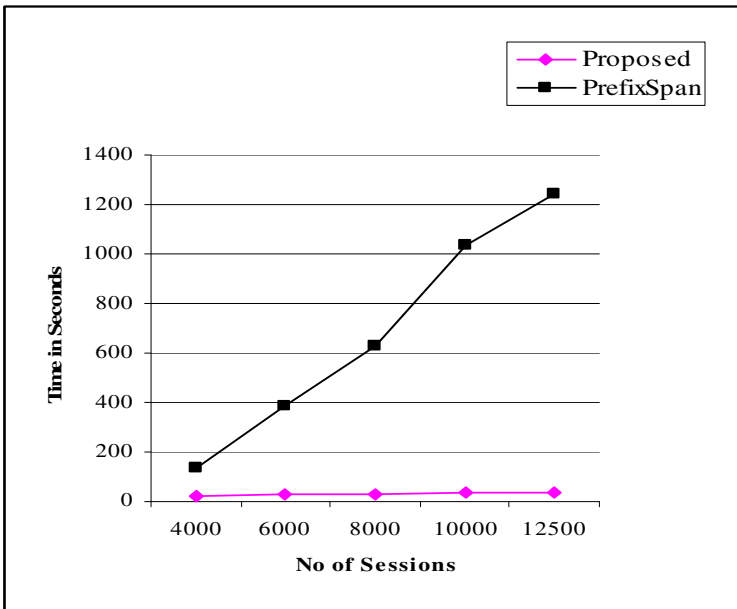


Fig. 3. Performance Analysis for min-sup=0.01

5 Conclusion

In this paper, a prefix based minimized tree algorithm has been proposed to mine web access sequence patterns with forward references in a web log database. The algorithm avoids candidate generation and mines frequent patterns using prefix based maximal potential sequence patterns generated by it. This makes the algorithm to be faster and efficient when compared to an existing approach. As a future work, the algorithm can be enhanced to handle web log sequences with backward references.

References

1. Chen, M.S., Park, J.S., Yu, P.S.: Efficient data mining for path traversal patterns. *IEEE Transactions on Knowledge and Data Engineering*, 209–21 (1998)
2. Mobasher, B., Jain, N., Han, E.H., Srivastava, J.: Web mining: Pattern discovery from World Wide Web transactions, Tech Rep: TR96-050, <http://www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.57.4087> (last cited on 1996)
3. Lee, Y.S., Yen, S.J.: Incremental and interactive mining of web traversal patterns. *Information Sciences* 178(2), 287–306 (2008)
4. Pei, J., Han, J., Mortazavi-Asl, B., Zhu, H.: Mining access patterns efficiently from web logs. In: Terano, T., Chen, A.L.P. (eds.) *PAKDD 2000*. LNCS, vol. 1805, pp. 396–407. Springer, Heidelberg (2000)
5. Ezeife, C.I., Lu, Y.: Mining web log sequential patterns with position coded pre-order linked wap-tree. *Data Mining and Knowledge Discovery* 10(1), 5–38 (2005)
6. Agrawal, R., Srikant, R.: Mining Sequential Patterns. In: *Proceedings ICDE 1995*, pp. 3–14 (1995)
7. Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., Hsu, M.C.: Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach. *IEEE TKDE* 16, 1424–1440 (2004)
8. Srikant, R., Agrawal, R.: Mining Sequential Patterns: Generalizations and Performance Improvements. In: *Int'l Conf Extd. DB. Tech.*, pp. 3–17 (1996)
9. Zaki, M.: SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Lrng.* 40, 31–60 (2001)
10. Cooley, R.: Web usage mining: discovery and application of interesting patterns from web data, Ph.D. thesis, University of Minnesota (2000)
11. Facca, F.M., Lanzi, P.L.: Mining interesting knowledge from weblogs: a survey. *Data & Knowledge Engineering* 53, 225–241 (2005)
12. Han, J., Kamber, M.: *Data Mining Concepts and Techniques*. Morgan Kaufmann, San Francisco (2001)

Scalable, High Throughput LDPC Decoder for WiMAX (802.16e) Applications

Muhammad Awais¹, Ashwani Singh², and Guido Maser¹

¹ VLSI Lab, Politecnico di Torino, Italy

² Navtel Systems, Houville la Branche, France

Abstract. This paper presents a layered LDPC decoder for WiMAX applications. A novel architecture based on parallel check node is proposed providing scalability in terms of supporting multiple code rates, block lengths and parallelisms to realize a high throughput LDPC decoder. The proposed design is fully compliant to support all 114 codes defined by WiMax standard. The IP core has been implemented using 130 nm standard cell ASIC technology. The proposed decoder achieves a throughput of 240 Mbps at 300 MHz and occupies a chip area of 2.76 mm².

Keywords: Low Density Parity Check codes, Min Sum, Layered Decoding, Tree way approach, Flexible architectures.

1 Introduction

WiMAX (Worldwide Interoperability for Microwave Access) is becoming a key technical standard for fixed, portable and mobile data networks. The standard features Low density Parity Check Codes (LDPC) as an optional channel decoding scheme for forward error correction, supporting up to 70Mbps [2]. Current research in the field of forward error correction (FEC) is aimed at finding the best possible error correcting codes which could allow high throughput decoding with efficient VLSI implementation meeting area, power and throughput metrics. LDPC codes, a class of linear block codes proposed by Gallager [1] have gained huge attention in wireless communication domain. Near Shannon Limit error correcting capabilities, low error floor, affordable complexity and intrinsic parallelism make LDPC an eligible candidate for a number of wireless standards [3].

Traditional parallel architectures provide maximum throughput but result in huge complexity [4], partially (block-level) parallel have sufficient throughput and affordable complexity but limited to specific structured codes only [5,6,8,7], while fully serial architectures have the minimum complexity and throughput. State of art for LDPC decoders mainly rely on multiprocessor implementations and proper interconnects [21,22] to achieve flexible, high throughput iterative decoding. In [8], a highly flexible LDPC decoder for WiMAX application able to process all specified codes has been proposed, but it uses serial implementation of check node, which limits the achievable throughput to a large extent.

Realizing high throughput decoders e.g. for wireless backbone networks (supporting data rates up to few hundred Mbps) either asks for massive parallelism or increasing the clock frequency which results in significant area and power overhead. However, incorporating parallelism at check node level is still not fully explored and can bring significant increase in throughput with affordable complexity. The key contribution of this work is a novel “Tree-Way” approach for parallel Min Sum check node. The proposed check node architecture is generalized up to check node degree $dc = 32$, which is sufficient to cover a large number of LDPC codes. A modular channel memory design technique at block-level is proposed and as a case of study, the proposed architecture is exploited to design a fully scalable WiMAX decoder.

2 WiMAX LDPC Codes and Decoding

In this section, we discuss about a class of architecture-aware LDPC codes which solves the interconnect, memory overhead, and scalability problems associated with LDPC decoders. The entire H matrix is composed of specific pattern of blocks with different cyclic shifts, which allows structured decoding and reduces decoder implementation complexity. The H_{BASE} matrix defined as [9] :

$$H_{BASE} = \begin{bmatrix} \Pi_{0,0} & \Pi_{0,1} & \dots & \Pi_{0,N} \\ \Pi_{1,0} & \Pi_{1,1} & \dots & \Pi_{1,N} \\ \vdots & \vdots & \vdots & \vdots \\ \Pi_{M,0} & \Pi_{M,1} & \dots & \Pi_{M,N} \end{bmatrix}$$

is associated to a parity check matrix H . It has M_b block rows and N_b block columns. The H_{BASE} is expanded, in order to generate H matrix, by replacing each of its entries $\Pi_{i,j}$ with a Z -by- Z permutation matrix, where Z is the expansion factor. The permutation matrix can be formed by cyclically shifting right the Z -by- Z identity matrix. The complete H matrix can best be described by a Tanner graph [11], a graphical representation of the associations between code bits and parity checks. Each row of this H matrix corresponds to a Check Node (CN) while each column corresponds to a Variable Node (VN) in Tanner graph. The set of shifts in the base matrix are used to determine the shift sizes for all other code lengths of the same code rate.

WiMAX features six classes of codes with code rates 1/2, 2/3, 3/4 (A & B) and 5/6 (A & B). Each class comes with 19 different codeword lengths with Z ranging from 24 to 96 with granularity of 4. Table 1 collects some details about the different WiMaX LDPC codes.

LDPC decoding is done using well known Message Passing (MP), Belief Propagation (BP) [10] or Sum Product (SP) [12] algorithms , in an iterative way which involves bilateral exchange of Log-Likelihood Ratio (LLR) messages between

VNs and CNs. The aim of BP algorithm is to compute the a-posteriori probability (APP) that a given bit in the transmitted codeword $c = [c_0, c_1, \dots, c_{N-1}]$ equals 1, given the received word $y = [y_0, y_1, \dots, y_{N-1}]$. The Min Sum (MS) Algorithm is an area efficient, sub optimal approximation to the SPA. It is also an iterative, soft decoding algorithm. Let α_{ij}^n represents the message sent from variable node VN_j to check node CN_i in n^{th} iteration, β_{ij}^n represents the message sent from check node CN_i to variable node VN_j in n^{th} iteration, $\mathcal{M}(j) = \{i : H_{ij} = 1\}$ i.e. set of parity checks in which variable node VN_j participates, $\mathcal{N}(i) = \{j : H_{ij} = 1\}$ i.e. set of variable nodes that participate in parity check i , $\mathcal{M}(j) \setminus i$: the set $\mathcal{M}(j)$ with check i excluded, $\mathcal{N}(i) \setminus j$: the set $\mathcal{N}(i)$ with variable j excluded. The conventional Min-Sum algorithm formulation is described as follows;

Algorithm 1. The Standard Min Sum Algorithm

1 : Initialization: For $j \in \{1, \dots, N\}$

$$\alpha_{i,j}^0 = \ln \frac{P(VN_j = 0|y_j)}{P(VN_j = 1|y_j)} = \frac{2y_j}{\sigma^2} \tag{1}$$

2 : Horizontal Scan : $\forall CN_i, i \in \{1, \dots, M\}$ do

$$\beta_{i,j}^n = \prod_{j' \in \mathcal{N}(i) \setminus j} \text{sign}(\alpha_{i,j'}^{n-1}) \cdot \min_{j' \in \mathcal{N}(i) \setminus j} \{|\alpha_{i,j'}^{n-1}|\} \tag{2}$$

3 : Vertical Scan $\forall VN_j, j \in \{1, \dots, N\}$ do

$$\alpha_{i,j}^n = \alpha_{i,j}^0 + \sum_{i' \in \mathcal{M}(j) \setminus i} \beta_{i',j}^n \tag{3}$$

4 : Decoding For each bit, compute its a-posteriori LLR

$$\alpha_j^n = \alpha_j^0 + \sum_{i' \in \mathcal{M}(j)} \beta_{i',j}^n \tag{4}$$

Estimated codeword is $\hat{C} = (\hat{c}_1, \hat{c}_2, \dots, \hat{c}_N)$ where element \hat{c}_j is calculated as

$$\hat{c}_j = \begin{cases} 0 & \text{if } \alpha_j^n > 0 \\ 1 & \text{elsewhere} \end{cases} \tag{5}$$

If $H(\hat{C})^T = 0$ then stop and output \hat{C} as decoded codeword.

Modifying the Variable node update rule (3) as

$$\alpha_{i,j}^n = \alpha_j^n - \beta_{i,j}^n \tag{6}$$

Table 1. H_{BASE} parameters for different WiMaX LDPC codes

Code Rate	1/2	2/3	3/4	5/6
H_{BASE} matrix	12×24	8×24	6×24	4×24
Type	irregular	irregular	irregular	irregular
Number of block rows [M]	12	8	6	4
Maximum row-column weight	7 – 6	11 – 6	15 – 6	20 – 4

We can merge horizontal and vertical scans in to a single horizontal scan where the check node messages $\beta_{i,j}^n$ are computed from $\alpha_j^{(n-1)}$ and $\beta_{i,j}^{(n-1)}$. This technique is called “Layered Decoding” [13], or “Turbo Decoding Message Passing (TDMP)” [9]. Layered decoding is based on the principle of using intermediate results directly in the next sub-iteration so that updated information is available to check node. This results in 50% decrease in number of iterations to meet a certain BER (equivalent to 2x increase in throughput) and significant memory savings as compared to standard two phase message passing.

3 Check Node Parallelism: A Tree-Way Approach

In Min Sum decoding, out of all LLRs of a CN, only two magnitudes are of interest i.e. the minimum and the second minimum. The state of the art for many applications consists of serial Min Sum check nodes incorporating on the fly calculation of running minimum, second minimum and sign product [14]. The interconnection network is also serial providing one input to all check nodes in each clock cycle. Incorporation of parallel processing inside the check node is a key contribution of this work whereby a novel check node based on “Tree-way” implementation, receives all variable to check node messages in parallel and writes back the updated extrinsic information simultaneously to all connected variable nodes. This results in significant decrease in check node latency and helps in achieving high throughput with smaller value of clock frequency and parallelism. Figure 1a shows an application of this scheme for check node degree $d_c = 8$. Each VN i.e. (I_1, I_2, \dots, I_8) is represented as a leaf node and tree is traversed performing min calculation at branch nodes until all VN extrinsic values are derived at the root nodes (e_1, e_2, \dots, e_8) . One of the key contribution of this paper is that for a parallel check node architecture we generalize the tree network connectivity and the data flow for any value of d_c up to 32 and present a fairly simple control mechanism for it. For the sake of architecture uniformity odd d_c values are considered as their even counterpart with extra VN intrinsic value initialized at $+\infty$ (i.e $d'_c = d_c$ if d_c is even; else $d'_c = d_c + 1$). Figure 1b shows the different stages of VN extrinsic calculation for proposed “Tree-Way” scheme (magnitude only).

Direct VN Comparison (DVC) Stage: As seen in Fig. 1b, for $d_c = 8$, the intrinsic values (I_1, I_2, \dots, I_8) are fed parallel to 4 compare select (CS) units.

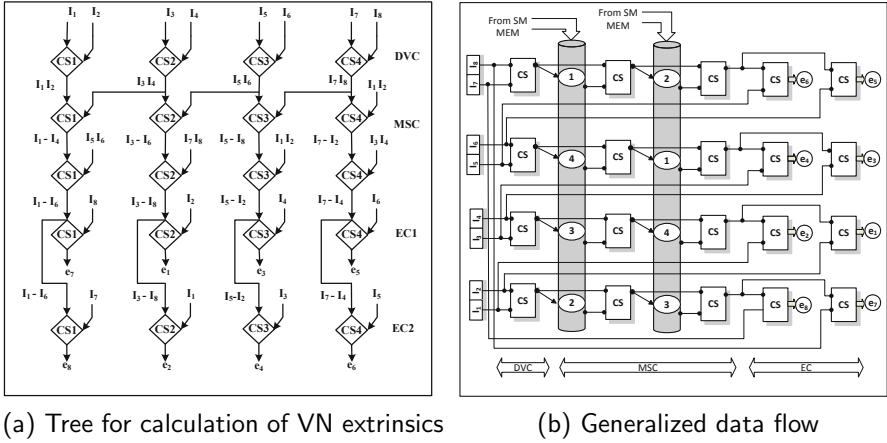


Fig. 1. Proposed “Tree-way” scheme for Parallel Check Node

For all values of d_c there is only one direct comparison stage. The outputs of DVC and each subsequent stage are passed on to next stage through a local shuffling network known as ACS network as well as are stored in switch matrix (SM) memory for use in later stages.

Multiple Shuffled Comparison (MSC) Stage: The shuffle network implements a circular shifting permutation. The rotational shift depends on d_c and the sub stage of the shuffled comparison stage. There are multiple shuffled stages depending on d_c and equal to N_{cc} , where N_{cc} is the required number of clock cycles for check node update. For different values of d_c Table 2 provides the information on N_{cc} values and shift associated with each shuffled stage.

Table 2. Clock Cycle requirement and Shift Permutation

d_c	N_{cc}	Permutation	d_c	N_{cc}	Permutation
5,6	4	1	19,20	7	1,2,4,8
7,8	5	1,2	21,22	7	1,2,4,8
9,10	5	1,2	23,24	8	1,2,4,8,10
11,12	6	1,2,4	25,26	7	1,2,4,8
13,14	6	1,2,4	27,28	8	1,2,4,8,12
15,16	7	1,2,4,6	29,30	8	1,2,4,8,12
17,18	6	1,2,4	31,32	9	1,2,4,8,12,14

As it can be seen from Fig. 1b, the input to the shuffle network is either the output from the immediate previous stage or output of a much earlier stage stored in the SM memories. Irrespective of the input source of the shuffle network, the shift associated with a stage is fixed.

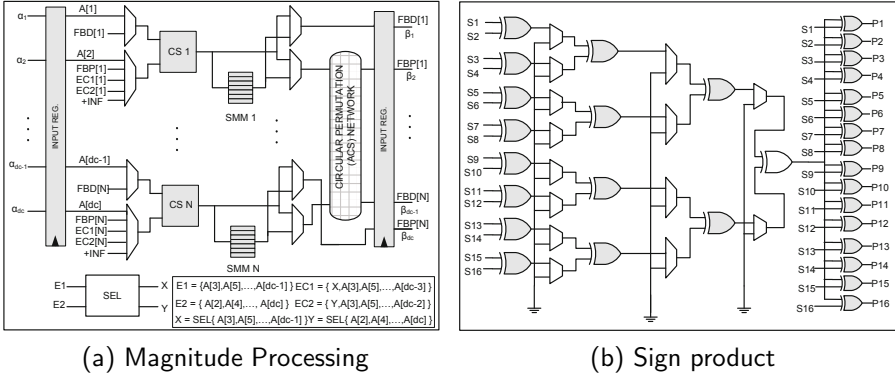


Fig. 2. Reconfigurable Tree Way Check Node : Generalized Data Path

Extrinsic Calculation (EC) Stage : The last two stages for any value of d_c are extrinsic calculation stage. As seen in Fig. 1b, input to these stages is the output from the last MSC stage and the shifted VN intrinsic values. This shift is circular over d_c and equal to 1 and 2 for EC stage 1 and 2 respectively.

4 Tree Way Processing Element Architecture

Figure 2a shows the generic data path for proposed “Tree-way ”processing element (magnitude). The α_n are the incoming LLR values, while β_n are the updated outgoing LLRs values where $(n = 1, 2, \dots, d_c)$ and d_c is the maximum check node degree of a given irregular H matrix. The Input Reg consists of d_c parallel registers to store α_n values which are input to the DVC and EC(1,2) stages. For compare select unit CS[i], $i = 1, 2, \dots, N$, the first input is either the intrinsic value $A[x]$, $x = 1, 3, \dots, d_c - 1$ or the direct feed back FBD[i] of previous stage. The second input is selected from five possible options i.e. the intrinsic value $A[y]$, $y = 2, 4, \dots, d_c$, the permutation feedback for previous stage FBP[i], the intrinsic values contained in EC1[i] and EC2[i] vectors and infinity. The output of each compare select unit is stored in its corresponding SM memory. There are $N = d_c/2$ CS units and SM memories and size of each memory is equal to number of stages (or Ncc) which depends on d_c . The inputs to ACS network come either directly from CS unit, or from SM memory. ACS network is implemented to allow for all possible permutations for a particular degree as given in Table 2. An $N \times N$ barrel shifter supports the implementation of all possible permutations for a degree d_c .

There are few implementations of parallel check nodes in literature but they are not optimal in terms of complexity, achievable frequency and scalability. In [15], a parallel check node architecture has been reported which has a complexity of $d_c(1 + 0.5(\log_2 d_c - 1))$ comparators. The authors in [16] reported a parallel check node based on divided group comparison technique. The reported check

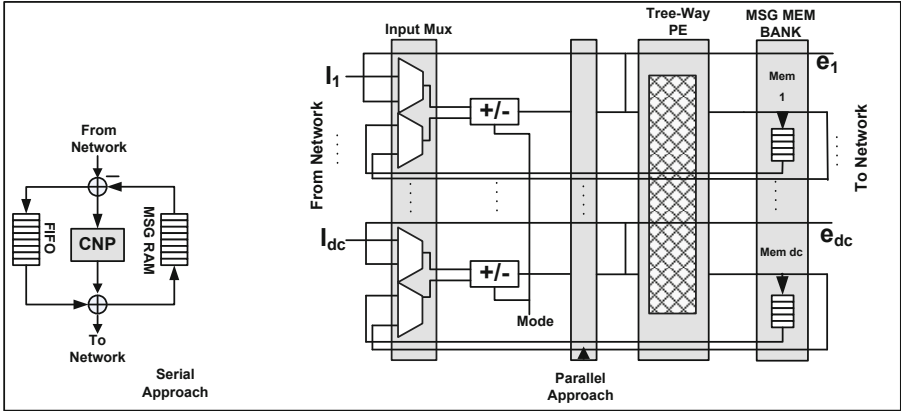


Fig. 3. Serial [7] v.s Proposed Layered Decoding Check Node

node has a huge complexity and critical path timing of three adders and three multiplexers for $d_c = 12$. The overall design could achieve a throughput of 86 Mbps at 125 MHz occupying 4.94 mm^2 area on chip using 130 nm technology.

Data path reuse is a distinguishing feature of our proposed “Tree-way” check node, which results in affordable complexity of $d_c/2$ comparators and critical path of one adder and four multiplexers for $d_c = 12$ thus achieving high clock frequency. The sign product tree is based on the property that product of all sign bits except the sign bit s_i is the exclusive or of all sign bits (including the sign bit s_i) with the sign bit s_i . Figure 2b shows the “Tree-way” implementation of sign product block. This architecture is also fully scalable to support multiple CN degrees as in case of irregular LDPC codes.

4.1 Proposed Layered Decoding / TDMP Architecture

Key idea behind layered decoding is that there is no processing inside variable node. It just acts as a memory unit to store the soft output estimate of previous sub-iteration. A layered data path architecture has been reported in [7]. The check node is serial with latency equal to d_c and consists of a FIFO, message ram and Check node Functional Unit (CFU). To ensure correct information input to CFU, the corresponding edge messages are immediately subtracted from the message RAM and result is passed to CFU. After processing, the same locations in the message RAMs are updated by newly calculated extrinsic messages. The output of CFU is added to corresponding input by passed by a FIFO. Thus the correct a-posteriori information is passed back to channel memory which always holds the updated VN information. This serial Check Node architecture is de-coupled such that the CFU is replaced by our proposed parallel tree way processing element. As shown in Fig. 3, single message RAM of $k \times d_c$ words used in serial architecture has been replaced by d_c number of message RAMs each consisting of k words, where k is the number of sub iterations in which

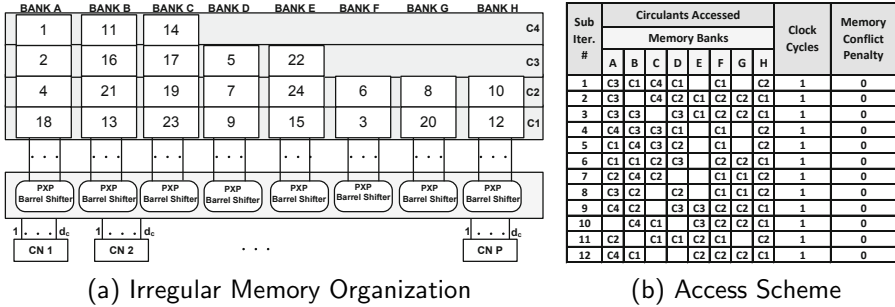


Fig. 4. WiMax 1/2 Rate LDPC : Channel Memory Organization

a variable node is accessed. The architecture shown is generalized and can be implemented for any value of d_c .

5 Block Level Memory Organization

As discussed before, the state of art for Min Sum decoding consists of serial interconnection network and check node with read and write latency of d_c clock cycles. To fully exploit the inherent advantage of proposed check node, a fully parallel network is desired which could move d_c messages to all P check nodes with minimum latency. To do so, an important task is how to organize memory to support parallel access up to d_c values by every check node with minimum number of clock cycles.

In this section a technique is presented to design memory organization supporting parallel access for proposed check node up to degree d_c . The objective is to break a single memory bank in to d_c number of parallel memory banks and distribute circulants (sub matrices) among them such that data access conflicts between circulants required in the same sub-iteration are minimum. This technique can be described as follows.

Given H_b matrix with dimensions $M_b \times N_b$, let W_r be the maximum row weight of H_b , if $W_r \bmod 2 = 0$ then $d_c = W_r$ else $d_c = W_r + 1$. The memory organization will consist of d_c memory banks with each bank containing N_b/d_c circulants. Proceed as follows

1. For $i = 1$ to N_b repeat
2. Select a column C_i and highlight all rows corresponding to non negative entries of column C_i
3. Check all other columns in highlighted rows. Columns which have negative entries in all highlighted rows of C_i are data independent from C_i and are not accessed in the same sub-iteration when C_i is accessed and can be placed in same memory bank with C_i .

In this way data dependencies between all circulants are noted down in an iterative manner. The N_b groups of variable nodes are stored in d_c memory

banks. For a regular memory organization, all memory banks have equal number of circulants. Each bank contains P single port memories each consisting of $(Z/P) \times (N_b/d_c)$ words. An irregular memory organization has unequal number of circulants for memory banks and is desirable since, it results in less data dependencies among the circulants. Applying this technique to 1/2 rate WiMax [2], we obtain a possible memory organization as shown in Fig 4. One implication of proposed technique is that the number of memory banks and complexity of interconnection network grows linearly with d_c . An alternate approach could be such that, for code rates with degrees greater than 8, the passing of VN to CN messages can be accumulated in multiple clock cycles using the same number of 8 memory banks. This avoids to a great extent, the possible interconnection network complexity and frequency reduction due to parallel nature of check node architecture. For example, to support WiMAX code rate 5/6 with $d_c = 20$, the proposed check node has read and write latency of 3 clock cycles with 8 bank memory scheme.

6 Synthesis Results

This section deals with ASIC implementation of proposed Layered LDPC decoder which features parallel "Tree-way" check node. The parameterized VHDL IP core of proposed decoder is synthesizable for all code rates, block lengths and parallelisms as proposed by WiMAX standard. Synopsys Design Vision Tool has been used to carry out synthesis of proposed IP core on 130 nm Standard Cell ASIC technology. The operating frequency has been set to 300 MHz. Table 3 shows the synthesis results of proposed decoder architecture and achievable throughput starting from the smallest code size with lowest code rate up to largest high-rate code. Total number of 15 iterations have been selected to meet a satisfactory error performance with 7 bit quantization. Parallelism values of 13-24 have been selected to realize a full mode architecture for moderate throughput applications. The proposed decoder achieves throughput well above 70 Mbps as specified by WiMax standard. However, the design is fully compliant to support higher parallelism for achieving gigabit throughputs. For example, parallelism of 96 results in throughput of 1.7 Gbps at 15 iterations with area consumption of 4.5 mm^2 at 300 MHz. The synthesis results show the maximum area of the decoder as in case of code rate 5/6 with $d_c = 20$, $Z = 96$ and $P = 24$. The total area is dominated by check node logic which is due to parallelism at check node level and huge flexibility to support multiple codes.

7 Comparison with Related Work

Table 4 compares the proposed work with some already published decoders. However, close comparison with the other similar implementations is difficult because of differences in design choices e.g. technology used, number of decoding modes (combination of code lengths and code rates), operating frequency, quantization, iteration count and parallelism. To simplify the comparison, area

Table 3. Throughput and Area Synthesis Results for Proposed Decoder IP

Throughput Results $P = 16 - 24$				Area [mm^2], 130nm @ 300MHz $P = 16 - 24$	
Code Rate	Max. Iter.	Code Size	Throughput Mbps	Quantization (bits)	7
				CNP	1.79
1/2	15	576-2304	74-137	Memory	0.94
2/3	15	576-2304	86-160	Network	0.03
3/4	15	576-2304	112-192	Controller	0.004
5/6	15	576-2304	130-240	Total	2.764

Table 4. Area and Throughput Comparison

Work	[19]	[20]	[7]	[18]	This Work
Tech.(nm)	130	65	65	90	130
Code	WiMax	WiMax	WiMax	WiMax	WiMax
Dec. Modes	19	114	114	114	114
Freq.(MHz)	83.3	400	400	400	300
Iterations	8	20	25	8-12	15
Quantization (bits)	8	7	6	7	7
Parallelism	4	27	24-96	3-4	16-24
Area (mm^2)	8.29	0.5	1.337	0.679	2.764
Scaled Area(mm^2)	8.29	2	5.345	1.30	2.764
Throughput (Mbps)	60-222	27.7-237.8	96-399	66.67-200	74-240
TAR (Mbps/ mm^2)	7.2-26.8	13.8-119	17.9-74.6	51.2-153.8	26.77-86.83

of each decoder has been scaled up to 130 nm process with a scaling factor of 2 and 4 respectively for 90 nm and 65 nm processes. On the contrary, the frequency has not been scaled as a universal conversion method between different technologies is not available. A parameter called throughput to area ratio (TAR) [17] defined as $TAR (Mbps/mm^2) = \text{Throughput} / \text{Area}$ has been included in the table to evaluate the efficiency of proposed decoder.

The authors in [19] presented a programmable multi-mode LDPC decoder for WiMAX. The decoder achieves a maximum throughput of 222 Mbps at the area cost of $8.29 mm^2$ using 130 nm process. In [7], the authors presented a full mode layered decoder for WiMAX which attains maximum throughput of 399 Mbps, occupying $1.337 mm^2$ area on 65 nm process. In [20], a flexible ASIP for multi-standard applications has been proposed. The decoder achieves a throughput of 237 Mbps and occupies an area of $0.5 mm^2$ on 65 nm process. The scaled area of ASIP yields a minimum and maximum TAR ($Mbps/mm^2$) of 13.8 and 119. In [18], the authors presented a full-mode WiMAX LDPC codec occupying area of $0.679 mm^2$ using 90 nm process. The decoder proposed in this work has a minimum TAR of $26.77 Mbps/mm^2$, greater than [19], [20] and [7]. The maximum TAR of proposed decoder is $86.83 Mbps/mm^2$ which is greater than [19] and [7] and comparable to ASIP decoder in [20] considering the fact that the ASIP decoder has been synthesized at 400MHz clock frequency using

a faster 65 nm technology. Finally, the performance of decoder presented in [18] is inferior to proposed decoder since it achieves the throughput of 200 Mbps at 400MHz with 12 iterations against 240 Mbps of our proposed decoder at 300MHz with 15 iterations.

8 Conclusion

LDPC codes have been proposed in a number of next-generation wireless standards. Implementing a fully scalable architecture while satisfying area, speed and power metrics is still a challenging task. In this paper we presented a fully scalable architecture for WiMax LDPC decoding based on a novel parallel “Tree-Way” Check Node realization. In addition, a comparison between the proposed and already published implementations is also presented. Even if state of the art low complexity techniques have been applied however area figures are still high when scaled to 130nm technology. The overall cost of proposed decoder is mainly determined by cost of huge flexibility and can be improved to a greater extent by supporting only a limited number of codes.

References

1. Gallager, R.G.: Low Density Parity Check Codes. *IEEE Trans. Inf. Theory* IT-8(1), 21–28 (1962)
2. IEEE Standard for Local and Metropolitan Area Networks-Part 16: Air Interface for Fixed broadband wireless access system, *IEEE Std. 802.16* (2004)
3. MacKay, D.J.C., Neil, R.M.: M Neil: Near Shannon Limit Performance of Low Density Parity Check codes. *Electronics Letters* 32, 1645–1646 (1996)
4. Blanksby, A.J., Howland, C.J.: A 690-mW 1-Gb/s 1024-b, rate-1/2 low-density parity-check code decode. *IEEE Journal of Solid-State Circuits* 37, 404–412 (2002)
5. Urard, P., Yeo, E., Paumier, L., Georgelin, P., Michel, T., Lebars, V., Lantreibecq, E., Gupta, B.: A 135Mb/s DVB-S2 compliant codec based on 64800b LDPC and BCH codes. In: *IEEE Solid-state Circuits Conference (ISSCC)*, San Francisco, USA, vol. 1, pp. 446–609 (2005)
6. Kienle, F., Brack, T., Wehn, N.: A synthesizable IP core for DVB-S2 LDPC code decoding. In: *Proceedings of IEEE Conference on Design Automation and Test in Europe (DATE)*, Munich, Germany, pp. 1530–1535 (2005)
7. Brack, T., Alles, M.: T Lehnigk-Emdem, F. Kienle, N. Wehn, L Insalata, F. Rossi, M. Rovini, L. Fanucci: Low Complexity LDPC Code decoders for Next Generation Standards. In: *Proceedings of the Conference on Design, automation and test in Europe*, Nice, France, pp. 1–6 (2007)
8. Brack, T., Alles, M., Kienle, F., Wehn, N.: A Synthesizable IP Core for WiMax 802.16e LDPC Code Decoding. In: *17th Annual IEEE Intl. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2006)*, Helsinki, Finland, pp. 1–5 (September 2006)
9. Mansour, M., Shanbhag, N.: A 640-Mb/s 2048-bit programmable LDPC decoder chip. *IEEE J. of Solid-State Circuits* 41(3), 684–698 (2006)
10. Gallager, R.: *Low Density Parity Check Codes*. MIT Press, Cambridge (1963)

11. Tanner, R.M.: A recursive approach to low complexity codes. *IEEE Trans. Info. Theory* IT-27, 533–547 (1981)
12. Kschischang, F.R., Frey, B.J., Loeliger, H.A.: Factor Graphs and the Sum-Product Algorithm. *IEEE Trans. Info. Theory* 47, 498–519 (2001)
13. Yeo, E., Pakzad, P., Nikolic, B., Anantharam, V.: High throughput low-density parity-check decoder architectures. In: *IEEE Proceedings of GLOBECOM*, vol. 5, pp. 3019–3024 (2001)
14. Bhatt, T., Sundaramurthy, V., Stolpman, V., McCain, D.: Pipelined block serial decoder architecture for structured LDPC codes. In: *Proceedings. IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2006, Toulouse, France*, vol. 4, pp. IV225–IV228 (2006)
15. Karkooti, M., Cavallaro, J.: Semi-parallel reconfigurable architectures for real-time LDPC decoding. In: *Proceedings of International Conference on Information Technology, Coding and Computing, Las Vegas, USA*, vol. 1, pp. 579–585 (2004)
16. Shih, X.-Y., Zhan, C.-Z., Wu, A.-Y.: A Real-Time Programmable LDPC Decoder Chip for Arbitrary QC-LDPC Parity Check Matrices. In: *IEEE Asian Solid State Circuit Conference, Taipei, Taiwan*, pp. 369–372 (November 2009)
17. Masera, F.Q.G., Vacca, F.: Implementation of a flexible LDPC decoder. *IEEE Trans. on circuit and systems II, Express Briefs* 24(6), 542–546 (2007)
18. Wang, Y.-L., Ueng, Y.-L., C.-L., Yang, C.-J.: Processing Task Arrangement for a Low -Complexity Full-Mode WiMax LDPC Codec. *IEEE Trans. on circuit and systems I* 58(2), 415–428 (2011)
19. Shih, X.-Y., Zhan, C.-Z., Lin, C.-H., Wu, A.-Y.: An 8.29 mm^2 640-Mbps 2048-bit programmable LDPC decoder design for Mobile WiMax system in 0.13 μm CMOS process. *IEEE J. of Solid-State Circuits* 33(3), 672–683 (2008)
20. Alles, M., Vogt, T., Wehn, N.: FlexiChap: A reconfigurable ASIP for convolutional, turbo and LDPC code decoding. In: *Proc. Turbo codes related topics*, pp. 84–89 (September 2008)
21. Quaglio, F., Vacca, F., Castellano, C., Tarable, A., Masera, G.: Interconnection framework for high-throughput, flexible LDPC decoders. In: *Proceedings of the conference on Design, automation and test in Europe, DATE 2006*, vol. 2, pp. 1–6 (March 2006)
22. Vacca, F., Masera, G., Moussa, H., Baghdadi, A., Jezequel, M.: Flexible Architectures for LDPC Decoders Based on Network on Chip Paradigm. In: *Proc. 12th Euromicro Conference on Digital System Design, Architectures, Methods and Tools, DSD 2009*, pp. 582–589 (August 2009)

Unique Mechanism of Selection of Traffic Flow Templates for Mobility IP Protocols Using Multihoming and IP Flow Mobility on the NGMN

Gustavo Jiménez and Yezid Donoso

System and Computation Engineering Department
Universidad de los Andes, Bogotá, Colombia
{ga.jimenez55,ydonoso}@uniandes.edu.co

Abstract. Seamless mobility, simultaneous multiple access over heterogeneous networks and other features are some of the 3GPP specifications along with other organizations have consolidated since Release 8 in the called Evolved Packet System (EPS), this represents the core for the integration of new and old standards and all existing networks (e.g. WiFi, UMTS, LTE, etc.) In this paper, we focus on the study of the performance of mobility IP protocols (MIP) on environment that allows multi access data network connectivity (MADNC), multihoming and providing services based on IP Flow Mobility (IFOM). We present through qualitative and quantitative analysis the advantages and the outstanding problems on an environment such as that evaluated. We implement a simple and effective mechanism for differentiation of services and application QoS policy on the EPS. Thus, the mechanism allows Mobile Nodes to refer to the Traffic Flow Templates (TFT) defined on the access network and use these as routing filters within Binding Update messages, using a new packet Traffic Flow Template Reference Mobility Option (TFTR).

Keywords: EPS; MIP; IFOM; multihoming; multi access; TFT; routing filters.

1 Introduction

EPS (Evolution Packet System) [1], was defined in the 3GPP Release 8 and provides a standardization of protocols and definitions for the NGMN (Next Generations Mobile Networks). This consists of two platforms: LTE (Long Term Evolution) designed for the evolution of radio interface layer as an access control technology of broadband access networks [16] and EPC (Evolved Packet Core) focused on the evolution of core network: providing support and seamless mobility between different access networks [2],[18]; service continuity between heterogeneous access systems; support for the selection of access system based on user preferences and conditions of access networks (e.g. UMTS, GSM, WiMAX, WiFi) or the policies defined by the operator,; control and QoS policy-based and Charging [19].

The above features are achieved because mobile devices can communicate using multiple interfaces (multihoming): However, the mobility IP protocols specifications

that have been submitted since 3GPP Release 8 are not support the fact that a subscriber can communicate using multiple access networks simultaneously [3] [4]. According to this, a device can set one or multiple simultaneous connections to different PDN (Packet Data Networks), however all traffic generated and exchanged between them and the mobile device is routed through the same network. When IP Flow Mobility (IFOM-3GPP Release 10) is implemented a device can communicate using both interfaces at same time without any service interruption [11]. As result, when a subscriber is under the coverage of a WiFi network, he could download or redirect some traffic (e.g. best effort, ftp, videos) in this access, while download some of the traffic (e.g. VoIP flow) in the 3GPP access.

The aim of this work is to study performance the mobility IP protocols (MIP) defined by 3GPP and the IETF (Dual Stack Mobile IPv6 (DSMIPv6) and Dual Stack Proxy Mobile IPv6 (DSPMIPv6)) on scenarios that allow analyze the Multihoming and IFOM, and provide a solution that improve the performance of MIP to problems such as packet loss and the high consumption of bandwidth [13]. The results will be analyzed using standard metrics, in order to obtain the performance indicators of our proposal.

The rest of this paper is organized as follow. In section 2, we describe some related works. In section 3, we present the qualitative and quantitative analysis of MIP's implementing IFOM, multihoming and multi access data networks connectivity. In section 4, we describe our proposal. In section 5, we present a performance analysis by comparing results obtained of our proposal with others models. Finally, we give some conclusions and describe related future work.

2 Related Works

At present, many works have been developed to analyze the mobility IP protocols (MIP's): Studies of latency, packet loss, behavior at high speeds, micromobility, among others. However, these studies do not assess the effect of implementing IFOM in the performance of MIP's. In [17] by example, the author analyzed the current problems of MIP proposed by the IETF in [5] and [6] when support IFOM. The authors propose number improvements. adding new types packages (Binding Identification and Flow Identification) and extensions to the Binding Cache to allow multiple Care of Address (CoA) registrations and routing of IP flows based on routing filters, however, although most of these ideas were later taken by the IETF and disseminated in [8], this work do not realized a quantitative analysis of the proposed solutions, so that does not provide information on how this solution affects the performance of mobility protocols.

In [20] the authors discussed a new mechanism to support multihoming on 802.11b (WiFi). The author makes some extensions on MIPv4 (Mobile IP v4) to support multiple CoA's registrations. They also make use of metrics to define the route selection strategies for the routing and balancing of the flow as the signal to noise ratio (SNR) and a metrics proposed, which bases its calculations for the selection of the Default Gateway in three factors: the deviations between the arrival times of some messages as the binding update, the bandwidth and the ability of FA (Foreign Agents) to support the traffic carried by the MN (Mobile Nodes).

In [21], [22], [23], [24], [25] and [26] the authors focus their works on the study of mobility IP protocols and mechanisms to support service continuity between heterogeneous networks, for which discusses different metrics such as transfer rates, packet loss (FTP, voice, etc.), throughput, handovers latency and compare their results with other existing mobility protocols such as HMIPv6 (Hierarchical MIPv6), FMIPv6 (Fast MIPv6, etc. However no study IP Flow Mobility and multihoming scenarios, in addition, the authors only show the performance of the algorithms for one stage of the four proposed in this paper.

3 Qualitative and Quantitative Analysis of Mobility Protocols

In the following chapters we will analyze the characteristics and performance of MIP (DSMIPv6 and DSPMIPv6), we developed a qualitative and quantitative analysis to identify the main different between the protocols such as features, advantages, disadvantages and the study of the effects of implementing IFOM and Multihoming on the performance of these, allowing us to obtain metrics (packet loss, latency) that are compared with scenarios without supporting of these characteristics to determine the improvement factor of the MIP in each aspects evaluated.

3.1 Qualitative Analysis

Table 1 is the result of feedback from other research and some IETF RFC and Drafts:

Table 1. Qualitative analysis of mobility IP protocols

	DSMIPv6	DSPMIPv6
Performance	Suboptimal routing problems.	Signaling reduction.
Security	Bidirectional tunnel between HA and each MN.	MN's share bidirectional tunnel in a MAG.
Deployment	Tunnels may be affected by NAT and Firewalls	Required optimal mechanisms for tracking UE.
Mobility Type	Host Based	Network Based
Route Advertisement	Broadcast –Subnet Prefix Shared	Unicast -Domain Prefix Shared
Mobile scope	Global Mobility	Local Mobility
Handover Management	Limited by the UE Characteristics	Managed by Network
Air Interface Overhead	Signaling Traffic High	Signaling Traffic Low
Required Infrastructure	HA (Home Agent) /PDN-Gw (Packet Data Network Gateway)	LMA (Local Mobility Anchor) and MAG (Mobile Access Gateway)
MN Modification	Yes, Client Software	No
Handover Latency	Bad, many messages between the UE and the entities	Good, the UE does not receive mobility signaling messages
Support IFOM	No, required extensions as FID, BID and routing filters	No, work in progress
Support Dual Stack	Yes, [7].	Yes, work in progress.
Multihoming	Support with modification to the UE.	Yes [14].

3.2 Quantitative Analysis

The proposed scenario is taken from [3], [4], consists of two access systems. The first a 3GPP (UMTS) and the second an untrusted non-3GPP network (WiFi) provided by a home WiFi hotspot, both connected to Internet and this connected with a Correspond Node (see Fig 1a). The UE (User Equipment) has two interfaces, each one with the ability to connect with one of the aforementioned access systems and services available on the environment proposed following these steps: the UE on the WiFi network, open a web browser (IPF1) and start to watch a video clip (IPF2) hosted on the internet. All traffic generated by these two flows is routed through the WiFi access, the UE calls simultaneously (VoIP) using a client application hosted in this, routing this traffic (IPF3) through 3GPP access. The selection of routes for the transmission of services takes the policies defined by the user (in the UE) and the operator (QoS and BW). Then UE start to download a FTP file (IPF4), that causes that the WiFi access is saturated such that the flow generated by the video clip (IPF2) is transferred to 3GPP access for the low QoS requirements of this. After when the file is downloaded, IPF2 is returned to the WiFi access. Finally, when the UE loses WiFi connectivity; all the flows are routed to the 3GPP access immediately.

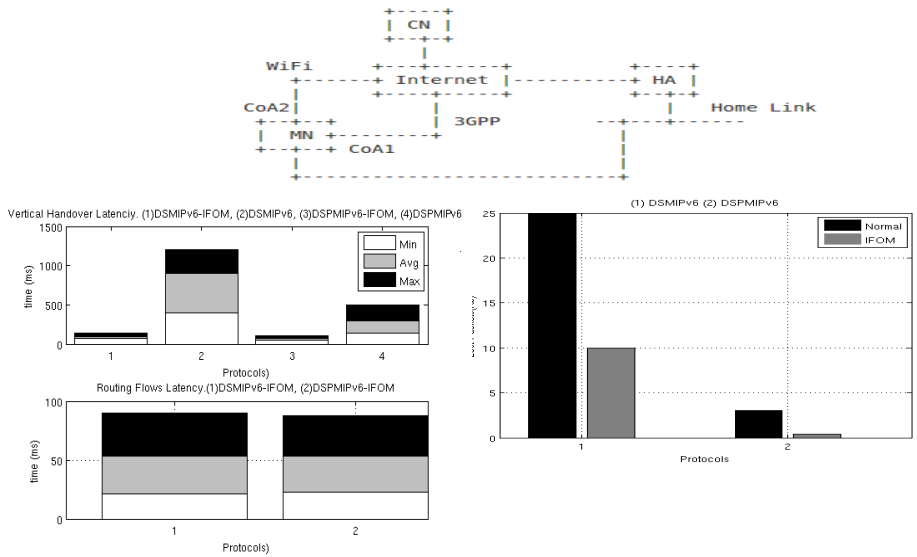


Fig. 1. a) IP Flow Mobility Scenario Proposed (up), b) Latency (down left), c) Lost packets (down right) analysis mobility IP protocols with IP flow mobility

Among the most important characteristics of the scenario we have: 50 meters of coverage of base station (NodeB), a bandwidth of 128kbps. By the other side, 10 meters from the access point coverage in wireless network, with a bandwidth 512kbps; simulation time 120s. At the 30 sec initiate the start to download the FTP

file ending at 70 sec, so that channel congestion will be given at this time; UE average speed is 1m/s and its movement starts at 80s; the VoIP call generates packages of 50kB/s, the video packet generates packages of 3KB that are transmitted every 5fps, the web pages rate are 20Kb/s and size of FTP file downloaded is 15MB; base Station and the hotspot are connected to the CN and the internet (wired) via an Access Router by duplex 1Gbps links with 10ms delays; and the points or results shown in the graphs shown below are the result of averaging 500 simulations in order to ensure a confidence level of 95% compared to the average value.

To compare the results, a scenario without support IP Flow Mobility was raised, on this only one interface is used in each stage, i.e. all the flows are transferred by an access. By the other side, for building of the scenarios and simulations of the UE behavior and the DSMIPv6 and DSPMIPv6 protocols, we used ns-2 (Network Simulator) [29] with extensions to support MIPv6, multihoming and PMIPv6 developed by [30], [31] and [32], furthermore some modifications were made to extend the code and allow some of the proposals of the IETF in [3], [4], [5], [6] and [8] such as: multiple CoA's registration, Multi access connectivity and extending to the Binding Cache and the BU messages to handle the use of BID, FID and to manage the IP Flows. The metrics evaluated in the performance tests were:

- Latency: is the time interval between the last packet received by the MN on old path and the first package received on new path.
- Packet Loss: represents the number of packets lost, during the handover and routing of flows.

Fig 1b shows the difference between the latency generated by protocols during the Vertical Handover and WiFi saturation. A notable difference is seen in the response times of 1.06s and 394ms for DSMIPv6 and DSPMIPv6 respectively, when we analyze the latency during the handover. This difference is generated by reducing the time required to update the Binding Cache (BC) using two interfaces, since many of these processes can run in parallel. Similarly, the differences presented between the times of both protocols are caused by the reduction of signaling traffic for each access, thus avoiding congestion of messages and therefore less competition for available resources. However, it should be noted that the use of two interfaces produces an increased amount of signaling messages in the EPC (Evolved Packet Core), because each one of these is treated independently, each request message generated configuration parameters (Subnet, Gateway, DNS, TTL etc.), and new messages in resolving conflicts of selection and routing traffic through other interfaces such as the selection of overlapping of domains or namespaces.

As mentioned in Table 1, the characteristics of DSPMIPv6 show that the performance of this is better than DSMIPv6. Above graph does show that the response times and packet loss of by DSPMIPv6 are lower due to reduced signaling (control messages) and not use of tunnels for packet transmission in access networks. This means that the available services are not affected by the use of the channel generated by the protocols, and that the packet loss can be reduced proportionally with the number of interfaces.

A very important parameter to measure is the packet loss; Fig 1c shows the result of this parameter. The packet loss is observed only on the services supported on UDP (TCP retransmits packets loss), so the figure shows the packet loss only for IPF2 and

IPF3 (one interface) and IPF3 (two interfaces) during the vertical handover and WiFi saturation. A first observation is that 82% of the lost packet was generated at the time of UE move and generate the Handover between heterogeneous networks using DSMIPv6, by the other side; DSPMIPv6 obtained a 94% loss for the same cause, this shows that the connectivity achieved in a system with IFOM is better since it enables immediate routing of flows in case of failure of access systems which are being transmitted. However, all these losses are generated by the Binding Update message (BU) and the update times of the FID in the Binding Cache of Home Agent (HA) and Correspondent Node (CN). This is a problem for seamless mobility, much of these losses are caused by lack of proper mechanisms to optimize the routing packets between the entities responsible for managing the mobility in the protocols. IETF has developed extensions in [27] and [28] or the use of micromobility, which are not discussed in this document.

4 Proposed Model

At this point, the focus of the paper has been assessing and improves the performance of IP mobility protocols when use multiple interfaces (multihoming) and implement IFOM (IP Flow Mobility). However, in a multi service scenario, it is important that the EPS provides an efficient solution to guarantee QoS and ensure that the user experience on each service is acceptable.

Currently, the specifications in the 3GPP Release 10 for QoS policies for each flow are not well defined between the UE and the PCC (Policy and Charging Control). Instead, we have is a set of policies for routing flows on each interface (routing rules and routing filters) that are sent with the FID (Flow Identifier Mobility Option) in the Traffic Selector Sub-Option in the BU (Binding Update) messages [4] (These specifications are described by a binary format presented in [12]), and set of polices for routing on each access system are implemented using TFT mechanisms or other mechanisms (non-3GPP accesses) to ensure proper treatment (QoS) on flows.

Table 2. Binding cache on home agent supporting TFTR

Home Address	Care of Address	Binding ID	TFTRID
HoA1	CoA1a	BID1	TFTRID1 TFTRID2
HoA1	CoA2a	BID2	TFTRID1

Given the above and to implement a simple and effective mechanism for service differentiation and application of QoS policies, we noticed that many times the selection mechanisms (TFT) are identical in both stages (access and interface selection). For that we propose the implementation of a mechanism for that the Mobiles Nodes (MN's) using the Traffic Flow Templates (TFT) defined on the access network and use this as filter for the routing messages within the Binding Update

(BU). We propose use a new packet called TFTR (Traffic Flow Template Reference Mobility Option) and a TFTRID (Traffic Flow Template Reference Identification) which replaces the current FID (Flow ID Mobility Option) as shown in Table 2. This modification offers advantages such as: reduced number of BU’s messages because the TFTRID only be changed when removal, upgrade or addition of flows and not when QoS features (e.g. data transfer rate, bandwidth) are modified, size reductions of the BU messages. To create the structure was taken as an example the scheme of Traffic Sub-Option Selector in [12].

The mechanisms used to update, remove and add the TFTRID are based on the procedures proposed in [8] and showed in Fig 2, for the establishments of the TFTRID: (1) The UE requests to ANDSF a list of available access from your location. (2) ANDSF returns a list of available access. (3) The UE initiate a service, e.g. IMS voice call and is received by the P-CSCF. (4) The P-CSCF is responsible for extracting of the information provided by the service related with QoS policies. (5) The P-CSCF communicates with the PCRF and with the information extracted, service policy defined by the operator, the subscription information requested (SPR) and other data’s, create a rule PCC and a QoS rule. (6) QoS rule is sent to BBERF by the PCRF for further activation and Bearer Binding. (7) The PCRF sends PCC rule to locate in the PDN-GW. This is responsible for the activation of the rule and the creation or modification of the EPS Bearers, PDP contexts or other with the policies defined in PCC rule, and ensuring that the traffic carried by the service will be provided the requested QoS characteristics. (8) The information is transmitted to the UE in an EPS Bearer. This indicates the TFTRID provided (Bearer ID) to the access (9) The UE sends BU messages to add this information to the Binding Cache in the HA or LMA (PDN- GW). And (10) initiates the traffic flow while the HA is responsible for monitoring that the traffic is transmitted by the access networks.

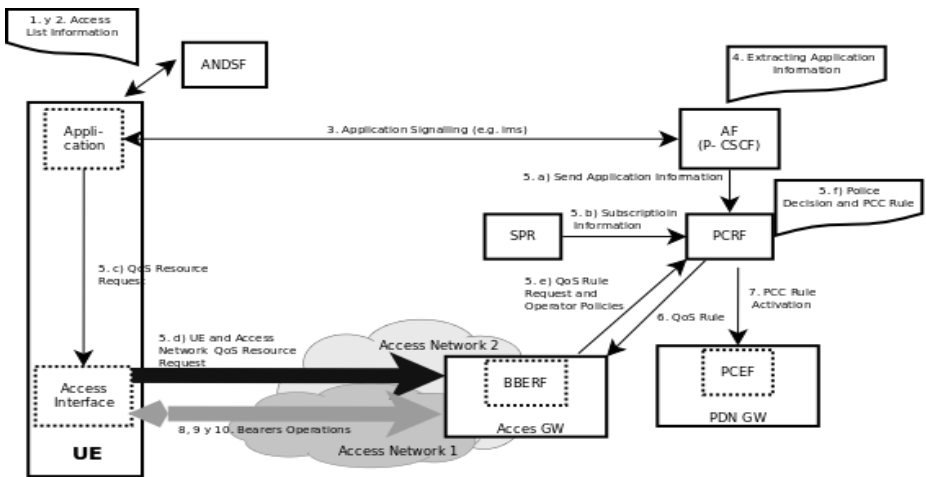


Fig. 2. Implementation TFTRID Procedure In EPS

5 Experimental Results

We have a particular interest to quantify the improvements of implement the TFT. For this based on the scenario and metrics proposed in section 3 and calculating the total cost generated by the signaling traffic, the packet loss rate and bandwidth consumed.

Some results founded show us that the latency problems presented in section 3 shows a reduction of approximately 78% for both protocols (the times was in the order of hundred milliseconds). These result is achieved due to the decrease in the sizes of the Binding Updates messages, since the Traffic Sub-Option Selector defined by the IETF for the FID in IPv4 has a size of 33 bytes while that the proposed (TFTRID) has a size of only 4 bytes.

Unlike the earlier scenarios proposed in Section 3 which analyzed the behavior of both mobility IP protocols (DSMIPv6 and DSPMIPv6) in face of the IP Flow Mobility and multihoming. In this section we evaluate the performance of the protocols versus the degradation of performance metrics proposed: we simulate with ns-2 (Network a MN (Mobile Node) operating under interference generated by a variable number of MN (maximum 100) that follow a random mobility model called RWP (Random Waypoint Mobility), which will examine both the scalability of the proposed environment with a more realistic group of users within a network. To compare the results of the use of TFTRID versus previous models defined by the 3GPP and the IETF in [3] , [4] ,[5] and [6] seen in the section 3 we will be calculated normalized values using the following function:

$$MPModel_{TFTRID} \text{ vs } MPModel_{Original} = (MPModel_{TFTRID} - MPModel_{Original}) / MPModel_{Original} \tag{1}$$

Where the $MPModel_{TFTRID}$ represent the metric of our proposed model and $MPModel_{Original}$ correspond to the same metric evaluated on the models presented previously. These calculations will determine clearly the increase or decrease between each of the performance metrics of the models.

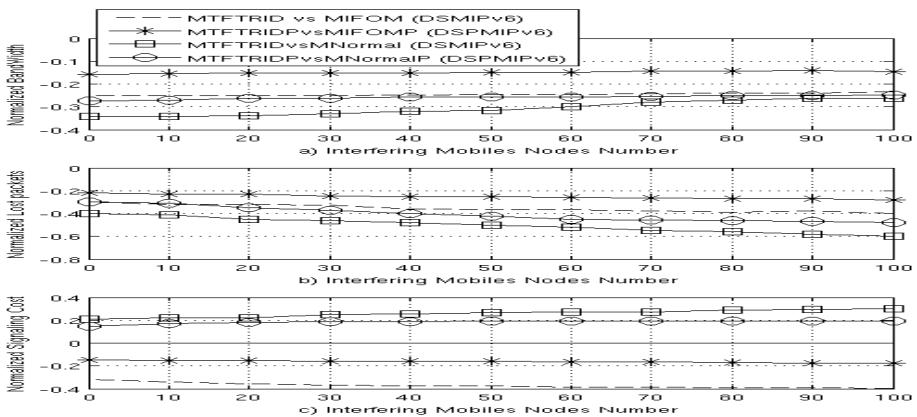


Fig. 3. a) Bandwidth b) Lost packets c) Signaling cost analysis proposed model

Fig 3a shows the behavior of the bandwidth through wireless access to the increasing number of Mobile Nodes interfering. we can see that the proposed model (MTFTRID) implemented on both protocols reduce by 25% (DSMIPv6) and 15% (DSPMIPv6) consumption of the bandwidth compared to the protocols implemented only IP Flow Mobility and 34% (DSMIPv6) and 27% (DSPMIPv6) compared with the original model (one interface). This reduction is achieved largely by reducing the signaling traffic generated by the messages of BU (Binding Update) and BA (Binding Acknowledgment). Fig 3b shows such as packet loss rate is reduced by 30% (DSMIPv6) and 20% (DSPMIPv6) regarding IFOM. This reduction also can be expected given the reduction in latency during the handover evidenced. Compared to the standard model (an interface), we could observe that the reduction is 40% to 60% using DSMIPv6 as a protocol for IP mobility and 30% to 42% if the chosen DSPMIPv6. The above results allow us to ensure that the protocol behavior DSPMIPv6 in high mobility scenarios intra-domain provides better performance than the DSMIPv6. By the other side, Fig 3c shows as on one hand the signaling traffic is reduced by 40% (DSMIPv6) and 15% (DSPMIPv6) with respect to the model implements only IFOM, which is an expected result because the reducing the size of packages and reduced BU constant updates on the Binding Cache. However, traffic is still higher by 30% (DSMIPv6) and 15% (DSPMIPv6) when compared with the original model that does not support multiple CoA's registration policies.

6 Conclusion and Future Works

We have presented qualitative and quantitative results that show that the implementation of IFOM and multihoming improves the performance of both mobility IP protocols. We identified that the mechanisms for route selection and implementation of QoS policies are evaluated twice: First for the mobile nodes and after in the access systems. Our propose implemented the TFT Reference Mobility Option (TFTR) as replacement of the Flow ID Mobility Option for use as a mechanism for packet filtering on the EPS; this allowed the reduction of signaling messages (Binding Update) and their size. The result of using the TFTR shows that evaluated performance metrics such as latency are significantly reduced; improving the user experience on the access networks and provides seamless connectivity to an undetectable level for them. We reduced the cost of signaling for reduction of the amount of Binding Update messages generated and their sizes. However, when compared with the original model is still higher the traffic flow.

For futures studies we consider the simulation of scenarios to analyze the impact of the Dual Stack in the performance of MIP and IFOM. Identify and specify the necessary changes in the EPS to support allocation of QoS policies per-flow in the PCC. Finally, implementation of a prototype to study the proposal made by using Mobile IPv6 for Linux, as outlined in [33] or a tool as OpenEPC [18].

References

1. 3GPP TS 22.278 V.10.1: Service requirements for the Evolved Packet System (EPS). Rel 10 (March 2010)
2. 3GPP TS 23.327 V.9.0.: Mobility between 3GPP-Wireless Local AreaNetwork (WLAN) interworking and 3GPP systems. Rel 9 (December 2009)

3. 3GPP TR 23.861 V1.3.0.: Multi access PDN connectivity and IP flow mobility. Rel 9 (September 2009)
4. 3GPP TR 23.261 V1.0.0.: IP Flow Mobility and seamless WLAN offload Stage 2. Rel 10 (March 2010)
5. IETF Network WG: Mobility Support in IPv6. RFC 3775 (June 2004)
6. IETF Network WG: Proxy Mobile IPv6. RFC 5213 (August 2008)
7. IETF Network WG: IPv4 Support for Proxy Mobile IPv6. RFC 5844 (May 2010)
8. IETF Network WG: Multiple Care-of-Addresses Registration. RFC 5648 (October 2009)
9. IETF Internet Engineering Task Force: Current Practices for Multiple Interface Hosts. Internet Draft (work in progress), draft-ietf-mif-current-practices-03, August 2010 (Expires: February 2011)
10. IETF Network WG: Multiple Interfaces and Provisioning Domains Problem Statement. Internet Draft (work in progress), draft-ietf-mif-problem-statement-09.txt, October 2010 (Expires: April 2011)
11. IETF MEXT WG: Flow Bindings in Mobile IPv6 and Nemo Basic Support. Internet Draft (work in progress), draft-ietf-mext-flowbinding-10.txt, September 2010 (Expires: March 2010)
12. IETF Network WG: Traffic Selectors for Flow Bindings. Internet Draft (work in progress), draft-ietf-mext-binary-ts-05.txt, October 2010 (Expires: April 2011)
13. IETF Network WG: PMIPv6 Multihoming Support for Flow Mobility. Internet Draft, draft-ietf-mext-flowbinding-10.txt, February 2010 (Expires: August 2010)
14. IETF Network WG: Multihoming Extensions for Proxy Mobile IPv6. Internet Draft, draft-bernardos-mif-pmip-02, March 2010 (Expires: September 2010)
15. Fokus, F.: OpenEPC Product Info- Understanding NGMN and Related Technologies-LTE, EPC and IMS. Tutorial 4, Page(s):171 (2009)
16. Yom, P.: LTE Update. IEEE Communication Magazine, 78 (February 2010)
17. Ahmed, T., Antoine, S.: Multi Access Data Network Connectivity and IP Flow Mobility in Evolved Packet System (EPS). In: Wireless Communications and Networking Conference (WCNC 2010), pp. 1–6. IEEE, Los Alamitos (2010)
18. Fokus, F.: Web Page Framework y Software OpenEPC. Disponible en, <http://www.openepc.net/en/openepc/index.html>
19. Olsson, M., Suitana, S.: SAE and the Evolved Packet Core: Driving the Mobile Broadband Revolution, 1st edn., p. 444. Elsevier Ltd, Amsterdam (2009)
20. Åhlund, C., Brännström, R.: Multihoming approach to Mobile IP, p. 4. CiteSeerx (2009)
21. Esaki, H.: Multi-Homing and Multi-Path Architecture Using Mobile IP and NEMO Framework, p. 6 (2004)
22. Jaeho, J., Jinsung, C.: A Cross-layer Vertical Handover between Mobile WiMAX and 3G Networks. In: IWCMC 2008, pp. 644–649 (2008)
23. Marques, H., Ribeiro, J.: Simulation of 802.21 Handovers Using ns-2. Hindawi Journal of Computer System, Network and Communications, 11 (2010)
24. Diab, A., Mitschele-Thiel, A.: Comparative Analysis of Proxy MIPv6 and Fast MIPv6. In: Mobiwac 2009, p. 9.0 (2009)
25. Seung-II, H., Youn-Hee, H.: Empirical Performance Evaluation of IETF Mobile IPv6 and Proxy Mobile IPv6. In: Mibility 2008, p. 7 (2008)
26. Lei, J., Xiaoming, F.: Evaluating the Benefits of Introducing PMIPv6 for Localized Mobility Management. In: IWCMC 2008, pp. 74–80 (2008)
27. Galli, S., McAuley, A., Morera, R.: An analytical approach to the performance evaluation of mobility protocols: the overall mobility cost case. In: PIMRC 2004, vol. 4, pp. 3019–3024 (2004)

28. Galli, S., Morera, R., McAuley, A.: An analytical approach to the performance evaluation of mobility protocols: the handoff delay case. In: VTC 2004, vol. 4, pp. 2389–2393 (2004)
29. The network simulator-ns-2: Disponible en (June 2009)
<http://www.isi.edu/nsnam/ns>
30. Motorola Lab Paris: Mobiwan: ns-2 extensions to study mobility in Wide-Area IPv6 Networks. Disponible en, <http://www.inrialpes.fr/planete/mobiwan>
31. Wang, Q.: NS2-MIUN- A Multi-homing Extension of Wireless Node Implementation in NS-2. Disponible en (2009),
<http://www.miun.se/personal/qinghua.wang/resources.htm>
32. HyonYoung, C.: Proxy Mobile IPv6 for Ns-2. Disponible en, <http://commani.net>
33. Diab, A., Mitschele-Thiel, A.: Comparative Analysis of Proxy MIPv6 and Fast MIPv6. In: Mobiwac 2009, p. 9 (2009)

Elliptic Curve Cryptography for Smart Phone OS

Sharmishta Desai¹, R.K. Bedi², B.N. Jagdale³,
and V.M. Wadhai⁴

¹ PG Student, IT Department, MITCOE, Pune, India

² Assistant Professor, Computer Department, MITCOE, Pune, India

³ Assistant Professor, IT Department, MITCOE, Pune, India

⁴ Principal, MITCOE, Pune, India

Abstract. Mobile Technology is growing rapidly. Usages of smart phones are increased for critical financial applications. This leads to many security issues as well. Implementing security features into such critical financial applications can minimize the transaction risk. Traditionally RSA, DH public key cryptography algorithms has been used. However ECC has proven results for smaller key size requirement which is more useful for resource constrained devices that take less memory, less bandwidth and less power consumption. In our paper, we have proved ECC's strength with respect to RSA. This paper contributes on implementation of ECC over GF (2^m) for smart phone OS which is used in mobile devices. Our experiment shows that ECC takes less computation time efforts than RSA when key size becomes greater than 512 bits which is advantageous on mobile or smart phones. In our implementation memory consumption is reduced as we are computing elliptic curve points dynamically when we need it and cipher text size is also reduced. We are avoiding cryptanalytic attack by eliminating same cipher text pattern generation. An experiment study is conducted on android OS which is one of the popular smart phone OS to show the effectiveness of proposed algorithm and also addressed cryptanalytic attack.

Keywords: ECC, RSA, DSA, ECDH, Cryptography.

1 Introduction

Google's Head of Android Security Said. "As smart phones are becoming popular, they are going to get some unwanted attention from criminals. Smart phones OS will become a major security target." [15]. Mobile devices are the storage of lot of sensitive data. A different approach is needed for ensuring security of data. In wireless communication for passing data over the air, there is a need of strong security with less key size [7].

The European Network & Information Security Agency (ENISA) has warned about the growing security risks associated with smart phones such as iPhone, BlackBerry, Android and Windows Mobile [16]. As these devices are limited in terms of memory, power or bandwidth, computation intensive algorithm execution like RSA will create heavy load on memory and processor [8]. So, the widely used RSA can be replaced

with ECC as it gives more security with less key size. Less key size will consume less memory and less power.

Elliptic Curve Cryptography is a public key cryptography given by two scientists Miller [2] and Kobitz [1] in 1990. Security of ECC lies in the discrete logarithm problem. It is a full exponential algorithm which is powerful against brute force attack [4]. There are many attacks possible on cipher text like timing attack, side channel attack and cryptanalytic attack. Table 1 gives comparison between ECC and RSA [17].

Table 1. Key Sizes in Bits with Equivalent Security Levels

Times to break in MIPS years	ECC	DH/DSA/RSA	RSA/ECC Key Size Ratio
10 ⁴	106	512	5:1
10 ⁸	132	768	6:1
10 ¹¹	160	1024	7:1
10 ²⁰	210	2048	10:1
10 ⁷⁸	600	21000	35:1

The rest of the paper is organized as below: Section 2 explains elliptic curve theory. Section 3 gives literature survey. Section 4 gives detailed explanation of our research work. Section 5 gives our experimental results. Section 6 proposes future work. Section 7 concludes our paper.

2 Elliptic Curve Cryptography Theory

Elliptic Curve Cryptography is used to generate keys using elliptic curve [18]. It is also used to generate digital signature and for converting plain text into cipher text [5].

Elliptic Curve over field k is given by equation

$$Y^2 = X^3 + aX + b \tag{1}$$

$a, b \in k$

By changing values of a and b , we get different curves. These a and b should follow following constraint, $4a^3 + 27b^2 \neq 0$.

Elliptic Curves are defined over two field's prime field F_p and binary field $F(2^m)$ [18]. The equation over prime field p is given by

$$Y^2 \text{ mod } p = X^3 + aX + b \text{ mod } p \tag{2}$$

The elements of binary field are the numbers between 0 to $p-1$. The equation over binary field is given by equation

$$Y^2+XY =X^3+aX+b \quad (3)$$

The elements of binary field $F(2^m)$ should be considered as numbers having maximum m bits. In our paper we have implemented ECC over binary curves. Two main operations involved in ECC are point doubling and point tripling that we will see in Research Methodology Section in detail.

3 Related Work

B.Muthukumar, Dr. S.Jeevanantharr explained Design of an Efficient Elliptic Curve Cryptography Coprocessor [8]. It has explained point doubling, point addition and point multiplication operation. Multithreading Elliptic Curve Cryptosystem implemented by Uma S.Kanniah and Azman Samsudin from Universiti Sains Malaysia. They have used two parallel mathematical algorithms, Karatsuba and Montgomery, for elliptic curve point multiplication[9]. But these two algorithms are complex to implement and slower for large number multiplication.

Concurrent Algorithm For High-speed Point Multiplication In Elliptic Curve Cryptography implemented by Jun-Hong Chen, Ming-Der Shieh and Chien-Ming Wu, Taiwan employed the nonadjacent form of a binary sequence to reduce the number of 1's in an operand so as to decrease the total number of addition in ECC encryption/decryption[10]. But It need an extra memory space to store an intermediate point, but it can achieve 100% hardware utilization.

Hai Yan and Zhijie Jerry Shi has given software implementation of ECC over 8-bit processor[11]. They have explained implementation on different word size processors.

Tohari Ahmad1, Jiankun Hu2, Song Han has explained An Efficient Mobile Voting System Security Scheme based on Elliptic Curve Cryptography [12]. In this paper, ECDH algorithm is combined with AES and compared with ECC. A method is proposed for mobile voting application. Implementation of Diffie-Hellman Key Exchange on Wireless Sensor Using Elliptic Curve Cryptography is explained by Samant Khajuria, Henrik Tange[14]. In this paper Diffie-Hellman key exchange is explained using Kobitz curve and TNAF (τ - adic non-adjacent form) with partial reduction modulo.

4 Research Methodology

In Elliptic Curve Cryptography, to convert plain text into cipher text, following steps are executed.

1. Convert Text into ASCII Format.
2. Generate points on Elliptic curve
3. Generate keys of Users
4. Encrypt Text

The detailed Elliptic Curve Cryptography Algorithm is given below.

In this paper we modified the work proposed by [4] to ECC implementation over $GF(2^m)$ for smart phones OS.

Generate Points of a Curve

```
Algorithm gen_points (a,b,p){
x=0
while (x<p){
Put values of a, b and x in equation  $y^2+xy=x^3+ax+b$ 
Find roots of the equation  $y^2+xy=x^3+ax+b$ 
//All values of(x,y) gives different points on elliptic
curve.}}
```

Generate Keys of a User

Suppose there are two users A and B. Following algorithm is used for generating keys.

```
Algorithm Generate_keys(){
Step 1:User A will select any random number  $K_A$  as a
private key.
Step 2: Select generator point G from the curve points
such that Point G is having small x and y coordinates.
Step 3: To generate public key  $k_{Ap}$  multiply  $K_A$  with G
using point_mult() algo.
Follow steps 1 to 3 to generate keys( $k_B$ ,  $k_{Bp}$ ) for user
B.}
```

Point Multiplication in ECC

To multiply any number K with point $p(x,y)$ we repetitively apply point doubling and addition operations.

```
Algorithm Point_mult(){
For doubling a point(2p) use following formulae
 $S = [(3x^2 + a)/2yp] \bmod p$ 
Then 2p has coordinates (XR, YR) given by:
```

$$X_R = (S^2 - 2x) \bmod p$$

$$Y_R = [S(x - X_R) - y] \bmod p$$

To determine $3P$, we use $P + 2P$, treating $2P=Q$. Here P has coordinates (x,y) , $Q=2P$ has coordinates (XQ, YQ) .

$$s = [(yQ - y) / (XQ - x)] \pmod p$$

$$P + Q = -R$$

$$XR = (s^2 - x - XQ) \pmod p$$

$$YR = (S(x - XR) - y) \pmod p$$

Encrypting Text

```

Algorithm Encrypt_Text(){
Convert text into its ASCII format
Select any point pm from generated points of a elliptic curve
Multiply ascii value with pm to get another point pm1 using Point_mult algo
Cipher text will be {kG, pm1+k*kAp}
    
```

Decrypting Text

```

Algorithm Decrypt_text (){
Take Cipher text will be {kG, pm1+k*kAp}
Calculate pm=pm1+k*kAp-kbkG}
    
```

The above implementation of ECC is shown below using use case diagrams in Fig 1 and Fig 2. First elliptic curve is selected by taking different values of a , b and m . Then points and keys are generated. Lastly using keys and points text is converted into cipher text using point multiplication method.

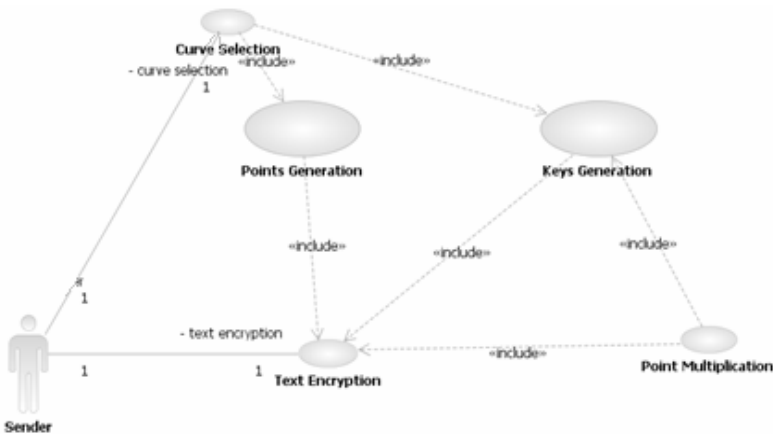


Fig. 1. Use case diagram For Sender

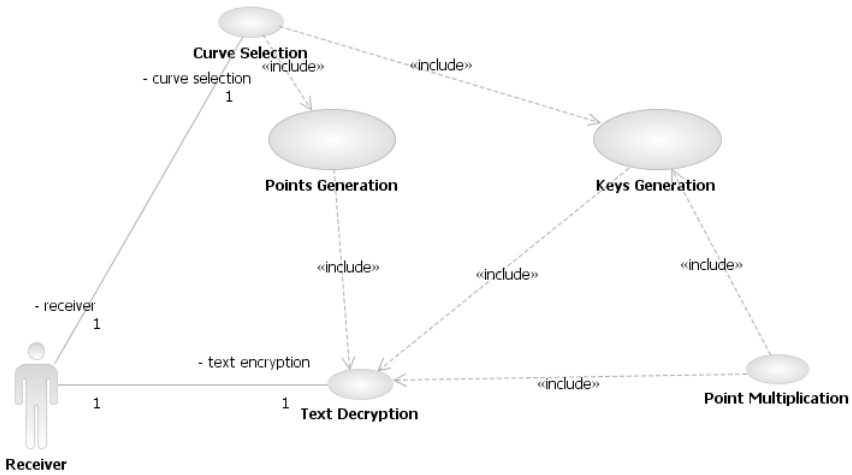


Fig. 2. Use case diagram for Receiver

In above implementation of ECC, We have experimented some new things.

After point generation of elliptic curve, instead of storing all points of the curve, in our implementation we have computed coordinates of the point whenever we need it. It saves the memory space required for ECC.

All the integers are considered in terms of polynomial. For example, number 7 $(111)_2$ is considered as x^2+x+1 .

We have converted whole plain text into ASCII value and then it is converted into point of a curve. The benefit of this method is avoiding repetition of cipher text block. So, cryptanalytic attack is not possible. Also size of cipher text is reduced resulting less storage space requirement.

ECC has following four advantaged over RSA which we have achieved in our implementation.

1. Less Storage Space
2. Less Processing power because of small key size
3. Less bandwidth consumption because of less key size and small cipher text.
4. Less time for encryption as we is described in Section 5.

5 Experimental Results

We have implemented ECC for Android which is one of the popular smart phone OS. Android is developed by Google for different mobile devices and for note book computers .It is widely used for many banking or financial purposes. It supports java enabled applications. In our experiments we have checked execution time required for ECC and RSA with variable text size and key size on android OS. The results are taken on SDK Platform Android 2.1 Update 1 Version. ECC Development environment is given in Fig.3. Machine configurations are Intel Core2 Duo CPU, 1.18GHz, 0.99GB RAM. Eclipse IDE is used for developing ECC in java. Same java

program is executed on Android Mobile and Jad & Jar file of ECC implementation can be executed on Windows Mobile. Windows mobile support C# language but java Jad and Jar file can be executed using JBlend software.

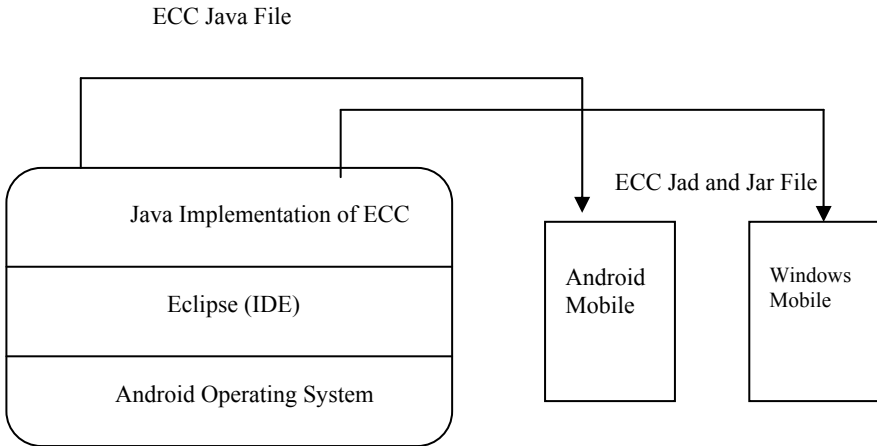


Fig. 3. ECC Development Environment

In Fig 4 and Fig 5, ECC is compared with RSA by varying key size and data size respectively.

In Fig 4, we can see RSA’s execution time is less than ECC when key size is small.

When we increase key size, RSA’s execution time also increases which is greater than ECC. For any security algorithm, key size is more means security is more. So, key size cannot be compromised.

In Fig.5, RSA’s execution time is compared with ECC by varying Data Size. There is no greater effect on RSA if we increase Data size.

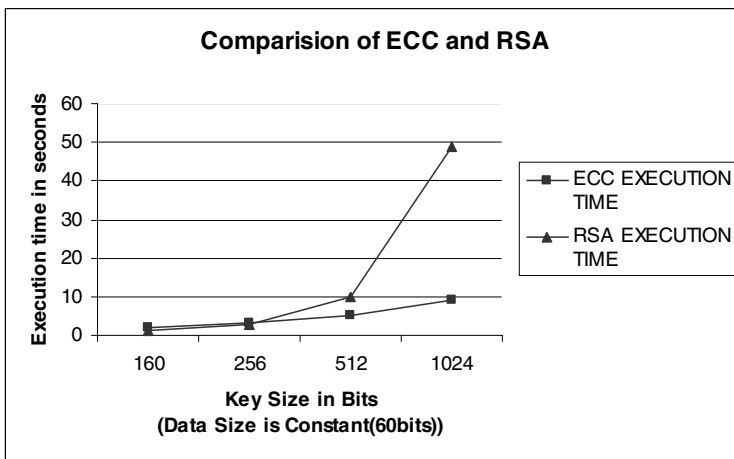


Fig. 4. Performance Comparison of ECC with RSA by varying Key Size on Android 2.1

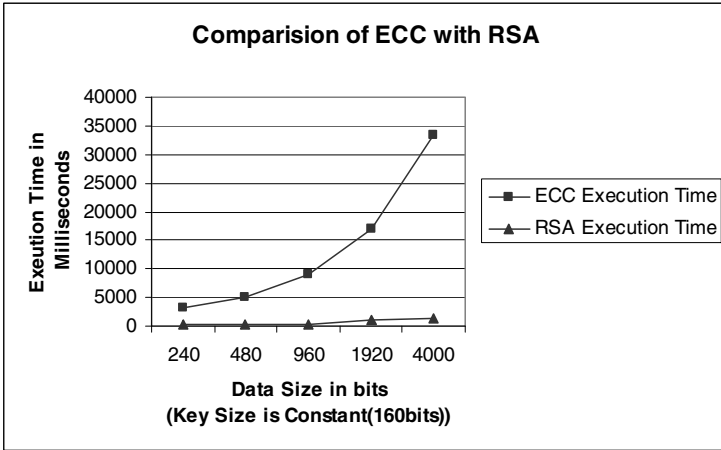


Fig. 5. Performance Comparison of ECC with RSA by varying Data Size on Android 2.1

In Fig.6, ECC is evaluated by varying Field Size. More is the field size, more is the execution time and security is also more.

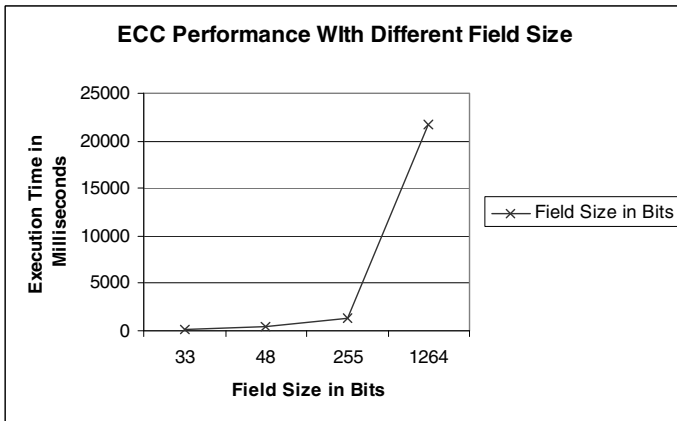


Fig. 6. Performance Evaluation of ECC By varying Field Size on Android 2.1

In Fig 7, ECC key generation time is checked. More is the key size, more time it will take but more will be the execution time.

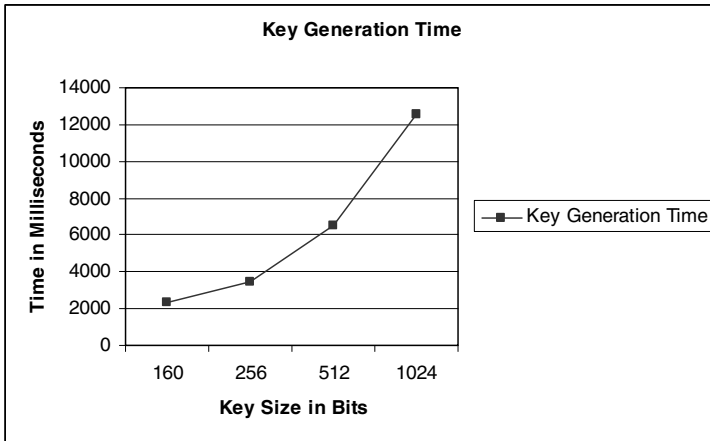


Fig. 7. ECC Key Generation Time Comparison on Android 2.1

6 Future Work

In our paper, we have implemented ECC for Android OS. In future, we can check performance of ECC on different smart phone OS like Windows Mobile, Blackberry or iPhone etc. Our implementation of algorithm is compatible on any machine, any operating system or on any network type. ECC can be replaced with Hyper ECC to get same security with minimum number of points.

7 Conclusion

In our paper, algorithm for implementing ECC is explained. We have implemented ECC in Java and tested on Android 2.1 Update 1 OS. In our ECC implementation we have reduced storage size required for storing points of elliptic curve by dynamically generating them. We avoided cryptanalytic attack by eliminating repetition of cipher text patterns. We reduced size of cipher text, so the storage required for this will be saved. Also the bandwidth required for transferring cipher text over the air will be reduced. Our implementation is compatible with any OS, any machine and with any network type. We took results for both ECC and RSA on Android OS. In our experiment results, it is observed that ECC's performance is always better than RSA when key size is increased more than 512 bits. In cryptography, large key size is used to increase security of the data. As smart phones are becoming popular in the world for its personal information storage or for doing different financial tasks online, Security is the need for them. So, RSA can be replaced with ECC to get more security.

References

1. Koblitz, N.: Elliptic curve cryptosystems. *Mathematics of Computation* 48, 203–209 (1987)
2. Miller, V.S.: Use of elliptic curves in cryptography. In: Williams, H.C. (ed.) *CRYPTO 1985*. LNCS, vol. 218, pp. 417–426. Springer, Heidelberg (1986)
3. Yan, H., Zhijie Jerry, S.: *Studying Software Implementations of Elliptic Curve Cryptography*. IEEE, Los Alamitos (2006)
4. Vigila, M., Muneeswaran's, K.: *Implementation of Text based Cryptosystem using Elliptic Curve Cryptography*. IEEE, Los Alamitos (2009)
5. Kong, H., Zeng, Z., Yan, L., Yang, J., Yao, S., Sheng, N.: *Combine Elliptic Curve Cryptography with Digital Watermark for OWL Based Ontology Encryption*. IEEE, Los Alamitos (2009)
6. Aydos, M., Yanik, T., Kog, C.K.: High-speed implementation of an ECC based wireless authentication protocol on an ARM microprocessor. *IEEE Proc. Commun.* 148(5), 273–279 (2001)
7. Lauter, K.: The Advantages of Elliptic Cryptography for Wireless Security. *IEEE Wireless Communications*, 62–67 (February 2006)
8. Muthukumar, B., Jeevanantharr, S.: *Design of an Efficient Elliptic Curve Cryptography Coprocessor*. IEEE, Los Alamitos (2009)
9. Kanniah, U.S., Samsudin, A.: *Multithreading Elliptic Curve Cryptosystem*. IEEE, Los Alamitos (2007)
10. Chen, J.-H., Shieh, M.-D., Wu, C.-M., Taiwan.: *Concurrent Algorithm For High-speed Point Multiplication In Elliptic Curve Cryptography*. IEEE, Los Alamitos (2005)
11. Yan, H., Shi, Z.J.: *Software implementation of ECC over 8-bit processor*. IEEE, Los Alamitos (2006)
12. Ahmad, T., Hu, J., Han, S.: *An Efficient Mobile Voting System Security Scheme based on Elliptic Curve Cryptography*. IEEE, Los Alamitos (2009)
13. Jagdale, B.N., Bedi, R.K., Desai, S.: *Securing MMS with High Performance Elliptic Curve Cryptography*. *International Journal of Computer Applications* 8(7), 17–20 (2010)
14. Khajuria, S., Tange, H.: *Implementation of Diffie-Hellman Key Exchange on Wireless Sensor Using Elliptic Curve Cryptography*. IEEE, Los Alamitos (2009)
15. <http://www.pcworld.com>
16. <http://topnews.co.uk>
17. <http://www.certicom.com>
18. Stallings, W.: *Cryptography and Network Security*, 4th edn. Prentice Hall, Englewood Cliffs (2006)

An Improved Secure Authentication Protocol for WiMAX with Formal Verification

Anjani Kumar Rai¹, Shivendu Mishra¹, and Pramod Narayan Tripathi²

¹ Department of Computer Science and Engineering,
Motilal Nehru National Institute of Technology, Indian Institute of Information
Technology, Allahabad, India

² Department of Electronics Engineering, Allahabad, India
{anjani.it12,2009is17,Pramod.n.tripathi}@gmail.com

Abstract. Privacy and Key management protocols (PKM) is used in WiMAX for providing authentication and key management. Basic PKM protocol provides one way authentication between SS and BS results many flaws. However, PKM protocol version 2 (PKMv2) solves the major security problems but new flaws have emerged. This paper analyzes the PKM protocol and its later versions using AVISPA which is a push button tool for the automated validation of security protocol. A secure authentication protocol has also been proposed and analyzed, results show that proposed protocol does not have any security flaws.

Keywords: AVISPA, privacy and key management protocol, IEEE 802.16, SPAN, WiMAX.

1 Introduction

IEEE 802.16 standard offers large bandwidth and high transmission speed to specify air interface of Wireless Metropolitan Area Network (Wireless MAN). IEEE 802.16-2004 [1] [2] is an amendment in IEEE 802.16, also known as WiMAX which is a forum promoting the IEEE 802.16 standard. In IEEE 802.16; MAC layer is divided into three parts: privacy sublayer; common part sublayer and convergence sublayer. Privacy sublayer uses Privacy and Key Management version 1 (PKMv1) protocol for secure authentication and distribution of session key information from the authenticator (BS) to supplicant (SS). Privacy and Key management protocol in IEEE 802.16 provides one way authentication in which BS authenticates SS using X.509 certificate of SS. In paper [4], authors have enhanced the PKMV1 authentication protocol using nonce in which SS authenticates BS and BS authenticates SS. S. Xu, M. Matthews and C.T. Huang also analysed the PKMV1 and its nonce version; they proposed a solution using timestamp [5]. IEEE 802.16e [3] amendment adds mobility functionality in IEEE 802.16. IEEE 802.16e defines PKMV2 protocol by providing mutual authentication between SS and BS. Papers [5] [6] [7] [8] [9] published the attacks on PKMv2 with proposed solution.

There are many tools available for verification of the security protocols. CasperFDR [10] [11] and Automated Validation of Internet Security Protocols and Applications (AVISPA) [12] are the well known advanced tools. Paper [6] made formal analysis of PKM and other versions using BAN logic. Since BAN logic has several deficiencies like inability to handle secrecy properties, authors [13] have modeled the PKM and its other version using CasperFDR. The weakness of CasperFDR is that it can not model the intruder as a legal user and certificate can not assign to users as their initial knowledge.

AVISPA is a push button tool for the automated validation of security protocol, AVISPA also addresses the problems associated with CasperFDR. A modular and expressive formal language called HLSPL (High Level Protocols Specification Language) is used by AVISPA to specify the security protocol and their properties. A large number of protocols, including several variants of generic protocols like Kerberos and EAP have already modeled in HLPSSL [14] [15]. AVISPA tool consist different modules such that hlpssl2if translator that converts user written hlpssl scripts into IF (intermediate format) specification. There are four different verification back end tools use to analyze the IF specification, one of them is OFMC (On-the-Fly Model-Checker) [16]. OFMC employs symbolic techniques to perform bounded analysis and protocol falsification. OFMC provides a translation which is use to find attack (if exist) in any protocol. Translation and checking are fully automatic and performed by OFMC without use of external tool. SPAN [17] interactively produces the Message Sequence Charts (MSC) [18] from an HLPSSL specification which provides the better understanding of the specification, check that it is executable and that it corresponds to what is expected. Attack Simulation is a mode of SPAN for automatic building of MSC attacks from the output of OFMC tool.

In this paper, PKM protocols and its later Versions are specified and verified using AVISPA and SPAN. Attacks are found on each of the protocols and results are discussed. This paper also proposes a secure authentication scheme which provides the mutual authentication between BS and SS. Our proposed protocol is also specified and verified, results show that proposed protocol does not have any security flaws.

The rest of the paper is composed as follows. In Section 2, we specify and analyze the original PKMv1, PKM Nonce Version [4] and PKMv2 authentication protocols. Section 3 discusses our proposed protocol with formal analysis. Section 4 concludes the paper.

2 Authentication Protocols in WIMAX

This section describes PKMv1, its nonce version [4] and PKMv2 authentication protocols. Protocols are also analyzed using AVISPA and SPAN, attacks are found in each of the protocols and results are discussed.

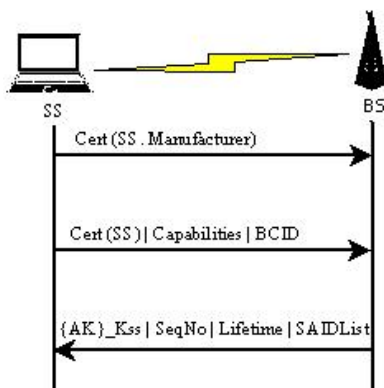


Fig. 1. PKMv1 Authentication protocol in WiMAX

2.1 PKMv1 Authentication Protocol

Fig. 1 illustrates the process of Privacy and Key Management version 1 (PKMv1) authentication protocol in IEEE 802.16 (WiMAX) consisting three steps. In step 1, SS sends X.509 certificate of SS manufacturer which is very informative and BS may ignore it. Following the step 1, the SS sends an authorization request to the BS, which contains the following information: Unique X.509 certificate of SS including its RSA public key, cryptographic algorithms that is supported by SS (Capabilities), the primary SAID (BCID). BS verifies the SS by X.509 certificate and sends authorization reply message which contains authentication key (AK) encrypted by public key of SS, authentication key sequence number, AK lifetime, list of security association IDs that the SS is authorized to access with their associated properties

2.2 Specifying and Verifying PKMv1 Authentication Protocol Using AVISPA/SPAN

Since message 1 in PKM and its other version is informative therefore this specification and later specifications will not include it, also we use two basic roles `pkm_init` played by SS and `pkm_res` played by BS. Each basic role consist the initial information known by the participant, its initial state and the transition by which state can change. After defining the basic roles, we have to define composed roles describing the sessions of the protocol. Finally a top level role containing global constant, a statement describing the initial knowledge of intruder and composition of one or more session is defined. Events witness (SS, BS, $m2$, $SS.Kss_{(inv(Kca))}$), request (BS, SS, $m2$, $SS.Kss'_{(inv(Kca))}$) and witness (BS, SS, ak , Ak'), request (SS, BS, ak , Ak') are used for authentication whereas secret (Ak' , ak , SS, BS) is used by an agent to allow that any value is shared between (only) agents.

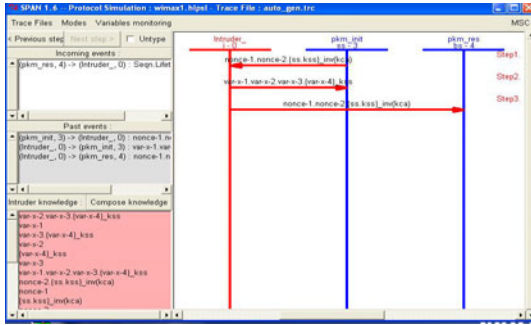


Fig. 2. SPAN animator screenshot for the attack on the event witness (SS, BS, ak, Ak'), request (SS, BS, ak, Ak') in PKMv1 protocol

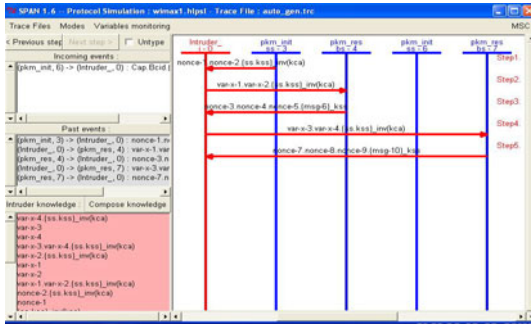


Fig. 3. SPAN animator screenshot for the attack on the event witness (SS, BS, m2, SS.Kss(inv(Kca))), request (BS, SS, m2, SS.Kss'(inv(Kca))) in PKMv1 protocol

After executing the protocol specification, attacks were found to each of the two events declared in the Specification. Using attack simulation mode of SPAN (Security Protocol Animator for AVISPA), we found the attacks from the output of OFMC tool.

Fig. 2 shows the attack on event witness (BS, SS, ak, Ak'), request (SS, BS, ak, Ak') in which SS authenticates BS on ak. When pkm_init (SS) sends the authentication request message to Pkm_res (BS). Message is intercepted by intruder and intruder fakes his own authentication reply message including the AK generated by him. SS will think that the AK is passed by BS and that it is only known by BS and him, but in fact, the intruder knows it. This attack is due to lack of mutual authentication. Intruder may also replay the intercepted message and get the response from BS.

Fig. 3 shows the attack on event witness (SS, BS, m2, SS.Kss(inv(Kca))), request (BS, SS, m2, SS.Kss'(inv(Kca))) in which BS authenticates SS on m2. In the first legitimate session pkm_init (SS) sends the authentication request to pkm_res (BS) which is intercepted by the intruder. Intruder copies the message

and replay the message in the next session which may causes denial of service attack to the victim SS. When we allow the intruder to play a role of legitimate SS then we did not find any new attack on the protocol whereas allowing the intruder to play a role of legitimate BS gives an attack in which intruder (BS) initiate a connection with a SS and gets authentication request message and sends this message to the BS in different session. This attack is similar to replay attack discussed previously.

2.3 Nonce Version of Privacy and Key Management Protocol

In [4], authors have modified the PKMv1 protocol by using nonce, their modified protocol is shown in fig. 4.

As shown in figure, message 1 is same as in previous protocol. In message 2, SS sends nonce (Ns) in the authorization request message. Following the authorization request message, BS sends authorization reply message including nonce (Ns), his own nonce (Nb), and pre authentication key to prevent the exposure of authentication key and signature of BS on the message 3.

2.4 Specifying and Verifying Nonce Version of Privacy and Key Management Protocol Using AVISPA / SPAN

Nonce version of PKM protocol is specified in the same way as PKMv1 protocol. In this specification we have added event witness (BS, SS, ns, Ns'), request (SS, BS, ns, Ns) in which SS authenticates BS on nonce ns. Since in this protocol BS sends pre authentication key in place of authentication key so event witness (BS, SS, ak, Ak'), request (SS, BS, ak, Ak') is changed by event witness (BS, SS, pre_ak, PreAk'), request (SS, BS, pre_ak, PreAk'). After executing the specification, we did not find any attack on the events witness (BS, SS, pre_ak, PreAk), request (SS, BS, pre_ak, PreAk') and witness (BS, SS, ns, Ns'), request (SS, BS, ns, Ns) because SS authenticates BS by adding BS's certificate in authentication reply message with the signature on the same message. Analyzing the attack on event witness (SS, BS, m2, SS. Kss'_(inv(Kca))), request (BS, SS, m2, SS. Kss_(inv(Kca))) in which BS authenticates SS on m2, results the replay attack in the same way as discussed in previous section.

This attack is still possible since nonce only assure SS that authentication reply is the response of corresponding authentication request whereas BS can not know that authentication request is sent recently or an old request also if nonces were reused then response corresponding a request could be replayed. PreAk in place AK (Authorization Key) is hardly more secure than AK generated by BS [5].

2.5 PKMv2 Authentication Protocol

In PKMv2 authentication protocol, one more message is added at the end of PKMv1 protocol. Fig. 5 shows the message flow in PKMv2 protocol. All the

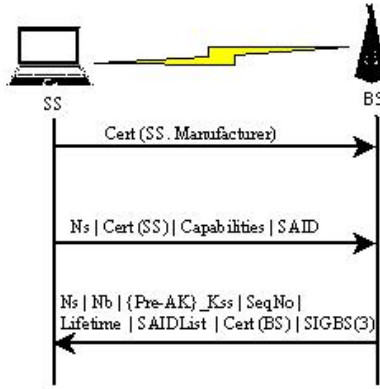


Fig. 4. Authentication protocol with nonce in WiMAX

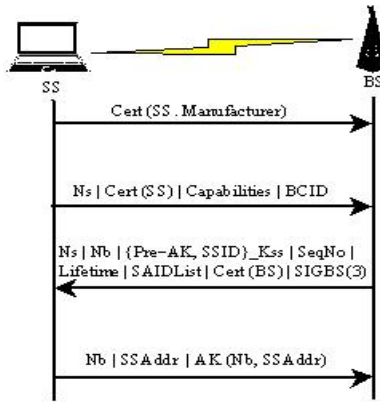


Fig. 5. PKMv2 authentication protocol in WiMAX

information is same as in PKMv1 and their nonce version, The SS Identifier (SSID) is unique identifier from the certificate of SS. SIGBS (3) is digital signature of BS enhances the authenticity of message 3 and AAID is the Authorized Association Identifier determines the selected security association. In step 4, SS acknowledges the message 3 (authorization reply message) with BS’s nonce (Nb) from message 3 and MAC (Physical) address of SS, which is encrypted by authorization key (AK).

2.6 Specifying and Verifying PKMv2 Authentication Protocol Using AVISPA / SPAN

Again we have excluded the message 1 and the protocol with rest of the message is modeled in HLPSL. Ak is added as symmetric key in both the basic roles

i.e. `pkm_init` and `pkm_res` and is shared between SS and BS. In each transition, one more event is added compared to transition of PKMv1 and its nonce version. SS and BS authenticates each other with the witness (BS, SS, ns, Ns'), request (SS, BS, ns, Ns) in which SS authenticates BS on ns and witness (SS, BS, nb, Nb'), request (BS, SS, nb, Nb) in which BS authenticates SS on nb whereas secrecy of authorization key is obtained with the event `secret` (PreAk', `pre_ak`, SS,BS).

Executing the specification using OFMC tool with SPAN, we did not find any attack on event witness (BS, SS, ns, Ns'), request (SS, BS, ns, Ns) but on the event witness (SS, BS, nb, Nb'), request (BS, SS, nb, Nb), we got an attack as discussed in [5] [6] [7] [8] [9] [13] and is known as Interleaving Attack. Using attack simulation mode of SPAN, screenshot showing the attack is shown in fig. 6. As shown in figure, Intruder initiates the protocol with a SS in a session labeled as 6 by sending a start message. SS sends the authorization request message (message 2 in protocol) to intruder or the intruder may intercept the authorization request message sent by SS. The intruder then forwards the authorization request message to BS.

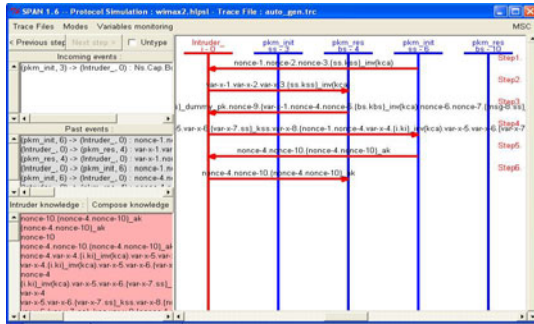


Fig. 6. SPAN animator screenshot for the attack on the event witness (SS, BS, nb, Nb'), request (BS, SS, nb, Nb) in PKMv2 protocol

However, the receiving instance of BS belongs to the different session, labeled with 3, which is the original session. In the same session intruder gets the response (authorization reply message) from the BS. Since the Pre Authorization key is encrypted with public key of SS therefore intruder can not acknowledge the response message. Now, intruder sends the authorization reply message impersonating as BS to SS in the previous protocol instance. This message contains same Authorization Key and nonce send by BS in original session. SS acknowledges the authorization reply message which contains nonce, MAC address of SS and the same information encrypted with Authorization Key. Intruder forwards this acknowledged message to BS and finish the protocol. This attack is interleaving attack in which intruder impersonates as SS to BS and vice versa.

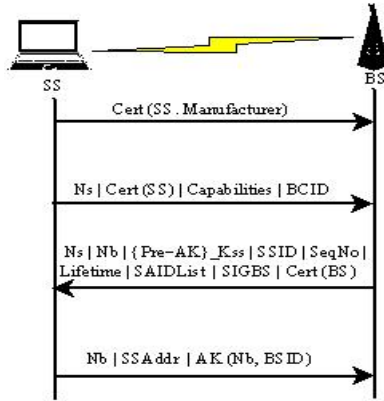


Fig. 7. Proposed authentication protocol in WiMAX

3 Proposed Protocol

In this authentication protocol, issue discussed in previous sections can be avoided. Fig. 7 illustrates the process of our proposed protocol which is an improvement of PKMv2 protocol. Here, authorization reply message (message 3) is changed in which SSID which was encrypted by public key of SS in PKMv2 protocol is sent without encryption because message is digitally signed by BS and encryption does not provide more security than without encryption. Certificate of BS is sent without signature on it since certificate is already signed by CA (trusted authority). In message 4, identity of BS (BSID) is added which is encrypted by authorization key (AK). Addition of BSID in message 4 avoids the interleaving attack discussed in previous section. Intruder (impersonating as SS) can not forward acknowledgement of authorization reply message (sent by legitimate SS to BS) because this message contains the identity of intended receiver (BS) which is encrypted by Authorization Key shared between SS and BS and hence intruder can not change the BSID. Also, SS address is not encrypted because AK is derived from SS address, if SS address is modified then BS can not derive the same AK and can not decrypt the message 4 [6].

3.1 Specifying and Verifying Proposed Protocol Using AVISPA / SPAN

Proposed protocol is also specified and verified using AVISPA / SPAN. In specification of proposed protocol some modification has been performed as discussed above and is specified in the same way as PKMv2. Fig. 8 shows the simulation of proposed protocol using protocol simulation button of SPAN. In protocol simulation message 1 is not included for the reason discussed in section 2.2. After executing the specification, we did not find any attack since message 3 in

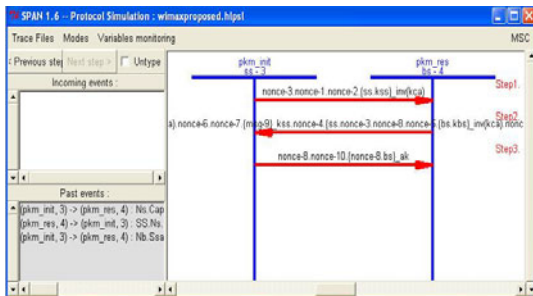


Fig. 8. Simulation of proposed authentication protocol using SPAN

```

wimaxproposed - Notepad
File Edit Format View Help
% OFMC
% Version of 2006/02/13
SUMMARY
SAFE
DETAILS
BOUNDED_NUMBER_OF_SESSIONS
PROTOCOL
C:\progra-1\SPAN\testsuite\results\wimaxproposed.if
GOAL
as_specified
BACKEND
OFMC
COMMENTS
STATISTICS
  parseTime: 0.01s
  searchTime: 0.37s
  visitedNodes: 139 nodes
  depth: 7 plies
    
```

Fig. 9. Result obtained for proposed authentication protocol using SPAN

fig. 8 (message 4 in fig. 7) contains the identity of BS encrypted by Authorization Key shared between SS and BS. Protocol is safe against all type of passive and active attack. Fig. 9 shows the result of proposed protocol using SPAN.

We know that the public key encryption/decryption is a costly process therefore modification of message 2 minimizes the authentication delay and encryption/decryption time since certificate is neither encrypted nor signed. Addition of BSID prevents the interleaving attack, multiplicity attack and replay attack.

4 Conclusion

In this paper, we specified and verified the PKM protocol and its later Versions using AVISPA / SPAN. Using attack simulation mode of SPAN, attacks were found and discussed on PKM protocol and its later versions. This paper also proposes a secure authentication protocol for IEEE 802.16 (WiMAX) network which is an improvement of PKMv2 authentication protocol. Proposed protocol provides mutual authentication between SS and BS and reduces the authentication delay and encryption/decryption time. Our proposed protocol is also specified and verified, no new vulnerability or attack has been surfaced. Future works will be focused on

the performance test of the protocol. Moreover, we intend to design the Authentication Protocol using Identity-Based Cryptosystem so that we can minimize the certificate management overhead associated with PKI.

References

1. IEEE std 802.16 2004: Air interface for fixed broadband wireless access system, IEEE (2004)
2. IEEE 802.16 and WiMax: Broadband Wireless Access for everyone, Intel White Paper (2004)
3. IEEE std 802.16e2005: Air interface for fixed broadband wireless access system amendment: Physical and medium access control layers for combined fixed and mobile operation in licensed bands, IEEE (2006)
4. Johnston, D., Walker, J.: Overview of IEEE 802.16 Security. IEEE Security & Privacy (2004)
5. Xu, S., Matthews, M., Huang, C.-T.: Security Issues in Privacy and Key Management Protocols of IEEE 802.16. In: Proceedings of the 44th ACM Southeast Conference (ACMSE 2006) (March 2006)
6. Xu, S., Huang, C.T.: Attacks on PKM protocols of IEEE 802.16 and its later versions. In: ISWCS 2006: Proceedings of the 3rd International Symposium on Wireless Communication Systems (September 2006)
7. Tian, H., Pang, L., Wang, Y.: Key management protocol of the IEEE 802.16e. Wuhan University Journal of Natural Sciences 12(1) (January 2007)
8. Sidharth, S., Sebastian, M.P.: A Revised Secure Authentication Protocol for IEEE 802.16 (e). In: International Conference on Advances in Computer Engineering (2010)
9. Yuksel, E.: Analysis of the PKMv2 protocol in IEEE 802.16e 2005 using static analysis. Informatics and Mathematical Modelling (2007)
10. Formal Systems (Europe) Ltd.: FDR2 user manual: Failure divergence refinement (May 2000)
11. Lowe, G.: Casper: A compiler for the analysis of security protocols. Journal of Computer Security 6, 53–84 (1998)
12. Avispa a tool for Automated Validation of Internet Security Protocols, <http://www.avispa-project.org>
13. Xu, S., Huang, C.-T., Matthews, M.M.: Modeling and Analysis of IEEE 802.16 PKM Protocols using CasperFDR. In: IEEE ISWCS (2008)
14. D6.2: Specification of the Problems in the High-Level Specification Language, <http://www.avispa-project.org>
15. Rai, A.K., Kumar, V., Mishra, S.: An Improved Password Based EAP Method for WiMAX with Formal Verification. In: IJCA Proceedings on International Conference and workshop on Emerging Trends in Technology (ICWET), vol. 8, pp. 29–35. Published by Foundation of Computer Science, USA (2011)
16. Basin, D., Mödersheim, S., Viganò, L.: An On-the-Fly Model-Checker for Security Protocol Analysis. In: Sneekenes, E., Gollmann, D. (eds.) ESORICS 2003. LNCS, vol. 2808, pp. 253–270. Springer, Heidelberg (2003)
17. SPAN a Security Protocol Animator for AVISPA, <http://www.irisa.fr/lande/genet/span>
18. Harel, D., Thiagarajan, P.S.: Message sequence charts. UML for Real: Design of Embedded Real-time Systems (2003)

Secured Fault Tolerant Mobile Computing

Suparna Biswas¹ and Sarmistha Neogy²

¹ West Bengal University of Technology, India

² Jadavpur University, India

sarmisthaneogy@gmail.com

Abstract. Checkpointing algorithms in mobile computing systems suffer from security attacks. Checkpoint data being stored in stable storage or being transferred over the wireless network is the source of information leakage. Secured Checkpointing comes with computational and storage overhead. In resource constrained mobile computing systems overhead optimization is a challenge. In this paper we propose a secured fault tolerant algorithm which combines movement based checkpointing algorithm with public key cryptography.

Keywords: Checkpoint, recovery, computation data, cryptography, authentication, confidentiality, nonrepudiation etc.

1 Introduction

Mobile hosts save and transfer checkpoint data to the mobile support stations through wireless channels which may be needed during recovery from failure. It is possible that some unauthorized malicious nodes can read or access data by “sniffing” when the data is being transferred. Most of the relevant works reported in [1],[2],[3],[4],[5],[6] have addressed only the checkpointing and rollback recovery satisfying constraints of mobile computing systems. But the security threats related to unauthorized access of stored checkpoint data, information leakage through checkpoint and computation data in transmission or other possible attacks have not been paid concern. This leads us to include security issues with checkpointing technique so that a secure, fault tolerant mobile computing system suitable for business applications related to financial transaction or active military unit with high confidential data can be formed. The work reported in [8] is one of its kind. This aspect reflects that a checkpointing algorithm appropriate for secure mobile checkpointing should be “mobility aware low overhead secure checkpointing”. High mobility rate of mobile hosts causes frequent hand-off. Hence checkpointing is triggered by threshold value of hand-off count of a mobile host. Here it is worth to mention that threshold of hand-off count can be set per system per application basis as two deciding factors considered to set threshold are mobility rate of mobile hosts. Security threats regarding authenticity, confidentiality, data integrity, nonrepudiation etc. are frequent and obvious in mobile computing systems. As checkpointing itself incurs overhead so security solutions should be simple but powerful. Investigating relevant works [8], [9],[14], on low overhead public key cryptography suitable for

resource constrained small portable devices, we found that ECC-160 point multiplication outperforms RSA 1024 private key operation and is comparable in performance with RSA 1024 public key operation.

2 Review of Related Works

Some of the relevant works are described here. In [1], [2] low overhead checkpointing and recovery schemes for mobile computing systems are reported. In [7], low overhead checkpointing algorithm is combined with shared secret key algorithm to make checkpointing secure in mobile computing system. In [8] and [9] performance analysis of low overhead public key cryptography technique suitable for small devices based on Elliptic Curve Cryptography is presented.

Sapna E. George [1] et.al describes a checkpointing and logging scheme based on mobility of mobile hosts. A checkpoint is saved when hand-off count exceeds a predefined optimum threshold. Optimum threshold is decided as a function of mobile host's mobility rate, failure rate and log arrival rate. Recovery probability is calculated and recovery cost is minimized in this scheme.

In [2] T.Park et.al has presented an efficient movement based recovery scheme. This scheme is a combination of message logging and independent checkpointing. Main feature of this algorithm is that a host carrying its information to the nearby mobile support station can recover instantly in case of a failure. If a mobile host is moving inside a "certain range", recovery information remains in host mobile support station otherwise it moves recovery information to nearby mobile support station. In this scheme two movement-based schemes are suggested -distance based and frequency based.

In [7], Nam et.al. proposed a secure checkpointing technique. In this paper every mobile node is authenticated to ensure that only checkpoint and recovery nodes get access to the content of checkpoint. Secured checkpointing technique's overhead is only 1.57 times to that of conventional checkpointing schemes. This proves such technique's feasibility. The proposed system is designed to be suitable for Batch RSA or Batch DH-key algorithm. For key management they have used key agreement model .They have addressed three possible attack areas in a local checkpointing host – data sniffing in volatile memory content, access to retained data after termination of process in swap area, cryptographic keys or any secured information stored in binary programs in the local disk. While the checkpoint is transferred from local node to stable storage, possible security attacks are data sniffing, man-in-the middle attack etc. Stable storage is a shared space which has high probability of attacks like sniffing, access to content and modification of data. Secure incremental checkpointing and secure probabilistic checkpointing are selected as base checkpointing technique for low overhead. But this algorithm suffers from temporary blocking of process execution at checkpoint node.

N.Gura et. al in [8] implemented and analyzed RSA-1024 and ECC-160 to find out that ECC-160 outperforms RSA-1024 in performance, memory usage and code size . ECC-160 performance is comparable with RSA 1024 and can be better if key size increases. Strength of security of RSA lies in difficulty of factoring large integers while in ECC strength of security increases with hardness of elliptic curve discrete logarithm problem.

In [9], Wander et.al in their work describes energy requirement analysis of RSA 1024 and ECC 160 public key cryptography for wireless sensor networks. Energy consumed in RSA 1024 public key and private key operations is 4.9 times than that of ECC-160. Because of longer key size, certificates are also larger causing more communication cost in RSA 1024. Handshake energy consumed in RSA1024 is 2.7 times than that of ECC-160. In this work they compared between RSA and ECC algorithms from the point of view of key size and energy required. In ECC-160, key size, computation and handshake energy consumptions are much less than in RSA 1024.

Our study leads to following observations:

- i) most of the checkpointing and recovery protocols are not secured
- ii) Symmetric key cryptography combined with checkpointing [1] is not secured enough because if the key gets compromised security measures will be jeopardized.
- iii) Public key cryptography based on ECC is suitable for small computing devices because of low key size [8] [9].

Hence we combine a movement based checkpointing algorithm with low overhead public key cryptography e.g. ECC-160 to implement a Secured Fault Tolerant Mobile Computing System.

3 System Model

Mobile computing system considered here comprises n number of mobile hosts and m number of mobile support stations where $n \gg m$. Mobile hosts are connected through wireless network and mobile support stations are connected through wired network. Mobile hosts send a hello message periodically to current mobile support stations to inform their connectivity. Communication links connecting mobile hosts and mobile support stations are assumed to be FIFO. Messages take arbitrary but finite amount of time during transmission. There are no synchronized clocks or shared memory among nodes. Two types of checkpoints are saved: i) Migration Checkpoint - ,mobile hosts save before planned disconnection ii) Permanent checkpoints – mobile hosts save if number of handoff exceeds threshold .During current checkpoint interval, messages sent, received are saved into log file. Mobile hosts refresh log files to control its size so that log file search does not incur overhead. Proposed algorithm is non-blocking i.e. mobile hosts can compute, send messages during checkpointing. Mobile hosts move randomly in intercell or intracell movement pattern [10].

3.1 Data Structures and Notations

MSS = mobile support station

MH = mobile host

MCS = Mobile Computing System

CMSS = Current MSS of an MH

old_MSS = MSS that an MH leaves due to hand-off

new_MSS = MSS to which an MH joins after hand-off

T_count = an integer variable to count time

h_c = handoff count
 h_{th} = hand-off threshold
 mh_i , $i = 1 \dots n$, n = number of Mobile Hosts
 mhi_s = mobile host that sends computation data or checkpoint data
 mhi_r = mobile hosts that receives computation data or checkpoint data
 mss_j , $j = 1 \dots m$, m = number of mobile support stations
 m_ch = migration checkpoint
 $m_ch_{i,j,k}$ = migration checkpoint saved by MH_i which is connected to MSS_j during k th checkpoint interval
 per_ch = permanent checkpoint saved by mobile host if h_c exceeds h_T .
 pub_mh_i = public key of mhi
 pri_mh_i = private key of mhi
 pub_mss_j = public key of mss_j
 pri_mss_j = private key of mss_j
 $(m_ch_{i,j,k})_{pub_mhi}$ = migration checkpoint encrypted with mhi 's public key
 $per_ch_{i,j,k}$ = local checkpoint saved by mh_i which is connected to MSS_j during k th checkpoint interval
 $(per_ch_{i,j,k})_{pub_mss_j}$ = local checkpoint encrypted with mss_j 's public key
 $((per_ch_{i,j,k})_{pub_mss_j})_{pri_mhi}$ = signed encrypted checkpoint data with mhi 's private key
 mp = movement pattern
intercell = MH moves across cells during a checkpoint interval
intracell = MH moves within a cell during a checkpoint interval
 $D_flag = 0$, MH connected
 = 1, MH disconnected
 $P_dis = 0$, MH disconnected, unplanned
 = 1, MH disconnected, planned
 = -1, MH failed
 $mh_dep []$ = during current interval, list of mhs from which computation messages received
 $mss_mh_list []$ = list of mhs connected to mss during current checkpoint interval
MH_Structure: MH_i , $i = 1 \dots n$
 MSS_j , $j = 1 \dots m$
Log_file $h_c = 0$, h_{th} = threshold
at sender : $T_c = 0$, T_{th}
 $mh_dep []$ Receiver_MH_id
MSS_Structure: MSS_id , $j = 0 \dots m$ chkpt_interval
 $Mss_mh_list []$ m_e
 log_file
 chkpt
 GCCS []
Checkpoint_file: MH_id
 status
 Data
 chkpt_interval

4 Basic Concepts of Present Work

Mobile support station and mobile host communicate through insecure wireless channel which is vulnerable to security attacks. Malicious nodes can act as original node, can sit in between sender and receiver to see and steal information so that data can be modified and resend as original sender violating authentication, data integrity, confidentiality, nonrepudiation. Mobile hosts fail due to many reasons e.g. battery power shortage, physical damage, memory crashes, processor failure etc. A malicious node can become a legitimate mobile node by accessing authentication information and can jeopardize the application running. Mobile Computing system can fail due to denial-of-service-attack [11]. Hence a system cannot be fault tolerant unless it is secured. In present work each node sends periodically a hello message (mh_id, interval) to its current mss till the mh disconnects or fails. Thus it is ensured that only authentic nodes can participate in computing and communication. Each mh encrypts and signs checkpoint data, computation data, before sending to current mobile support station. As a result, no malicious node can see or modify original data packet. Even if a malicious node is able to have a copy of original data packet, he/she will not be able to decrypt it because decryption key is only with receiver. Thus data confidentiality, data integrity and nonrepudiation is implemented. Each mobile support station and mobile host verify signature before processing any checkpoint data or computation data.

4.1 Working Example

In figure1 mhs compute, send, receive computation and communication message, saves checkpoint as h_c exceeds h_{th} , coordinates [12] with dependant mhs in mh_{i_dep} . Each mss maintains and updates connected mh information in mss_mh_list . Let us assume that mh_1, mh_2, mh_3, mh_4 all are initially connected to mss_1 and that is stored in $mss_mh_list_{initial}$ as shown in fig.1. Mh_1 moves in intercell movement pattern randomly, handoffs and saves checkpoint as h_c exceeds h_{th} . Each mobile host saves list of dependant mobile hosts [13] during current checkpoint interval as shown in fig. 1 and coordinates with them while checkpointing.

Description of Symbols in fig.1:

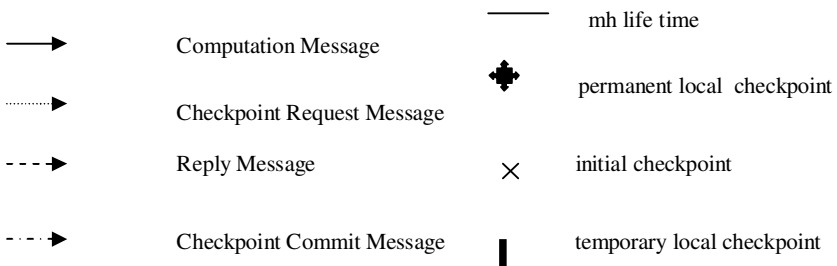


Fig. 1. mhs communicate, saves checkpoint, maintains dependence list etc.

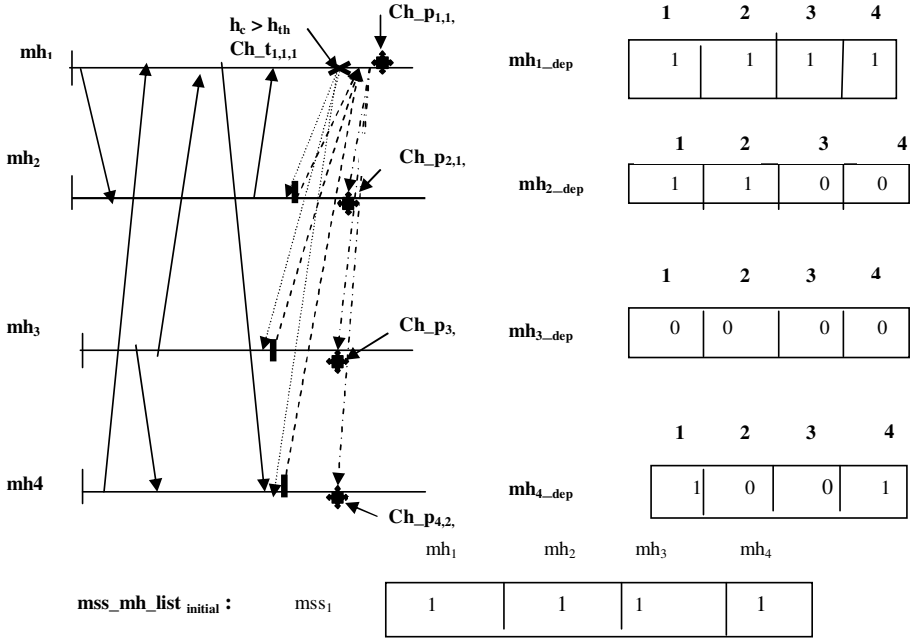


Fig. 1. (continued)

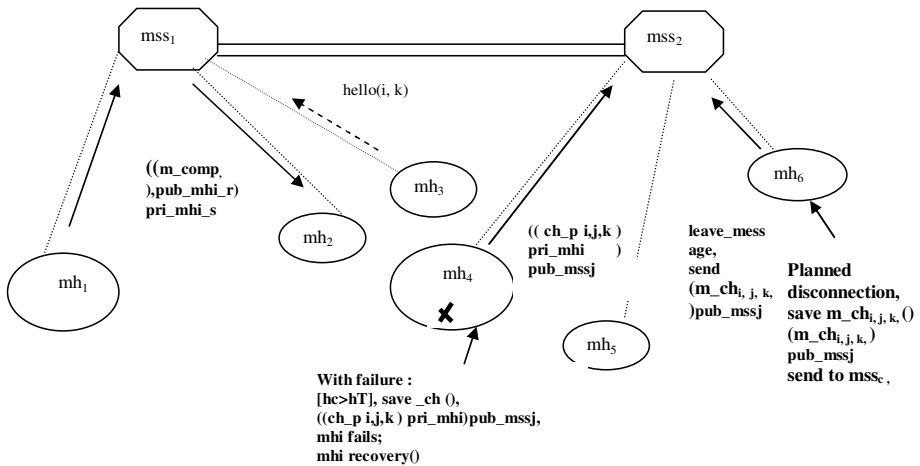


Fig. 2. mhs compute, send computation message, hello message

In Figure2, mh₁ saves checkpoint as h_c exceeds h_{th}, encrypts in local memory as it is working in failure free condition. It encrypts, signs and sends computation data to mh₂ through current mss. Mh₄ works in failure prone environment and hence saves checkpoint data if h_c exceeds h_{th}, and encrypts, signs and sends to mss₂ which is its

current mss. Mh_4 fails and recovery operation starts. Mh_6 saves migration checkpoint, encrypts and sends along with leave message to mss_2 before disconnecting. Mss_2 deletes mh_6 from its mss_mh_list . Mh_3 and mh_5 either compute or are in doze mode [13]. In figure3 mh_4 sends recovery message to mss_1 . Mss_1 broadcasts mh_4 's recovery message to all mss. Mss_2 replies with mh_4 's last saved checkpoint, mh_4 starts computation. Mh_6 joins mss_1 , mss_1 broadcasts join message to all mss, mss_2 replies with mh_6 's $m_checkpoint$, mh_6 gets added in mss_1 's mss_mh_list .

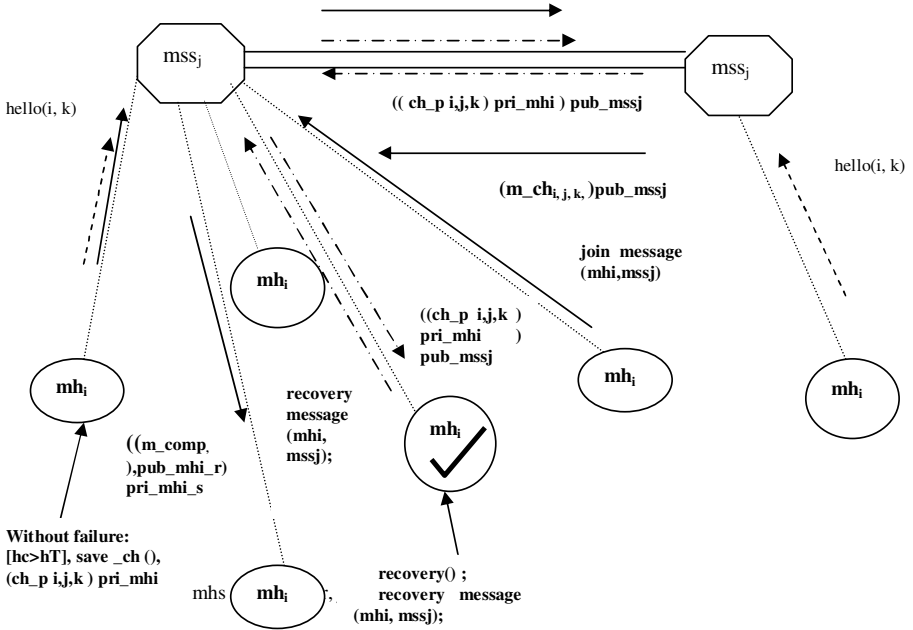


Fig. 3. mhs fail, recover, join in a secure manner as described in section 3. mhs fail, recover, join

5 Algorithm: Secured Fault Tolerant Mobile Computing

5.1 Checkpointing

/ mhs compute, send and receive encrypted computation data, communication messages, move randomly within or across cells, save checkpoints, forward encrypted and signed checkpoints to mss_c s. */*

```

 $h_c = 0, t_c = 0, h_{th} = 3, t_{th} = 3$  unit,  $m=2, n=10;$ 
for ( $j = 0 ; j < m , j ++$ ) {
for ( $i = 0 ; i < n , i ++$ ) {
    if ( $mh_{i,j}$  sends hello message periodically to  $mss_j$ );
         $mss_j$  adds  $mh_i$  in  $mss\_mh\_list$  during each interval;
    
```

```

        mhi,j_status = connected;

else
    mhi,j_status = disconnected;
if (( mp_mhi,j = intercell ) && ( hand-off ==TRUE ))
{
    hci,j ++;
else
    tci,j ++;
}
if ( ( hci,j > hth ) || ( tci,j > tth ) )
    save checkpoint ;
if (operation mode == failure free)
{
    mhi encrypts checkpoint with pri_mhi ;
    saves to local memory ;

else ( operation mode == with failure)
{
    secured computing ( );           // module 5.1.1
    secured checkpointing ( );       // module 5.1.2
    recovery ( );                     // module 5.1.3
    mhi starts computing saved state onwards.

```

5.1.1 // secured computing () // mhi computes, saves computation data, signs encrypts and sends to other mhi, which receives, decrypts and computes with it

```

{
    mhi computes ( );                // saves computation data
    E(comp_data) = (comp_data)pub_mhj_r ; // mhi_s encrypts m_cpq,i,k with
mhi_r's
                                public key
    forward E(comp_data) to mhi_r ; // through mssc;
    D(E(comp_data))pri_mhi_r = comp_data; // mhi_r decrypts E(comp_data) with
                                private key
}

```

5.1.2 // secured checkpointing ()

```

{
    E(ch_pi,j,k) = (ch_pi,j,k)pub_mssj ; // mhi encrypts ch_pi,j,k with mssj's public key
    SE((ch_pi,j,k)pub_mssj)pri_mhi ; // mhi signs (ch_pi,j,k)pub_mssj with
                                private key
    forward S(E(ch_pi,j,k)) to mssj ; } // mhi recovers in mssnew

```

5.1.3 // recovery ()

{

5.1.3.1 verify signature () $E(ch_{p_{i,j,k}}) = V(SE(ch_{p_{i,j,k}}))_{pub_{mhi}}$; // mss_j verifies mh_i's sign**5.1.3.2 decrypt ()** $(ch_{p_{i,j,k}}) = D(E(ch_{p_{i,j,k}}))_{pri_{mssj}}$; // mss_j decrypts checkpoint**5.1.3.3 encrypt ()** $E(ch_{p_{i,j,k}}) = (ch_{p_{i,j,k}})_{pub_{mhi_s}}$; // encrypts with mh_i's public keysends to mh_i through mss_{new}; }**6 Performance Analysis**

Theoretical study and analysis of performance of our proposed secured checkpointing algorithm is based on following assumptions: Each mh saves checkpoint if handoff exceeds threshold mss search cost of a failed mh during current checkpoint interval is bounded by h_{th} . Size of each mh's computation data, checkpoint data will be almost equal since it is assumed that the nature and volume of computation do not vary much. Hence costs of saving checkpoints, writing it to local memory or stable storage are almost constant. Due to advancement in wireless communication, transmission cost of checkpoint and computation data is less and almost constant. Cost is estimated in terms of required energy unless stated otherwise.

6.1 Secured Fault Tolerant Computing Cost (C_{sftc}) $C_{sftc} = \text{Base checkpoint cost (} C_{ch} \text{)} + \text{Cryptography cost (} C_{cryp} \text{)} = C_1 + C_2$ **6.1.1 Base Checkpoint Cost (C_{ch}) : C₁**

Base checkpoint here is movement based and each mh save checkpoint only if handoff count exceeds threshold. Thus checkpoint rate of mh is proportional to handoff rate which is a factor of mobility rate and movement pattern [10]. If a mh fails, it recovers in a new mss.

Table 1. Cost components of C₁

Cost Components	Subcomponent description	notations
C _{ch}	Cost of permanent checkpoint + Cost of migration checkpoint + Cost of failure Recovery	C _{ch_p} + C _{mch} + C _{rec}

C_{ch_p}	Cost to write checkpoint to memory of mh + Cost to transmit over network + Cost to write checkpoint in stable storage of mss	$C_{ch_{write_mh}} + C_{trans} + C_{ch_{write_mss}}$
C_{ch_m}	Cost to write checkpoint to memory of mh + Cost to transmit over network + Cost to write checkpoint in stable storage of mss	$C_{mch_{write_mh}} + C_{trans} + C_{mch_{write_mss}}$
C_{rec}	Cost of searching mss after last saved checkpoint and before failure + Cost of transferring checkpoint data scattered in number of mss $< h_{th}$ before failure	$C_{search_mss} + C_{trans}$

Table 1. (continued)

6.1.2 Cryptography Cost (C_{cryp}) : C_2

Table 2. Cost components of C_2

Cost Components	Subcomponent description	notations
(C_{cryp})	cryptography cost of	$C_{ch_cryp} + C_{comp_cryp}$
	checkpoint + cryptography cost of computation data)	
C_{ch_cryp}	Cryptography cost of (permanent +migration)checkpoint	$C_{pch_cryp} + C_{mch_cryp}$
C_{pch_cryp}	Public key encryption + digital signature + signature verification + private key decryption	$C_{E_pch} + C_{sign_pch} + C_{sign_verify_pch} + C_{D_pch}$
C_{mch_cryp}	Public key encryption+ private key decryption	$C_{E_mch} + C_{D_mch}$
C_{comp_cryp}	Public key encryption+ private key decryption	$C_{E_comp} + C_{D_comp}$

Based on the assumptions for a particular application running in a MCS, C_1 is almost constant. Hence, $C_{sftc} \propto C_2$.

This findings help us to chose a low overhead public key cryptography suitable for small devices e.g. ECC-160. Implementation, performance analysis and comparison of public key cryptography algorithms is beyond scope of our work. We adopt the results obtained in [8], [9] which are described in following tables respectively.

Table 3. Time required to generate and verify signature [8]

Public key algorithms	Signature generation (t)	Signature verification (t)
RSA - 1024	10s	400 ms
ECC – 160	150 ms	630 ms

Table 4. Energy required to generate and verify signature (mJ), key exchange (mJ) [9]

Public key algorithms	Signature		Key Exchange	
	generate	verify	Client	Server
RSA - 1024	304	11.9	15.4	304
ECDSA – 160	22.82	45.09	22.3	22.3

Our choice of ECC-160 point based multiplication algorithm for cryptography is justified with above results.

6.2 Comparative Study of Secure Checkpointing Protocols

Proposed secured checkpointing protocol is compared with the work described in [7] on the basis of following factors:

Table 5. Comparison of proposed work and the work reported in [7]

Algorithms	Proposed	Secure Checkpointing [7]
Type of Cryptography	Assymmetric key	Symmetric key
Recovery node	Each mobile host saves checkpoint, fails and recovers in any mobile support station	Each checkpoint node has a dedicated recovery node which is initialized in the beginning
Key selection	Each node has a unique private-public key pair	Each pair of checkpoint node and recovery node share a secret key
Secure communication	A node with a private-public key pair can communicate securely with any node in the system	A node with a shared secret key can communicate securely with the node that shares the same key
Strength of security	Each node is sole owner of its private key	Secret key is shared over the network, hence can be stolen.
Security attacks prevented	Authentication, Confidentiality, Integrity, Nonrepudiation	Authentication, Confidentiality, Integrity

7 Conclusion

Secured Fault Tolerant Mobile Computing has application in different important domains e.g. military unit, financial and banking sector, disaster management, any monitoring and control unit, highly confidential collaborative research work distributed over the countries or over the regions within a country etc. These applications require secured uninterrupted communication and data transmission. Here checkpointing is done in a secured manner so that security attacks in turn can not increase failure rate of computing nodes. Combining cryptography with checkpointing ensures secure computation and communication, reduces failure rate of mobile computing system over a period time but all of these come at a cost of additional overhead. Challenge is to reduce cryptography cost as much as possible. The proposed work shows that without adding much cost overhead secured checkpointing is achievable.

References

1. George, S.E., Chen, I.R., Jin, Y.: Movement-Based Checkpointing and Logging for Recovery in Mobile Computing Systems. In: *MobiDE*, pp. 51–58 (2006)
2. Park, T., Woo, N., Yeom, H.Y.: An Efficient recovery scheme for fault-tolerant mobile computing systems. *Future Generation Computer System* 19(1), 37–53 (2003)
3. Men, C., Xu, Z., Wang, D.: An Efficient Handoff Strategy for Mobile Computing Checkpoint System. In: Kuo, T.-W., Sha, E., Guo, M., Yang, L.T., Shao, Z. (eds.) *EUC 2007*. LNCS, vol. 4808, pp. 410–421. Springer, Heidelberg (2007)
4. Ciciani, Q.F., Baldoni, B., R.: Checkpointing Protocols in Distributed Systems with Mobile Hosts: a Performance analysis. In: *Workshop on Fault-Tolerant Parallel and Distributed Systems*, pp. 742–755 (2006)
5. Agbaria, A., Sanders, W.H.: Distributed Snapshots for Mobile Computing systems. In: *Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications (Percom 2004)*, pp. 1–10 (2004)
6. Jiang, T.Y., Li, O.H.: An Efficient Recovery Scheme for Mobile Computing System. In: *International Conference on Machine Learning and Cybernetics*, vol. 4, pp. 2031–2036 (2003)
7. Cao, G., Singhal, M.: Mutable Checkpoints: A New Checkpointing Approach for Mobile Computing Systems. *IEEE Transactions on Parallel and Distributed System* 12(2), 157–172 (2001); ISSN: 1045-9219
8. Gura, N., Patel, A., Wander, A., Eberle, H., Shantz, S.C.: Comparing Elliptic Curve Cryptography and RSA on 8-bit CPUs. In: Joye, M., Quisquater, J.-J. (eds.) *CHES 2004*. LNCS, vol. 3156, pp. 119–132. Springer, Heidelberg (2004)
9. Wander, A.S., Gura, N., et al.: Energy Analysis of Public-Key Cryptography for Wireless Sensor Networks. In: *3rd IEEE International Conference on Pervasive Computing and Communications, Percom 2005* (2005)
10. Biswas, S., Neogy, S.: A mobility-based checkpointing protocol for mobile computing system. *IJCSIT Journal* 2(1), 135–151 (2010)
11. Yim, H.-b., Jung, J.-i.: IP traceback algorithm for doS/DDoS attack. In: Kim, Y.-T., Takano, M. (eds.) *APNOMS 2006*. LNCS, vol. 4238, pp. 558–561. Springer, Heidelberg (2006)

12. Elnozahy, E.N., Alvisi, L., Wang, Y., Johnson, D.B.: A Survey of Rollback Recovery Protocol in Message Passing System. *ACM Comput. Surv.* 34(3), 375–408 (2002)
13. Prakash, R., Singhal, M.: Low Cost Checkpointing and Failure Recovery in Mobile Computing Systems. *IEEE Transactions on Parallel and Distributed Systems* 7 (October 1996)
14. Piotrowski, K., Langendoerfer, P., Peter, S.: How public key cryptography influences wireless sensor node lifetime. In: *Proceedings of the fourth ACM Workshop on Security of ad hoc and Sensor Networks, SASN (2006)*; ISBN:1-59593-554-1

A Survey of Virtualization on Mobiles

Suneeta Chawla, Apurv Nigam, Pankaj Doke, and Sanjay Kimbahune

Tata Consultancy Services Limited, TCS Innovation Labs Mumbai, India
{suneeta.chawla, apurv.nigam, pankaj.doke,
sanjay.kimbahune}@tcs.com

Abstract. In this paper we cover the landscape of virtualization, with a focus on mobile phones. We present the evolution of virtualization from the 70's to date. We draw parallels between virtualization systems on mainframes, on commodity systems and mobile phones. One clear pattern that emerges is the business need of virtualization and the adoption of virtualization software eventually culminating in hardware support by the processor for virtualization. This last event also plays a key role in standardization and mass adoption of the virtualization software. The mobile virtualization area is quite dynamic and nascent today. However, given the rate of innovation, we expect standardization of mobile virtualization in the coming few years with novel applications not seen before. Some of the research prototypes which we deliberate upon are game changers for the computing industry. In this survey, we also draw attention to key challenges and research areas in mobile virtualization. This paper is an attempt to weave together a holistic picture of mobile virtualization research and industry, and summarize it.

Keywords: Virtualization, Hypervisor, Virtual Machine Monitor, VMM, Mobiles, ARM, KVM, Xen.

1 Introduction

Virtualization is the concept of creating virtual replicas of physical resources [34], [50]. These days, mobile phones have become more popular and powerful. [33] With research and innovations in mobile computing, virtualization is becoming a feasible technology on mobiles also. To understand mobile virtualization we first provide an overview of the formal and informal definitions of the terms used in virtualization of machines. We refer primarily to the work of Robert P. Goldberg [1], [2], [3], [4] to provide these definitions and the theoretical underpinnings which have served as a strong foundation to this area, to date.

“A virtual machine is a very efficient simulated copy (or copies) of the bare host machine.” [2].

“A virtual machine monitor (VMM) is any control program that satisfies the three properties of efficiency, resource control, and equivalence. Then functionally, the environment which any program sees when running with a virtual machine monitor present is called a virtual machine. It is composed of the original real machine and the virtual machine monitor.” [2].

Informally, the three properties can be described as follows [49]:

Efficiency: All the non-sensitive instructions which may not harm the hardware can directly run on the hardware without any intervention from the VMM.

Resource Control: The VMM is responsible for allocation of resources. Whenever a VM needs any resource, a request is sent to the allocator of VMM.

Equivalence: A program executes in presence of VMM in the same way as it would do in absence of VMM.

The systems which satisfy above properties are said to achieve “classical virtualization”.

A VMM can also be defined as software which implements the function ‘f’ which maps virtual resources ‘v’ to real resources ‘r’ to implement a many-is-to-one relation [3]. In the literature, a VMM is also known as Hypervisor.

A kernel is ‘a collection of facilities of “universal applicability” and “absolute reliability”--a set of mechanisms from which an arbitrary set of operating system facilities and policies can be conveniently, flexibly, efficiently, and reliably constructed.’[5]

As more and more complex systems were being developed and new processors designed, the kernel started becoming large [47]. The challenges posed by the maintenance, verification and engineering of such a complex piece of software led researchers to think of an alternative approach. This alternate approach considered a very simple layer of software in lieu of the complex kernel. Such a piece of software based on certain basic primitives came to be known as a microkernel. Early versions of OS based on microkernel had some issues which were solved in the subsequent versions. Thus, a microkernel is a kernel based on the principle of minimality of code and concepts [6], [41], [42], [48].

During the run of this course the notion of a hypervisor (VMM) which would run on the bare metal and yet support multiple OS was also founded. One can see certain parallels between a microkernel and a hypervisor. While a microkernel supports the construct of entire OS divided across two modes of OS, the hypervisor acts as a common denominator for multiple OS.

These definitions provided us a foundation to understand virtualization. While we give great importance to the theoretical work in virtualization, we also cover significant ground in the engineering area to reflect upon the practical systems which are currently being built and used in the industry. Among the engineering systems, we investigate and present the prominent software in both open source as well as commercial systems [43].

2 Background

In this section we explain key virtualization concepts which will act as a consistent and rigorous framework to describe the body of work in mobile virtualization.

Goldberg classifies VMM as Type-I, Type-II and Hybrid VMM [1]. The VMM which runs on bare hardware and is responsible for system scheduling and resource allocation is known as Type-I VMM. There is another category of VMM which does

not run on bare hardware and requires presence of host OS is known as Type-II VMM. The host OS is responsible for resource allocations and Type-II VMM is responsible for creating an environment for multiple VM's to execute. QEMU is an example of Type-II hypervisor. It is a generic and open source machine emulator and virtualizer [7]. There is a third category of VM called Hybrid VM in which all privileged instructions are interpreted by VMM.

A substantial amount of work has been carried out on desktop virtualization. The most popular desktop architecture is x86. This architecture supports four distinct privilege levels known as CPU ring. These rings are arranged in hierarchical order of privileges ranging from ring 0 with highest privileges to ring 3 with least privileges. The ring 0 is closest to hardware and all the privileged instructions need to be executed in this ring. The advantage of having different privilege is that all the components do not have same rights to access the resources like memory, I/O and sensitive instructions. Traditionally, in non virtual environment OS kernel runs in ring 0, OS services in ring1 and ring2, applications in ring3. In processors with hardware virtualization support a new ring level was introduced called root ring. This ring resides below ring 0 and has highest privileges. [8]

Similarly the most widely used mobile architecture is ARM [35], [36]. It is based on RISC architecture. It has total 7 processor modes out of which 6 are privileged modes and 1 is unprivileged mode. FIQ, IRQ, Supervisor, Abort, Undef and System are privileged modes and User is unprivileged mode. The purpose of each mode is summarized in table below [9], [10].

Table 1. ARM Processor Modes

Mode	Purpose
Supervisor	For kernel/OS to operate
System	To process SWI by kernel
IRQ	For general interrupts
FIQ	For interrupts. Mode has its own set of registers.
Abort	Selected when bad memory locations are encountered.
Undef	Selected on occurrence of undefined exception / instruction
User	For application execution

There are a few challenges in virtualizing a processor. One of them is the execution of sensitive instructions, which is a responsibility of the VMM. Sensitive instructions are those which are control sensitive or behavior sensitive [2]. If a VM needs to execute a sensitive instruction it should trap to VMM.

In case of x86 architecture, there are 17 unprivileged sensitive instructions which cannot be trapped [46]. To compensate for this limitation either VMM should scan all instructions before execution so that a forceful trap to VMM is done or changes need to be made in the guest OS. With the introduction of hardware support in x86 processor, hardware extensions are available to trap all sensitive instructions [11].

Similarly, ARM too is not classically virtualizable, because it also has sensitive instructions which do not generate trap when executed in user mode [12], [44], [45]. To overcome this issue of non-virtualizable Instruction Set Architecture of ARM, Trust Zone technology has been introduced in ARM, which also provides a good security framework. Trust Zone creates two virtual processors out of a single physical

processor, one of which hosts a secure environment and the other non-secure environment. Monitor mode is used for context switching between the two environments. This new release has features for interrupt and emulation support.

Due to architectural limitations, these processors do not support classical virtualization [13]. To compensate these limitations, three types of virtualization techniques have been introduced.

Full Virtualization: Full virtualization is a virtualization technique in which there is complete simulation of hardware of the host machine [14], [15]. The virtual machines that are formed as a result of full virtualization reflect every minute feature of the underlying hardware. Any operating system that is intended to run on bare metal hardware must be able to run unmodified on the virtual machines (Figure1). Fully virtualized virtual machines have all the instruction sets, interrupts, device access and memory access of the underlying hardware. VMware uses binary translation to virtualize the non virtualizable sensitive instructions [51].

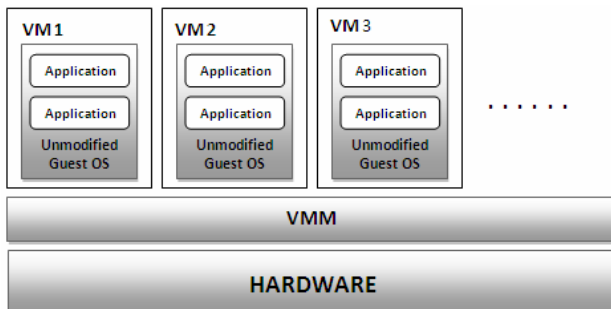


Fig. 1. Full Virtualization

Para Virtualization: Paravirtualization refers to a technique to implement virtualization by making certain changes in the kernel of guest operating system (Figure 2). The kernel changes involve replacing the non virtualizable instruction set with hypercalls that communicate directly with the hypervisor. The performance speed of paravirtualized guests is more than fully virtualized guests due to hypercall interface which is available to them [14], [15].

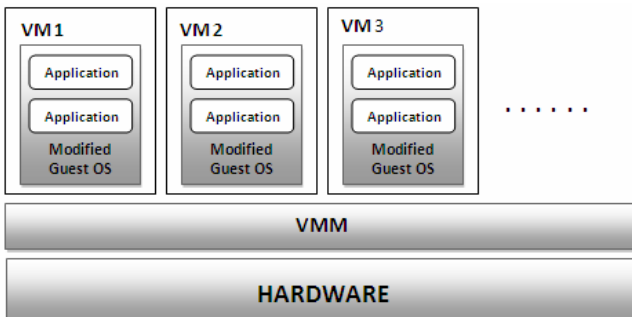


Fig. 2. Paravirtualization

Hardware Assisted Virtualization. It is a virtualization technique which requires a processor with hardware support for virtualization [15]. It is also known as pure virtualization as the VMM is able to trap all the sensitive instructions and the guest OS can run unmodified. In 2005/2006 Intel and AMD both introduced hardware support to their processors. In case of desktops, hypervisor runs in the root mode below ring 0 and the guest OS runs in the ring 0. In 2010, hardware virtualization was also introduced to ARM processors. This virtualization can cause overheads as each sensitive instruction needs to be emulated.

3 The Evolution of Virtualization

Virtualization has evolved over decades from its conception with Mainframes [1]. With the advancement of desktop processors in desktop, virtualization was introduced to desktops in the 80's and 90's. Today progress is happening in the direction of embedded systems and mobile phones. There is a vast potential of mobile virtualization to revolutionize the way we use mobile phone. All the developments which we saw in case of mainframes and desktops are making their way into mobiles.

We see the work in porting of Xen [16], [17], [18], [19], [20] and KVM [21], [22] which are main open source solutions in desktop virtualization, to mobiles. Also companies like VMware which had revolutionized desktop virtualization have devised a mobile virtualization platform.

We can notice mobiles virtualization following the same roadmap as it has been for desktop. The benchmark developments in the field of mainframe, desktop and mobile virtualization are summarized in table below.

Table 2. Chronology of virtualization developments

	Mainframe	Desktop	Mobile
1960's:	IBM Watson Research Lab worked on the architecture of VMs.		
1970's:	Comprehensive formalization of virtualization by Goldberg		
80's till late 90's		Personal computers were becoming popular	
1990's		Stanford University's research program in SimOS and DISCO resulted in the founding of VMware (1998)	
1997		Virtual PC released by Connectix in June	
2001		Xen 0.x released	

Table 2. (continued)

2004	KVM Qumranet founded	
2005- 2006	Processors with hardware virtualization support -Intel 13 November 2005 -AMD 23 May 2006	
2006 November 2008		Xen on ARM 2006[17] VMware launched Mobile Virtualization Platform(MVP) KVM on ARM [21],[22]
July 2010 September 2010		Hardware virtualization introduced to mobile processors. -ARM Cortex A15 September 2010 -Intel atom processor Q3 2010

4 Virtualization on Mobiles

The past few years have seen a prolific rise in the area of virtualization on mobile, right from more powerful processors, to the development of sophisticated and innovative research in virtualization software. This section highlights some of the key advancement, namely: Approaches for running two OSs in parallel on a single mobile phone, Virtualization on a desk phone, KVM on ARM, a secure virtualization processor architecture called VIRTUS, cloud computing for mobile, virtualized interfacing with I/O devices and storage of data in virtual containers.

4.1 Running Two OS in Parallel on Same Mobile

Brakensiek et al (Nokia Research Center) [23] conducted experiments to demonstrate Linux and Symbian OS executing in parallel. They have chosen target hardware as ARM MPCore prototype board. They have used L4 Fiasco microkernel together with L4 server functions as the virtualization layer. The result of the experiment is that both Linux and Symbian kernel tasks runs as individual L4 task. They have concluded that paravirtualization is feasible.

OK Labs has successfully ported a Microvisor to the most common embedded CPUs and systems based on TI OMAP 3530. This microvisor, no more than a few tens of kilobytes, is able to load and execute Android and Linphone in a paravirtualized manner [24], [25], [39], [40].

Seehwan Yoo et al. [26] have implemented a VMM (hypervisor) called MobiVMM for mobile phones. They have used OMAP 2430 and paravirtualized Linux 2.6.21. MobiVMM consumes very less memory. To make it more efficient it has been designed in such a way that it will automatically terminate virtual machines which are not running for a time. User is also interrupted to disable certain service in

case battery level reaches below a threshold. Currently they have demonstrated running two Linux OS in parallel. Their work is in progress and they plan to port other OSs like Windows mobile on MobiVMM, the Xen Hypervisor for mobile.

4.2 Phone Virtualization

Avaya Labs [27] have introduced the concept of phone virtualization. In addition to providing basic phone functionality multiple guest OS can run on desk phone. Many organizations provide smart phones to workers which are connected to enterprise applications. They have come up with an idea to have virtualization on desk phone so that desk and smart phone have same environment. Advanced functionalities like VOIP, conferencing etc can be made available to desk phones. By doing this telephony infrastructure cost is reduced.

The data between smart phone and desk phones can be synced. If multiple OSs resides on desk phone then on an incoming call, caller number is compared with address book entries for each virtualized phone. Using this information appropriate smartphone is booted dynamically.

Virtualization also provides an isolation layer to protect the phone functionality because of bugs in third party application. They have conducted experiment to virtualize desk phones using OKL4 (microvisor by OKLabs). They have used paravirtualized android kernel on ARM 1176 processor and tested on QEMU emulator.

4.3 Mobile Virtualization Using Xen

Xen Hypervisor [16] is an open source paravirtualization technology originally based on x86 architecture and has been ported to the ARM architecture [17], [18], [19], [20], [38] used for mobile phones.

x86 architecture supports four distinct CPU rings. Xen runs at ring0, Guest OS kernels run at ring1, and applications run at ring3, respectively. Hypercalls are used to execute sensitive instructions by unprivileged processes.

Majority of the smart phones use ARM architecture. If XEN hypervisor runs in one of the 6 privileged modes, then Guest OSs and their applications have to share the single available unprivileged mode. Thus, techniques are required to protect Guest OS and user applications running in the same mode. To achieve this Sangwon Seo [19] has implemented two logical modes - kernel mode and user mode on one physical mode. This enables the Guest OSs and user applications to run in two separate logical modes of different privileges. This way, OS level virtualization is achieved.

Super and Smith [19] ported Xen on a netbook to exhibit the usefulness of virtualizing the wireless network connection. They performed the experiment on an Intel Atom processor netbook whose computing environment is comparable to modern smartphones'.

LeMay et al. [20] have attempted to port the Xen hypervisor, as well as the associated Linux guest operating systems, to the ARM architecture. They have used CodeSourcery ARM cross-compiler with support for the ARM EABI. They have devised a mechanism for running a Linux kernel to run on QEMU. Their work on porting Xen is in progress.

One of leading smartphone company Samsung is working on Xen ARM Project. The project is led by Sangbum Suh from Samsung. They are working on porting Xen hypervisor to the ARM processor.

4.4 Mobile Virtualization Using KVM

The work at Columbia University presents how KVM can be ported on ARM to run unmodified OS [20], [21]. KVM requires hardware support for virtualization. The solution is capable of running even the closed Oss like Symbian as guest OS.

4.5 Virtualization - A Warrior for Security in Mobile Phones

Data privacy and security of content is more important on mobile since it is a more personal device than a personal computer. Virtualization provides a solution to address mobile security challenges [23], [37].

Inoue et al. [28] have proposed a processor virtualization architecture called VIRTUS. It provides a dedicated domain for pre-installed applications and virtualized domains for downloaded native applications.

A domain is an environment in which a group of applications could be executed isolated from other applications. When the domains are used for execution of applications, then they enable to prevent the interference of other applications in their own address space. The amount of resources that is available to a particular domain is also fixed, so there is control over the amount of resources an application running in a domain could utilize. So the protection of native applications that are crucial for the operation of mobile phone, from downloaded applications is done by running both categories of applications in different domains where they can't interfere with each other.

VIRTUS requires the modification of OS kernel to include the code of new kernel modules and interrupt handlers to perform new functionalities. VIRTUS components master, slave and processor separation logic are responsible for performing VMM functions of domain scheduling, domain setting and domain separation.

4.6 Cloud Computing for Mobile Devices

Cloud computing has been used in context of web, but the same can be the future of mobile phones. Cloud computing provides solution to the problems faced by mobiles phones like resource constraints. Cloud computing can provide infrastructure where processing and storage of data takes place external to mobile device Ex, mobile Gmail, Google maps. Huerta-Canepa and Lee [29] have suggested a way to provide external resource via virtual cloud. They have created a virtual framework where resources are shared from the near by mobile devices.

Security can also be taken care by using cloud computing concept. Running an anti virus application on mobile takes a toll on its resources, so Oberheide et al. [30] has suggested a way where the application runs outside mobile phone i.e. in the cloud. Their architecture involves a host agent which runs on mobile phone and sends files to network, and network service which has multiple detection engine scans the file and sends back the result.

4.7 Chameleon: A Capability Adaptation System for Interface Virtualization

Sang-bum Suh et al. [31] have introduced a system, called Chameleon, which enables the coordination between mobile phone and I/O devices in changing environment and dynamically install capability adaptors. If a mobile user wants to browse multimedia gallery of videos stored in mobile then he can use home TV. When he moves to some other location like office he can use monitor of computer and continue watching the clipping without remembering the point from where he needs to resume. In such a case, the responsibility to coordinate between the different environments and save the states lies with his mobile device.

In proposed system the mobile platform (guest) always moves with the user and there is stationary environment system (host) that can provide resources locally to the user. So as the new guest arrives at the stationary host, it has to transfer its state to the new host and continue its activities using new resources of the new host. Since the host which owns the resources knows them better so it is always good that the hosts adapt to new generic guests instead of guests adapting to the changing hosts.

Authors have chosen Xen as their base to implement capability framework. They have split device driver model of Xen to make the capability adaptation happen easily. The guest domains in XEN have front end (FE) device drivers and the host domain has the back end (BE) device drivers. When guests need to access the devices, then FE drivers which only know about the BE drivers contact them. BE drivers actually know the details of hardware devices with which they are interacting. In this way hardware requests of guests are met through BE drivers without the guest actually knowing the specifications of devices.

Capability adaptation is implemented by adding capability adaptors (CA) between the front end and back end drivers. FE issues requests to CA instead of BE directly so that CA can do adaptations on the requests based on its information about the FE and BE. This system allows dynamic selection of capability adaptors at run-time and makes the system more flexible.

4.8 VStore - Efficiently Storing Virtualized State across Mobile Devices

Georgia Institute of Technology, Atlanta and Motorola Labs [32] have devised a mechanism for storage management and content protection for mobiles. Data is kept in virtual containers. These containers are distinct from the containers containing the guest operating system and applications. Virtual containers have embedded information about privileges for accessing and sharing content. The advantage of storing such information in isolated containers is to guarantee data privacy until user himself releases data and not causing any performance overhead on client applications. Using VStore, the end user gets the notion that all the content is present at a central location in spite of their actual locations which may be on LAN, networked servers or networked peers.

5 Conclusion and Future Work

Analysis of the research domain of virtualization on mobiles, as can be seen above indicates a rapid pace towards innovation. We foresee commercial systems built using

these technologies in couple of years. The challenges in this area would be in terms of power and security. Future work would see a convergence of virtual networks, high speed networks, distributed systems, non local VM etc. These merely reflect the oft repeated phrase that “network is the computer” and mobile serves as digital touch point. Future work would encompass about VM migration across devices and clouds.

Acknowledgments. We thank Arun Pande, Ananth Krishnan, Sachin Adawadkar and the Development team at TCS Innovations Lab, for their invaluable support.

References

1. Goldberg, R.P.: Architectural Principles for Virtual Computer Systems. Harvard University, Harvard
2. Popek, G.J., Goldberg, R.P.: Formal Requirements for Virtualizable Third Generation Architectures. Magazine, Communications of the ACM 17 (1974)
3. Goldberg, R.P.: Architecture of Virtual Machines. In: AFIPS National Computer Conference, New York (1973)
4. Goldberg, R.P.: Survey of Virtual Machine Research: Honeywell Information Systems and Harvard University
5. Wulf, W., Cohen, E., Corwin, W., Jones, A., Levin, R., Pierson, C., Pollack, F.: HYDRA: The Kernel of a Multiprocessor Operating System. Magazine, Communications of the ACM (June 1974)
6. Tsyban, A.: Formal Verification of a Framework for Microkernel Programmers
7. QEMU, Open Source Process Emulator, http://wiki.qemu.org/Main_Page
8. Intel® 64 and IA-32 Architectures Software Developer’s Manual
9. Ville, P.: ARM Architecture (2002)
10. ARM Architecture overview, http://wiki.osdev.org/ARM_Overview
11. Robin, J.S., Irvine, C.E.: Analysis of the Intel Pentium’s Ability to Support a Secure Virtual Machine Monitor. In: SSYM 2000 Proceedings of the 9th Conference on USENIX Security Symposium (2000)
12. Varanasi, P.: Implementing Hardware-supported Virtualization in OKL4 on ARM: Thesis submitted for B.Sc Honours, Computer Science at The University of New South Wales, School of Computer Science and Engineering (November 2010)
13. Adams, K., Agesen, O.: A Comparison of Software and Hardware Techniques for x86 Virtualization. In: ASPLOS 2006, San Jose, California, USA, October 21-25 (2006)
14. Reames, P., Chan, E., David, F., Carlyle, F., Campbell, R.: A Hypervisor for Embedded Computing. Illinois Journal of Undergraduate Research 2 (2007)
15. Understanding Full Virtualization, Paravirtualization and Hardware Assist: VMWare Inc (2007)
16. Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I., Warfield, A.: Xen and the Art of Virtualization. In: SOSP 2003, October 19-22, University of Cambridge Computer Laboratory, Cambridge (2003)
17. Ferstay, D.R.: Fast Secure Virtualization for the ARM Platform: The University of British Columbia (March 2006)
18. Seo, S.: Research on System Virtualization using Xen Hypervisor for ARM based secure mobile phones. In: Seminar Security in Telecommunications, Berlin University of Technology, Korea Advanced Institute of Science and Technology (2010)

19. Super, K., Smith, J.M.: XenITH - Xe. In: *The Hand*. Technical Report, Department of Computer & Information Science (2010)
20. LeMay, M., Jin, D., Reddy, S., Schoudel, B.: Porting the Xen Hypervisor to ARM
21. Nilsson, A., Dall, C., Albert, D.: *Android Virtualization*. Columbia University (2009)
22. Dall, C., Neih, J.: *KVM for ARM*. Columbia University
23. Brakensiek, J., Droge, A., Botteck, M., Hartig, H., Lackorzynski, A.: Virtualization as an Enabler for Security in Mobile Devices. In: *Proceedings of the 1st workshop on Isolation and Integration in Embedded Systems IES*. ACM, New York (2008)
24. Heiser, G.: *Virtualization for Embedded Systems*. Technology White Paper, Open Kernel Labs Inc. (2007)
25. McCammon, R.: *Streamlining Android Migration with Virtualization*: Open Kernel Labs
26. Yoo, S., Liu, Y., Hong, C.H., Yoo, C., Zhang, Y.: *MobiVMM, A Virtual Machine Monitor for Mobile Phones*. In: *Proceeding MobiVirt Proceedings of the First Workshop on Virtualization in Mobile Computing*. ACM, New York (2008)
27. Acharya, A., Buford, J., Krishnaswamy, V.: *Phone Virtualization Using a Microkernel Hypervisor*. In: *Internet Multimedia Services Architecture and Applications (IMSAA)*, pp. 1–6. IEEE, Los Alamitos (2009)
28. Inoue, H., Sakai, J., Eda, M.: Processor virtualization for secure mobile terminals. In: *ACM Transactions on Design Automation of Electronic Systems*, vol. 13(3). ACM, New York (2008)
29. Canepa, G.K., Lee, D.: *A Virtual Cloud Computing Provider for Mobile Devices*. ACM Press, New York (2010)
30. Oberheide, J., Veeraraghavan, K., Cooke, E., Flinn, J., Jahanian, F.: *Virtualized In-Cloud Security Services for Mobile Devices*. In: *MobiVirt Proceedings of the First Workshop on Virtualization in Mobile Computing*. ACM, New York (2008)
31. Suh, S., Song, X., Kumar, J., Mohapatra, D., Ramachandran, U., Yoo, J.H., Park, I.: *Chameleon: A Capability Adaptation System for Interface Virtualization*. SW laboratories, Samsung Electronics, Korea and Georgia Institute of Technology, Atlanta
32. Seshasayee, B., Narasimhan, N., Bijlani, A., Pai, A.: *VStore - Efficiently Storing Virtualized State Across Mobile Devices*. In: *MobiVirt Proceedings of the First Workshop on Virtualization in Mobile Computing*. ACM, New York (2008)
33. Want, R.: *When Cell Phones Become Computers*. IEEE CS, Los Alamitos (2009)
34. Buzen, J.P., Gagliardi, U.O.: *The evolution of virtual machine architecture: Honeywell Information Systems*. Harvard University, Cambridge
35. Fornaeus, J.: *Device Hypervisors*. ACM Press, New York (2010)
36. Bhardwaj, R., Reames, P., Greenspan, R., Nori, V.S., Ucan E.: *A Choices Hypervisor on the ARM architecture*
37. Cox, L.P., Chen, P.M.: *Pocket Hypervisors: Opportunities and Challenges*. In: *Mobile Computing Systems and Applications*, pp. 46–50. IEEE, Los Alamitos (2007)
38. Armand, F., Gien, M., Maigne, G., Mardinian, G.: *Shared Device Driver Model for Virtualized Mobile Handsets*. In: *MobiVirt, Proceedings of the First Workshop on Virtualization in Mobile Computing*. ACM, New York (2008)
39. Heiser, G.: *The Role of Virtualization in Embedded Systems*. In: *Proceeding IES Proceedings of the 1st Workshop on Isolation and Integration in Embedded Systems*. ACM, New York (2008)
40. Liedtke, J.: *Toward Real Microkernels*. Magazine, *Communications of the ACM* 39(9) (September 1996)

41. Black, D.L., Golub, D.B., Julin, D.P., Rashid, R.F., Draves, R.P., Dean, R.W., Forin, A., Barrera, J., Tokuda, H., Malan, G., Bohman, D.: Microkernel Operating System Architecture and Mach. *Journal of Information Processing* 14(4) (March 1992)
42. Engler, D.R., Kaashoek, M.F., O'Toole Jr, J.: Exokernel: An Operating System Architecture for Application-Level Resource Management. In: *SOSP 1995 Proceedings of the fifteenth ACM Symposium on Operating Systems Principles* (1995)
43. Rosenblum, M., Garfinkel, T.: *Virtual Machine Monitors - Current Technology and Future Trends*. IEEE Computer Society, Los Alamitos (2005)
44. Liebergeld, S.: *Efficient Virtualization on ARM Platforms*: Technische Universitat Dresden (2009)
45. Schaik, C.V., Heiser, G.: High-Performance Microkernels and Virtualisation on ARM and Segmented Architectures: Open Kernel Labs and National ICT Australia
46. Adams, K., Agesen, O.: A comparison of software and hardware techniques for x86 virtualization. In: *ASPLOS-XII Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM, New York (2006)
47. Hansen, P.B.: The Nucleus of a Multiprogramming System. *Magazine, Communications of the ACM* (April 1970)
48. Liedtke, J.: On micro-Kernel Construction. In: *SOSP 1995 Proceedings of the Fifteenth ACM Symposium on Operating Systems Principles* (1995)
49. Rose, R.: *Survey of System Virtualization Techniques* (March 8, 2004)
50. Seawright, L.H., MacKinnon, R.A.: A Study of Multiplicity and Usefulness. *IBM Systems Journal archive* 18(1) (March 1979)
51. Sugerman, J., Venkitachalam, G., Lim, B.H.: Virtualizing I/O Devices on VMware Workstation's Hosted Virtual Machine Monitor. In: *Proceedings of the USENIX Annual Technical Conference* (2001)

Mobile Peer to Peer Spontaneous and Real-Time Social Networking

Abhishek Varshney¹ and Mohammed Abdul Qadeer²

¹ Department of Computer Science and Information Systems,
Birla Institute of Technology and Science,
Pilani 333031, Rajasthan, India

² Department of Computer Engineering,
Zakir Husain College of Engineering and Technology,
Aligarh Muslim University, Aligarh 202002, India
abhishekvrshny@gmail.com, maqadeer@gmail.com

Abstract. In this paper, we try to present a social networking model based on a hybrid combination of mp2p and traditional client-server technologies. The handheld devices used today are much better in performance in terms of processing power and memory than many of the web servers of early times. These handheld devices are also equipped with so many multimedia features that they have become the primary device for people to capture day-to-day occurrences. Moreover, a handheld device is always carried by a person with him wherever he goes. We would try to develop a hybrid model of client-server and peer to peer technologies which would provide spontaneous and instantaneous updates to the peers in the trusted group without actually “pushing” of updates by the user, at the same time, overcoming the limitations of traditional client-server approach of the social networking websites of today.

Keywords: mobile peer to peer network, mp2p, real-time sharing, spontaneous social networking, ubiquitous networking.

1 Introduction

Social Networking has greatly influenced the way people communicate over the internet. With the emergence of social networking websites such as Facebook, Myspace and Orkut, people continue to spend more and more of their internet time on social networking. According to a report from the Nielsen Company, in August 2009, 17 percent of all time spent on the Internet was at social networking sites, up from 6 percent in August 2008 [1]. On the other hand, people tend to carry most of their multimedia content like pictures, videos, sound clips etc. in small mobile devices like smart phones and PDAs [2]. The increase in functionality of today’s handheld devices with the advantage of the mobility can make a handheld, a significant resource centre with an overall change in the way people communicate with each other.

The fact that the people always carry their handhelds with them has inspired us to explore the scope and implications of next generation mp2p social networking which

would alleviate the dependence on conventional server based community system for communication [3]. The popularity of Web 2.0 server based community can be accounted to the fact that users continue to share their daily experiences and information with their friends and wider audience in a great manner. The daily experiences that users tend to share can be in the form of photos, videos, blogs or just status updates. Almost all the social networking services are based on centralized servers where the information and content needs to be updated continuously by the user himself for others to see and access [4] [5]. Server-based services require central service provider and system which users, both content sharers and watchers, can then access. Interoperability between different services is almost non-existent. The information shared in these types of services is not real time and if the real time feature is made available, it would increase the technical complexity of the service to a great extent. In mp2p social networking, the daily experiences and information is shared at the same time they occur. The handhelds have become excellent means of recording the daily experiences via location information, camera, notes, and other means.

There have been several works related to mp2p social networking in the past. The social networks in p2p infrastructure have already been proposed and applied to personal computers by Sonja Buchegger et al [6]. Mehdi Mani et al [2] developed an application for mp2p social networking in Java. We would try to take forward the concept of mp2p social networking by connecting peers and sharing information through http access, thus, making the experience similar to that of hosting and accessing a social networking website from a remote server. The paper makes an attempt to extend the domain of peer to peer social networks to mobile and handhelds. With the development of PAMP (Personal Apache MySQL PHP) for Symbian OS based handhelds by Nokia [7] and httpd4mobile [8], it is not just possible but also feasible to run http server on handheld devices which, in turn, makes it possible to host personal websites on mobile devices.

2 Limitations with Server Based Community Services

The following limitations of the conventional server based community services have inspired us to propose a mp2p social networking model:

2.1 Distributiveness of Services and Service Providers

There does not exist any service that connects the people and helps them communicate. For example, there are various services like Facebook, Orkut, Twitter etc. Each of them has a unique feature that distinguishes them from the rest. For instance, one of the features provided by Facebook is a mini-feed where all the updates from other users appear. Apart from it there is a feature called Wall, which can be thought to be a guest book for the friends who visit the user's profile. On the other hand, Twitter is a micro-blogging service that enables its users to send and read messages known as tweets.

It therefore happens many a times that the user has to subscribe to more than one of the services in order to enjoy the features provided by one of the service but not by the other.

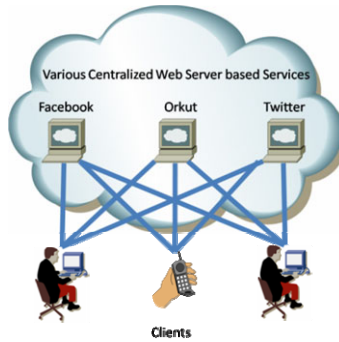


Fig. 1. Distributiveness of server based Social networking services

2.2 Security and Privacy

Security and privacy are topics of concern in centralized server based services. The users of these services post quite a lot of their personal data on these services, like, Facebook, Orkut, or StudiVZ. These services have full control of the users' data. As a result, the users are very much prone to get advertisements in their mailboxes. There can be instances where the personal contact details and other private data can be sold to third parties. The third party companies may use the data either for sending advertisements to the users or for study purposes.

Apart from it there have been a few cases in which the users' privacy has been compromised. One of such cases was Facebook Beacon. With Facebook Beacon, it was possible to track down the buying habits of a user on various online shopping websites like Amazon. Facebook Beacon would then update the user's mini feed with the collected data. The purchase details of the user were thus visible to all his friends. This is definitely a breach of user's privacy.

In spite of these serious security issues, people continue to use these social networking websites because they have, in a way, become addicted to this type of social life which has become a substitute for traditional mailing lists, discussion forums and even real life communication. The online social networks of today have become the most convenient way to discuss topics, spread information about events, etc.

Even, there are now discussions on the ownership rights of the data and information posted by the user. It is a matter of debate that whether the data posted by the user belongs to the user or to the social networking website, like Facebook, MySpace or twitter [9]. Thus, security and privacy turn out to be major issues with server based community services.

2.3 Non Spontaneity and Manually "Pushing" the Updates

There does not exist any mechanism at all in the conventional server based community services that can fetch the data automatically from the user's end and make it accessible to the trusted group. Therefore, it requires the cumbersome task from the user's end to continuously update his profiles on the different social networking sites he has subscribed to.

2.4 The Rigidity of the Services Provided

The user remains bounded to the layout of the services provided by the social networking website. He can just change the color or theme of the layout but not actually the complete layout in which the user wishes to see the information. For example, a user may want only the text updates from his friends to appear for him while another may give priority to the latest photos clicked by his friends. There may be complex cases too. For instance, a user may prefer to see the location update from his children while the updates of meetings and engagements from his boss. These types of features are not yet provided by any of the existing services.



Fig. 2. Non flexibility in the layout of information of existing server based social networking services

2.5 Necessity of an Internet Connection to Access the Services

In a centralized server based community service, internet connectivity is required for all transactions [10]. A user cannot interact with peers in a Local Area Network or in his vicinity without an uninterrupted internet connection. For example, while going to a multiplex cinema, a user would like to ask about the reviews of a particular movie through his handheld from those who are coming out of the previous show and are in the Wifi range of his handheld. Thus, an active internet connection is a vital requirement to connect to an online social network.

3 Working of Mobile Peer to Peer Social Network

In mp2p social networking, handhelds act as both, server and client, to interact with other peers. With the development of PAMP (Personal Apache MySQL PHP) for Symbian OS based handhelds by Nokia [7], it is now possible to even host personal websites on a handheld with PHP or Python as the scripting language and using MySQL as the Database Management System. The Python modules could access the internal features of the phone like camera, SMS inbox, Call Logs, GPS data, Calendar, Contacts, etc and make it available over the network to the users connected and authorized to receive and interact with the data. A similar type of application based on Java also exists, known as httpd4mobile [8].

It is therefore possible to allow http access to a handheld using any of the above mentioned server technologies available for handheld devices. A user can install 3rd party frontend in his mobile or handheld and view the updates from his trusted group in the fashion he desires and even restrict or allow the data from his own handheld that he wants to share. He no longer remains bounded to the services, features and the layout provided by the service provider (Facebook, Orkut etc, in this case). Thus, the data from his handheld is shared in the way he desires.

The working of peer to peer social networking has been well defined by Sonja Buchegger et al [10]. Peer to peer communication is established by a two tiered architecture. The peers that wish to connect with each other form one tier. The other tier comprises of the Look up service that contains the data of the peers like their IP addresses, etc which is used to find peers. The way the first connection is established between peers is known as bootstrapping problem. There can be two approaches to solve this problem. In the first approach, the peer that wish to connect looks for the IP addresses of peers through a website hosted on a peer that is supposed to be stable and “always on”. In the second approach, caching may be used, in which the IP addresses of neighbors are stored and they are used for connection. But both of these approaches suffer serious drawback due to the dynamic IP systems. The IP addresses are dynamic and thus change periodically and so when the user tries to connect next time, he may not get the peers at the desired IP addresses. Sonja Buchegger et al [10] used the concept of Distributed Hash Tables (DHT). Distributed hash tables (DHTs) are a class of decentralized distributed systems that provide a lookup service similar to a hash table: (key, value) pairs are stored in the DHT, and any participating node can efficiently retrieve the value associated with a given key. Responsibility for maintaining the mapping from keys to values is distributed among the nodes, in such a way that a change in the set of participants causes a minimal amount of disruption. This allows DHTs to scale to extremely large numbers of nodes and to handle continual node arrivals, departures, and failures. The entire data of the users, friends, communities, posts etc is stored in the form of DHTs which correspond to URL of the peer where it can be found. The Dynamic DNS helps in resolving the URLs to corresponding IP addresses. DHT offers fast look up service but it is limited to only exact match searches. A more versatile method or any improvement in existing method of using DHT has to be used for solving bootstrapping problem more precisely.

Another issue in a peer to peer infrastructure can be the establishment of a reliable and efficient search engine. Since each hand held in the proposed model is a server, so every device can have its own search engine. Each search engine indexes the list of URLs in the DHT and generates the result. With the proposed model, it is possible to limit the search to the devices in any particular location, to active or non active devices or even to a single device, thus increasing the response time.

4 Features/Advantages/Implications of mp2p Social Networking

4.1 Spontaneity and Realtime Sharing of Data and Information

The handhelds and mobile phones are an integral part of one’s life today. A person always carries his handheld with him. Also, the multimedia features of a mobile phone

like camera, sound recorder etc become the first medium to capture any memorable or important event in daily life. The multimedia data like images, videos etc are instantly available for the friends to view as soon as they are captured. The calendar and the scheduler from the handheld can help a friend know about the current and future engagements, meetings etc of the user. It's just the permissions that the user needs to grant and associate it accordingly with different people in the trusted group. The location of the user from his GPS data can also be instantaneously viewed without the need of "Pushing" the information manually. But, at the same time, there can be limitations with the GPS system like, if the GPS is not active, it may take a few minutes to track the satellites or there can be instances when the request for location comes and the user is indoors where it is not possible to get the location. The PAMP can prove to be of great importance in this regard where the GPS data is continuously monitored and any update is quickly stored in the database. Any query regarding the location of the user can be handled by providing the latest data in the MySQL table to the request. This system also possesses some limitations like that of battery consumption and the probable exploitation of the location history stored in the database [7].

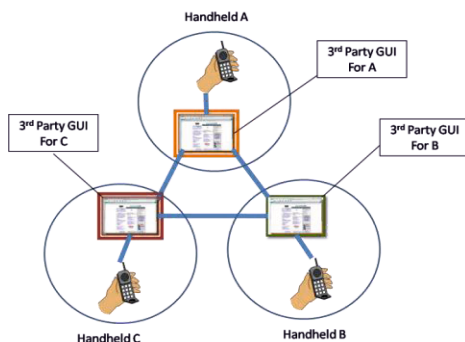


Fig. 3. The flexibility to decide one's own layout of the information in mp2p services

4.2 The Flexibility to Decide One's Own Layout of the Information

In the current scenario, the data is stored in the centralized servers and they are revealed to the users in the way the social networking website provides. But in peer to peer social networking, there is no centralized server based data storage; rather, the entire data is stored in a distributed manner in peers itself. The data is stored in one's own handheld device. It's just the front end that a user needs to display the information. Therefore, there will be a scenario in which the 3rd party layouts can be installed in the handheld devices (they can be in the form of web pages) and the user can decide what data to share and what not. At the same time, he can manage the priority and the format in which the information is visible to him.

4.3 Direct Exchange

In peer to peer social networking, the entire system is decentralized and is not based on a centralized server. The internet connection is not required to access the social

networking services. The users can exchange information and connect to each other even without an internet connection. The data is stored in the handheld devices and is available for the peers to share. Peer to peer social networks can even be established in Local Area Networks without internet connectivity making use of local devices like home access routers or ubiquitous computing devices. The handheld can connect to each other through adhoc networks over wifi and share information in a local environment without being actually connected to internet. This feature of mp2p social networking allow users to remain connected to other users at least to some extent rather than being disconnected completely in absence on an active internet connection.

4.4 Security

In peer to peer social networks, the data can be encrypted to ensure privacy. The data is available for access only to those peers to whom the user provides access. The user can set the permissions individually for each user so as to determine which user has access to what type of data. Also, since this service is decentralized, there is no control over the service terms and the users' data by any single entity. In addition, decentralization prevents the network services from being entirely inaccessible when the server is switched off. It also prevents the use of users' data for study purposes by various organizations to infer relationships or behavior [10].

4.5 Trust

The computational problem of trust is to determine how much one person in the network should trust another person to whom they are not connected and how they find friends in an online social network. Since, each hand held device can be associated with a unique mobile number or a unique IMEI number; therefore it can be additionally used for finding friends in a mobile peer to peer network with these numbers associated with individual handhelds in the Distributed Hash Table (DHT). Once a trusted group is formed, users can create communities in the same way as created on Facebook, Orkut etc. The users who join the communities get a list of all the members of the community stored in their database. Any new user, thus trying to find or join a community can easily search it through any of the existing members. Optimization techniques in the design of the database have to be applied so that the size of database on any individual handheld does not become too large. The experience could be personalized by displaying recommended results for each user in the same way as on existing social networks.

4.6 Location Based Services

A location-based service (LBS) is an information and entertainment service, accessible with mobile devices through the mobile network and utilizing the ability to make use of the geographical position of the mobile device. The location based services can start a new era of social networking. Here can be listed a few practical examples that may illustrate the significance of location based services:

- A friend who wants to know a user's current location will know it automatically without disturbing the user and without requiring the user to update his location.
- A person after boarding off at the airport can make a query to the handheld devices in his vicinity and search for a co-passenger who is headed towards the same direction or locality to hire a taxi mutually.
- A family planning to go to a picnic to a nearby beach can make a query to the handheld devices on the beach (connected through internet) to know about the weather and the sailing conditions.
- The government can broadcast a warning to all the handhelds on a sea shore about an upcoming storm.
- Location based reminders can also prove to be an essential service. Users can set reminders as when they reach a particular location, an action can be carried out. This action may vary from activating an alarm with a note for the user to sending a message/mail to some other remote device.

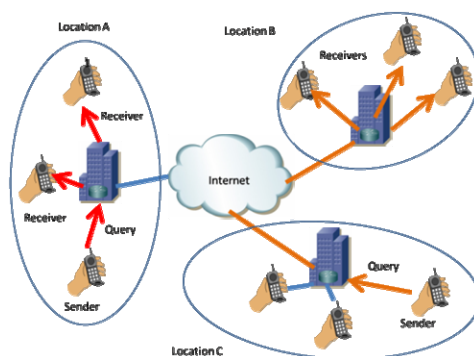


Fig. 4. Location based services in mobile peer to peer social networking services

5 Implementation of Mobile Peer to Peer Social Networking Service

We are currently working to make peer to peer mobile social network work flawlessly. We are trying to make use of PAMP (Personal Apache MySQL PHP) [7] for Symbian OS based handheld devices. It is possible to access the internal functions of the handheld, usually written in open C or open C++, through python modules. The python modules can then be integrated to work within a browser based application. With the Apache and MySQL for handheld devices, it is possible to achieve the same level of experience in browsing through mp2p social networking service as that with traditional web based social networking websites. At the same time, we are working with httpd4mobile to achieve the same for Java based mobile devices [8].

The network topology that we propose to implement is shown in figure 5. It uses handheld devices with PAMP running on each device. This topology uses client-server model at the micro level where, in addition to being a client, each handheld device is also a server in itself. For the look up service, we propose to use a

Distributed Hash Table and a Dynamic DNS (DDNS) service. Although, it may seem to be a contradiction to the theory of peer to peer networking, but it does not pose any security threat as data remains stored in handhelds only. Since, the IP of the device changes periodically, the DDNS keeps track of the changing IP addresses. Each handheld device is provided with a unique URL which, in a way, becomes the identity of the user and his device. The DHTs stores the data related to the users and the corresponding URL from where the data could be retrieved. The DDNS stores the IP addresses related to each URL so that when peers try to find and connect to a user, they get connected to it. This form of topology may be considered as a hybrid form of peer to peer and client server topologies. We are currently in a phase to make an impressive interface for our proposed network in PHP and incorporate the features of social networking websites like, messaging, finding friends, posting comments, making communities, etc.

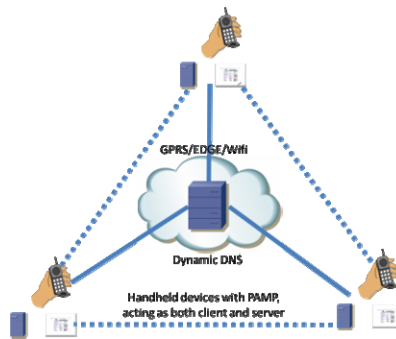


Fig. 5. The proposed network topology

6 Conclusion and Future Work

In the present work, we explored the benefits and the implications of mp2p social networking. The security of users' information, spontaneity of service, instant sharing and flexibility of service and direct exchange of information without requiring an active internet connection are some of the advantages that make it more powerful and effective than the conventional server based social networking services.

But there are still some glitches that need to be worked upon to make peer to peer social networking work flawlessly. Running PAMP on mobile devices faces some challenges like RAM memory and battery consumption. Also, the processing of PAMP may not be as good as that of current web servers. But, with the exponential rise in the technical advancement of mobile technology, this problem may soon be eliminated. Also, the lack of seamless handover between WLAN and cellular technologies is still a problem.

We, therefore, expect to bring a full-fledged mp2p social networking service in a short time.

References

1. Press Release of nielson Company,
[http://blog.nielsen.com/nielsenwire/
wp-content/uploads/2009/09/InternetSpend_SocialNetworks.pdf](http://blog.nielsen.com/nielsenwire/wp-content/uploads/2009/09/InternetSpend_SocialNetworks.pdf)
(retrieved on 16th November 2009)
2. Mani, M., Nguyen, A.-M., Crespi, N.: What's up 2.0: P2P Spontaneous Social Networking. In: Proceedings of the 2009 IEEE International Conference on Pervasive Computing and Communications (2009)
3. Muilu, P., Syrjänen, T.: Mobile Peer-to-Peer Media and Presence Community. In: IEEE CCNC 2008 Proceedings (2008)
4. Sarvas, R., Viikari, M., Pesonen, J., Nevanlinna, H.: MobShare: Controlled and Immediate Sharing of Mobile Images, Multimedia. ACM Press, New York (2004)
5. Futurice Media Sharing service, <http://www.kuvaboxi.fi>
6. Buchegger, S.: Delay-Tolerant Social Networking. In: Position paper for Extremecom 2009, Sweden August 14-18 (2009)
7. Wikman, J., Nurminen, J.K.: Open Source Web Application Development Stack for Symbian-based Mobile Phones. In: The Second International Conference on Next Generation Mobile Applications, Services, and Technologies, NGMAST (2008)
8. httpd4mobile, <http://www.tools4mobile.eu/httpd4mobile.html>
(retrieved on November 16, 2009)
9. Schiöberg, D.: A Peer-to-peer Infrastructure for Social Networks. Diplom Thesis, TU Berlin, Berlin, Germany(December 17, 2008)
10. Buchegger, S., Schiöberg, D., Vu, L.-H., Datta, A.: PeerSoN: P2P Social Networking — Early Experiences and Insights. In: Proceedings of the Second ACM EuroSys Workshop on Social Network Systems (2009)
11. Vu, L.H., Aberer, K., Buchegger, S., Datta, A.: "Enabling Secure Secret Sharing in Distributed Online Social Networks". In Proceedings of Annual Computer Security Applications Conference (ACSAC) 2009, Hawaii, December 7-11 (2009)
12. Buchegger, S., Datta, A.: A Case for P2P Infrastructure for Social Networks - Opportunities and Challenges. In: Proceedings of WONS 2009, The Sixth International Conference on Wireless On-demand Network Systems and Services, Snowbird, Utah, USA, February 2-4 (2009)
13. Xu, B., Wolfson, O.: Data management in mobile peer-to-peer networks. In: Ng, W.S., Ooi, B.-C., Ouksel, A.M., Sartori, C. (eds.) DBISP2P 2004. LNCS, vol. 3367, pp. 1–15. Springer, Heidelberg (2005)
14. Buchegger, S.: Delay-Tolerant Social Networking. In: 1st Extreme Workshop on Communication (ExtremeCom 2009), Laponia, Sweden (2009)
15. <http://www.opendht.org/> (retrieved on November 16, 2009)
16. <http://bamboo-dht.org/> (retrieved on November 16, 2009)
17. <http://wiki.opensource.nokia.com/projects/PAMP>
(retrieved on November 16, 2009)

Analysis of a Traffic Classification Scheme for QoS Provisioning over MANETs

Chhagan Lal, V. Laxmi, and M.S. Gaur

Department of Computer Engineering, Malaviya National Institute of Technology, Jaipur, India
{chhagan, vlaxmi, gaurms}@mnit.ac.in

Abstract. Differentiated services (DiffServ) is a networking framework. It classifies traffic using differentiated service code point (DSCP) field in packet header at network layer. Each traffic class is mapped with some priority queue that differs from other priority queues in terms of provided throughput and delay. In this paper, we modify the existing DiffServ model in Qualnet simulator so that it suits the dynamic nature of MANETs. We analyze the enhanced DiffServ architecture by configuring each node in such a way that it works as ingress as well as core node. The effectiveness of the traffic classification scheme in terms of network scalability and data rate is analyzed and compared against the best effort traffic method over MANETs for video, audio and plain data. Simulation results obtained for DiffServ enabled networks are compared with DiffServ disabled ones, evaluation metrics being delay and packet delivery ratio. Simulation result shows a reasonable improvement in all QoS metrics, if DiffServ is enabled in network.

1 Introduction

Mobile ad hoc networks (MANET) are collection of self configurable, self-organizable and self-maintainable mobile hosts that communicate with each other through wireless channel in multihop fashion with no centralized control. The infrastructure less, inexpensive and quick deployment of MANET finds its use among many applications such as VOIP, video conferencing, online gaming and other multimedia applications. Despite the large number of routing solutions available in MANETs, their practical implementation and use in real world is still limited. Providing required QoS guarantees in wireless multihop networks is more challenging than wire-line networks due to its dynamic topology, distributed nature, interference, multihop communication and contention for channel access. For multimedia and other delay or error sensitive applications that attract a mass number of users towards the use of MANETs, best effort routing protocols may not be adequate. As a result, the research focus has shifted from best-effort services to QoS adaptation.

In wired networks two models are used for providing required level of QoS to multimedia applications: Integrated services (IntServ) [1] and Differentiated services (DiffServ) [2]. IntServ is an architecture that provides fine-grained QoS to the applications having constrained delay or bandwidth requirements. IntServ has to be configured on each node from source to destination. An application flow is admitted when all the nodes from source to destination are capable for providing the required QoS. When sufficient resources are available for a flow admission, a resource reservation protocol

(RSVP) [3] is used to reserve resources at each intermediate node. RSVP uses out-band signaling for resource reservation. On the other hand, Diffserv is a coarse-grained QoS system that provides no hard QoS guarantees. Diffserv classifies traffic either based on service requirements, users or any other criteria. In Diffserv network, nodes check the DSCP value in each packet and provide different levels of service to packets via scheduling and priority queuing.

IntServ is not suitable for MANETs as RSVP uses out-band signaling that consumes the scarce bandwidth. Additionally the signaling overhead increases with an increase in mobility of nodes. This may lead discharge of battery at nodes and adversely impact their life. As the network scales in terms of traffic flows, the information needed to maintain the flow also increases (scalability problems). Diffserv is a light weight model that enhances the best effort services without making any reservations. It uses per hop behaviors (PHB) [4] at each node to specify the type of service for each incoming packet. Within PHB, there are two standard groups: Assured Forwarding (AF) and Expedited Forwarding (EF). Within AF, the group is further divided into four independent classes with three drop precedence levels. AF offers different levels of forwarding resources in each DiffServ node. In the case of network congestion, the relative importance of the packet determines its drop precedence among other packets in its AF class. EF, on the other hand, known as "Premium" service, gives the best service your network can offer. EF is defined as a forwarding treatment where the rate of packet flow from any DiffServ node is whatever rate ensures highest priority and no packet loss for in-profile traffic. Thus Diffserv architecture is a good candidate for a QoS model in MANETs.

The reminder of this paper is structured in following manner. In Section 2 we discuss related work carried out in this field. An overview of traffic classification model used in this paper is given in Section 3. Simulation and performance evaluation are presented in Section 4. Finally, Section 5 concludes the paper with directions for future work.

2 Related Work

When looking through the literature for the approaches towards QoS provisioning over MANETs, it can be summarized that most of the proposed solutions are inspired from two QoS models. Some uses an approach similar to IntServ i.e for each QoS required flow resources are reserved prior to the flow via call admission control (CAC) and flow states are maintained at each node along the path. Other method for QoS adaptation is derived from DiffServ i.e. traffic is divided into different priority classes.

In [5], an admission control scheme is used before using DiffServ model for QoS provisioning. The author makes DiffServe model simple and suitable for MANETs by classifying the traffic as QoS class and best-effort class. Before admitting a QoS flow the bandwidth and end-to-end delay for the path from source to destination is calculated. If it satisfies the application needs, the flow is admitted. The author does not mention the way in which available bandwidth at each node is calculated (an NP-hard problem) in MANETs. Also the presented solution works only for reactive routing protocols. The QoS traffic used has low packet size, i.e. 64 bytes and the effectiveness of proposed scheme is neither analyzed with respect to network scalability nor is it compared with the basic DiffServ model.

In [6], a node-disjoint multipath routing protocol is used for load balancing and fault tolerance while QoS support is provided using basic DiffServ model. Traffic load is routed over multiple paths towards destination in case of network congestion, notified by intermediate nodes through setting of congestion notification bit. The effectiveness of the proposed approach is not measured against multimedia traffic. Also the packet drop due to queue overflow and link break can be falsely treated as congestion.

In [7], authors present a novel approach for QoS sensitive traffic by combing the advantages of IntServ and DiffServ models. They couple DiffServ model with routing protocol to perform stateless and controlled reservation for QoS traffic. AODV forwarding table is replaced by QoS label switched forwarding and best effort forwarding table at each node. The effectiveness of the scheme is compared against basic diffServ model. The proposed scheme exposes some limitations such as low simulation time, network scalability is not considered, packet size is low as compared to one required for multimedia application packets and mobility speed is kept low during simulation.

Providing QoS providing over MANETs is an active research area, a large number of solutions have been proposed in literature. A survey on QoS solutions are presented in [8] [9]. In this paper we analyze and compare enhanced DiffServ model derived from basic Diffserv architecture that is suitable for MANETs against the traditional system used for data transmission.

3 Traffic Classification Model for QoS Provisioning

DiffServ is a computer networking architecture used to provide IP QoS, where QoS is a set of service requirements (e.g., bandwidth, link delay) to be met by the network during transmission of traffic. This architecture is flexible and allows for either end-to-end QoS or intra-domain QoS by implementing complex classification and mapping functions at the network boundary or access points. Within the DiffServ network, packet behavior is regulated by classification and mapping.

Traditional DiffServ model used in wireline networks is not suitable for MANETs. We incorporate several modifications in basic DiffServ model to make it suitable for MANETs. Unlike traditional DiffServ, we configure each node in such a way that it works as an ingress as well as egress node as per requirement. The reason for doing this is that in MANETs we do not have a way to select ingress and egress nodes due to high mobility and undefined boundaries. To decrease the overhead caused by IP output queue maintenance we use only three queues at each node to forward input traffic. This greatly reduces the processing speed resulting in low processing time in each queue and thereby, decreasing end-to-end delay.

QoS elements configured for traffic classification with mapping function at each node are shown in (refer Fig. 1). Nodes are configured with four basic components as shown in (refer Fig. 2).

Our enhanced Diffserv model is worked as follows. When a packet arrives at input queue of a node it is passed to a service selector. This function selects whether this packet is sent to classification services or to mapping function. Selection is based on the information present in DSCP field of packet header. If the DSCP field of incoming

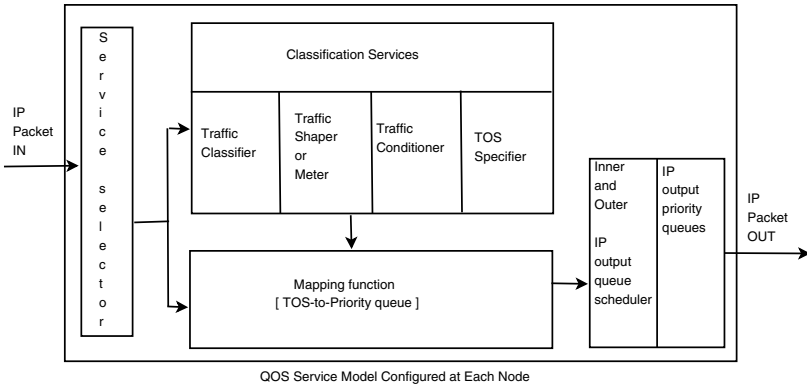


Fig. 1.

packet is set i.e. packet is already passed through some edge node and it is not the best effort traffic packet. The packet is sent to mapping function so that it can be mapped to its corresponding priority queue defined in PHB file. Otherwise, it is sent to classification function to check whether there is any service class defined for this packet in classification file. If no entry for that packet is specified in classification file then it belongs to best effort traffic. After that, the packet is sent to mapping function. Mapping function checks packets DSCP field to put it in the appropriate output queue. Due to service selector function each node is capable of dynamically changing their behavior and works either as ingress or egress as per requirement. We find that source node always work as ingress as well as egress node and intermediate nodes only forward the packets using PHBs (per hop behavior) so they work as core nodes. In case of best effort traffic, packets are mapped with default output queue.

We configure a classification file at each node consisting of four functions: (i)Traffic classifier, it classify the incoming traffic into different QoS classes, (ii)Traffic shaper, it defines the bandwidth and burstiness characteristics for the given class of traffic, (iii)Traffic conditioner, it defines the action taken when packets are determined to the Out-Profile, (iv)TOS Specifier, it defines the standard PHB services on a particular condition class. With the use of mapping function we map each incoming packet based on its DSCP value to corresponding output priority queue. This function is configured in PHB manner at each node to provide specified level of services to incoming packets. We use two schedulers named strict priority as inner scheduler and weighted round robin (WRR) as outer to schedule the packets in their mapped priority queues. To forward multimedia traffic we use RED queue and best effort traffic is forwarded using FIFO queue.

4 Simulation and Performance Evaluation

We have performed analysis by using scalable network simulator Qualnet version 5.0. Results obtained after simulation are used to show the effectiveness and limitations of

DiffServ model when used for QoS provisioning while transmitting multimedia traffic over MANETs. The result obtained using DiffServ are compared against the results obtained while using traditional system of routing.

4.1 Simulation Scenario and Parameters

We randomly select a source-destination pair in the network. All simulations are performed on different seed values and final results are generated using average values obtained for times using different randomization. Source node sends three type of traffic i.e video, audio and best-effort traffic simultaneously towards destination using Ad-hoc on-demand distance vector (AODV) [10] as underline routing protocol. Video/Audio traffic is characterized by their high packet size and low inter-packet time as compared to best-effort traffic. The parameters used during simulations are shown in Table 1.

Table 1. Simulation Parameters

Parameters	Values
Simulation time	300 sec
Scenario Dimension	600x600 to 1800x1800 meter
Number of nodes	25 to 55
Application protocol	CBR
Transport protocol	TCP,UDP
Routing protocol	AODV
QoS Specification	Yes
QoS model	DiffServ
Packet size	160 to 1760 bytes
Inter-packet time	20 to 50 ms
IP Output queues	Three (FIFO, RED)
MAC Specification	802.11
Mobility model	Random way-point

4.2 Simulation Results

In this section, we analyze the results obtained from various simulations. Evaluation results for DiffServ model and traditional routing model are compared in terms of various QoS metrics. Effects on end-to-end delay and packet delivery ratio are measured against increase in network size and network load for multimedia traffic and best-effort traffic.

1. **Effect of Network Scalability:** Effect of network scalability is determined for end-to-end delay and packet delivery ratio. Results obtained by simulation are shown in Figure 2 and 3. Traffic Characteristics is chosen in such a way to statistically encapsulate behavior of three kinds of traffic, i.e. video, audio and raw. As shown in

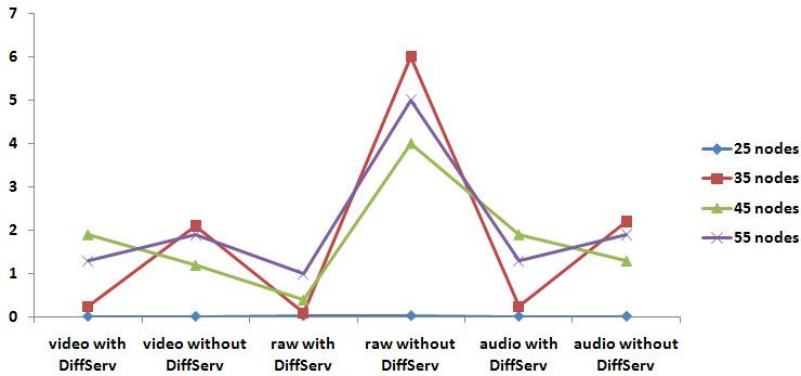


Fig. 2.

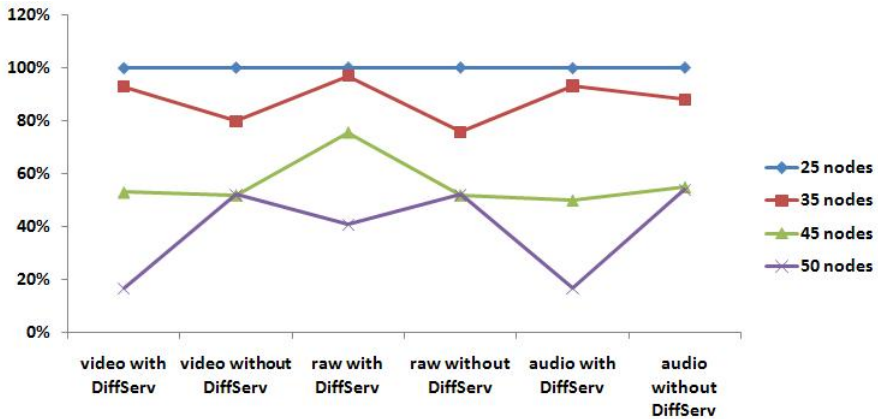


Fig. 3.

Figure 2 raw traffic end-to-end delay is greatly improved while using DiffServ. The video/audio traffic delay improves with Diffserv model but improvements are slow and smooth increase with network scaling as compared with traditional routing model.

As we can see in Figure 3, packet delivery ratio is case of video/audio traffic with DiffServ is good in small and moderate size networks. As the network size increase, due to delay or jitter constraints, packets are dropped instead being stored at intermediate nodes in case of congestion or no route to destination.

2. **Effect of Increase in Network Traffic:** Multimedia traffic has large packet size and low inter-packet time. This increases network load significantly and affects other

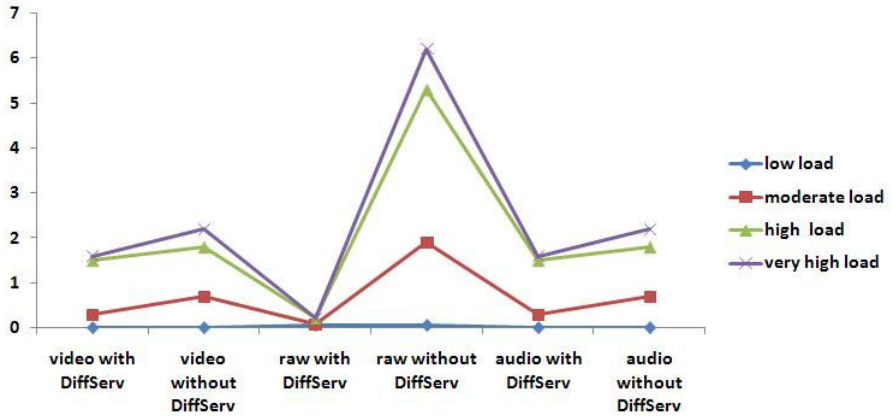


Fig. 4.

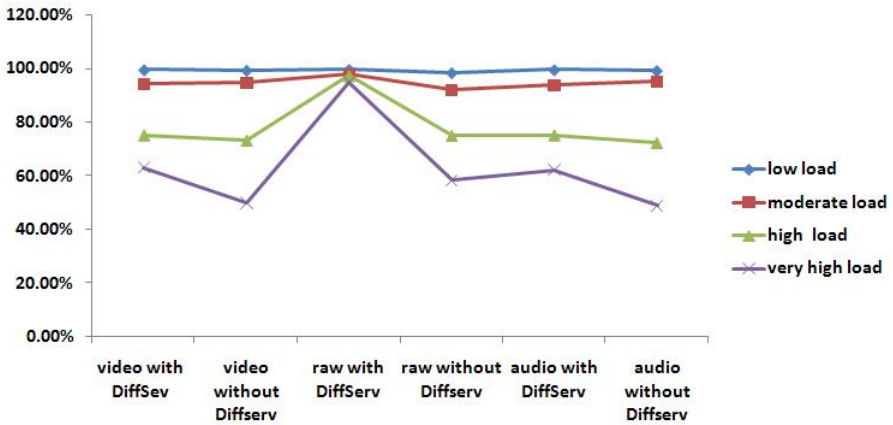


Fig. 5.

types of traffic. We use one source to generate video traffic and one to generate audio traffic. We increase the traffic in network by increasing size of data packets used for video and audio traffic by 500 bytes and 200 bytes respectively for each simulation round. The results obtained on end-to-end delay and packet delivery ratio are shown in Figure 4 and 5.

As shown in Figure 4, end-to-end delay over DiffServ is acceptable at low and moderate traffic for all three types of traffic. We observe that raw data is greatly benefited in terms of QoS metrics while using DiffServ as QoS model as compared to traditional routing model. As the network load increases further the DiffServ EF traffic class is overloaded and begin to dropping packets rapidly.

5 Conclusions

In this paper, we analyze DiffServ model for multimedia traffic over MANETs in terms of QoS metrics. Results obtained from various simulations are compared with traditional routing system. The comparison shows great improvement in all QoS metrics, if DiffServ model is used. Simulation result shows the effect of increase in network size and traffic on end-to-end delay and packet delivery ratio. We observe that DiffServ model is effective with moderate network size and traffic. As network capacity and size increases, the service classes used for QoS sensitive traffic overloads itself and due to this QoS flow performance decreases. For raw traffic like E-mail, FTP, HTTP DiffServ model shows improved results even when network scales or its traffic density increases.

An admission control scheme can be used to stop a service class being overloaded. This discards the flows whose admission affects the ongoing flows and degrades network performance. A method that integrates DiffServ model functionality with routing protocols in such a way that can scale well and maintain performance is required. Effects of Multi-protocol Label Switching (MPLS) combined with DiffServ need be analyzed over MANETs and shall be explored as future work.

References

1. Braden, R., Clark, D., Shenker, S.: Integrated services in the Internet Architecture-an overview. IETF RFC1663 (1998)
2. Blake, S.: An Architecture for Differentiated Services. IETF RFC2475 (December 1998)
3. Braden, R., Zhang, L., Berson, S., Herzog, S., Jamin, S.: Resource reservation Protocol (RSVP) Version 1 Functional Specification. RFC2205 (September 1997)
4. Black, D., Brim, S., Carpenter, B.: Per Hop Behavior Identification Codes (June 2001)
5. Haq, M.A., Matsumoto, M., Bordim, J.L., Kosuga, M., Tanaka, S.: Admission control and simple class based QoS provisioning for mobile ad hoc network. In: Vehicular Technology Conference, VTC 2004-Fall (2004)
6. Li, X., Cuthbert, L.: Stable Node-Disjoint Multipath Routing with Low Overhead in Mobile Ad Hoc Networks. In: Proceedings of the IEEE MASCOTS 2004, Netherland (October 2004)
7. Farooq, M.O., Aziz, S.: Stateless and Controlled Reservation Based DiffServ Model for Mobile Ad Hoc Networks. In: Wireless and Mobile Communications, ICWMC (2008)
8. Hanzo-II, L., Tafazolli, R.: A Survey of QoS Routing Solutions for Mobile Ad Hoc Networks. IEEE Communications Surveys Tutorials (2007)
9. Hanzo II, L., Tafazolli, R.: Admission control schemes for 802.11-based multi-hop mobile ad hoc networks: a survey. IEEE Communications Surveys Tutorials (2009)
10. <http://www.ietf.org/mail-archive/web/manet/current/msg00155.html> (2003)

Modeling and Verification of Chess Game Using NuSMV

Vikram Saralaya, J.K. Kishore, Sateesh Reddy, Radhika M. Pai,
and Sanjay Singh

Department of Information and Communication Technology
Manipal Institute of Technology, Manipal University, Manipal, India
sanjay.singh@manipal.edu

Abstract. Gaming industry has always relied on testing their product by playing it extensively. However, testers have their own limitations. When such a product is deployed, extreme gamers find those bugs that were overlooked by the testers. Hence testing is a best-effort service and does not assure that a particular product is working bug free. Application of formal methods to games is a vast area, but less explored. It has been applied to some of the simple games like Tic-Tac-Toe, Rush-Hour etc. Formalizing a chess game is complex since the game can enter a countably infinite number of states. In this paper we build a model which takes a sequence of moves (called as "Notation" in Chess Community) as input and verify that standard rules of the game are not violated. Specifications are written using LTL (Linear-Time Temporal Logic). We have used NuSMV (extension of Symbolic Model Verifier) as a model checking tool to verify the LTL specifications.

1 Introduction

Games have always been treated as non-critical applications. Until recently the gaming industry never felt the necessity to provide a completely bug-free product to its customers. Thus it mainly relied on the traditional testing approach which is economical and easy to incorporate. However it comes at the cost of annoying customers with occasional bugs. Many major players in the industry have now shifted towards formalizing their products.

Mathematical technique for modeling, specifying and verifying any system is termed as formalization. Formalization of games has three basic stages. The game has to be represented in terms of a model. This model should encapsulate every aspect of its real time counterpart. A proper understanding of the domain as well as formalization concepts are required to accomplish this task. The second step is to specify the properties that we expect from the model. These can be specified using an appropriate temporal logic. The final step is verifying whether the model satisfies its specification.

Formalization of games has been considered seriously in a few research communities. Formalization of a game like Chess has been considered from a purely mathematical and Game Theory point of view. Khoussainov et al. in [1] speaks

about formalization of finite perfect-information games with reference to board games including Chess. Zheng Zhang in [2] explains the formalization and model checking applied to Tic-Tac-Toe game. Tic-Tac-Toe is very simple in comparison to Chess, but provides useful insights into some important modeling concepts. Collette et al. in [3] discusses formalization of a complex puzzle like Rush Hour. Its complexity can be compared to that of chess but they both traverse an entirely different path to achieve their goals. Walter Storm in [4] discusses modeling of Sudoku puzzle using simulink verifier. A similar approach will be followed in this paper for modeling the game of Chess but using NuSMV as the verifier. Formalization of Chess game by representing its notation in terms of state is discussed by Yuri in [5]. However, to the best of our knowledge no work has been reported in the literature to verify a Chess game by modeling it in terms of a Finite State Machine (FSM).

Formalization of a game like chess can be viewed from multiple perspectives. One approach is to verify whether the given sequence of moves from the initial configuration of the game is played according to the rules. This approach essentially frees us from the burden of generating the moves which involves some sort of intelligence to be incorporated into the model.

We use NuSMV as the model verifier to verify that our model meets its specification. The NuSMV either outputs *true* or *false* depending on whether the sequence of moves represented in terms of the model conforms to the rule of game or not. Throughout the paper we use the term *valid* to represent this. We have used a parser to generate the sequence of moves from the given standard Chess notation as input. We vary this input and check the result for correct and incorrect sequences.

This paper is organized as follows. Section 2 gives an introduction to formalization of games by considering a general game G . Section 3 discusses the detailed formalization and modeling of the Chess game. Section 4 is about the properties a given Chess game should satisfy and how these can be formally specified using LTL. Section 5 briefly explains about the verification results obtained. Finally section 6 concludes the paper.

2 Formalization of Games

Let us consider a generic game. The configuration of the game is the assignment of values to its state variables at any point in time. Let the game be played between two players and has a set of finite number of configurations [6]. This set can further be divided into two mutually exclusive sets. From the first set of configuration it is the first player who has to take action or make his move. We call this set as P1. Similarly the second player will have to move from configuration P2. We also group some of these configurations as winning configurations for respective players. A configuration is a winning configuration if irrespective of the opponents moves the game can be won for the respective player by choosing an appropriate path. Let W_{P_1} represent winning configuration for player 1. Similarly W_{P_2} represent the winning configuration for player 2.

Thus $P1, P2, W_{P1}$ and W_{P2} can be expressed as:

$$W_{P1} \subseteq P1 \cup P2$$

$$W_{P2} \subseteq P1 \cup P2$$

It is clear that winning configuration of one player cannot be the same for the other. Hence we can write:

$$W_{P1} \cap W_{P2} = \emptyset$$

Let the initial configuration be s_0 . From this configuration it is player 1's turn to take action. His action will result in a transition to another state $s_1 \in P2$. This process repeats wherein a move by one player will result in a transition to a configuration from which the other player has to take action. The resulting sequence can be represented as:

$$s_0 s_1 s_2 \dots s_i \dots s_n$$

The game stops when :

$$s_i \in W_{P1} \cup W_{P2}$$

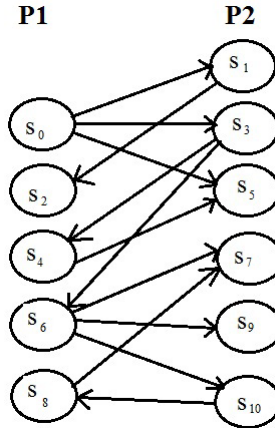


Fig. 1. Model of a general game G which dynamically determines the transitions

Transition through different configurations of game are shown in Fig.1. A point worth considering is that, every node has one or more outgoing edges. The actual edge that may be traversed depends on the intelligence we incorporate into our model and its decision making capability. On the contrary, consider Fig.2 where the transitions are clearly predetermined because every node has exactly one outgoing edge. An appropriate model can be chosen depending on the application to be modeled.

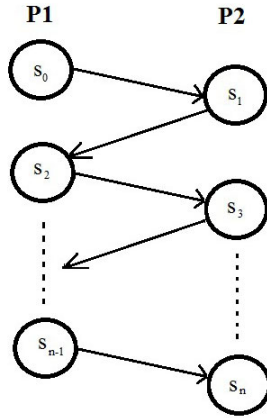


Fig. 2. Model of a predetermined general game G

3 Modeling the Chess Game

Chess is a finite game, played between two players who take turns alternatively. Player 1 starts the game from the initial configuration by making his move. Hence general modeling techniques discussed in the previous section applies to chess game as well. More specifically, we have P1 replaced with White’s configurations and P2 with Black’s configurations.

We choose to implement the second modeling option specified in the previous section in which the transitions are predetermined as shown in Fig.2. This suits our application need in which the sequence of moves are the input to the model.

The entire modeling process from an abstract level perspective is shown in Fig.3. Our main focus is to verify whether the input sequence of moves take only valid transitions. To achieve this objective we first need to encode and embed this sequence into our model. This is not a static one time process which is usually

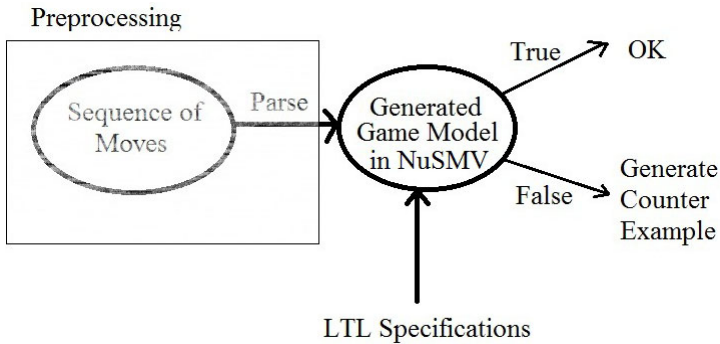


Fig. 3. Abstract view of our approach to model a chess game in NuSMV

the case in many formalization models. The dynamism comes from the fact that we may have to verify a different sequence of moves every time. Hence some sort of preprocessing is required. To simplify this task we have designed a simple parser which generates the model for a given sequence of moves represented using standard chess notation. In standard Chess notation King is K, Queen is Q, Rook is R, Knight is N, Bishop is B and default is a Pawn. The complete description about notation and rules can be found in [7]. The parser effectively translates input move sequence into a set of states and transitions among them.

The FSM representation of chess game \mathcal{C} can be given as:

$$\mathcal{C} = \{V, Q, \pi, \delta, q_0\}$$

where

- V : set of state variables
- Q : set of states
- π : labeling of states in terms of state variable values
- δ : transition function
- q_0 : initial state $\in Q$

The definition of V , Q and π is given in Table 1.

Table 1. Definition of V , Q and π

$V = \{$ <div style="text-align: center; padding: 5px 0;"> white.king.rank, white.king.file, white.queen.rank, white.queen.file, white.bishop1.rank, white.bishop1.file, white.bishop2.rank, white.bishop2.file, white.knight1.rank white.knight1.file, white.knight2.rank, white.knight2.file white.rook1.rank, white.rook1.file, white.rook2.rank white.rook2.file, white.pawn[i].rank (i=1 to 8), white.pawn[i].file (i=1 to 8), black.king.rank, black.king.file, black.queen.rank, black.queen.file, black.bishop1.rank, black.bishop1.file, black.bishop2.rank, black.bishop2.file, black.knight1.rank black.knight1.file, black.knight2.rank, black.knight2.file black.rook1.rank, black.rook1.file, black.rook2.rank black.rook2.file, black.pawn[i].rank (i=1 to 8), black.pawn[i].file (i=1 to 8), move, notationNumber, pieceToMove, capture, rankChange, fileChange </div> $\}$ $Q = \{ s_0, s_1, s_2, \dots, s_n \mid n = (2 * \text{number of moves}) - 1 \}$ $\pi : \{ Q \rightarrow D(V) \mid D(V) \text{ contains range of each } v \in V \}$

The initial state of the game is represented by positioning each of the pieces in their initial positions. Each piece has a rank (horizontal) and file (vertical)

Table 2. Initial Position of Pieces

Piece	Rank	File	Piece	Rank	File
white.king	1	5	white.queen	1	4
white.bishop1	1	3	white.bishop2	1	6
white.knight1	1	2	white.knight2	1	7
white.rook1	1	1	white.rook2	1	8
white.pawn[1]	2	1	white.pawn[2]	2	2
white.pawn[3]	2	3	white.pawn[4]	2	4
white.pawn[5]	2	5	white.pawn[6]	2	6
white.pawn[7]	2	7	white.pawn[8]	2	8
black.king	8	5	black.queen	8	4
black.bishop1	8	3	black.bishop2	8	6
black.knight1	8	2	black.knight2	8	7
black.rook1	8	1	black.rook2	8	8
black.pawn[1]	7	1	black.pawn[2]	7	2
black.pawn[3]	7	3	black.pawn[4]	7	4
black.pawn[5]	7	5	black.pawn[6]	7	6
black.pawn[7]	7	7	black.pawn[8]	7	8

attribute which takes a value in the range 1 to 8 that mimics the 8X8 squares on the chess board. The initial position of the pieces is shown in Table.2

The state variable *move* keeps track of whether it is white's or black's turn to take action in the current configuration. To model the chess moves we require five state variables as stated in Table.3

Table 3. State Variables Required to Model a Single Move and their Range

Variable	Range
notationNumber	1..500
rankChange	-7..7
fileChange	-7..7
pieceToMove	{none,king,queen,rook1,rook2, bishop1,bishop2,knight1, knight2,pawn1,pawn2,pawn3, pawn4,pawn5,pawn6,pawn7,pawn8}
capture	{none,king,queen,rook1,rook2, bishop1,bishop2,knight1, knight2,pawn1,pawn2,pawn3, pawn4,pawn5,pawn6,pawn7,pawn8}

The *notationNumber* variable is used to hold the current move number such that a *n* move game will have *notationNumber* ranging from 1 through 2*n*. The *pieceToMove* variable holds the piece that needs to change its position in the next state. The two other variables *rankChange* and *fileChange* does maintain the respective offsets by which the piece has to displace. Finally *Capture* suggests

whether any opponents piece were captured by the *pieceToMove* during its movement. We use these to decide the next state of the different elements in the model.

Example 1. Let us model a short sequence of moves which is commonly referred to as *scholar's mate*. These moves are given as input to our parser. The sequence of moves is shown below :

1. e4 - c5
2. Bc4 - Nc6
3. Qh5 - Nf6
4. QXf7#

The parser takes each move in turn and represents them using the five variables : *notationNumber*, *pieceToMove*, *rankChange*, *fileChange* and *capture*.

Consider the first move *e4*. The variable *pieceToMove* is assigned to value *pawn5* because *e4* represents a pawn in the fifth file which remains in the same file but is displaced to fourth rank. The parser keeps track of the previous position of every piece. This helps in the calculation of offset. Since previously the pawn was at second rank the *rankChange* will take a value $4 - 2 = 2$ and *fileChange* = 0. This pawn move results in no capture of enemy piece and hence *capture* = *none*. The move being considered is the first and hence *notationNumber* = 1. On the similar lines we can translate the other moves into our model representation.

4 Specification and Model Checking of Chess Game

A model is never right or wrong in itself. It either satisfies the property desired out of it or it does not. To verify this, we need to specify the properties in a way that is understandable by the system. Consider our model \mathcal{C} with the states s_i . Then we have to check whether the model satisfies a given property ϕ in a particular state as shown below.

$$\mathcal{C}, s_i \models \phi$$

The property ϕ can be specified in a suitable logic. For our model we need to quantify over different states in a path that change with time. To accomplish this temporal aspect we prefer the LTL [8]. Throughout the paper all our specifications will be coded in the syntax of LTL.

The desired properties of a good Chess model are the following :

1. Pieces move according to the standard rules of the game
2. No two consecutive states in a path are identical. This also assures that the players take alternative turns
3. The game ends when a winning configuration is reached

The second and the third property can be specified using the existing model parameters itself. But to specify the first property related to piece movement, we need details about the previous position of these pieces. This information cannot be extracted directly from the model we discussed in the previous section. Thus a new set of attributes to store the previous position of the pieces is accommodated in the model. These are represented by prefixing them with a "prev" keyword. With these necessary information at our disposal we discuss the LTL specifications for some of the pieces like King, Queen and Knight in the following subsections. The specifications for others can be derived in a similar manner.

4.1 King Movement

The king in a chess board has a very restricted movement with the maximum offset, which is the change in values of rank or file in either direction for a piece. It is 1 for the King in any direction. Thus either of rank and file changes can take a value from the set $\{-1, 0, +1\}$. This property is expected to be true globally in all the states and represented using the LTL specification ϕ_1 in Table 4.

4.2 Queen Movement

The queen can move along the rank or file in which it lies previously. Additionally diagonal movement can be permitted which is specified by constraining the magnitude of the offset along the rank to be equal to that along the file. The LTL specification for the Queen's movement is given by ϕ_2 in Table 4.

4.3 Knight Movement

This tricky piece can move in "L" shape and hence involves simultaneous change in rank and file in its every move. The LTL specification for the Knight's movement is given by ϕ_3 in Table 4.

The final step is to verify whether the model satisfies the specification mentioned above using NuSMV. If in any state of our model one of the piece shows anomalous behavior with respect to its movement then a counter example for the same will be generated.

5 Verification Results and Discussion

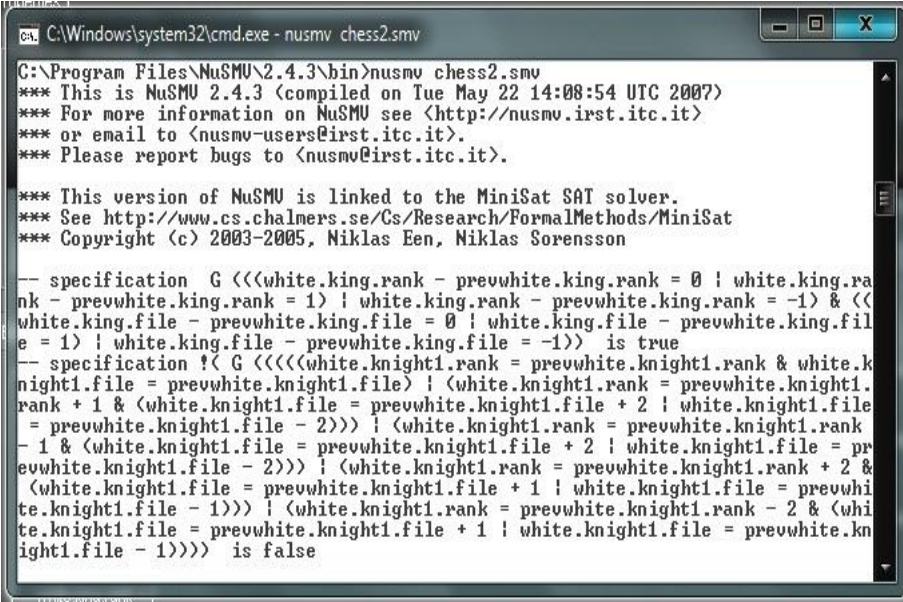
We have verified the model using NuSMV version 2.4.3 [9]. The model verifier has to evaluate a large number of states and hence takes approximately 8 seconds to run on a 3 GHz processor. The result we obtained by verifying the knight and the the king movement is shown by the screenshot in Fig 4.

It can be seen that the second specification returns false(last line in Fig 4). This is expected since we have purposefully specified it as !G(valid knight moves).

Table 4. LTL Specification for King, Queen and Knight in NuSMV

Piece Movement	Property LTL Specification
King	$\phi_1 : G($ (white.king.rank - white.prev_king.rank = 0 & white.king.rank - white.prev_king.rank = 1 & white.king.rank - white.prev_king.rank = -1) (white.king.file - white.prev_king.file = 0 & white.king.file - white.prev_king.file = 1 & white.king.file - white.prev_king.file = -1))
Queen	$\phi_2 : G($ (white.Queen.rank = white.prev_Queen.rank) (white.Queen.file = white.prev_Queen.file) (white.Queen.rank - white.prev_Queen.rank = white.Queen.file - white.prev_Queen.file white.Queen.rank - white.prev_Queen.rank = -1 * white.Queen.file - white.prev_Queen.file))
Knight	$\phi_3 : G($ (white.knight.rank = white.prev_knight.rank & white.knight.file = white.prev_knight.file) (white.knight.rank = white.prev_knight.rank+1 & (white.knight.file = white.prev_knight.file + 2 white.knight.file = white.prev_knight.file - 2)) (white.knight.rank = white.prev_knight.rank-1 & (white.knight.file = white.prev_knight.file + 2 white.knight.file = white.prev_knight.file - 2)) (white.knight.rank = white.prev_knight.rank+2 & (white.knight.file = white.prev_knight.file + 1 white.knight.file = white.prev_knight.file - 1)) (white.knight.rank = white.prev_knight.rank-2 & (white.knight.file = white.prev_knight.file + 2 white.knight.file = white.prev_knight.file - 2))

NuSMV returning *false* to this specification means that it is not the case that valid knight moves have been violated at least in one of the states. For the other specification which is $G(\text{valid king moves})$ NuSMV returns *true*. This is in accordance with the desired property for the king. An important aspect to bear during the design is that the state space of Chess game is very large. Thus more the number of specifications we verify at a time the more slower the NuSMV gets. Verification of one or two SPEC's at a time will help the cause.



```

C:\Windows\system32\cmd.exe - nusmv chess2.smv

C:\Program Files\NuSMV\2.4.3\bin>nusmv chess2.smv
*** This is NuSMV 2.4.3 (compiled on Tue May 22 14:08:54 UTC 2007)
*** For more information on NuSMV see <http://nusmv.irst.itc.it>
*** or email to <nusmv-users@irst.itc.it>.
*** Please report bugs to <nusmv@irst.itc.it>.

*** This version of NuSMV is linked to the MiniSat SAT solver.
*** See http://www.cs.chalmers.se/Cs/Research/FormalMethods/MiniSat
*** Copyright (c) 2003-2005, Niklas Een, Niklas Sorensson

-- specification G <<(white.king.rank - prevwhite.king.rank = 0 ! white.king.ra
nk - prevwhite.king.rank = 1) ! white.king.rank - prevwhite.king.rank = -1) & <<
white.king.file - prevwhite.king.file = 0 ! white.king.file - prevwhite.king.fil
e = 1) ! white.king.file - prevwhite.king.file = -1>> is true
-- specification !( G <<<<(white.knight1.rank = prevwhite.knight1.rank & white.k
night1.file = prevwhite.knight1.file) ! (white.knight1.rank = prevwhite.knight1.
rank + 1 & (white.knight1.file = prevwhite.knight1.file + 2 ! white.knight1.file
= prevwhite.knight1.file - 2))>> ! (white.knight1.rank = prevwhite.knight1.rank
- 1 & (white.knight1.file = prevwhite.knight1.file + 2 ! white.knight1.file = pr
evwhite.knight1.file - 2))>> ! (white.knight1.rank = prevwhite.knight1.rank + 2 &
(white.knight1.file = prevwhite.knight1.file + 1 ! white.knight1.file = prevwhi
te.knight1.file - 1))>> ! (white.knight1.rank = prevwhite.knight1.rank - 2 & (whi
te.knight1.file = prevwhite.knight1.file + 1 ! white.knight1.file = prevwhite.kn
ight1.file - 1))>> is false

```

Fig. 4. Screenshot : Verification of King and Knight moves in the model

6 Conclusion

Gaming industry is an area that is getting popular day by day. On one side there is a conventional way of testing a software product to make it bug free while on the other side we have the formalization techniques being widely accepted by many industries. Thus formalization of games has a bright prospect in the coming future. Modeling and verification of a game assures the stake holders that their product is bug-free. In this paper we have tried to model and verify the Chess game for any given sequence of moves. Using NuSMV as the model verifier tool, we could check the validity of different moves that is played during the course of a game. The system verifies the moves given as input using the standard Chess notation. However, there is further scope for improvement. We have not incorporated any sort of intelligence in choosing a move from a given configuration. A similar approach can be followed to model a standard commercially available Chess game engine like Deep Blue. Some heuristics need to be used to avoid the state explosion problem that is likely to arise.

References

1. Khoussainov, B., Nerode, A.: Automata Theory and its Applications. Birkhauser, Boston (2001)
2. Zhang, Z.: Playing tic-tac-toe game using model checking. Technical report, University of Illinois, Chicago (November 2004)

3. Collette, S., Raskin, J.-F., Servais, F.: On the symbolic computation of the hardest configurations of the RUSH HOUR game. In: van den Herik, H.J., Ciancarini, P., Donkers, H.H.L.M.(J.) (eds.) CG 2006. LNCS, vol. 4630, pp. 220–233. Springer, Heidelberg (2007)
4. Storm, W.: Solving sudoku using simulink design verifier. Technical report, Lockheed Martin Aeronautics Company, Bethesda (September 2009)
5. Khomskii, Y.: Infinite games. Technical report, University of Sofia Bulgaria, Summer Course (July 2010)
6. Hurd, J.: Formal verification of chess endgame databases (2005)
7. FIDE: Fide handbook e.i.01b. appendices (2010), <http://www.fide.com/fide/handbook.html?id=125&view=article>
8. Huth, M., Ryan, M.: Logic in Computer Science, Modelling and Reasoning about Systems, 2nd edn. Cambridge University Press, Edinburgh, UK (2005)
9. Cavada, R.: Nusmv: a new symbolic model checker (2010), <http://nusmv.fbk.eu/>

SMMAG: SNMP-Based MPLS-TE Management Using Mobile Agents

Muhammad Tahir¹, Dominique Gaiti², and Majid Iqbal Khan¹

¹ COMSATS Institute of Information Technology, Park Road, Islamabad, Pakistan
{muhammad_tahir, majid_iqbal}@comsats.edu.pk

² ERA, Institut Charles Delaunay, Université de Technologie de Troyes, Troyes, France
dominique.gaiti@utt.fr

Abstract. This paper proposes a model, SMMAG, which utilizes intelligent mobile agents for Traffic Engineering (TE) and network management activities of Multi-Protocol Label Switching (MPLS) network using commands from Simple Network Management Protocol (SNMP). It describes a mechanism for TE and network management by intelligently gathering descriptive statistics from the devices. It also helps in analyzing collected data which further can be used for decision making by a network administrator and/or mobile agent.

Keywords: MPLS, Traffic engineering (TE), SNMP, Mobile agents.

1 Introduction

In today's competing technologies, MPLS [1] has got a unique importance. It is an emerging technology which provides many advantages like simplified forwarding, efficient explicit routing, QoS routing, Traffic Engineering etc. over the traditional IP forwarding, ATM, Frame Relay etc. MPLS is a technique in which packets are assigned some predefined labels and forwarding takes place based on these labels. Devices in the MPLS network, called LSRs, build a special path called Label Switched Path (LSP) over which these labeled packets travel. Computing these paths while keeping in view the bandwidth and administrative constraints comes under the category of MPLS Traffic Engineering [1].

MPLS Network may contain different devices that need to be managed, e.g. hubs, switches and especially routers. To manage data flowing through these devices, it is necessary to consider Simple Network Management Protocol (SNMP) [2], [3]. SNMP commands (such as GET, GET-NEXT, SET & TRAP) collect data from the device and send it to management stations for decisions. These SNMP commands collect data using MIB (Management Information Base). MIB is a database used to manage different objects (entities) of a network. These SNMP commands can be applied using programming techniques for example static code for one time execution, Remote Procedure Calls (RPCs), and writing mobile intelligent codes by establishing the monitoring sessions between the device and the monitoring entity. Mobile code (or mobile agent) provides more flexibility (for example learning from the actions performed etc.) over static codes. We can execute SNMP queries by using agents over

these managed devices. Thus agents provide us a mechanism to introduce some in-network decision-making capabilities for complex network management tasks. They can move from one location to another in a network to perform some predefined tasks or intelligently define their own tasks based upon the knowledge learned.

Our objective in this paper is to propose a model that utilizes intelligent mobile agents for Traffic Engineering (TE) and network management activities of Multi-Protocol Label Switching (MPLS) network using commands from Simple Network Management Protocol (SNMP). Rest of the paper is organized as follows; Section 2 literature review and our problem statement; then detailed explanation of the SMMAG model and its working is represented in Section 3, Section 4 discusses how SNMP-based MPLS data can be read from MIB and Section 5 presents conclusion including limitations and future work.

2 Literature Review

This section presents literature that has utilized intelligent mobile agents for network management and other activities. In [4] authors proposed the idea of utilizing agents in MPLS Network for connection establishment (i.e. the creation of MPLS Tunnels and LSPs). The work done by [5] represents different broad areas of network management like fault management, accounting management, configuration management, performance management and security management using mobile agents. Their ultimate purpose is to generate “plug and play” networks. For each management activity, different solutions are proposed. Agents named discovery agents and networking agents are proposed. CORBA architecture is used for real world implementation.

White et al., [6] proposed a scheme to deal with Point-to-Point (P2P) and Point-to-Multipoint (P2MP) connection requests. Authors talk about three types of agents: explorers, allocators and de-allocators for distributed control and management. In [7] mobile agents are introduced for following network operations (not exactly representing the network management): to find/trace a network user, to search for a file/software from the network and to trace the memory and network nodes on Linux network. Although there are five agents that have different tasks but only one agent is active at one time, so there is no inter-agent communication.

Adhicandra et al, [10] argue that under congestion scenarios high network management traffic generated by protocols such as, SNMP can result in extreme load on network management station. Therefore the authors propose to transform centralized management activities to distributed using mobile agents. The results showed that when the number of managed devices is small then static agents are better option compared to mobile agents and vice versa.

Another study [11] provides a mechanism to enable QoS routing using mobile software agents. In particular authors utilize mobile agent along with wave paradigm to establish QoS enable multipoint to point routing tree. An important observation made by the authors is that mobile agents result in reduced traffic only when they operate in isolation. However, if agents operate in cooperative manner then they result in increased traffic. Andrzej Bieszczad et al., [12] discuss various application scenarios of mobile agents including their advantages and disadvantages while in [13] authors analyze the use mobile agents for establishing fault tolerant systems.

2.1 Problem Statement

MPLS Network management is different from a traditional IP Network management because it involves different additional features like configurations in LERs (Label Edge Router) and LSRs (Label Switch Router), management of IP to FEC (Forwarding Equivalence Class) mapping, management of label distribution protocols mechanisms, management of explicit path mechanism and different aspects of QoS routing. If we consider TE management aspect of an MPLS network, we need to focus on above mentioned features. Moreover, we need to automatic the process of information gathering (from the MPLS-based devices) and processing. Here the importance of agents and SNMP become true which are portable and can collect information seamlessly. We have focused on the management of MPLS-TE activities which has the following parts:

- The management of an MPLS Network for LSR/TE-MIBs configurations
- Label manipulation features
- Label distribution protocols detection
- Bandwidth collection information and
- Explicit Path configuration issues

Currently, this list of items is managed and monitored, manually, by network administrators. In proposed model we have considered these items in TE-management perspective. Our ultimate objective is to minimize the human intervention in traffic engineered of MPLS network.

3 SMMAG: The Proposed Model

Our objective is to provide a simple approach for the collection and manipulation of TE management information by using mobile agents. The information include, reading and writing different information on the router interface without the need of human intervention, like adjusting the values of labels at different LSRs, adjusting and reading information about the Label Space (per-platform or global, use of label stacks, MPLS tunnel statistics and admin status of interfaces and tunnels etc).

Our proposed model can be divided into sub-modules. First sub-module defines agent structure (named as Performance Agent Model (PAM)). The second sub-module groups Performance Agents (PAs) into Agent Group Model (AGM). Finally our model describes the overall behavior. It not only describes the actions performed by the agents but also the cooperation among them. SMMAG also illustrate the decision activities taken by agents on their own to complete some pro-active tasks.

3.1 Performance Agent Model (PAM)

The main purpose of Performance Agent is to manipulate the specific data (i.e. tables) of a MIB. It performs specific SNMP operations (i.e. GET, SET, etc) and exchange results with other PAs. The environment for a PA is the specific router for which it is deployed.

PAM can be divided into following components:

1) *Task Manager*: Task Manager is the main unit of PA (Fig. 1). It contains number of tasks (SNMP commands) to be applied on the specific MIB. It is used to extract/update rules from the Semantic Unit (SU) and apply them on certain tasks. It also updates Repository Unit (RU) and communicates with the communication module to read or write results to the Black Board (BB).

2) *Repository Unit (RU)*: RU contains the facts (results) and history of the specified tasks performed by the PA. It contains four types of information: the type of action performed by the PA, the action performed on which table entry of a MIB, result and the status of the action performed. Type of action could be a SNMP operation, table entry represents Object Identifier (OID) (e.g. *mplsInterfaceLabelMinIn*) of a MIB, result contains the value obtained, and the status represents the action performed (indicated as 1) or not performed (indicated as 0).

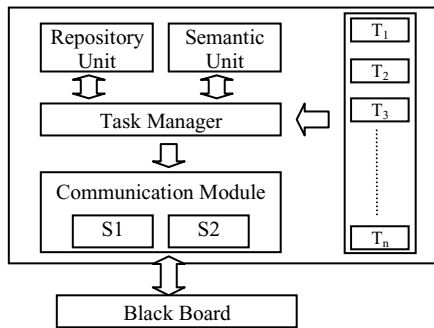


Fig. 1. Performance Agent sub-module and its interaction with Black Board

3) *Semantic Unit (SU)*: SU contains the word and sentence structures (logical expressions) to define the actions performed by the Task Manager. It contains the rules (IF...THEN) which are used by the task manager and are updated using the facts and history of the RU. So SU provides the computational (analytical) and decision power to the PA.

Decision Scenario-1: Consider Task Manager has performed a task (e.g. *check the serial interface status of the router*) four times in a short span of time. This results that interface was “administratively down” for the first and third time and was “up” for the second and fourth time. These results are stored in RU. Now using this history from the RU, SU may add/modify its logical structure for the *same task* going to be performed fifth time. Now, SU may say (actually infer) that “try to check the interface status at $t=t+\Delta t$ times (t : original time interval for continuous checking of interface, Δt =small amount of time added to original time). Because this interface was in monitoring state by the administrators at time t , so there could be the possibility that it would be “up” at $t=t+\Delta t$. This decisive power could result in interface status as “up” this time.

4) *Black Board (BB)*: Black Board is a simple agent, which acts like a synchronized variable, where different PAs can read or write results and communicate to each

other. BB contains the data provided by Communication Module (CM) for other agents. For example, if the type of data *pasted* on BB is *mplsInterfaceAvailableBandwidth*, then any other PA can use this information.

5) *Communication Module (CM)*: It reads and writes data from/to BB. CM has storage S1, which contains a specific *value* (PA identity and a one-bit flag). CM updates S1 whenever it needs certain information from other PAs. CM writes this value (PA identity and flag=0) to BB. Other agents can read this request and may provide the results. Whenever another agent provides a result it updates the value of the flag to 1 (updated). CM also has another storage named S2, which contains one-bit flag field. This field represents that the result is for an external agent (value=1) or for the same agent (value=0).

3.2 Agent Group Sub-Module (AGM)

AGM is the combination of Managing Agent, BB and one PA of each LSR. Managing agent is responsible for getting results from BB and sending it back to Network Management Station (NMS). Figure 2 represents the Agent Group Model which describes the inter-agent cooperation and information retrieval.

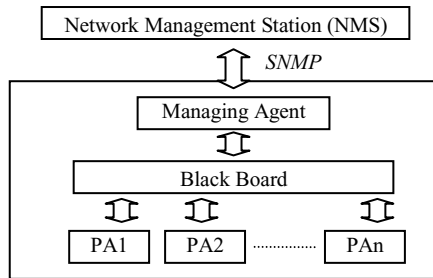


Fig. 2. Inter-agent cooperation and information retrieval

3.3 Working of SMMAG

Figure 3 represents the overall structure of the proposed model. We consider NMS as a device (computer) and software to take care of the TE management activities using SNMP. One AGM is applied for each LSR. In the proposed model we have considered five MPLS routers and one NMS for all of them. Now to represent everything together: first of all NMS is instantiated which in turn invokes the agent system for the creation and navigation of agents. The agents created are one-hop mobile agents, which travel from NMS to specified device and perform their actions and send results back to NMS and then terminate. There are some agents that may reside there for a long duration of time to get the TRAP results (SNMP value) in case of network failure i.e. Node/Link failure.

In this proposed model the combination of centralized and distributed approach has been used (hybrid approach). It is centralized in the sense that one management station is central, to collect all the statistics, from the agents. It is distributed in the

sense that agents are distributed to each device and perform their actions. They learn from the environment (device), stores their experience in the database (Repository Unit) and *reason* on certain actions before being performed. Also they react according to the situations (for example, a TRAP is initiated by a device which is captured by the agent residing there and sending back the information to NMS). The agent may reason on this situation because this TRAP could mean that the device has gone through fluctuation for a very small amount of time and the system may return to normal state immediately. So that agent may send the history information and as well as the latest statistics.

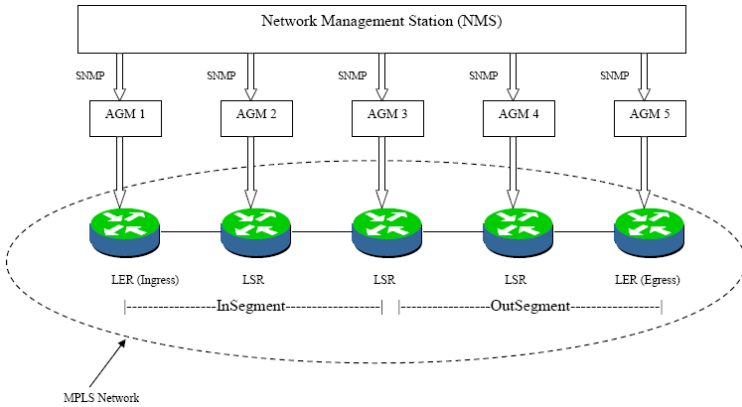


Fig. 3. SMMAG: MPLS-TE Management Model

4 Reading MIBs for SNMP-Based MPLS Data

This section describes how SNMP actually works to collect MPLS relevant data from specific MIBs. A few examples are illustrated.

4.1 Performance Agents for LSR-MIB

Different kinds of PAs can be deployed for LSR-MIB [8]. For example:

1) *Performance Agent for Configuration Table:* This PA works on the MPLS-enabled interface and can perform different tasks. It can read minimum and maximum label values that the LSR can send and receive through this interface by querying following OIDs: `mplsInterfaceLabelMinIn`, `mplsInterfaceLabelMaxIn`, `mplsInterfaceLabelMinOut` and `mplsInterfaceLabelMaxOut`. It can also indicate that what type of label space is being used by this interface by querying following OID: `mplsInterfaceLabelParticipationType`. If the result is 0 then its per-platform label space and if the result is 1 then its per-interface label space. Similarly, it can collect the statistics about the total bandwidth and available bandwidth on this interface by querying the following OIDs: `mplsInterfaceTotalBandwidth` and `mplsInterfaceAvailableBandwidth`. All this information is stored in the RU for history and for future reasoning purposes.

2) *Performance Agent for the cross-connect table*: This PA can be used to associate the InSegment Labels with OutSegment Labels. Its tasks may include checking the LSPID for the OutSegment using *mplsXClspld* object.

Decision Scenario-2: After checking the LSPID, the agent can take decision to get related information (from different agents), like, to which outgoing interface this LSP belongs to and what is the outgoing label for this interface. All this information are bundled in one unit and sent to NMS to have a useful view of the TE management (this information will be collected by different agents but an overview is represented by one agent: Management Agent).

Similarly, to check the administrative status of this segment *mplsXCAdminStatus* object is used. If the returned value is 1 then it represents “up” status of the segment, 2 represents down status and 3 represent testing status.

4.2 Performance Agents for TE-MIB

There are different kinds of PAs for TE-MIB [9] depending upon the tables used for this MIB. These are:

1) *Performance Agent for the MPLS Tunnel Table*: The tasks followed by this PA could be: *mplsTunnelName*, *mplsTunnelDescr*, *mplsTunnelSignallingProto*, *mplsTunnelHoldingPrio*, *mplsTunnelLocalProtectInUse*, *mplsTunnelPathInUse*, *mplsTunnelAdminStatus* (the value of 1 represents the tunnel status to “up”, 2 for “down” and 3 for “testing”) etc.

2) *Performance Agent for the Tunnel Hop Table*: This agent specifies the hops that constitute an explicitly defined path. Its tasks could be: *mplsTunnelHopAddrType*, *mplsTunnelHopLspId*, *mplsTunnelHopType* (the value of 1 represent strict hop type and 2 represents loose hop type).

Decision Scenario-3: If the admin status is set to down, then agent from its previous knowledge (from RU and SU) may decide to collect more information like tunnel name, tunnel description, protocol used, tunnel admin status and tunnel LSPID from other agents (like PA for tunnel table and PA from the tunnel Hop Table). While looking at this information, NMS comes to know the characteristics of the new established tunnel and may decide future actions (e.g. getting the available bandwidth information etc).

The other characteristics can be studied by consulting the specific MIBs [8][9].

5 Conclusion and Perspectives

The overall problem discussed in this paper is the TE-management of an MPLS network while considering different features, like labels mechanism, explicit path mechanism, label distribution, bandwidth collection information and different configuration of LSR/TE-MIBs. The techniques used for MPLS-TE management were briefly discussed which includes: intelligently collecting data from the device, by using agents and updating its knowledge and useful view of data to NMS.

Our proposed model is not only a monitoring model but also it represents the decisive, learning and reasoning power, which are the key characteristics for any intelligent agent. So the agents proposed are one-hop mobile intelligent agents. The main idea behind this proposed model was not only to collect TE management information but also to manipulate and represent it in an intelligent way so that the NMS and agents can take some decisions based on this information.

Also there exist some limitations like history size of RU and SU is not specified. History size is dynamic because it depends upon the number of actions assigned to an agent and the cooperation between different agents. This cooperation is also dynamic because it depends upon the agent and type of information it wants to collect from other agent in a specific time.

Currently the agents perform only tasks specified at their creation time. As future work if we increase the number of agents, this will in turn increase the complexity of collecting statistics, the cooperation among agents and the network traffic behavior. Furthermore, Communication Module may be extended by providing the flexibility of storing more data and sharing of more than one piece of information to other agents at one time. We plan to use a MAS simulation environment for example Madkit to validate our proposed model.

References

1. Davie, B., Rekhter, Y.: *MPLS: Technology and Applications*. Morgan Kaufman Publishers, San Francisco (2000) ISBN: 1-55860-656-4
2. Case, J., et al.: *SNMP: Simple Network Management Protocol*, RFC 1157 (1990), <http://www.faqs.org/rfcs/rfc1157.html> (last retrieved March 08, 2011)
3. Mauro, D.R., Schmidt, K.J.: *Essential SNMP*, 2nd edn. O'Reilly Media, Inc., Sebastopol (2005)
4. Yucel, S., Saydam, T.: Connection management in MPLS networks using mobile agents. In: Lorenz, P. (ed.) *ICN 2001*. LNCS, vol. 2094, pp. 329–338. Springer, Heidelberg (2001)
5. To, H.H., Krishnaswamy, S., Srinivasan, B.: Mobile agents for network management: when and when not! In: Liebrock, L.M. (ed.) *Proceedings of the 2005 ACM symposium on Applied computing (SAC 2005)*, pp. 47–53. ACM, New York (2005)
6. White, T., Pagurek, B., Oppacher, F.: Connection management by ants: an application of mobile agents in network management. In: *Proceedings International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, Las Vegas, Nevada, USA (1998)
7. Manvi, S.S., Venkataram, P.: A method of network monitoring by mobile agents. In: Kumar, A., Reddy, V.U. (eds.) *Int. Conf. on Communications, Control, and Signal Processing, CCSP 2000*, Viva Books, India, pp. 214–218 (2000)
8. Srinivasan, C.: *Multiprotocol Label Switching (MPLS) Label Switching Router (LSR) Management Information Base (MIB)* (2004), <http://www.rfc-archive.org/getrfc.php?rfc=3813> (last retrieved March 08, 2011)
9. *CISCO MPLS-TE MIB*, <http://cisco-press-traffic-engineering.org.ua/1587050315/ch08lev1sec2.html> (last retrieved March 08, 2011)

10. Adhicandra, I., Pattinson, C., Shagouei, E.: Using Mobile Agents to Improve Performance of Network Management Operations. In: Postgraduate Networking Conference (PGNET 2003), Liverpool, UK (2003)
11. Gonzalez-Valenzuela, S., Leung, V.C.M.: QoS routing for MPLS networks employing mobile agents. *IEEE Netw.* 16, 16–21 (2002)
12. Bieszczad, A., Pagurek, B., White, T.: Mobile agents for network management. *IEEE Communications Surveys*, Fourth Quarter 1998 1(1) (Fourth Quarter 1998)
13. Tanaka, Y., Hayashibara, N., Enokido, T., Takizawa, M.: Fault-Tolerant Distributed Systems in a Mobile Agent Model. In: Bressan, S., Küng, J., Wagner, R. (eds.) *DEXA 2006*. LNCS, vol. 4080, Springer, Heidelberg (2006)

Face Detection and Eye Localization in Video by 3D Unconstrained Filter and Neural Network

Pradipta K. Banerjee¹, Jayanta K. Chandra¹, and Asit K. Datta²

¹ Department of Electrical Engineering
Future Institute of Engineering and Management
Kolkata-700150

pradiptak.banerjee@gmail.com

² Department of Applied Optics and Photonics
University of Calcutta, Kolkata
asitdatta@gmail.com

Abstract. Frequency domain 3D-filter designing for automatic face detection and neural network based searching algorithm for eye localization of detected faces in video sequences is proposed. A series of spatio-temporal volumes are constructed from the video sequences of faces by concatenating the frames of a single complete cycle of face position is used to design a 3D unconstrained correlation filter by classical Fourier approach. The Unconstrained Optimal Trade-off Synthetic Discriminant Function (UOTSDF) filter is generalised here into a video filter of 3D spatio-temporal volume. After extracting the facial region by 3D correlation filter in frequency domain of the video frames, a neural network is employed to locate the eyes. The novelty of the face detection in video by frequency domain analysis and fast eye searching by parallel neural net of Generalised Regression Neural Network (GRNN) is validated with the benchmark database like VidTIMIT video database.

1 Introduction

Due to unlimited variations in pose and expression, the research in face detection in video scenes becomes more challenging in recent years. Numerous contributions in face detection by frame based methods can be found in [4, 10, 3, 20, 18] where the temporal information, the most important part of video scenes, is ignored. This is the main drawback of these approaches while handling the video sequences. Instead of detecting each frame, temporal approach exploits temporal relationships between the frames to detect multiple human faces in a video sequence. To detect the frontal faces and tracking non-frontal faces with online adaptive face models is reported in [21] and the edge concentration problem is overcome in [14] by edge pixel counting to track facial features in video sequences. In [8, 13, 22] face detection methods by fully using the temporal information provided by video are proposed. Cascade AdaBoost approach for real time face detection in video scenes is proposed in [18] and the color information based human face validation is made in [12]. In majority cases, the analysis

and processing for face detection are carried out on the spatial representation of the face image i.e., the intensity and / or color values of the video frames. However, there are significant results that support the use of frequency domain representation of face images. While proposing frequency domain operations on face image, one of the major approach is related to the use of correlation filters (CFs). Varied types of CFs offer many attractive qualities particularly where missing data and poor data quality are common problems. There are many contributions in the field of frequency domain face recognition using CFs [17, 16, 11]. However a serious problem occurs due to in-class variability of multi-view faces dataset is larger than that of front-view faces dataset in video sequences. Face detection of video sequences has many problems as, in general, the persons face is exposed to very different illumination conditions, different size scales, abrupt change in pose, rapid head orientation, different face expressions, and specially in many occasions significant parts of the face are occluded and only limited face information is available.

In this paper a combined approach of frequency domain correlation filter for face detection and neural network based eye localization in video scenes has been proposed. In general the correlation filters used for face detection and/or recognition are 2D filters. Unlike [15], the proposed work uses UOTSDF-filter rather than MACH-filter [2], for generalised face detection technique and no specific class of action is recognized. The main objective in this study is fast face detection in video scenes for eye localization which is done by a parallel neural network architecture of GRNN for quick response. To localize the face part of the video frames a generalised 3D UOTSDF filter is proposed. The main reason of using UOTSDF instead of MACH is as follows: MACH uses the average similarity measure (ASM) between intraclass images while maximizing the average correlation height (ACH) in the correlation plane and hence it is mainly used for classification purposes. But for a generalised face detection system the exact location of the faces is needed and hence the use of shift invariant property of correlation filter is more logical than a correlation filter classifier. As UOTSDF filter has built in shift invariant property and the criteria of ASM is ignored in its design methodology, it is generalised here into a video filter of 3D spatio-temporal volume for face detection in video scenes.

Section 2 describes the theoretical development of the proposed 3D UOTSDF filter. The applications of the filter in face detection from video is described in Section 3. After detecting the face part it is desirable to train the GRNN for eye localization and it is discussed in Section 4. The effects of the proposed system is refined with the benchmark database in Section 5. The concluding remarks are given in Section 6.

2 3D UOTSDF Filter

While several correlation filters [5, 11, 17, 16] are designed in recent years, the UOTSDF filter is used in this effort because of its ability of peak sharpness while providing noise tolerance, its simplicity in computation and especially the built-in shift invariance i.e. if the test image is shifted with respect to the training

images, then the correlation peak will be shifted by the same amount in the response surface.

In general, the UOTSDF filter combines the training images into a single composite template by optimizing three performance metrics: the Average Correlation Height (ACH), the Average Correlation Energy (ACE) and the Output Noise Variance (ONV) from a given series of instances. It is desired from the filter's response that the ACH should be high enough for the certain class of images while suppressing the sidelobe regions around the peak to make it sharp as well as reducing the ONV for noise tolerant.

This procedure results in a two dimensional template of 2D UOTSDF filter in frequency domain that may express the general shape or appearance of training face sequences. The importance of the general shape of training sequences is that when the designed filter (2D UOTSDF) is correlated with testing sequences in the frequency domain via a 2D fast Fourier transform (FFT) (2DFFT is an efficient algorithm to compute 2D discrete Fourier transform (DFT)), it results in a surface where the highest peak corresponds to the most likely location of the face in the frame is found due to the shift invariance property of the correlation filters. A simple and straightforward way to locate the face in a single video frame can be obtained by simple correlation of 2D UOTSDF with the successive 2D video templates. But in order to fully encompass the information of both space and time contained in a video sequence, the 3D UOTSDF filter is synthesized by the informations obtained from spatio-temporal volumes of consecutive face video sequences.

2.1 Formulation of 3D UOTSDF

In this subsection the 3D filter generation is discussed by using the temporal derivative of each pixel resulting in a volume for each training sequence. A series of spatio-temporal volumes i.e. some video files are taken from the face video sequences and concatenated the frames of a single complete cycle to synthesize the 3D filter. From a set of spatio-temporal volumes the temporal derivatives of each pixel of each video sequences are calculated by sobel operator [6]. It is a differential operator computing an approximation of the gradient of the image intensity and it's very fast to apply since it's based on a small window (3×3 kernel) to convolve with the whole image. The Eq. (1) shows the temporal derivative operation of one video sequence with the Sobel kernel.

$$\hat{g}(x, y) = \hat{f}(x, y) * \begin{pmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} \quad (1)$$

where $\hat{f}(x, y)$ is one of the frames in video scenes and $\hat{g}(x, y)$ is the resulting edge image after temporal derivative operation. Bold uppercase and lowercase symbols represent frequency domain matrix and vectors respectively while light lowercase with a hat represents the space domain matrices. These edge images are then stored in a 3D matrix to construct the spatio-temporal volumes and

the set of spatio-temporal volumes are then processed in frequency domain by 3D FFT for further operation of synthesizing the 3D UOTSDF filter. The 3D FFT operation of the spatio-temporal volume is given by:

$$\mathbf{G}(u, v, w) = \sum_{t=0}^{T-1} \sum_{y=0}^{C-1} \sum_{x=0}^{R-1} \hat{g}(x, y, t) e^{-j2\pi[\frac{ux}{R} + \frac{vy}{C} + \frac{wt}{T}]} \quad (2)$$

where, $\mathbf{G}_{3D}(u, v, w)$ is the resulting volume in frequency domain obtained from the volume $\hat{g}(x, y, t)$ corresponding to the temporal derivative of the input sequence. C is the number of columns, R , the number of rows and T the number of frames in one video training set. After having the resulting volume in the frequency domain the 3D matrix $\mathbf{G}_{3D}(u, v, w)$ is lexicographically ordered and the resulting column vector \mathbf{g}_i of dimension $T \times C \times R$ (where $i = 1, 2, 3, \dots, T_s$. T_s is the total number of frames used in the whole training set) is obtained.

Having obtained the column vector \mathbf{g}_i the UOTSDF filter is synthesized to satisfy the considerations of maximizing ACH while minimizing both ACE and ONV. ACH is expressed as

$$ACH = \left| \frac{1}{T_s} \sum_{i=1}^{T_s} \mathbf{h}_{1D}^+ \mathbf{g}_i \right| = |\mathbf{h}_{1D}^+ \mathbf{m}| \quad (3)$$

where, \mathbf{m} is the average of the T_s vectors $\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3, \dots, \mathbf{g}_{T_s}$. and \mathbf{h}_{1D} is the desired UOTSDF filter vector. The ONV is expressed by the energy function E_1 as

$$E_1 = \mathbf{h}_{1D}^+ \mathbf{O} \mathbf{h}_{1D} \quad (4)$$

where \mathbf{O} is a diagonal matrix containing the elements of the input noise power spectral density along its diagonal. To suppress the sidelobe regions for getting a distinct peak value in the correlation plane ACE is minimized and given by the energy function E_2 as

$$E_2 = \mathbf{h}_{1D}^+ \mathbf{D} \mathbf{h}_{1D} \quad (5)$$

where, \mathbf{D} is a diagonal matrix containing the average power spectral density of the training spatio-temporal vectors along its diagonal. The closed form solution [9] of the UOTSDF filter vector which maximizes the square of ACH instead of constrained peak values [1] of all training images to a specific value (to reduce the computational time) can be given in a single line equation as,

$$\mathbf{h}_{1D} = (\alpha \mathbf{O} + \beta \mathbf{D})^{-1} \mathbf{m} \quad (6)$$

where, \mathbf{h}_{1D} is the synthesized UOTSDF filter vector. The scalar parameters α ($0 \leq \alpha$) and β ($\beta \leq 1$) are user specific non negative optimal tradeoff parameters, considered as relative weights of noise tolerance and peak sharpness. Having obtained the 1D UOTSDF filter (\mathbf{h}_{1D}) it is now reshaped in the reverse order by arranging the vector elements into a volume containing R rows, C columns and T frames with proper care. Thus a 3D UOTSDF filter \mathbf{H}_{3D} in frequency domain

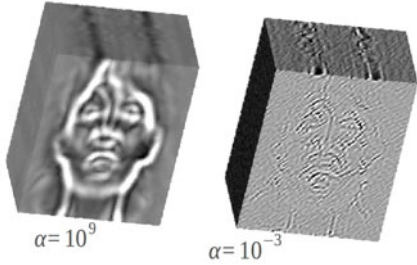


Fig. 1. Shows the 3D UOTSDF filter for two different settings of α . Left one for high α and right one for low α

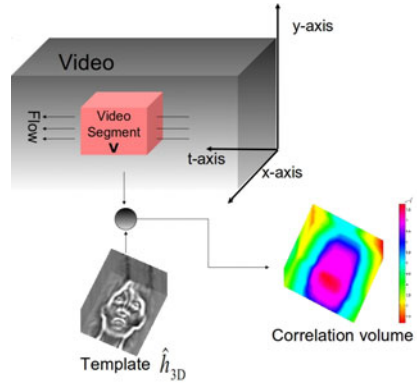


Fig. 2. Shows the space time correlation in video

is generated and the corresponding spatial domain filter \hat{h}_{3D} can be obtained by 3D inverse Fourier transform according to the following equation

$$\hat{h}_{3D}(x, y, t) = \sum_{u=0}^{R-1} \sum_{v=0}^{C-1} \sum_{w=0}^{T-1} \mathbf{H}_{3D}(u, v, w) e^{j2\pi[\frac{ux}{R} + \frac{vy}{C} + \frac{wt}{T}]} \quad (7)$$

where, $x = 0, 1, 2, \dots, R - 1$, $y = 0, 1, 2, \dots, C - 1$, $t = 0, 1, 2, \dots, T - 1$. Fig. (1) shows the volumetric representation of 3D UOTSDF filters obtained for two different setting of α . In this study a low value of $\alpha = 0.001$ is taken for getting a distinct peak in correlation surface.

3 Face Detection in Video

Having obtained the video filter it is now required to correlate this 3D template with a small video clips . The detail correlation process in space-time domain for video is shown in Fig(2). In this study the benchmark video database VidTIMIT [19] is used. This database is comprised of video sequences consist of the person moving their head to the left, right, back to the center, up, then down and finally return to center. The video filter \mathbf{H}_{3D} is synthesized with five small clips of each video having dimension $(71 \times 61 \times 30)$. In the testing stage a test '.avi' file of dimension $(128 \times 128 \times 222)$ is taken. The test file contains the face video with different pose and head orientation. The objective of the proposed strategy is to find the location of the face in the video scenes and the process is shown in Fig(2). The shift invariant property of the correlation filter makes it easy to locate the point of interest in the target image and this is reflected in the correlation plane where the maximum value of the response is obtained. Fig(3) shows the different locations of the distinct peak in the correlation volumes. The positions of the peak vary according to the movement of face in video. Having obtained the faces in video (see Fig(4)) the next step is to find the location of

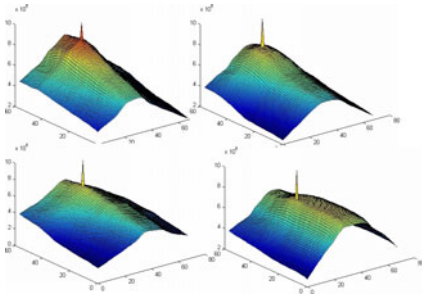


Fig. 3. Shows the distinct peak obtained in the correlation volumes. Some of the video frames are shown for validation

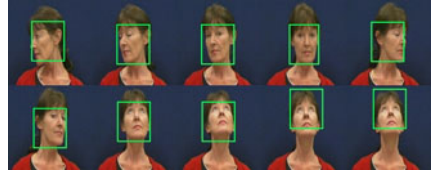


Fig. 4. Shows the face detection and localization in video sequences with varying pose and head orientations. Out of 222 frames are shown where pose variations are clearly visible alongwith the face detection by rectangular boxes

eyes in faces. To do so some training is needed and hence a parallel neural net architecture of GRNN [7] is employed. One of the main reasons in selecting the GRNN is that it is especially advantageous due to its ability to converge to the underlying function of the data where only few training samples are available. The additional knowledge needed to get the fit in a satisfying way is relatively small and can be done without additional input by the user. This makes GRNN a very useful tool to perform predictions and comparisons of system performance in practice such as in the paradigm of pattern recognition.

4 Training of GRNN for Eye Localization

4.1 Eye Templates for Training

In the previous section the exact location of the face in video sequence is detected and hence the searching region of eyes will be consequently reduced which is one of the main advantages of the proposed system. As for example for a test video file of dimension $128 \times 128 \times 222$, the face has been localized in the region of dimension $54 \times 47 \times 222$ (where 222 stands for the number of frames in the video clips). Hence the searching region is reduced remarkably. The eye templates \hat{e}_T are cropped from each gray scale frames of the training video clips. The dimension of each templates are fixed to 15×15 . For a supervised learning method negative training is needed so that the system can classify the authentic and impostor patterns.. This negative training is also obtained by extracting non-eye templates $\hat{n}e_T$ from the training video frames. The dimension of non-eye templates are of same size as eye-templates. The templates are then lexicographically ordered to form the pattern vectors e_T^i and ne_T^i of eye and non-eye respectively. The pattern matrix $\hat{p}_{d^2 \times T_n}$ is generated by concatenating the column vectors e_T^i and ne_T^i (where, $i = 1, 2, \dots, T_n$) side by side, where d^2 stands for the total number of pixels in an eye-template and T_n indicates the total number of training patterns. Training of GRNN needs the target vectors \hat{t} associated with each input pattern. In this

study the target vectors associated with the eye templates are set to '1' and that of for non-eye templates set to '0'. Thus a target matrix $\hat{t}_{d^2 \times T_n}$ is formed where $T_n/2$ number of columns containing values =1 and the rest $T_n/2$ columns are comprised of zero.

GRNN consists of four layers of neurons Fig(5). The first layer is the input layer, which is fully connected to the next layer and is just used to distribute the input vector to each node in the next layer. The second layer is the pattern layer and may contain T_n nodes, where T_n is the number of samples within a training data set and each node represents the input vector, p_j , associates with the j^{th} sample (or observation) in training data. The signals of the pattern neuron i , going into the third layer (summation layer), where the denominator neurons are weighted with the corresponding values of the training samples. The

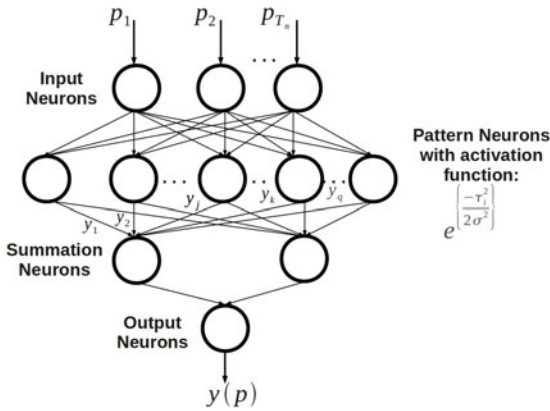


Fig. 5. GRNN architecture

network computes the most probable value of an output 'y' given only training vectors 'p'. The expected value of 'y' given by the input vector 'p' is given by

$$E\left[\frac{y}{p}\right] = \int_{-\infty}^{\infty} y P(p, y) / \int_{-\infty}^{\infty} P(p, y) \tag{8}$$

The probability estimator $P_E(p, y)$ is based on sample values p_j and y_j of random variables p and y is given as

$$P_E(p, y) = \frac{1}{(2\pi)^{(d^2 \times 1)/2} \sigma^{d^2 \times 1}} \frac{1}{T_n} \sum_{j=1}^{T_n} e^{-\frac{(p-p_j)^T (p-p_j)}{2\sigma^2}} e^{-\frac{(y-y_j)^2}{2\sigma^2}} \tag{9}$$

Now for an unknown sample u_i , the distance between the training sample and the unknown vector is given as $\tau_i^2 = (u_i - p_i)^T (u_i - p_i)$. If p_r is the point of

prediction, given by $p_{ri} = \exp\left[\frac{-r_i^2}{2\sigma^2}\right]$, the approximation for y for a given u_i can now be evaluated as

$$y = \sum_{i=1}^{T_n} p_{ri} u_i / \sum_{i=1}^{T_n} p_{ri} \quad (10)$$

GRNN, implements the above theory in a neural network. The network acts by taking a weighted average between target vectors whose design input vectors are closest to the new input vector.

5 Results of Eye Detection

In the testing stage for a test video file, the face detection has been made in prior to reduce the searching zone for eyes. After face localization a small template of size $15 \times 15 (d^2 = 225)$ is taken at the starting point $(1, 1)$ for each frame. Then for each frame this template is slid over the face region by shifting an amount of one pixel first in row wise and then column wise. The extracted vector (of dimension 225×1) is then fed to the trained GRNN as unknown vector. For each location GRNN predicts a value in between 0 to 1 according to the Eq. (10). After proper training the hope is that there are two locations in the face images where GRNN should provide the maximum values. These two locations are nothing but the eyes. The above procedure is tested over the whole database with different person's video clips. Fig(6) shows the precise eye location of one person of the database while the faces are with expression variations. To detect the eyes in frontal faces is much easier than pose variation as in case of expression variations the frontal faces are observed. To validate the robustness of the proposed strategy of face detection and eye localization the next experiment is performed over the video sequences where pose variations and different head orientations are prominent. Fig(7) shows the proposed strategy successfully detects the eyes as well as faces even in drastic change in pose and head orientations in video sequences. Table(I) shows the accuracy rate of both face and eye detection of five subjects in database.

Table 1. Shows both face and eye detection rate in videos of five persons under three different sessions from VidTIMIT database

Subjects		(1)	(2)	(3)	(4)	(5)
	Session					
Face detection Rate(%)	s1	98	98	100	100	98
	s2	98	98	100	98	96
	s3	98	96	96	96	98
Eye detection Rate(%)	s1	98	96	96	94	98
	s2	94	90	88	90	92
	s3	89	88	95	92	90

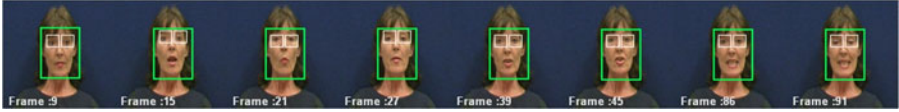


Fig. 6. Shows both the face detection and eye localization in video sequences. Some of the sequence frames are given from the resulting video file. The faces in the video frames show different expression variations

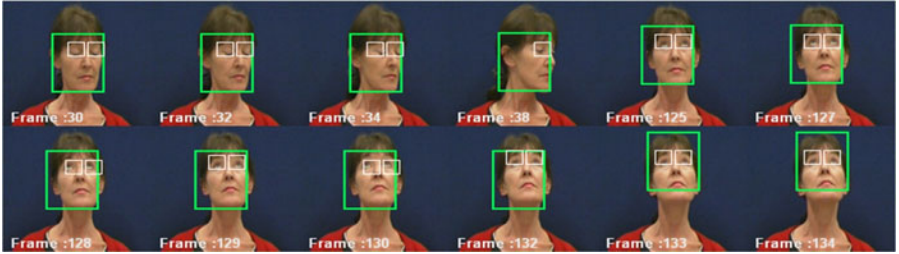


Fig. 7. Shows the precise eye localization in video faces under drastic pose variations

6 Conclusions

In this study the detection of faces in video sequences can also be treated as face tracking in video, i.e. the movement of faces for successive frames can be detected easily with the shift of peak value in each frame of the correlation volume. In similar way it can be said that the proposed system helps to track the eyes in a video file irrespective of certain degree of pose and expression variations. The reduced computational time is obtained due to the face localization by the video filter's shift invariance property. As the face acquires a small part of a video scene, the searching region for GRNN to evaluate the eye location is reduced. For a video file of 300 frames, both face detection and eye localization is obtained within 6 sec, which is very much acceptable for real time applications. In this study the face videos with constrained background are tested. The application of the proposed strategy has not been performed over the real time applications where faces are with unconstrained background and illumination variations. So further studies about this problem is necessary. How to extend the proposed strategy for real time applications with lesser time needs further investigations.

References

- [1] Mahalanobis, A., Kumar, B.V.K.V., Casassent, D.: Minimum average correlation energy filter. *Applied Optics* 26 (1987)
- [2] Mahalanobis, A., Kumar, B.V., Song, S., Sims, S., Epperson, J.: Unconstrained correlation filter. *App.Opt.* 33, 3751–3759 (1994)

- [3] Heisele, B., Poggio, T., Pontil, M.: Face detection in still gray images. Tech. rep., Artificial Intelligence Laboratory, MIT (2000)
- [4] Moghaddam, B., Pentland, A.: Probabilistic visual learning for object representation. *IEEE Trans. Pattern Analysis and Machine Intelligence* 19(7), 696–710 (1997)
- [5] Kumar, B.V.K.V.: Minimum variance synthetic discriminant functions. *J. Opt. Soc. Am.* 3 (1986)
- [6] Catalano, G., Gallace, A.: B.Kim, Pedro, S., F.Santoro: Optical flow. Tech. rep. (March 2009), <http://www.cvmt.dk/education/teaching/f09/VGIS8/AIP/>
- [7] Specht, D.F.: A general regression neural network. *IEEE Transaction on Neural Networks* 2, 568 (1991)
- [8] Maio, D., Maltoni, D.: Real-time face location on grayscale static images. *Pattern Recognition* 33, 1525–1539 (2000)
- [9] Figue, J., Rfrgier, P.: Optimality of trade-off filters. *Applied Optics* 32(11), 1933–1935 (1993)
- [10] Rowley, H.A., Baluja, S., Kanade, T.: Neural network based face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence* 20(1), 23–38 (1998)
- [11] Heo, J., Savvides, M., Abiantun, R., Xie, C.: Face recognition with kernel correlation filters on a large scale database. In: *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. II, p. 181 (2006)
- [12] Yang, J., Waibel, A.: A real time face tracker. In: *Proc. of the Third IEEE Workshop on Applications of Computer Vision*, pp. 142–147 (1996)
- [13] Mikolajczyk, K., Choudhury, R., Schimd, C.: Face detection in a video sequence -a temporal approach. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 96–101 (2001)
- [14] Silva, L., Aizawa, K., Hatori, M.: Detection and tracking of facial features. In: *Proc. of SPIE Visual Communications and Image Processing*, Taiwan
- [15] Rodriguez, M.D.: Ahmed, J., M.Shah: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: *IEEE conference on Computer Vision and Pattern Recognition*, pp. 1–8 (June 2008)
- [16] Savvides, M., Kumar, B.V., Khosla, P.K.: Robust shift invariant biometric identification from partial face images. In: *Proc. of SPIE Defense and Security Symposium*, vol. 156 (2004)
- [17] M.Savvides, Venkataramani, Kumar, B.: Incremental updating of advanced correlation filters for biometric authentication systems. In: *Proc. of IEEE Int. Conf. on Multimedia and Expo*. vol. III, p. 229 (2003)
- [18] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: *Proc. Conf. Computer Vision and Pattern Recognition*. pp. 511–518 (2001)
- [19] Sanderson, C., Paliwal, K.: Polynomial features for robust face authentication. In: *IEEE International Conference on Image Processing (ICIP)*. vol. 3, pp. 997–1000 (2002), <http://itee.uq.edu.au/~conrad/vidtimit/>
- [20] Xu, T.-Q., Li, B.C., Wang, B.: Face detection and recognition using neural network and hidden markov models. In: *Proceedings of the 2003 International Conference on Neural Networks and Signal Processing*, pp. 228–231 (2003)
- [21] Liu, Z., Wang, Y.: Face detection and tracking in video using dynamic programming. In: *Proc. International Conference Image Processing* (2000)
- [22] Zhang, Z., Potamianos, G., Liu, M., Huang, T.: Robust multi-view multi-camera face detection inside smart rooms using spatio temporal dynamic programming. In: *7th International Conference on Automatic Face and Gesture Recognition*, pp. 407–412 (April 2006)

Secret Image Sharing Using Steganography with Different Cover Images

Noopa Jagadeesh, Aishwarya Nandakumar, P. Harmya, and S.S. Anju

Centre for Cyber Security, Amrita Vishwa Vidyapeetham, Coimbatore, India
{noopajagadeesh, aishwarya.nk12,
harmya.gopalan, anjuss.cys}@gmail.com

Abstract. A novel approach to secret image sharing based on a (t, n) -threshold scheme with steganography is proposed. A secret image is first processed into n -shares which are then hidden in n -user selected different cover/camouflage images. Any t out of n participants can cooperate to reveal the secret data. The important essential of secret image sharing approaches is that the revealed secret image must be lossless. In the earlier visual secret sharing (VSS) scheme, the secret image can be shared by generating n random like images, called shadows or shares. The produced shadows can be transmitted instead of the original secret image. Once involved participants stack more than t shadows, the secret image can be revealed by visual perception without computation. But the generated shadows are often meaningless. A malicious intruder may be attracted to such meaningless shadows delivered over an insecure channel. To handle such meaningless threat, the steganography approach is utilized to embed the shadows in different cover images, called stego images. From visual perception, the content of the stego image is meaningful and can conceal the shadows from intruders. This scheme is a novel image sharing technique that satisfies all of the essentials of the traditional secret sharing scheme. The use of n different cover images further prevent the intruders from imagining a secret being transferred when compared to the shares being embedded into n similar cover images.

Keywords: secret sharing, steganography, modulo operator, visual cryptography.

1 Introduction

The continuing improvements in computer technologies and the increase in Internet usage are responsible for the increasing popularity of network-based data transmission. In many important applications, such as the communication of commercial affairs or military documents, the images must be kept secret. Many image-protection techniques, such as data encryption and steganography have been proposed to increase the security of secret images. However, one common defect of all these techniques is their policy of centralized storage, in which an entire protected image is usually maintained in a single information carrier. If a cracker detects an abnormality in the information carrier in which the protected image resides, he/she try intercepting it, attempting to decipher the secret inside, or simply ruin the entire

information carrier (and once the information carrier is destroyed, the secret image is also lost forever). Further the transmission of secret through the same cover image further increases the suspicion of the intruder that some data is been transmitted. This brings the importance of sharing the secret using different types of cover images. Secret image sharing proposed is a protection mechanism that does not suffer from these problems. It works by splitting the secret image into n shadow images that are transmitted and stored separately. One can reconstruct the original image if at least a preset number of these n shadow images are obtained; but knowledge of less than t shadow images is insufficient for revealing the secret image.

A well-known principle in the analog world is the term reduced trust, meaning that in order to keep a secret, the less knowledge or power each entity has the better. This is the basic philosophy, and *secret sharing* or *secret splitting* or *shared control* is a method to achieve this in the digital world. The secret sharing mechanism [1-2] has been widely applied to share a secret key. In this mechanism, each participant has a private shadow; some authorized participants with integrated shadows can cooperate to recover the secret key. The purpose of secret sharing is to recover the secret key while some shadows are lost, distorted, or stolen. In 1979, Blakely and Shamir introduced the (t, n) -threshold secret sharing system. In this scheme, a dealer can encode and divide secret data into n shadows. The dealer then distributes these shadows to the involved participants. With any t out of n shadows, authorized participants can cooperate to reveal the secret data accurately.

Utilizing the (t, n) -threshold concept, Noar and Shamir designed a secret image sharing technique known as visual secret sharing (VSS). Using the VSS technique, the secret image can be shared by generating n random like images, called shadows or shares. The produced shadows are transmitted instead of the original secret image. Once involved participants collect and stack more than t shadows, the secret image can be revealed by visual perception without computation. The VSS technique, however, is applied to binary images due to this stacking property. These generated shadows often lead to problems of meaningless [3, 4]. A malicious intruder may be attracted to such random-like shadows delivered over an insecure channel.

To handle such meaningless threat, the steganography approach is utilized to camouflage the shares in cover images, called stego images (shadow images) [5-9]. The proposed scheme uses n different cover images instead of a same cover. From visual perception, the content of the stego image is meaningful and can conceal the shadows from intruders.

The rest of this paper is organized as follows. A brief introduction of various secret sharing schemes is given in Section 2. The proposed scheme is given in Section 3. The experimental results and analysis in Section 4. Finally, the conclusions and future work in Section 5.

2 Secret Sharing

Secret sharing schemes are normally set up by trusted authority, which computes all shares and then distributes them to participants via secure channels. The trusted authorities that setup the scheme is called a dealer. The participants hold their shares until some/ all of them decide to pool/stack their shares and recreate the secret. The

combiner, who on behalf of the cooperating group computes the secret, does the recovery of the secret. The combiner can be a mutually trusted participant who collects all shares, calculates the secret, and distributes it secretly to the active participants. The various secret sharing schemes are

2.1 (t, t) Threshold Schemes

The secret can be recovered only when all participants cooperate. Let the secret integer k be given. The dealer chooses a modulus p that can be any integer greater than k . Its value determines the security parameter. Next the dealer selects randomly, uniformly and independently $(t - 1)$ elements. S_1, S_2, \dots, S_{t-1} from Z_p . The share S_t is given by

$$S_t = k - \sum_{i=1}^{t-1} S_i \pmod{p} \tag{1}$$

The shares are distributed securely to the participants from the set $P = \{P_1, P_2, \dots, P_t\}$. At the pooling time, the combiner can reconstruct the secret only if he/she is given all shares as

$$k = \sum_{i=1}^t S_i \pmod{p} \tag{2}$$

Obviously, any $(t - 1)$ or fewer shares provide no information about the secret k .

2.2 (t, n) Secret Sharing

A (t, n) Shamir scheme is constructed by the dealer Don. First Don chooses n different points $x_i \in GF(p)$, for $i=1, 2, \dots, n$. These points are public. Next Don selects at random coefficients a_0, a_1, \dots, a_{t-1} from $GF(p)$. The polynomial $f(x) = a_0 + a_1 x + \dots + a_{t-1} x^{t-1}$ is of degree at most $(t-1)$. The shares are $S_i = F(x_i)$ for $i=1, 2, \dots, n$, and the secret is $k = F(0)$. The share S_i is distributed to the participant $P_i \in P$ via a secure channel and is kept secret. When t participants agree to cooperate, the combiner Clara takes the shares and tries to recover the secret polynomial $F(x)$. She knows t points on the curve $F(x)$.

$(x_i, F(x_i)) = (x_i, S_i)$ for $i=1, 2, \dots, t$. These points produce the following system of equation:

$$\begin{aligned} S_1 &= a_0 + a_1 x_1 + \dots + a_{t-1} x_1^{t-1} \\ S_2 &= a_0 + a_1 x_2 + \dots + a_{t-1} x_2^{t-1} \\ &\dots\dots\dots \\ S_t &= a_0 + a_1 x_t + \dots + a_{t-1} x_t^{t-1} \end{aligned} \tag{3}$$

The system has a unique solution for $(a_0, a_1, \dots, a_{t-1})$ since the corresponding Vandermonde determinant is different from zero. The Lagrange interpolation formula allows us to determine the polynomial $F(x)$ of degree $(t-1)$ from the t different points

$$(x_i, S_i) .$$

2.3 Modular Scheme

Assume that every participant $P_i \in P$ is assigned a public modulus P_i , $i = 1, 2, \dots, n$. The modulo can be primes or mutually co primes. Let the modulo be such that $P_1 < P_2 < \dots < P_n$. The dealer selects at random an integer S such that $0 < S < \prod_{i=1}^t P_i$. The secret $k \equiv S \pmod{P_0}$. Next the dealer distributes shares $S_i \equiv S \pmod{P_i}$ to the participants P_i ($i = 1, 2, \dots, n$) via secure channels. Assume that there is t or more participants who want to recreate the secret. The combiner takes their shares $S_{i1}, S_{i2}, \dots, S_{it}$ and solves the following system of congruence using Chinese remainder theorem. The secret $k \equiv S \pmod{P_0}$.

2.4 Proactive Secret Sharing

There is a need for scheme that allows servers to generate a new set of shares for the same secret from the old shares without reconstructing the secret. Such a scheme is called as an *proactive secret scheme* (PSS). In reality, compromise to a server are very hard to detect, especially when the attacker/intruder simply steals certain secret information without modifying anything on the victim server. To strengthen the security of a replicated service, we can invoke PSS periodically (at regular intervals). Before the start of execution of PSS, every server checks the integrity of its code and state, and there by trying to remove any attackers that might exist in that server at that point in time.

To show how a proactive scheme can be achieved, let's study an example first. Let us first assume that an adversary can only break into a server and have access to information stored or collected by that server. The adversary can't change the code of the server. Suppose we have a simple (2,2) sharing scheme. To generate two shares for secret S , we randomly select S_1 and S_2 , so that $S_1 + S_2 = S$. We want the two servers with shares S_1 and S_2 to change their respective shares to S_1' and S_2' , so that these two shares remain an (2,2) sharing of the same secret S and these two shares are independent from the old shares. The proactive secret sharing can be performed through the following steps:

1. Server 1 generates two sub shares S_{11} and S_{12} from its share S_1 using the same secret sharing scheme as the one used to generate S_1 and S_2 from S ; that is, server 1 randomly selects two sub shares S_{11} and S_{12} , so that $S_1 = S_{11} + S_{12}$. Server 2 does the same thing to S_2 : It randomly generates two sub shares S_{21} and S_{22} , so that $S_2 = S_{21} + S_{22}$.

2. Server 1 sends S_{12} to server 2 through a certain secure channel. Server 2 sends S_{21} to Server 1.

3. Server 1 has both S_{11} and S_{21} and can add them to get a new share $S_1' = S_{11} + S_{21}$.

Server 2, on the other hand, has both S_{12} and S_{22} and can generate a new share $S_2' = S_{12} + S_{22}$. Now we show that S_1' and S_2' constitute a (2, 2) sharing. The sum of these two shares is the sum of all the four sub shares, which is the sum of S_1 and S_2 , which is S .

These two shares are independent from the old ones because these sub shares are generated randomly. Also, no servers know the secret during the entire process. Server 1 generates S_{11} and S_{12} and learns S_{21} from server 2, but server 1 never knows S_{22} and thus does not know S_2' or S . Server 2, on the other hand, never knows S_{11} , and thus does not know S_1' or S .

3 The Proposed Scheme

3.1 (t, n) Sharing Procedure

The dealer firsts selects a prime number m and assigns a unique key K_i for each participant, where $i = 1, 2, \dots, n$. To share the secret image S , the dealer converts S into the m -ary notational system. For instance, we assume that the chosen m is equal to 7. If two continuous secret pixels in S are 83 and 110, then the converted digits become $(1, 4, 6)_7$ and $(2, 1, 5)_7$.

Let us assume that the shared $(t-1)$ digits of S are S_1, S_2, \dots, S_{t-1} . Suppose that O_1, O_2, \dots, O_n be the chosen grayscale cover image with pixels $H \times W$, and P_{1i} is a pixel of O_1 . The dealer first computes the value of d as

$$d = P_{1i} \bmod m$$

If 157 is the first pixel (P_{11}) of first cover image O_1 , then value of d is 3.

With d and S_1, S_2, \dots, S_{t-1} , an invertible polynomial can be formulated as

$$F(x) = S_1 + S_2x^1 + \dots + S_{t-1}x^{t-2} + dx^{t-1} \bmod m \tag{4}$$

The dealer can thereby generate n shadows y_i by feeding the secret key K_i into $F(x)$.

$$y_1 = F(K_1), y_2 = F(K_2), \dots, y_n = F(K_n) \tag{5}$$

Suppose if dealer intends to make 3 shares then $F(x) = (1 + 4x + 2x^2) \bmod 7$ and $y_1 = 1, y_2 = 0, y_3 = 5$

In order to hide the values of y_1, y_2, \dots, y_n , n different cover images O_1, O_2, \dots, O_n are selected. For each of the cover images the following computations is done.

$$Q = \text{floor}(P_i / m) \times m, \tag{6}$$

$$P(\text{new})_i = Q + y_i, \tag{7}$$

Where P_i represent the i 'th pixel of each of the original cover images.

3.2 Secret Retrieval Procedure

Given any t out of n stego images O_i and the key K_i from the involved participants, the secret image S can be reconstructed. We first assume that $P(\text{new})_i$ is the corresponding pixel value of O_i . To extract the secret digits, authorized participants must derive the polynomial $F(x)$ from $P(\text{new})_i$. Thus, the participants utilize the modulo operation to obtain shadows y_i 's by computing

$$y_i = P(\text{new})_i \bmod m \tag{8}$$

With this obtained shadow y_i and secret key K_i , polynomial $F(x)$ can be reconstructed by Lagrange's interpolation formula:

$$F(x) = S_1 + S_2x^1 + \dots + S_{t-1}x^{t-2} + dx^{t-1} \pmod m \tag{9}$$

Authorized participants can thereby obtain the secret digits S_1, S_2, \dots, S_{t-1} by extracting the first $(t-1)$ coefficients of $F(x)$.

4 Experimental Results and Analysis

To demonstrate the performance of the novel (t, n) -threshold sharing scheme the size of grayscale cover images is set to 298×298 pixels. Figure 1 shows the shared secret image with 100×100 pixel



Fig. 1. The secret image

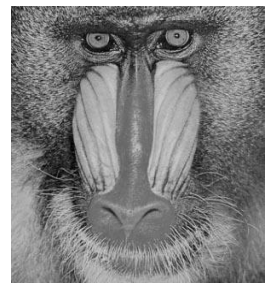
Table 1 lists the qualities of the stego images with various test cover images. This experiment is implemented in the case of the $(3,n)$ -threshold scheme. The prime number m is set as 7. In Table 1, we can see that the quality of the stego image is quite satisfactory for different secret images and stego images. We display three stego images in Figures. 2(a), 2(b) and 2(c). Judging from the visual perception of these three stego images, our scheme can successfully protect the shares from intruders. Authorized participants can later extract the secret image from these three different stego images.



(a)The stego image 1



(b)The stego image 2



(c)The stego image 3

Fig. 2. The resulting shares

Table 1. The PSNR of the shadow images for various images

Share1	Share2	Share3
Naturals=47.60	Lena=46.24	Baboon=45.20
Clown=46.53	Lena=46.21	Naturals=47.40
Pepper=45.63	Baboon=45.83	Naturals=47.00
Tiffany=43.38	Splash=37.43	Lena=47.42

5 Conclusion and Future Works

A common drawback of image sharing schemes using steganography approaches is that the revealed secret image is distorted. Although the distortion is small, it is unacceptable for significant secret content. In this approach, a novel sharing scheme that can reveal the lossless secret image and satisfy related sharing essentials is used. The secret image is revealed from any t of n different stego images. The lossless of the novel sharing scheme is a practical essential to preserve valuable secret images, such as military and medical images.

The future works include extending the secret sharing scheme for color images and sharing multiple secret. The (t,n) secret sharing scheme can be extended further in such a way that each authorized participant can be given a privilege/access level.

Acknowledgment. Special thanks to Chin-Chen Chang, Pei-Yu Lin, Chi-Shiang Chan authors of “Secret image sharing with revertible steganography”. We also extend our thanks to Cyber Security Department, Amrita Vishwa Vidyapeetham, Coimbatore.

References

1. Shamir, A.: How to share a secret. *Communications of the ACM* 22(11), 612–613 (1979)
2. Naor, M., Shamir, A.: Visual cryptography. In: De Santis, A. (ed.) *EUROCRYPT 1994*. LNCS, vol. 950, pp. 1–12. Springer, Heidelberg (1995)
3. Wang, R.Z., Su, C.H.: Secret image sharing with smaller shadow images. *Pattern Recognition Letters* 27(6), 551–555 (2006)
4. Chang, C.C., Lin, C.C., Lin, C.H., Chen, Y.H.: A novel secret image sharing scheme in color images using small shadow images. *Information Sciences* 178(11), 2433–2447 (2008)
5. Tsai, C.S., Chang, C.C., Chen, T.S.: Sharing multiple secrets in digital images. *The Journal of Systems and Software* 64(2), 163–170 (2002)

6. Thien, C.C., Lin, J.C.: Secret image sharing. *Computer & Graphics* 26(1), 765–770 (2002)
7. Lin, C.C., Tsai, W.H.: Secret image sharing with steganography and authentication. *The Journal of Systems and Software* 73(3), 405–414 (2004)
8. Wu, Y.S., Thien, C.C., Lin, J.C.: Sharing and hiding secret images with size constraint. *Pattern Recognition* 37(7), 1377–1385 (2004)
9. Chang, C.-C., Lin, P.-Y., Chan, C.-S.: Secret image sharing with revertible steganography., *International Conference on Computational Intelligence and Natural Computing* (2009)

A Secure Data Hiding Scheme Based on Combined Steganography and Visual Cryptography Methods

Aishwarya Nandakumar, P. Harmya, Noopa Jagadeesh, and S.S. Anju

Centre For Cyber Security, Amrita Vishwa Vidyapeetam, Coimbatore, India
{aishwarya.nk12, harmya.gopalan, noopajagadeesh,
anjuss.cys}@gmail.com

Abstract. Steganography and Visual Cryptography are two areas in which many studies and intensive research have been carried out. Even though the basic idea of both areas in hiding information are similar, steganography and visual cryptography makes use of different methodologies in order to protect their data. In this paper, a method is proposed for combining steganography and visual cryptography. In the proposed method, secret data is embedded using Matrix embedding technique using Hamming codes and shares are generated from this stego image using Random Grids method. The main advantage of Random grids method is there is no pixel expansion; hence embedded secret data can be fully recovered on stacking the shares. This idea of combining steganography and visual cryptography can present a great area for research especially in the world of forensics.

Keywords: Matrix Embedding using Hamming codes, Random Grids, Visual Cryptography, Steganography.

1 Introduction

Steganography is the art of sending information through files in a manner that the very existence of the message is unknown. Steganography is a form of security through obscurity where the security lies in the fact that only sender and receiver know the method in which the message is hidden. While Cryptography is about protecting the messages (their meaning), steganography is about hiding the message so that intermediate persons cannot see the message.

Visual cryptography is a technique which allows visual information (e.g. printed text, handwritten notes and pictures) to be encrypted in such a way that the decryption can be performed by the human visual system, without the aid of computers. Visual cryptography scheme eliminates complex computation problem in decryption process, and the secret images can be restored by stacking operation. In the (r,n) Basic VC scheme[4], the input image is transformed into n noise-like shares to ensure that the contained secret is unreadable. These shares can be printed on transparent slides and distributed to the participants. Any subset of r or more shares can decrypt the secret in the original image, but no information about the secret can be obtained from fewer shares. The decryption process in a VC scheme involves inspecting the stacked shares

with the unaided eye without computation. The ciphering model of VC is similar to a one-time pad in the sense that each image is decrypted with a different set of shares, and provides high security to the protected secrets.

Visual cryptography and steganography are somewhat similar in concept[5]. Both steganography and visual cryptography algorithms have been efficiently used for hiding data. We can further increase security, reliability and efficiency by combining steganography and visual cryptographic techniques. In the traditional visual cryptography schemes, pixel expansion takes place; which will lead to distortion of embedded data. So we need to find a visual cryptographic technique in which there is no pixel expansion.

Kafri and Keren[2] proposed a picture encryption idea using random grid (RG) as the basis of the generated shares. A characteristic of image encryption by RG is that the size of each generated RG is the same as that of the original secret image. They defined a RG as a transparency consisting of a two-dimensional array of pixels. Each pixel can be fully transparent or totally opaque, and the two choices are equally likely events.

For embedding the hidden data, we use matrix embedding[1]. It is a general coding method that can be applied to most steganographic schemes to improve their embedding efficiency, the number of message bits embedded per one embedding change. Because small number of embedding changes is less likely to disrupt the statistic properties of the cover image, schemes that employ matrix embedding generally have better steganographic security.

The rest of this paper is organized as follows. Section 2 describes Random Grids method. Matrix embedding technique using Hamming codes is described in section 3. The next section discusses the proposed method. Experimental Results are shown in section 5. Conclusion and Future work is described in section 6

2 Random Grids Method

A random grid (RG) is a transparency comprising a two-dimensional array of pixels. Each pixel p can be fully transparent ($p=1$, the light can transmit it) or totally opaque ($p=0$, the light will stop on it), and the choice between the alternatives is by a random process with an equal probability[2,3,7]. Hence correlation will not be there between the values of different pixels in the array. The transparent pixels allow the light to pass through while the opaque pixels stop it. Since probabilistically the number of the transparent pixels is equal to that of the opaque pixels in a random grid, the average light transmission of a random grid is $1/2$. When two RGs with the same dimensions are well-stacked together pixel by pixel, the light transmission rate of the superimposed image can be expressed in terms of three cases: (1) When the two RGs are independent, the average light transmission is $1/4$; (2) When the two RGs are identical, the average light transmission is $1/2$; and (3) When the two RGs are complementary to each other, the average light transmission is 0. The Main advantage of this method is, the size of the share is same as that of original image.

Taking the three light transmission rates of different configurations of two RGs mentioned above, three different algorithms are obtained [2,3,7].

We summarize the steps of the three algorithms below

Input: Binary Secret Image P
Output: Two random grids G1 and G2

2.1 Algorithm 1

- Step 1: Generate a random grid G1 which has the same dimension with P.
- Step 2: Fetch a not-yet-processed pixel p from P according to the scanning order (e.g., from left to right and from top to bottom).
- Step 3: Examine the value of p, and then proceed with one of the following sub steps:
 - Step 3.1 If p is a black pixel, copy the corresponding pixel (the pixel in G1 at the same location relative to p in P) of G1 to the corresponding pixel of G2.
 - Step 3.2 If p is a white pixel, assign the complement value of the corresponding pixel of G1 to the corresponding pixel of G2.
- Step 4: Repeat steps 2 and 3 until all pixels in P are processed.

2.2 Algorithm 2

- Step 1: Generate a random grid G1 which has the same dimension with P.
- Step 2: Fetch a not-yet-processed pixel p from P according to the scanning order.
- Step 3: Examine the value of p, and then proceed with one of the following sub steps:
 - Step 3.1 If p is a black pixel, copy the corresponding pixel of G1 to the corresponding pixel of G2.
 - Step 3.2 If p is a white pixel, generate a random number from {0,1} and assign it to the corresponding pixel of G2.
- Step 4: Repeat steps 2 and 3 until all pixels in P are processed.

2.3 Algorithm 3

- Step 1: Generate a random grid G1 which has the same dimension with P.
- Step 2: Fetch a not-yet-processed pixel p from P according to the scanning order.
- Step 3: Examine the value of p, and then proceed with one of the following sub steps:
 - Step 3.1 If p is a black pixel, generate a random number from {0, 1} and assign it to the corresponding pixel of G2.
 - Step 3.2 If p is a white pixel, assign the complement value of the corresponding pixel of G1 to the corresponding pixel of G2.
- Step 4: Repeat steps 2 and 3 until all pixels in P are processed.

An Example of Random Grid Method using Algorithm 3 of the secret mage shown in Figure 1 is given Figure 2.

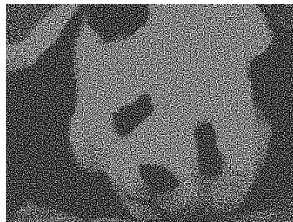


Fig. 1. Secret Image

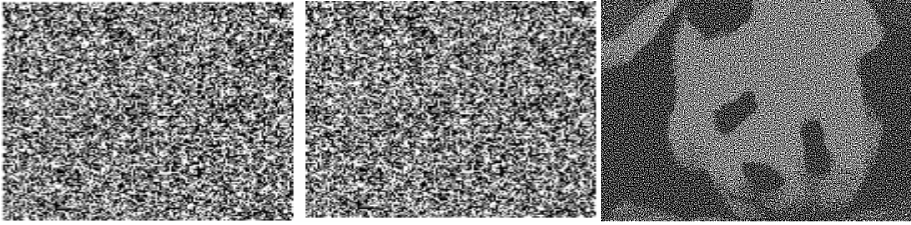


Fig. 2. Example of Random Grids Method

3 Matrix Embedding Using Hamming Codes

In this section, we describe binary Hamming codes [6] and explain how they can be used for matrix embedding [1]. Binary Hamming codes are $[2^p-1, 2^p-1-p]$ linear codes with parity check matrix H of dimensions $p \times 2^p-1$ whose columns are binary expansions of numbers $1, \dots, 2^p-1$.

For example, the parity check matrix H for $p = 3$ is

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

Let us assume that the cover object is an image comprising of N pixels. Most steganographic schemes assign a bit to each possible pixel value, for example, as the LSB of the grayscale value. The embedding then usually proceeds by changing the pixel values to match their assigned bits to the desired message bits. To do so, one might for example flip the least significant bit of the pixel grayscale value. Assuming the embedded message is a random bit-stream, the probability that each pixel will have to be changed is 0.5. Thus, on average we embed 2 bits per embedding change. We can also say that the scheme has 2 as the embedding efficiency.

The embedding efficiency can be improved using matrix embedding, by dividing the cover image into N/n subsets; each consisting of n pixels, where n is the length of an appropriately chosen code. For matrix embedding using the binary Hamming code, $n = 2^p-1$. We now show that we can embed p message bits in each subset by making at most one embedding change.

3.1 The Algorithm

Consider p message bits to be embedded in 2^p-1 pixels [1]. We denote by x the vector of LSBs of 2^p-1 pixels from the cover image. The sender and receiver share a $p \times 2^p-1$ binary H matrix that contains all non-zero binary vectors of length p as its columns. The sender modifies the pixel values so that column vector of their LSB's, y , satisfies $m = Hy$. If by chance the syndrome of the cover pixels already communicates with the correct message, i.e. $Hx = m$, the sender does not need to modify any of the 2^p-1 cover

pixels, set $y=x$ and proceeds to the next block of 2^p-1 pixels and embeds the next segment of p message bits. Figure 3 shows the algorithm for Matrix Embedding using Random Grids

When $Hx \neq m$, the sender looks up the difference $Hx = m$, as a column in H . Let it be the j th column and represent it as $H[.,j]$. By flipping the LSB of the j th pixel and keeping the remaining pixels unchanged,

$$Y[j]=1-x[j]$$

$$Y[k]=x[k], k \neq j$$

The syndrome of y now matches the message bits, $Hy=m$. This is because $Hy=Hx+H(y-x)=Hx-m+H[.,j]+m=m$, because condition $Hx-m=H[.,j]$ and in binary arithmetic $z+z=0$ for any z .

The recipient follows the same path through the image as the sender and reads p message bits from the LSB of each block of 2^p-1 pixels as the syndrome $m=Hy$.

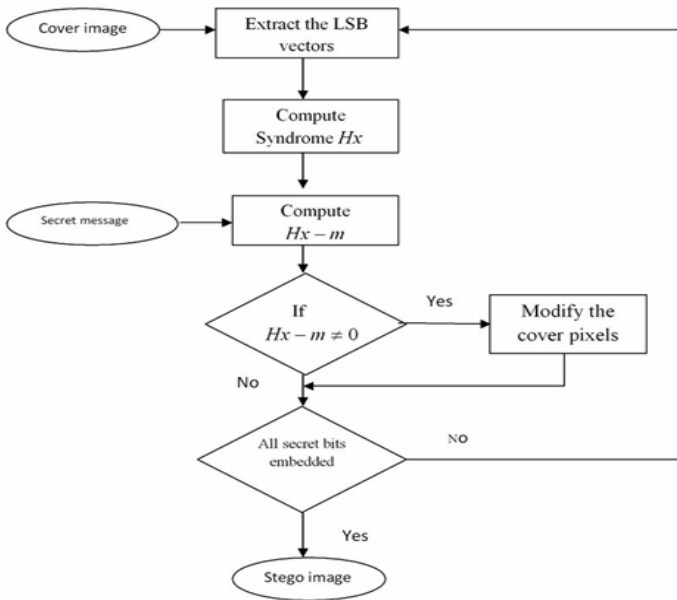


Fig. 3. Matrix embedding using Hamming codes

Thus, assuming m is a random bit-stream, we make on average $1-2^{-p}$ embedding changes per block during embedding. Therefore, the embedding efficiency e_p , defined as the number of random bits embedded per one embedding change is

$$e_p = p / (1 - 2^{-p}) \tag{1}$$

Since we are embedding p bits into 2^p-1 pixels, the relative message length is $p / (2^p-1)$.

4 Proposed Method

In the proposed method we used Matrix embedding using Hamming codes for embedding the secret data and Shares are generated using Random Grids Method. The system architecture is as shown in Figure 4. We embed the Secret data using Matrix Embedding using Hamming Codes. A (2,2) Visual cryptography scheme using Random Grids is used for generation of shares from the stego image. For reconstruction of stego image we stack the two shares; stacking refers to Boolean “OR” operation. Once the stego-image is recovered, we extract the secret data using extraction algorithm.

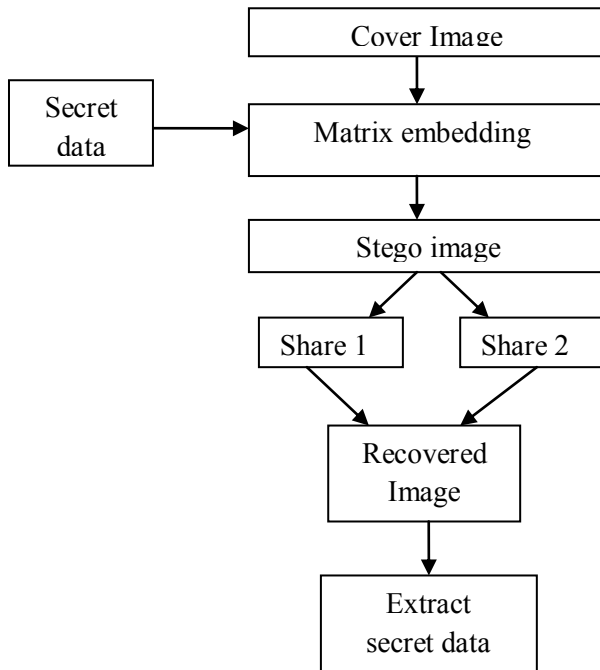


Fig. 4. System Architecture

5 Experimental Results

The results of simulation of proposed method in MatLab are given in the following section. The cover image used was a binary image of Lena of size 256 X 256 shown in Figure 5a .The secret data was “Hello World”. After embedding the stego-image obtained is as shown in Figure 5b.The shares of the stego-image is generated and the image is recovered by stacking the two shares which is shown in Figure 5c (a)-(b). The recovered image was given to extraction algorithm and secret data “Hello world” was successfully recovered.



Fig. 5a. Cover Image



Fig. 5b. Stego Image

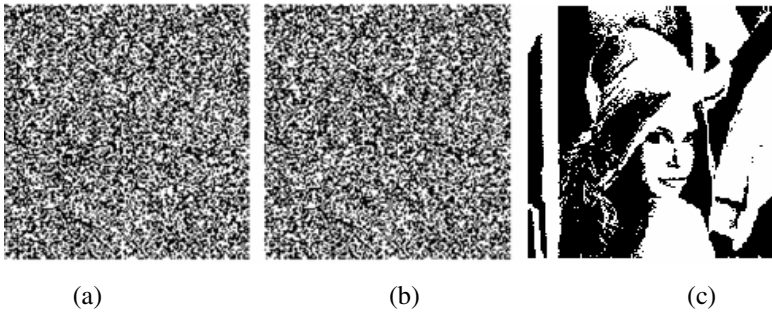


Fig. 5c. (a) Share1 (b) Share2 (c) Recovered Image

6 Conclusion and Future Work

In the paper we have presented a scheme that combines steganographic method with a visual cryptographic method was proposed. We discussed matrix embedding method using Hamming codes for embedding the data and Random Grids Method for generation of the shares. The image was recovered by stacking the shares and the secret data was successfully recovered. This idea of combining steganography and visual cryptography can present a great area for further exploration which would open up some more venues in the world of forensics and anti-forensics.

The Future work comprises of developing methods for grey scale and Color images. Also develop algorithms combining other steganographic and visual cryptographic methods that can increase efficiency, reliability and security for sending hidden data.

Acknowledgements. The authors sincerely thank Jessica Fridrich and David Soukal for their valuable contribution to the area of steganography. Special thanks to Kafri and Keren for their incredible work on Random grids.

References

1. Fridrich, J., Soukal, D.: Matrix Embedding for large payloads. *IEEE Transactions on Information, Security and Forensics* 1(3), 390–395 (2006)
2. Kafri, O., Keren, E.: Encryption of pictures and shapes by random grids. *Optical Letters* 12, 377–379 (1987)
3. Shyu, S.J.: Image encryption by random grids. *Pattern Recognition* 40, 1014–1031 (2006)
4. Naor, M., Shamir, A.: Visual cryptography. In: De Santis, A. (ed.) *EUROCRYPT 1994*. LNCS, vol. 950, pp. 1–12. Springer, Heidelberg (1995)
5. Abboud, G., Marean, J., Roman, Yampolskiy, V.: *Steganography and Visual Cryptography in Computer Forensics*. In: *Fifth International Workshop on Systematic Approaches to Digital Forensic Engineering* (2010)
6. Petitcolas, F.A.P., Anderson, R.J., Kuhn, M.G.: Information Hiding-A Survey. *Proceedings of the IEEE, special issue on protection of multimedia content* 87(7), 1018–1062 (1996)
7. Wang, R.Z., Lan, Y.C., Lee, Y.K., Huang, S.Y.: Incrementing visual cryptography using Random Grids. *Optics Communications* 283, 4242–4249 (2010)

Cognitive Environment for Pervasive Learners

Sattvik Sharma, R. Sreevathsan, M.V.V.N.S. Srikanth, C. Harshith,
and T. Gireesh Kumar

Amrita Vishwa Vidyapeetham, Coimbatore, Tamilnadu, India
{sattviksharma, sreevathsan.ravi, mvvnssrikanth,
smilealways.achu, gireeshkumart}@gmail.com

Abstract. We present a novel approach for taking the ordinary classroom teaching to a new level, by creating a cognitive environment which includes the use of an image sensing device, number of client/student systems and a master/faculty system connected over a network. An automatic face detector and recognizer are activated on all the client systems for taking the attendance. The detection process is based on the adaboost algorithm, which is a cascade of binary features to rapidly locate and detect faces; recognition is achieved using principle component analysis. Hand gesture detection for students to raise doubts is achieved using adaboost algorithm. The system can also detect whether the students are asleep by extracting the eye region alone and applying principle component analysis to classify whether eyes are closed or open. Kalman filter is used to track the detected eye in consecutive frames. Experimental results show that our system is a viable approach and achieves good detection and recognition rates across wide range of head poses with different lighting conditions.

Keywords: Adaboost, PCA, Kalman Filter, Eigen faces, Haar features, Haar cascade.

1 Introduction

Technology has brought about a revolution in our lives, to add to this it has been continuously evolving and penetrating in all the walks of life. Here in this paper we propose an innovative system which instills technology to our class rooms to make it more interactive and automates many processes which can reduce the burden of the teachers. In many human computer interactive systems, face and gesture recognition were achieved with the assistance of specialized devices (e.g., data glove, markers, etc.), but this paper focalizes on interacting with the system without any of the special devices [11]. The most impressive paper, Viola and Jones [7] introduces the concept of an "integral image", along with a rectangular feature representation and a boosting algorithm as its learning method to detect faces at 25 frames per second. This showed an improvement in computation time by an order of magnitude over previous implementations of face detection algorithms. The performance of the detector is further improved with help of kalman filter by Paulo Menezes et al. [12]. In the field of recognition, the work of Turk and Pentland [3] is PCA technique and is widely used.

The paper is designed as follows. The proposed system is categorized under auto-attendance module, student-teacher interaction module and auto-vigilance module which are covered under Section 2. The real time implementation of the system is covered under Section 3 and whose results are discussed in Section 4. Section 5 concludes the paper with the gist about future augmentation works.

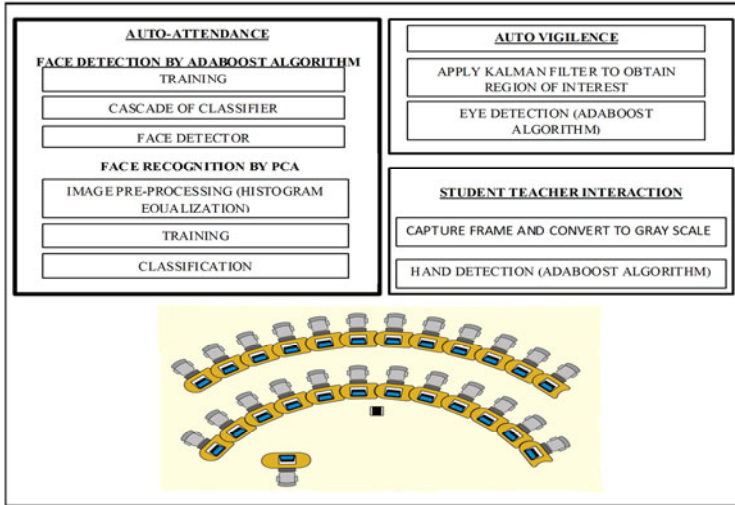


Fig. 1. System Design

2 Proposed System

The cognitive environment is based on the client server architecture where a classroom consists of a cluster of computers connected over a LAN. This establishes a two way communication between the client and server. The Smart class room system is implemented with an assumption that every personal computer in the network has the software installed and a webcam attached to it. The various modules of the system are shown in Fig 1. Starting with a convenient auto attendance, the attendance is achieved without any manual help. During the class session the system's auto-vigilance module checks the student's alertness in the class, without the teacher pointing it. The student-teacher interaction module is achieved through the gestures captured by the system. At the end of the session the system beeps an alarm to intimate the teacher about end of his/her session.

2.1 Automatic Attendance

The framework behind the automatic attendance module is shown in Fig 2. The face is detected using the Adaboost Algorithm and the cropped face is sent for recognition using PCA. If the face is recognized in ten consecutive frames, then the details of the recognized person is sent to the master system. This information is appended into a log file to prepare the attendance list.

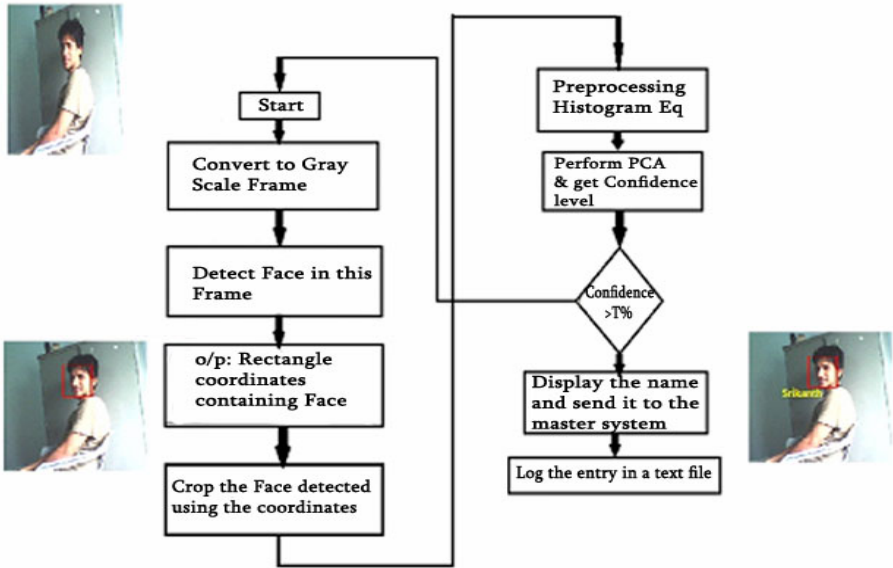


Fig. 2. Framework of Automatic Attendance Module

2.1.1 Face Detection

The primary task of this module is face detection which is done with the help of Adaboost algorithm[6]. The algorithm comprises of these concepts (1)Haar-like Features,(2)Integral image concept, (3)Learning Classifier Functions and (4)Classifier cascade. The actual face detection is performed by classifier cascade with the help of features obtained from training.

2.1.1.1 *Haar-Like Features.* The features are similar to the basis functions in Haar wavelets. The features used for face detection are simple Haar-like rectangular features as shown in Fig. 3.



Fig. 3. The two, three rectangular Haar-like features and how they are placed over the image

2.1.1.2 *Integral Image.* A major advantage of using these rectangular features is that they can be computed very quickly using the concept of integral image. The value in the integral image at the pixel (x, y) is the sum of all the pixels to the left and above (x, y) in the original test image:

$$ii(x, y) = \sum_{x \leq x', y' \leq y} i(x', y')$$

where $ii(x, y)$ is the integral image and $i(x, y)$ is the actual image [1][6].

2.1.1.3 Learning Classifier Functions. For all the training data, labelled as faces and non-faces, all the possible features are computed, along with their corresponding optimal thresholds that minimize their individual misclassification errors. Among all these features, the one that has the least error is selected as a “good” feature and its threshold acts as the separating boundary between faces and non-faces. Thus, a weak classifier consists of a feature (f), its threshold (θ), polarity (p), and the following hypothesis (h):

$$h_j(x) = \begin{cases} 1 & \text{if } p_j f_j(x) < p_j \theta_j \\ 0 & \text{otherwise} \end{cases}$$

The boosting algorithm during learning uses T hypotheses which are constructed each using a single feature. The final hypothesis is a weighted linear combination of the T hypotheses where the weights are inversely proportional to the training errors. Thus a strong classifier is prepared during training.

2.1.1.4 Classifier Cascade. In practice, no single strong classifier is used. Instead, a series of many such classifiers are learnt to form a cascade of classifiers. The simpler classifiers come earlier in the cascade and they can reject majority of non-face like sub-windows while retaining almost all the regions containing a face. The sub-windows that pass these earlier simpler classifiers are tougher to distinguish from faces and require more complex analysis. This is where the later stages of the cascade prove useful (Fig. 4). The final desirable false positive and detection rate governs the individual accuracy values for each of the stages.

The cascade of classifier works as per the given algorithm:

```

window_size = window_size0
scale = 1
faces = {} //no faces initially in the array
while ( window_size ≤ image_size)
do {
classifier_cascade = classifier_cascade0 scaled by
scale
dX= scale
dY= scale
for 0 ≤ Y < image_height - window_height do
for 0 ≤ X < image_height - window_height do
region_to_test = { 0 ≤ x < X + window_width;
0 ≤ y < Y + window_width}
if classifier_cascade(region_to_test) == 1
then

```

```

        faces = faces U {region_to_test}
    end if
X = X + dX
end for
Y = Y + dY
end for
scale = scalexC          /* C -scale factor, e.g. 2.1 */
end while
return faces }
    
```

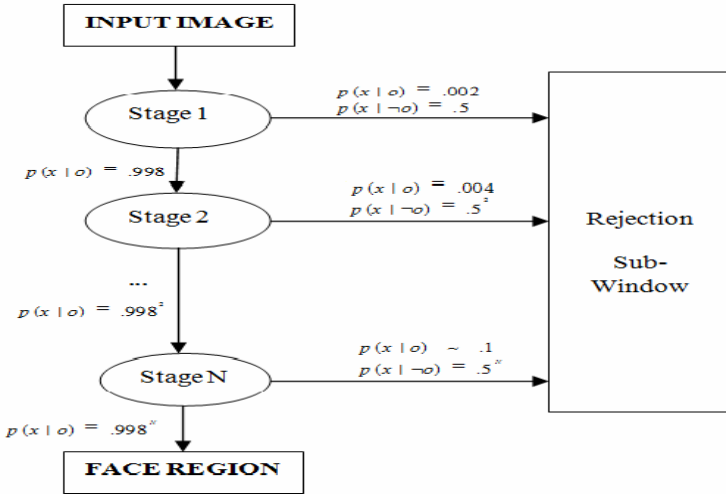


Fig. 4. Schematic description of the detection cascade where $p(x|o)$ denotes the chances that the region x contains the faces and $p(x|\neg o)$ denotes the chances that the region x does not contain faces

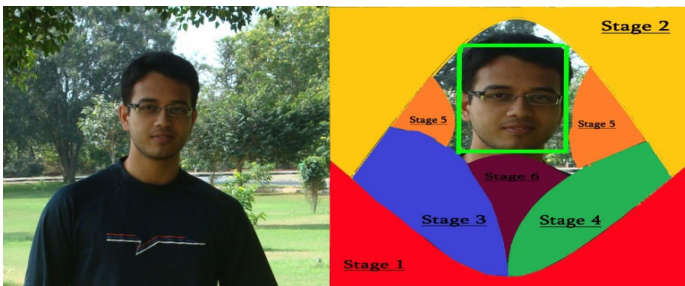


Fig. 5. The sample frame from video sent to cascade and the output at each stage of the cascade of classifier is shown and finally the detected face. The solid regions with different colours show the regions detected by the strong classifier at each stage respectively.

The output of the above algorithm is the face region present in the input image as shown in Fig. 5, which is cropped and sent for recognition.

2.1.2 Face Recognition

The PCA [3][4] used for face recognition comprises of training and classification phase. The training data is kept at client systems so as to reduce the traffic improving the performance.

2.1.2.1 Training. PCA computes the basis of a space which is represented by its training vectors[3][4]. These basis vectors, actually eigenvectors, computed by PCA are in the direction of the largest variance of the training vectors called as eigen faces. Each eigen face can be viewed a feature. During learning the following four steps are followed.

1. Load the training data.
2. Do PCA on it to find a subspace.
3. Project the training faces onto the PCA subspace.
4. Save all the training information
 - a. Eigen values
 - b. Eigen Vectors
 - c. The average training face image
 - d. Projected Faces
 - e. Person ID numbers.

2.1.2.2 Face Recognition. During recognition phase the following steps are followed:

1. Load all the training information
 - a. Eigen values
 - b. Eigen Vectors
 - c. The average training face image
2. Project the test image onto the PCA subspace.
3. Recognise the person among the training images by calculating the nearest neighbour in the subspace.

The Euclidean distance is calculated between the projected test and train examples. It is used to compute the Confidence level with which the system is able to recognize the person, this confidence level varies from 0.0 to 1.0. The person is said to be correctly recognized only if the confidence value lies greater than 0.9.

$$Confidence = 1 - \sqrt{\text{leastDistSq} / (nTrainFaces * nEigens)} / 255$$

The face recognition procedure is also self explanatory from the Fig. 6.

2.1.2.3 Automatic Vigilance. The automatic vigilance is achieved by detecting the eye and recognizing whether it is open or closed. The eye is detected using the same

adaboost algorithm and detected eye is recognized as open or closed eye through PCA. Though the same set of algorithms are used as in the previous module, eye detection and recognition differs in few ways from that of face and is discussed in following subsections. The eye is tracked for its status (opened or closed) continuously.

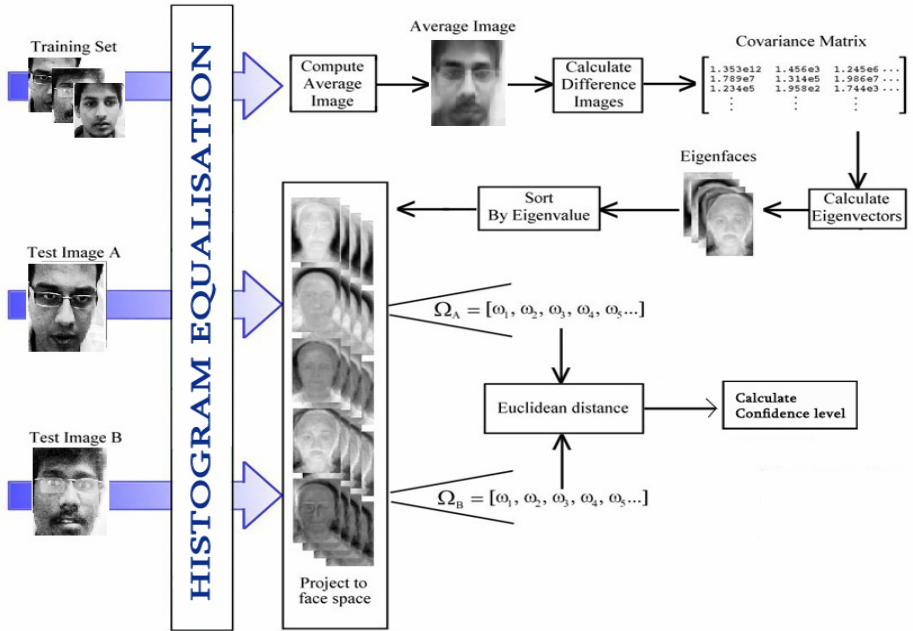


Fig. 6. PCA technique that trains and recognizes different persons

In order to track the expected position of the eye in the next frame, kalman filter is applied in successive video frames(Fig. 7.). The inclusion of a Kalman filter [12]. serves two purposes: increase the quality of the tracking and processing speed. The Adaboost eye detector is made to run across these promising regions only. Thus, the detection speed is improved. When the detected eye region is sent for recognition and if the status appears to be closed for 20 seconds, then the client is found to be sleeping and an alarm beeps in his system.

2.2 Student Teacher Interaction

The software helps the students to raise doubts in the class and have an interaction directly with the teacher. In this module we concentrate on human gestures like hand movement as a means to connect to the student with the teacher. The hand detection is carried out again with the help of adaboost algorithm. The various fist samples are considered during the training of Adaboost algorithm, which results the cascade of strong classifiers. The fist detector is prepared from this and it is made to run over every frame. The module gets activated in the client systems when the master issues the command to ask questions. During this module, if the client has some doubt,

he can raise his hand so that his fist gets detected. If the hand gets detected, the client passes his identity to the server. At the master's end, corresponding student identity is referred and displayed to the master. In this manner, the master gets to know the student who is raising the doubt and he can respond to it.

3 System Implementation

The paper is implemented on Visual Studio IDE with the help of OpenCV library (Intel Corporation). The online video frame is captured from the webcam which is attached to the laptop. The input frame is at first converted to grayscale so that processing speed can be improved. The API in the library which is used for the face detection is `cvHaarDetectObjects()`. The parameters to this functions are tuned accordingly for the system. The haar classifier cascade used are `haarcascade_frontalface_alt.xml`, `haarcascade_lefteye_2splits.xml`, `haarcascade_fist.xml` which are loaded using another API called `cvLoadHaarClassifierCascade()` and are released after the use through `cvReleaseHaarClassifierCascade()`, the scale factor is flag like `CV_HAAR_FIND_BIGGEST_OBJECT` and `CV_HAAR_DO_ROUGH_SEARCH`

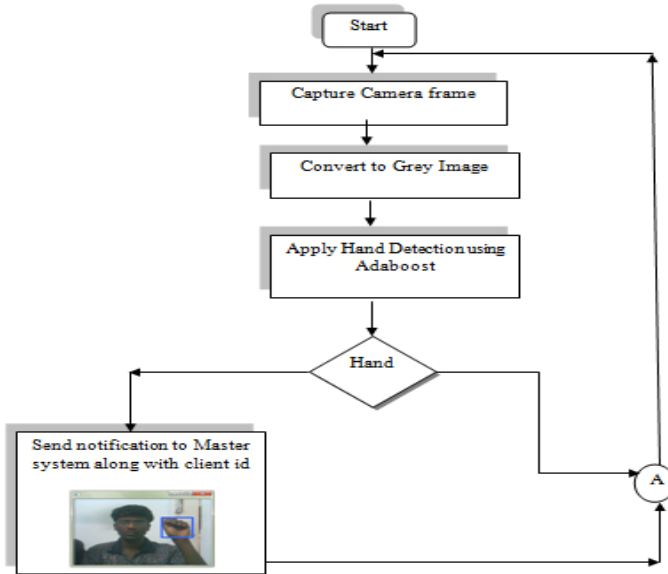


Fig. 7. Flow chart for hand detection in Student Teacher Interaction module

flags are set for face, eye or fist is to be detected at a quicker rate. The minimum feature size is 20 in both width and height. The faces obtained are cropped for recognition purpose. These cropped images undergo Histogram Equalization to remove noises present. Then, the PCA classifies these cropped images as per the

algorithm explained in section 2.1.2.2. The Confidence level is calculated and the output image has the rectangle imprinted over face. It also show the recognized person's name along with the confidence level only if the person has been trained already by our system. This information is logged into text file and is stored. The time at which the person appeared can also be logged.

4 Results and Discussion

The System is tested with three samples of three different skin tone under various conditions (Fig. 8.). The samples are made to move from the webcam ranging from 20 to 240 cms. At every scale, variation in poses, expressions and rotation of head (left-right, top-bottom) is tested. The frontal face detector could detect the faces at an angle bounded between 45° (left-right) and 60° (top-bottom). From the illumination perspective, the detector could not detect the faces only if the light source is in front of the camera. The detector is independent of the occlusions like spectacles. The detector fails when too much of light is reflected back from the spectacles of the person. It is inferred from above results that the face detection is skin colour, scale and rotation invariant under varying poses and expressions. When these detected faces are cropped and sent for recognition phase, different results are inferred after recognition. The samples are trained online before recognizing them. The result does not depend on facial expressions like happiness, sadness or anger. The system is capable of recognizing the same person irrespective of moustache and beard. The confidence value of recognition depends on different factors like poses, expressions, scale, rotations that were considered during training. The system is tested in environments like bedroom, classroom and corridors with controlled illumination.

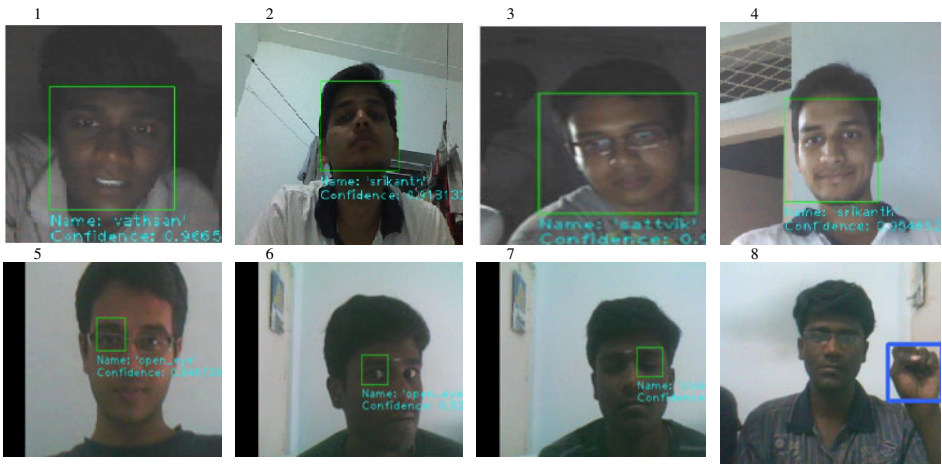


Fig. 8. 1,2,3,4 are the results of auto attendance module. 5,6,7 are the results of auto vigilance. 8 depicts results of student teacher interaction module.

5 Conclusion and Future Augmentation

In this paper we have presented a novel approach for making an environment in the class to address to various problems as discussed throughout the paper. As far as detection of face, eye and hand is concerned, the true detection rate is experimentally calculated to be around 98.2%. The recognition of the person is processed at a speed of 25-30 milliseconds and with confidence level ranging from 0.95 - 0.98. In future work will incorporate Scale Invariant Feature Transform (SIFT) along with Adaboost in order to track the eyes efficiently irrespective of rotations and scale in auto vigilance module. The PCA technique can be replaced with Probabilistic PCA to achieve a slightly higher recognition rates. The expression recognition can also be very useful in analyzing the activity of the students.

References

1. Shihavuddil, et al.: Development of real time Face detection system using Haar like features and Adaboost algorithm. *International Journal of Computer Science and Network Security* 10(1) (2010)
2. Senior, A.W.: Recognizing faces in broadcast video. In: *Proceedings of the IEEE workshop on Real-Time Analysis and Tracking of Face and Gesture in Real-Time Systems*, Kerkyra, Greece (1999)
3. Turk, M.A., Pentland, A.P.: Face recognition using Eigenfaces. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–591 (1991)
4. Monwar, M., et al.: A Real-Time Face Recognition Approach from Video Sequence using Skin Color model and Eigenface Method. In: *Canadian Conference on Electrical and Computer Engineering* (2006)
5. Rahman, N.A.b.A., Wei, K.C., John: RGB-H-CbCr Skin Colour Model for Human Face Detection. In: *Proceedings of The MMU International Symposium on Information & Communications Technologies* (2006)
6. Viola, P., Michael Jones, J.: Robust Real-Time Face Detection. *International Journal of Computer Vision* 57(2), 137–154 (2004)
7. Viola, P., Michael Jones, J.: Robust real-time object detection. In: *Second International Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing, and Sampling* (2001)
8. Sharma, S., Sreevathsan, R., Prashanth, M., Vrishod, K.S., Raghupathy, M., Gireesh Kumar, T.: Human Face Detection and Recognition using a Hybrid Approach. In: *proceedings of the International Conference on Computing* (2010)
9. Gupta, S., Sahoo, O.P., Goel, A., Gupta, R.: A New Optimized Approach to Face Recognition Using EigenFaces. *Global Journal of Computer Science and Technology* 10(1) (2010)
10. Romdhani, S., Torr, P., Scholkopf, B., Blake, A.: Computationally efficient face detection. In: *Proc. Int. Conf. on Computer Vision*, pp. 695–700 (2001)
11. Kubota, N.: Human Detection and Gesture Recognition Based on Ambient Intelligence. In: *Face Recognition, i-Tech, Vienna*, pp.558 (2007)
12. Menezes, P., Barreto, J.C., Dias, J.: *IAPR Workshop on Machine Vision Applications*, pp.48–51 (2002)

A Robust Background Subtraction Approach Based on Daubechies Complex Wavelet Transform

Anand Singh Jalal and Vrijendra Singh

Indian Institute of Information Technology, Allahabad, India
anandsinghjalal@gmail.com, vrij@iiita.ac.in

Abstract. This paper describes a simple and robust approach for background subtraction in Daubechies complex wavelet domain. A background subtraction approach exploiting noise resilience capability of wavelet domain combined with local spatial coherence and median filter in the training stage is proposed. The effectiveness of the proposed approach is demonstrated via qualitative and quantitative evaluation measures on both indoor and outdoor video sequences. The experimental results illustrate that the proposed approach outperforms state-of-the-art methods.

1 Introduction

Moving object detection from a video sequences is the primary step in many computer vision applications. These applications include object tracking, human computer interaction, vehicle traffic analysis and visual surveillance. Although a lot of studies have been conducted in recent years, the subject is still challenging. Some of the popular approaches proposed in the literature include background subtraction method, optical flow method and statistical learning method [1]. Algorithmic complexity is the major disadvantage of optical flow method. It requires higher time span than other methods. The requirement of training samples and higher computational complexity makes statistical learning methods infeasible for real time processing. The background subtraction approach is one of the very popular ways for extracting foreground objects from video sequences [1]. In this approach, the current frame and a reference frame is compared, to extract the moving object. However, accurate detection could be difficult due to a number of variations in environments such as illumination, shadow and camera jittering.

In recent years, many solutions have been proposed for background modelling and subtraction. Frame differencing [1] is a simple and easy way to extract moving object from a video sequence. In this approach image difference between consecutive frames is used and considerable differences in pixels value are considered as foreground region. However, Frame differencing methods suffer from fat boundary and thresholding problem. Wren et al. [2] proposed a method to model the background independently at each pixel location using a single Gaussian distribution. A recursive updating using a simple linear filter is used to estimate the Gaussian parameter. However it fails whenever some kind of variations occurs in the background. Stauffer and Grimson [3] proposed a method known as Gaussian Mixture Model (GMM), to

handle multi-modal distributions using a mixture of several Gaussians. The mean (μ), standard deviation (σ) and weight (w) is updated recursively to imitate the new observations for pixel value. The GMM is the most representative approach and has been widely used [4]. However, the major disadvantage of GMM is that it is computationally intensive and requires a tricky parameter optimization. Elgammal et. al. [5] exploited a nonparametric kernel density estimation to build a background PDF. The probability density estimation is performed using the recent historical samples without any assumption about background and foreground. The model is robust in nature and has good model accuracy as compare to Gaussian mixture model in the more complex scenes. However, the high computation cost limits its scope.

In recent years, wavelet domain is used for moving object segmentation [6] [7] [8]. In [6], the double-edge problem in the spatial domain is overcome, using a change detection method with different thresholds in four wavelet sub-bands. In [7], the authors proposed a method to extract moving object by using three consecutive frame differences in discrete wavelet transform (DWT) domain for frames at times $(n-1)$, (n) and $(n+1)$ and edge map of frame at time (n) . In [8] a real-time multiple objects tracking algorithm is proposed and the fake background motion is suppressed by performing the background subtraction method in 2-level real discrete wavelet transform. The first frame of the video sequence is assumed to be a background image. Guan [9] proposed a method for foreground segmentation and shadow suppression using HSV color space in Multi-Scale Wavelet domain. An optimal threshold is automatically computed to extract moving objects from video sequences. The moving objects are extracted using the hue value. Even though, these wavelet based methods show promising results. However, they are not adaptive in nature and tested against simple scenarios. Also discrete real wavelet transform (DWT) suffers from shift-sensitivity.

In the proposed work, we have taken advantage of the Daubechies complex wavelet transform properties to develop a robust background subtraction approach to extract moving objects. The Daubechies complex wavelet transform is approximately shift-invariant and has better directionality information with respect to DWT. The motivation is, the noise resilience nature of wavelet domain, as the lower frequency sub-band of the wavelet transform has the capability of a low-pass filter. In this paper, we have discussed a simple and effective approach of background modelling and subtraction by exploiting the low frequency sub-band characteristics of the object image in complex wavelet domain. Besides, this we have also exploited the local spatial coherence of the foreground pixels, to make the proposed method more robust against camera jittering and illumination changes.

The paper is organized as follows. Section 2 presents an overview of Daubechies complex wavelet transform. Section 3 describes the proposed approach. Experimental results are discussed in section 4 and finally the conclusion is presented in section 5.

2 Daubechies Complex Wavelet Transform

The discrete complex wavelet transform (CxWT) like discrete real wavelet transform (DWT) provides an efficient framework for representation and storage of images at multiple levels [10]. The wavelet transform divide an image into four sub-images.

These sub-images are label as approximation coefficients (LL), horizontal coefficient (LH), vertical coefficient (HL) and diagonal coefficient (HH). The approximation coefficient appears just like the compressed (filtered) original, while other coefficients contain the detailed information. One of the features of discrete wavelet transform (DWT) is that the spatial information is retained even after decomposition of an image into four different frequency bands [10].

Real DWT is non-redundant and an efficient tool to analyze signals. However, it suffers from the problem of shift sensitivity [5]. Complex wavelet transform (CxWT) can reduce these short comings. We have used Daubechies CxWT, as it is approximately shift-invariant and less redundant as compare to other complex wavelets [11].

Any function $f(t)$ can be decomposed into complex scaling function $\varphi(t)$ and a mother wavelet $\psi(t)$ as:

$$f(t) = \sum_k c_k^{j_0} \varphi_{j_0,k}(t) + \sum_{j=j_0}^{j_{\max}-1} d_k^j \psi_{j,k}(t)$$

where, j_0 is a low resolution level, $\{c_k^{j_0}\}$ and $\{d_k^j\}$ are known as approximation [$\varphi(t) = 2 \sum_n a_n \varphi(2t-n)$] and detail coefficients [$\psi(t) = 2 \sum_n (-1)^n \overline{a_{1-n}} \varphi(2t-n)$].

Where $\psi(t)$ and $\varphi(t)$ shares the same compact support $[-N, N+1]$ and a_n s are coefficients. The a_n s can be real as well as complex valued and $\sum a_n = 1$.

The Daubechies wavelet bases $\{\psi_{j,k}(t)\}$ in one dimension are defined through the above scaling function and multiresolution analysis of $L^2(\mathfrak{R})$. During the formation of solution if we relax the Daubechies condition for a_n [11], it leads to complex valued scaling function. We have used this symmetric Daubechies complex wavelet transform for tracking.

3 Proposed Approach

In background subtraction approach, we compare current frame with a reference frame known as background image. A significant difference indicates the presence of moving objects. However, if the reference is not modeled or updated adequately, this approach can be highly vulnerable to environment conditions like illumination and structural background changes. Also, with larger sizes of images, the above pixel by pixel operation may tend to slow down the overall computation. Hence, background subtraction in the wavelet domain at a higher level is employed in the proposed method. Performing the background subtraction in a wavelet domain provides a noise resilience capability to the system.

Since a background is defined as temporally stationary part of the video, so background scene represents stationary pixels in the video. Thus the background image can be computed exploiting the fact that moving objects reside in only some

portions of image frames and disappear over time. In the proposed algorithm an initial background model is obtained through a median filter and then it is recursively updated in the wavelet domain by adjusting parameters. The formulation of background modelling and subtraction in wavelet domain is as follows:

Let $F_n(x, y)$ corresponds to the intensity value at each (x, y) pixel location in the n^{th} image frame in the spatial domain and $W_\varphi^L F_n(k, l)$ represents the wavelet coefficient at (k, l) position in approximation subband at L^{th} level in the n^{th} image frame. So the reference image frame in wavelet domain is defined as:

$$W_\varphi^L B(k, l) = \text{median}(W_\varphi^L F_1(k, l), W_\varphi^L F_2(k, l), \dots, W_\varphi^L F_N(k, l)) \quad (1)$$

Where $W_\varphi^L B$ represents the wavelet coefficient at (k, l) position in the reference image.

Although the assumption is that the background is temporally stationary, we do allow certain amount of variation to make the algorithm noise resilience and robust to environmental changes.

The background subtraction is exploited to extract the moving object(s).

$$W_\varphi^L F_n^{diff}(k, l) = \left| W_\varphi^L F_n(k, l) - W_\varphi^L B_n(k, l) \right| \quad (2)$$

Where $W_\varphi^L F_n(k, l)$, $W_\varphi^L B_n(k, l)$, $W_\varphi^L F_n^{diff}(k, l)$ represent the wavelet coefficient at (k, l) position in approximation subband at L^{th} level for current frame, background frame and difference frame respectively. This background subtraction task is followed by thresholding to get the foreground object.

$$I_{obj+sha}(k, l) = 1, \quad \text{if } W_\varphi^L F_n^{diff}(k, l) \geq Th(k, l) \\ = 0, \quad \text{otherwise} \quad (3)$$

Where $I_{obj+sha}$ represents a binary mask containing object and shadow region.

The following updating strategies are applied to the background image using the knowledge of the pixel's classification in the current frame.

$$\left. \begin{aligned} W_\varphi^L B_{n+1}(k, l) &= (1 - \alpha)W_\varphi^L B_n(k, l) + \alpha W_\varphi^L F_n(k, l) \\ &\quad \text{If } (k, l) \text{ is background} \\ W_\varphi^L B_{n+1}(k, l) &= W_\varphi^L B_n(k, l) \\ &\quad \text{If } (k, l) \text{ is foreground} \end{aligned} \right\} \quad (4)$$

Where α is an adapting rate having value between 0 to 1; a smaller value of α tend to slow convergence, while a large value makes the modelling too sensitive.

A threshold describing a statistically significant change in the value of wavelet coefficient at each pixel position (k, l) is used for thresholding. An empirically determined value is used to initialize the threshold value. For each pixel the threshold is updated regularly using the following equations:

$$\left. \begin{aligned} Th_{n+1}(k,l) &= \sqrt{(1-\alpha)Th_n(k,l)^2 + \alpha(W_\phi^L F_n(k,l) - W_\phi^L B_n(k,l))^2} \\ &\quad \text{If } (k,l) \text{ is background} \\ Th_{n+1}(k,l) &= Th_n(k,l) \\ &\quad \text{If } (k,l) \text{ is foreground} \end{aligned} \right\} \quad (5)$$

Since foreground pixels have a propensity to appear in a sets of connected points as a blob. So we do not perform a pixel wise computation rather a weighted average of values $|W_\phi^L F_n(k,l) - W_\phi^L B_n(k,l)|$ is computed in a 3×3 neighborhood centered at (k,l) . So the eq. 2 and eq. 3 can be modified and a pixel is classified as foreground if:

$$|W_\phi^L F_n(k,l) - W_\phi^L B_n(k,l)| * C > W_\phi^L Th_n(k,l) * C \quad (6)$$

Where C is 3×3 Laplacian distribution mask used for convolution.

We can take mask bigger than 3×3 but that may cause degradation in detection close to the boundaries of the foreground objects. The reason of using Laplacian distribution mask is that in many practical applications, the Gaussian assumption may not hold completely, particularly indoor scenes and compressed video sequences [12].

4 Experimental Results and Discussion

To evaluate the performance of the proposed approach, we conducted experiments on several video sequences. Here, we are showing the results on two videos. The first video is recorded in our campus (outdoor environment), having the problem of camera jittering and noise in the background due to fluttering leaves and other local movements. The second video is Hall Monitoring video which is a commonly used benchmark test sequence especially for evaluating the effectiveness of background subtraction techniques. Due to the several lighting sources the Hall Monitoring video suffers from noise and variation in indoor illumination.

We have compared the performance of the proposed background subtraction approach with frame differencing [1] method and recently proposed improved Gaussian mixture model (IGMM) method [13]. Both qualitative and quantitative measures are used to compare the segmentations results. In order to provide a quantitative perspective, we used the false positive rate (FPR) and false negative rate (FNR) measures as given in [14]. These measures are defined as:

$$\begin{aligned} \text{FNR} &= \frac{\text{the number of foreground pixels wrongly classified}}{\text{the number of foreground pixels in the ground truth}} \\ \text{FPR} &= \frac{\text{the number of background pixels wrongly classified}}{\text{the number of background pixels in the ground truth}} \end{aligned}$$

The ground truth is generated by manually labeling the corresponding frames. The white and black colors represent the foreground and background pixels respectively. To ensure a fair comparison, the experimental results shown in Figs. 1-2 are computed without any morphological processing.

Fig. 1 shows the experimental results of different methods on video 1. The results demonstrate the robustness of the proposed approach against camera jittering and small background movements such as fluttering leaves etc. The reason that our method is able to handle these variations so well is because we are also considering the local spatial coherence to classify pixel as foreground or background. From the detection results, we can observe that the person in the fourth frame is detected quite accurately by the proposed method. While the other two methods unable to detect them accurately due to small size and camera jittering. It has been observed that IGMM approach produces very poor results after frame #430, due to the problem of camera jittering.

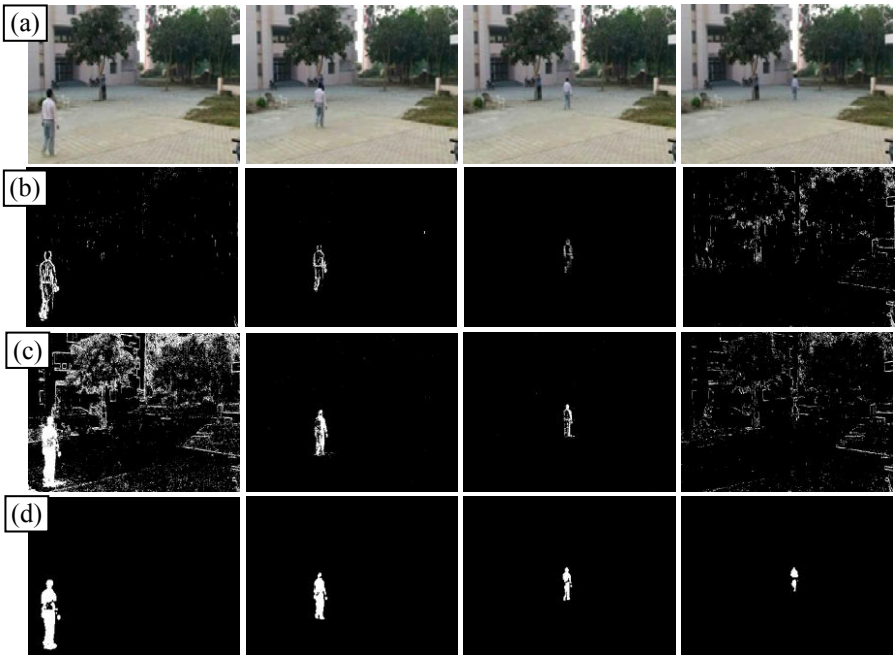


Fig. 1. Background subtraction results on Video 1 a) Original frames. b) Frame differencing results c) IGMM results d) Proposed method results.

In order to test the effectiveness of the proposed method in an indoor environment and lighting variations, we performed experiments on a well known indoor dataset i.e. ‘Hall Monitoring’. Fig. 2 displays the results of the different methods on this video and shows the capability of the proposed method to handle the noise due to variations in the illumination. On the other hand for frame differencing and IGMM method, the noise is still considered as the foreground as shown in figs. 2(b-c).

From the experiments, we can observe that frame differencing method is very sensitive to variations due to illumination changes as shown in fig. 2(b). It is also observed that in frame differencing method, the extracted moving object is not filled in the overlapping area as shown in figs. 1(b) and 2(b). Yet another problem in frame

differencing approach is selection of the threshold. Selecting a single threshold for all pixels makes it sensitive to different kinds of variations. Even though the IGMM method gives good results in general, but the method also shows poor performance in complex situations such as lighting variations, camera jittering as shown in fig. 1(c) and fig. 2(c).



Fig. 2. Background subtraction results on Video 2 (Hall Monitoring) a) Original frames. b) Frame differencing results c) IGMM results d) Proposed method results

Table 1. Comparison of false positive rate and false negative rate

#Frame	Video 1						#Frame	Video 2					
	Proposed Method		IGMM		FD			Proposed Method		IGMM		FD	
	FP	FN	FP	FN	FP	FN		FP	FN	FP	FN	FP	FN
5	.0031	.1909	.1393	.1180	.0061	.5687	25	.0164	.1477	.0617	.0884	.0452	.4747
143	.0016	.1129	.0018	.3582	.0013	.7024	51	.0104	.0622	.0233	.1387	.0424	.6390
303	.0007	.2532	.0006	.5487	.0003	.7755	117	.0074	.2382	.0096	.5228	.0418	.6397
432	.0004	.3324	.0352	.8171	.0304	.9724	227	.0082	.3218	.0116	.4685	.0432	.5584

Table 1 illustrates the FPR and FNR values for different methods. It is evident from Table 1 that the proposed background subtraction method gives less false positives than the comparison methods. The values of false negative are also very less in most of the image sequences in our method. In our method, most of the false positive and false negatives occur on the boundary area of the moving object, because

neighborhood pixels are considered in the classification. It should be noticed that the large false negative in case of FD is mostly due to misclassification in the inner areas of moving object. From the visual and numerical interpretation, as depicted in Figs. 1-2 and Table 1, it is clear that the proposed background subtraction method outperforms the other comparative methods, especially in conditions of camera jittering and illumination variations.

5 Conclusion

This paper addresses the problem of background subtraction in complex wavelet domain. In the proposed algorithm a temporal median filter is used to generate initial background model in a training stage and then foreground pixels are obtained by applying background subtraction scheme in the subsequent frames. The local spatial coherence of image pixels is exploited to makes the proposed approach more robust against illumination changes and camera jittering. We have demonstrated the robustness of the proposed approach in different video sequences with different kind of complexities.

References

1. Piccardi, M.: Background Subtraction Techniques: a Review. In: Proc. IEEE Int. Conf. Systems, Man, Cybernetics, pp. 3099–3104 (2004)
2. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.P.: Pfunder: Real Time Tracking of the Human Body. *IEEE Trans. Pattern Analysis and Machine Intelligence* 19(7), 780–785 (1997)
3. Stauffer, C., Grimson, W.E.L.: Adaptive Background Mixture Models for Real-Time Tracking. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 246–252 (1999)
4. Parks, D.H., Fels, S.S.: Evaluation of Background Subtraction Algorithms with Post-processing. In: Proc. IEEE Int'l Conf. Advanced Video and Signal-based Surveillance, pp. 192–199 (2008)
5. Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.S.: Background and Foreground Modeling using Non-parametric Kernel Density Estimation for Visual Surveillance. *Proceedings of the IEEE*, 1151–1163 (2002)
6. Huang, J., Hsieh, W.: Wavelet-based Moving Object Segmentation. *IEE Electronic Letters* 39(19), 1380–1382 (2003)
7. Huang, J., Hsieh, W.: Double Change Detection Method for Wavelet-based Moving-Object Segmentation. *IEE Electronic Letters* 40 (2004)
8. Cheng, F.H., Chen, Y.L.: Real Time Multiple Objects Tracking and Identification Based on Discrete Wavelet Transform. *Pattern Recognition* 39(6), 1126–1139 (2006)
9. Guan, Y.-P.: Spatio-temporal Motion-based Foreground Segmentation and Shadow Suppression. *IET Computer Vision* 4(1), 50–60 (2010)
10. Wang, Y., Doherty, J.F., Duck, R.E.V.: Moving Object Tracking in Video. In: Proceedings of 29th IEEE Int'l Conference on Applied Imagery Pattern Recognition Workshop, pp. 95–101 (2000)
11. Lina, J.-M.: Image Processing with Complex Daubechies Wavelets. *Journal of Mathematical Imaging and Vision* 7(3), 211–223 (1997)

12. Kim, H., Sakamoto, R., Kitahara, I., Toriyama, T., Kogure, K.: Robust Silhouette Extraction Technique using Background Subtraction with Multiple Thresholds. *Optical Engineering* 46(9) (2007)
13. Zivkovic, Z., Heijden, F.: Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction. *Pattern Recognition Letters* 27(7), 773–780 (2006)
14. Heikkila, M., Pietikainen, M.: A Texture-Based Method for Modeling the Background and Detecting Moving Objects. *IEEE Trans. Pattern Analysis Machine Intelligence* 28(4), 657–662 (2006)

File System Level Circularity Requirement

Mukhtar Azeem¹, Majid Iqbal Khan¹, and Arfan Nazir²

COMSATS, Institute of Information Technology,
Computer Science Department, Islamabad, Pakistan

¹{mazeem,majid.iqbal}@comsats.edu.pk,

²se_arfan@yahoo.com

Abstract. File system is an abstraction of storage media. This abstraction provides a friendly mechanism to the user to access the underlying storage elements. The user may be an end user of the computer who accesses files through graphical user interface (GUI) of an operating system or a programmer who accesses stored files through the system calls. Besides many other uses of a file, one is its use to store temporal data, i.e., the data that is valid for a specific period of time only. The data stored prior to a particular time instance becomes stale or useless and the space occupied by such data can be re-utilized. In such applications, after examining the required information, the data prior to a particular instance of time can be overwritten by new information. In this paper, we introduce the concept of circular file - file that can start overwriting itself after storing a particular amount of data or, indirectly, for a particular amount of time if time sensitive application like video surveillance is considered.

1 Introduction

A file system provides a method to organize, store, retrieve, and manage information on a permanent storage medium such as a disk. File system manages permanent storage and is capable of dealing with a large number of files and thus form an integral part of all operating systems. However, most of the computer users are vaguely familiar with the general concept of a file, directory and role of a file system. While using a general purpose operating system an end user accesses files through a graphical user interface (GUI) of the operating system that hides all the complexity of a file system and provides a user friendly interface.

The primary purpose of a file system is to provide a mechanism to store a named piece of data and to allow retrieval of that data using the name given to it. The named piece of data is generally referred to as a file. A file provides only the most basic level of functionality in a file system. Beside many other uses of a file, one is its use to store temporal data, i.e., data that is valid for a specific period of time only. The data stored prior to a particular time instance becomes stale or useless and the space occupied by such data can be re-utilized. Examples of such data and applications include but are not limited to the multimedia information for time shift recorders/players [4], patient monitoring system, video surveillance information, log files, etc. In such applications, after examining the required

information, the data prior to a particular instance of time can be overwritten by new information.

Many multimedia applications deal with large audio and video streaming data that require significantly large space for temporary storage. It is not uncommon for these temporary files to grow to several gigabytes in size. The MPEG2 compressed video bit rate, e.g., varies between 4-100Mbps [6], a few hours of which results in requirement of a large storage. Even though the disk sizes are increasing on a regular basis and one may argue that disk space is no longer a problem. However, one has to be aware of the simple fact that both the quality and size of such data is also on a rise. The size of the storage will never be infinite to store data forever.

Similarly, many high performance data acquisition systems can generate large amounts of data that may need to be processed in a relatively short period of time. However, a jitter in data generation rate and that of the processing speed makes it necessary to temporarily store such information until it is processed. In some cases, data must be archived for later offline analysis. In order to reduce the amount of data to a manageable level in such systems, the system may be designed to store only the data surrounding certain events. These events may include external triggers or user commands. In monitoring applications, for example, events are often related to some undesirable physical conditions such as temperature, pressure, etc. The data is temporally sensitive and is useless after certain period of time. This archived data can be used to better understand and then control the phenomena generating this data.

One application level solution to above problems that has been around for quite some time is the use of circular buffers. A classical example of the use of circular buffers is the type-ahead buffer of the keyboard in a PC [2]. The concept behind circular buffer is that it starts overwriting its own data after the buffer is full. So, it almost never refuses the write request. It has been successfully implemented for buffering of system log files [7], event recording, etc. With the computers starting to support multimedia information and also with the digital communication being fast and accessible to a common man, a similar buffering capability is required for multimedia information storage. This requires temporary storage of large amount of streaming data in the disk files. The streaming data is a continuous. It is not possible to store it continuously on a regular disk file since storage size is limited. Thus it is proposed that we must start overwriting files at the tail, when they grow beyond a particular size. This functionality can be achieved by implementing the circular buffer at the application level but the huge size buffers are not feasible. We expect that by incorporating circularity into file system one can free the application from repeating circular buffer-management and can achieve optimal performance. We will elaborate it further in the forthcoming sections.

In this paper, we discuss how the concept of circular buffers is ported for its implementation at file system level in the form of circular files. In proposed system, the user has to specify only the size of a file at the time of its creation and the system pre-allocates all of the required blocks at file creation time. This

guarantees that file will never run out of space during continuous write operation. Once created, the user does not need to worry about the circular nature of the file and can go on writing information in it just as in a regular file. However, once the file is full (the allocated capacity is consumed), it automatically starts overwriting itself without any user intervention.

Rest of the paper is organized as follows; Section 2 presents the circular buffer, its classical implementation with an example. Section 3 and 4 discusses the requirement of circular buffer for different multimedia operations. Section 5 and 5 the requirement of circularity in a file in the file system module of an OS to support circular files. Section 6 discussed our prototype implementation of filesystem and finally the Section 7 gives the future direction for further work.

2 Circular Buffer Requirements

The buffer is a named piece of RAM used to temporarily store information for later access. This buffer is normally accessed linearly like in an array with the lower index pointing to lower memory and higher index pointing to higher address in RAM as shown in Fig. 1.

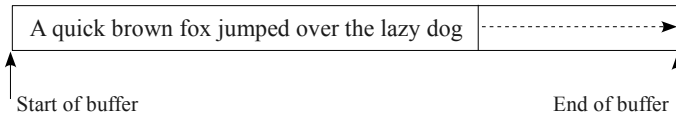


Fig. 1. Linear Buffer

However this is not the only use of RAM buffer. There are requirements such as the storage of scan codes coming from the keyboard when a key at the keyboard is pressed. This buffer is known as type-ahead buffer 5. The scan codes are stored in the type-ahead buffer till the time these are read by an application. If the buffer is full, the BIOS starts overwriting the oldest scan codes in the buffer. In this way the buffer is conceptually accessed like a circular buffer as shown in the Fig. 2.

Such a buffer is almost always seen wherever asynchronously received information is temporarily stored before being processed. A few examples in addition to the already introduced type ahead buffer include data received through RS-232 port and through Ethernet data port or USB port etc.

In the cases given above, information has to be temporarily stored to overcome the discrepancy between the reception and processing speed. Had there been no difference between the speed of information being received and speed with which it being processed, there would have been no need of an intermediate buffer. However, the data reception rate is always jittery, as shown in the Fig 3, and to overcome this jitter, we always need a buffer in between the device driver's Interrupt Service Routine (ISR) and the application 9.

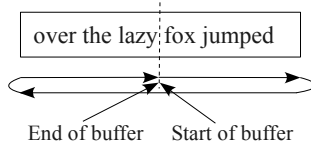


Fig. 2. Circular Buffer

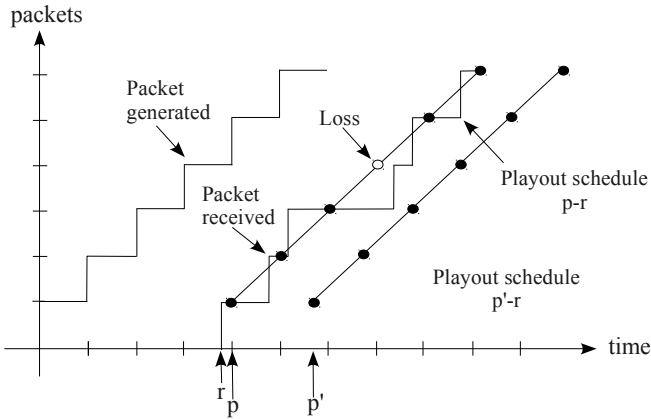


Fig. 3. Effect of Jitter on buffer requirement [9]

The Fig 3 above shows the effect jitter on the playout time at the receiver. It shows that starting the playout at p resulted a loss of packet. Had the playout started at p' there would not have been any loss. However to delay the playout of frames we need a circular rather than a linear buffer.

3 Circular Buffer Requirement for Multimedia Applications

These applications use the information that has become highly important in this communication era. The mode of data reception is interrupt driven and traffic is jittery. Additionally, the data rate is large, of the order of 4-100 Mbps [10]. The processing requirements for such type of information is also very high and Multimedia Digital Signal Processors (DSP) are employed for the purpose [3]. The DSP processor encodes/decodes the compressed video using Fourier Transformation. In this way, information is to reside in some intermediate buffer before being consumed by the application. Other examples that require this sort of processing are audio processing and satellite reception etc.

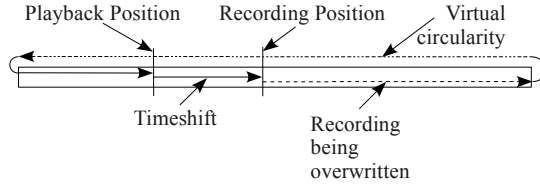


Fig. 4. Timeshift operation

The Personal Video Recorders (PVR), deal with compressed video as well as provide time shift feature [4]. The player receives live compressed MPEG video stream and displays it after decoding the stream. Depending on the option of the user or the design of the player, the received video is also recorded in the hard disk that is present in the set top box of the PVR. Time shift feature permits the user to pause a live stream being viewed. Therefore, the video can be resumed at a later stage. It also permits the user to rewind the live video that he was watching live. Both the operations technically require the same mechanism at the Operating System level. The amount of time for which the video can be paused varies and is dependent on the size of the storage present in PVR. Virtually, this feature shift the time since, e.g., a live video being received at 6PM, can be paused and viewed at 7PM. The Fig 4 shows the time shift action.

4 Circular Buffer Requirements for Multimedia Operations

QCIF (Quarter Common Interface Format) is among the lowest resolution standards. One frame is 144x176 pixels. A QCIF sized video with frame rate of 25 frames per second requires about 50 Gb ($60 \times 60 \times 25 \times 144 \times 176 \times 24 = 54743040000$ bytes) for the storage of 1 hour of uncompressed video. Therefore buffer size requirement for a multimedia stream is so large that use of RAM is not adequate. Permanent storage, which is often larger in size than RAM, is thus employed to store the incoming information compressed and uncompressed information. The access time of permanent storage like the hard disk is normally higher (between 2 to 7 milliseconds) as compared to that of RAM (between 10 to 30 nanoseconds). File system layout also adds to this delay. However, the video decoding process is relatively slow such that it cannot benefit from the fast RAM based storage and therefore the effect of additional delays due to the file based buffers can be ignored.

In addition, the video surveillance requires that recording for the past few hours be saved for analysis in case of an accident or other event of interest (like theft) etc., recording into files plays an important role. Since most of the time it is desired to record video of the past few hours only, the circular buffering functionality has an important role to play.

Since most of the file systems do not support the circular functionality as is required by the multimedia buffering operations, it leaves the responsibility of circularity management, to application developers. Inadequate file management for circularity may result in inefficient implementation which adversely effects temporarily sensitive applications like the PVR and surveillance systems.

It is therefore proposed that the support for circular operation of a file be implemented and provided at the kernel level. The benefit of kernel level support is an efficient implementation and additionally, in case of multiple applications utilizing circular files, the code redundancy is avoided. Additionally we propose that since most of the applications that we discussed here are related to embedded devices like the PVR and surveillance cameras etc, this functionality is expected to be extremely useful in embedded kernels.

5 Filesystem and Circular Buffer

This section discusses important elements of a file system. There are two types of information related to the file system, i) *on-disk* data structures and ii) *in-memory* data structures. Both these data structures are important in making a file system look like it is. The *on-disk* data structures contain the information that is stored on the disk permanently, like file name and file size, date and time of creation etc. These data elements are created when a file is opened for reading or writing. The *in-memory* information can be divided into two parts, a) per file information and b) per user or per process information. The per file information is a system wide information, created when first time a file is opened, remains alive till any instance of file is alive and is destroyed when the last instance of the file is closed, e.g., the number of instances of a particular file that are open at a particular instance of time. The per user/process information is created for every user that accesses the file. e.g., each user has a different file position indicating the position within the file where he/she is accessing the file. This information (being per user information) is destroyed when a user closes the file he/she was accessing.

Circular Vs. Linear File Management

Linear file management is straightforward. When file is opened for reading a single read pointer is maintained as the *in-memory* data element. By default, this read pointer is equated to 0 (zero) when the file is initially opened. As the read operation progresses, the read pointer is moved forward. The `fseek` or `lseek`, and other read operations move this pointer. Since every application that has opened a file for reading has its own associated read pointer, therefore every application can commence independent of the other. The write operations have no effect on the read pointer/s. The file continues to grow, conceptually, infinitely (till disk is full). The case is different for circular files. In case of circular file write pointer can commence only upto the size of the file after which it resets back

to 0 (zero). Similarly the read pointer can move upto the size of file, if there is sufficient data available for reading, after which it is reset back to 0 (zero). Read pointer cannot, however move beyond the write pointer since this indicates the end of available data. On the other hand the write pointer can go on incrementing even if it collides with the read pointer, since that is when it will overwrite the oldest data. It should push the read pointer forward in that case. The case is simple if a single read and write operation are permitted.

6 Prototype Implementation of Circular Files

Our prototype implementation for circularity at filesystem level was achieved using *ext2* by modifying the *ext2* code. The following instructions are kept generic and can be used as guidelines to those who want to implement circular files at any operating system. The regular files require two pieces of information to be stored as persistent disk data structures a) The start of file marker and b) The file size. Whenever a file is opened, user instructs the filesystem regarding the read/write position of the file using flags **SEEK_SET**, **SEEK_CUR** or **SEEK_END**. These flags create in-memory pointer for the read/write file position for the particular file instance. Thus each instance of the file has its own read/write file position. When a write operation is performed on a regular (non-circular) file, only the information regarding the size of file is to be updated on the disk. The file sharing algorithms handle the discrepancies in multiple file instances.

6.1 Circular File Handling

The following file handling functions were modified in the *ext2* file system in order to implement circular functionality.

- **File name of a Circular File:** The circular file creation requires the allocation of a fixed number of blocks to the file at the time of creation. The file starts overwriting once these preallocated blocks are all written. Our implementation used the file name of the to carry the file size at the creation time. Therefore the filename was a concatenation of the filename, textual size of the file and "cir" as extension. The filesystem calls extract the size of the file from filename. Additionally all the filesystem calls differentiate between the regular and circular files by the presence of "cir" as an extension.
- **Create a Circular File:** The file creation functionality of the filesystem identifies that it has to create a circular file and creates such a file of the size specified in the name as described above.
- **Open a Circular File:** The opening of a file that is circular requires to generate the in-memory variable representing the Start of file from the on-disk persistent variable as described in the previous section. This also happens when a file is newly created.

- **Write to a Circular File:** The write position is calculated at the time a file is opened and later updated after every write operation, circularly, using modulo division like the following:

$$\text{WriteFilePoistion} = (\text{WriteFilePosition}+1)\% \text{SizeOfFile}$$

This position marker is stored in the on-disk-variable. Since the regular files dont contain space for such a variable, an unused disk structure was used in ext2 structures. This is the varible used to create file position (in-memory variable) when the file is opened by a process.

- **Read a Circular File:** Reading function was also modified function reads from the file position provided by the calling function, know that start of file is created from the stored on-disk structure. The the increment/decrement in the read pointer is also a modulo operation. Reading ends when read position is less than current WritePosition.
- **Seek a Circular File:** The seek function is also based on the file position stored in the file structures using modulo operation.

Our prototype implementation permits only a single read and a single write operation at a time. We modified the ext2 driver and used the modified code as loadable kernel module. The implementation was tested by reading video frames from a USB webcam and writing the frames into a circular file. Whereas, another instance of that file was created by opening it for read operation. This second read instance was used by an application to display the read video frames. The application went on infinitely whereas the circular file did not grow beyond the specified limit.

We identified an unused disk data structure to store the start of file marker. In this way we did not need to modify the format of the ext2 formatted disk to support circularity. A disk already formatted for ext2 can be used to support circularity. Therefore a disk containing a circular file can be used with a standard file drivers not supporting circular functionality. Even the circular file data can be accessed through regular drivers. However, if a circular file is accessed through a regular driver the data will be accessed out of sequence, since there is no concept of keeping the Start of File marker in regular files.

7 Future Work

We used formatted file names to control the nature of file as well as its size. This is not a standard mechanism of specifying not standard parameters. There should not have been any change in the standard API. Rather the circularity specific functionality must be controlled through the ioctl function which provided standard mechanism for non standard functionality. The support of simultaneous multiple read along with a single write operations are complex as compared to linear files. This is since the start of file marker (an *on-disk* data element) must be pushed forward for every write operation after the file capacity is full

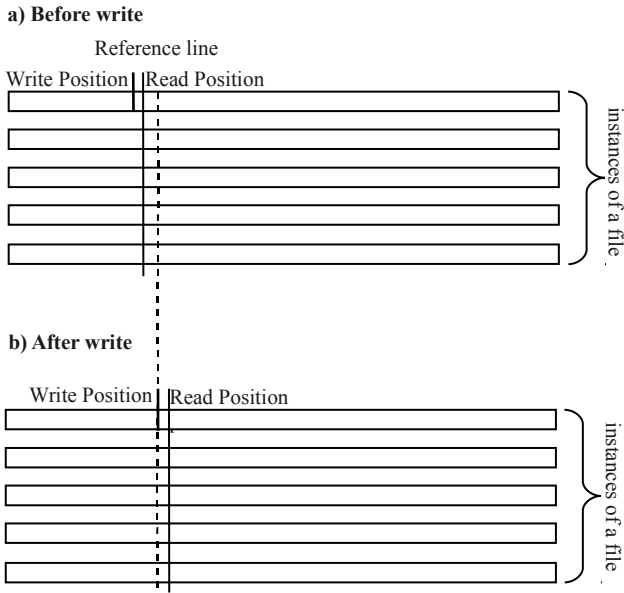


Fig. 5. Multiple instances of a file showing the changes in position of read pointer due to a write operation

and writes pointer is rotated, see Fig 5. Additionally if more than one instances of read operation are in progress, every write operation must update all the *in-memory* instances of the read pointer.

Non-standard Functionality Implementation

The non standard functionality to support circular files requires the use of *ioctl* or *fcntl*. The circular file require following functions to be implemented: a) File size passing and b) File type indication for circular file. The standard method for the implementation of these and other non standard operations on a file or a device is through *ioctl/fcntl* system call. Therefore *ioctl/fcntl* code related to filesystem must be modified to support circular files instead of using the filename to pass such information.

References

1. Kamariotis, O.: Bridging the Gap Between CBR and VBR for H.264 Standard. Transaction on Engineering, Computing and Technology 8 (October 2005)
2. Norton, P.: The Peter Norton Programmer's guide to the IBM PC (1985)
3. Smith, S.W.: The Scientist and Engineer's Guide to Digital Signal Processing. Arch. Rat. Mech. Anal. series 78, 315–333 (1982, 1999)

4. Kageyama, M., Ohba, A., Matsushita, T., Suzuki, T., Tanabe, H., Kumagai, Y., Yoshigi, H., Kinoshita, T.: Free time-shift DVD video recorder. *IEEE Transactions Consumer Electronics* 43(3), 469–474
5. Hyde, R.: *The Art of Assembly Language*, ch. 20, p. 1158
6. Halsall, F.: *Multimedia Communications, Applications, Networks, Protocols and Standards*, ch. 4, p. 235. Addison Wesley, Reading (2001)
7. Rosenblum, M., Ousterhout, J.K.: The Design and Implementation of a Log-Structured File System. *ACM Transactions on Computer Systems* 10(1), 26–52 (1992)
8. Peterson, L.L., Devie, B.S.: *Computer Networks, A Systems Approach*, ch.1, pp. 49–50. Morgan Kaufman Publishers, San Francisco (2003)
9. Kurose, J.F., Ross, K.W.: *Computer Networking*, ch.3, p. 588.
10. Zhu, W.W., Hou, Y.T., Wang, Y., Zhang, Y.Q.: End-To-End Modeling and Simulation of MPEG-2 Transport Streams over ATM Networks with Jitter. *CirSysVideo* 8(1), 9–12 (1998)

An Adaptive Steganographic Method for Color Images Based on LSB Substitution and Pixel Value Differencing

Azzat A. Al-Sadi and El- Sayed M. El-Alfy

College of Computer Sciences and Engineering,
King Fahd University of Petroleum and Minerals,
31261 Dhahran, Saudi Arabia
azzat.sadi@gmail.com, alfy@kfupm.edu.sa

Abstract. Data transmission in open access environments, such the Internet, has raised several security issues. Steganography has been one of the effective methodologies for concealing the existence of secret data by hiding it in a cover medium or carrier. Although several steganographic approaches have been proposed in the literature, this area is still active in research. In this paper, we propose and evaluate a novel steganographic approach that integrates the idea of pixel indicator with variations of two common steganographic schemes, namely least-significant bit (LSB) and pixel-value differencing (PVD). The aim is to increase the hiding capacity in color images without significant degradation in the quality of the cover image. Also it adds another level of security by varying the technique used for each color channel.

Keywords: Steganography; Pixel-Value Differencing; Least-Significant Bit; LSB Substitution; Pixel Indicator.

1 Introduction

With the rapid development of Internet technology, security has become extremely important. One of the security areas is steganography, which is the art and science of covered writing. Although steganography has been known for a long time, it becomes imperative in the recent age of information technology. Unlike data encryption, steganography provides a crucial means for concealing confidential data into cover media so that an unauthorized person will not be aware of the existence of this data at the first place [1, 2]. Steganography can use different cover media; such as text, image, audio and video. However, the enormous widespread applications of steganography are hiding data into digital images [15]. This is because of the simplicity of computation and the extensive use of images over the Internet with a variety of extensions.

In this paper, we focus on three methods: Least Significant Bit (LSB) replacement method [3, 4, 7, 10], Pixel-Value Differencing (PVD) method [5] and a combination of these two methods (PVD+LSB) [9]. Our proposed approach combines the idea of pixel indicator [13] with variations of LSB and PVD to hide data in color images. It differs from the existing techniques in two aspects. First, it has been designed to deal with color images rather than gray scale images. Second, it aims at increasing the

embedding capacity without much degradation of the image quality while confusing the steganalysis and making it harder to predict the hidden data.

The rest of this paper is organized as follows. Background information about related steganography methods is presented in Section 2. Related methods for color images are discussed in Section 3. Our proposed method for color images is described in Section 4. Experimental results are discussed in Section 5. Section 6 concludes the paper.

2 Background

2.1 LSB Substitution

This method is one of the earliest proposed steganographic techniques. The idea is to store a fixed number of bits of the secret data directly into the least significant bits of the pixels of the cover image. Because LSB is extremely simple to implement and incurs less processing time, it is commonly used. However, the insertion of fixed-length bits in least significant bits may cause noticeable distortion in the image because not all pixels can tolerate large changes in its data [6]. Furthermore, it is easy to attack. Hence, there is a tradeoff between the amount of secret data that can be embedded, the image distortion, and the security of the stego-image (the image after embedding the secret data). Several modifications have been proposed to enhance the performance of the original LSB; among them is PVD which will be explained in the next subsection.

To understand the operation of LSB replacement method, let’s consider an example in which we want to hide the word “Azzat” into a cover image using a simple LSB method of one bit only. Let’s assume the digital representation of the secret word “Azzat” is as follows:

01000001 01111010 01111010 01100001 01110100

where $A \equiv 01000001$, $z \equiv 01111010$, $a \equiv 01100001$, and $t \equiv 01110100$. Also assume the image bits is the stream below, where the least significant bit of each byte of the image is shown in bold and underlined.

01011010 00101011 10101011 10101010 11101011 11010100 01000111 11111001
 01011010 10101101 10010111 10101111 10101011 10100111 01010100 01011011
 10110111 11111011 00101011 10010101 10101000 01010100 10101010 11010101
 10100100 01011000 11011010 01010101 01001001 10110000 01000010 01010100
 10101011 10100111 10101000 01011000 10101010 00101011 11111011 10101011

After hiding the word Azzat, only some pixels will be affected, since some bits of the secret word will be as same as the cover image bits. For example, after hiding the letter “A”, the first image bytes will be

A= 0 1 0 0 0 0 0 1
 01011010 00101011 10101011 10101010 11101011 11010100 01000111 11111001

So, only three bytes have been changed (as indicated by the bold underlined bit). After the embedding process, the resulting stego-image will be as follows (where the changed bits are indicated in bold and underlined):

```

A  01011010 00101011 10101010 10101010 11101010 11010100 01000110 11111001
z  01011010 10101101 10010111 10101111 10101011 10100110 01010111 01011010
z  10110110 11111011 00101011 10010101 10101001 01010100 10101011 11010100
a  10100100 01011001 11011011 01010100 01001000 10110000 01000010 01010101
t  10101010 10100111 10101001 01011001 10101010 00101011 11111010 10101010

```

LSB has the following limitations

- Since LSB is simple and well known, it becomes vulnerable to security attacks.
- Increasing the amount of embedded data in each pixel results in more visual degradation in the image quality.
- Due to the uniform distribution of the embedded data over the whole cover image, the disruption of the image histogram becomes noticeable.

2.2 Pixel-Value Differencing Method

Wu and Tasi [5] proposed a Pixel-Value Difference (PVD) method to hide a secret data into 256 gray-valued images. This method relies on the idea that not all pixels can store the same number of bits of the secret data. Instead of inserting the secret bits directly to the end of each byte of the cover image (which is the way in which LSB works), they determine the number of bits to be embedded based on the differences between pairs of adjacent pixels. This allows the method to embed more data into the cover image without too much reduction in the stego-image quality.

PVD scans the cover image from the left-upper corner in zigzag and divides it into blocks with two consecutive non-overlapping pixels in each block. The differences of two-pixel blocks are used to categorize the smoothness and contrast properties of the cover image. The pixels around an edge area will have large differences whereas the pixels at the smooth area will have small differences. The larger the difference, the more bits to be hidden.

Wu and Tasi segmented the gray level width [0, 255] into smaller ranges. To facilitate binary data embedding, each range must be a power of 2. Ranges with small widths represent smooth areas while ranges with large widths represent edge areas. The range r_k is denoted by u_k and l_k which are upper level and lower level of this range respectively. Each range indicates the number of bits, n_k , that will be hidden in a pixel pair where $n_k = \log_2(u_k - l_k + 1)$. The ranges used in their paper are {8, 8, 16, 32, 64, 128} and {2, 2, 4, 4, 4, 8, 8, 16, 16, 32, 32, 64, 64}. However, after embedding the data, some pixel values may exceed 255 which are invalid values; consequently no secret data will be hidden in these pixels. Therefore, Wu and Tasi proposed a falling-off-boundary process to discover these pixels.

Assume P_i and P_{i+1} are pixel pair, g_i and g_{i+1} are their gray values. The difference d is calculated as $g_{i+1} - g_i$ and its value falls in the range from 0 to 255. The difference value of a pixel pair after embedding data, d' , can be determined by:

$$d' = \begin{cases} l_k + b_k & \text{for } d \geq 0 \\ -(l_k + b_k) & \text{for } d < 0 \end{cases} \quad (1)$$

where b_k is the decimal value of the secret bits to be embedded in this pixel pair. Let g'_i and g'_{i+1} be the gray values after embedding the secret bits and can be determined by:

$$(g'_i, g'_{i+1}) = \begin{cases} g_i - \left\lfloor \frac{d' - d}{2} \right\rfloor, g_{i+1} + \left\lfloor \frac{d' - d}{2} \right\rfloor & \text{if } d \text{ is odd} \\ g_i - \left\lceil \frac{d' - d}{2} \right\rceil, g_{i+1} + \left\lceil \frac{d' - d}{2} \right\rceil & \text{if } d \text{ is even} \end{cases} \tag{2}$$

2.3 PVD + LSB Method

Because the PVD method does not utilize the smooth area to hide large amount of secret data [8, 9], the capacity is still low. In order to achieve higher capacity, Wu et al. [9] used a combination of PVD and LSB. This method is based on the idea of using PVD when the difference between a pair of pixels is large (edge area), and using 3-bit LSB per pixel with readjusting equations when the difference is small (smooth area). The discrimination between the edge area and the smooth area is determined by calculating the difference between pixel pairs based on a threshold value, div , which is controlled by users and used as a secret key. Fig. 1 shows an example with $div = 15$. The following readjusting equations will keep the pixels in the same range after hiding the secret data:

$$(g'_i, g'_{i+1}) = \begin{cases} (g'_i - 8, g'_{i+1} + 8) & \text{if } g'_i \geq g'_{i+1} \\ (g'_i + 8, g'_{i+1} - 8) & \text{if } g'_i < g'_{i+1} \end{cases} \tag{3}$$

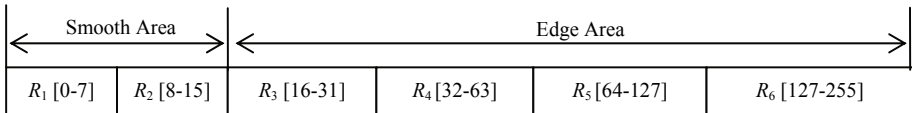


Fig. 1. An example of the division, $div = 15$

Table 1. The indicator table

First bit of P_i in Red channel	First bit of P_{i+1} in Red channel	Green channel	Blue channel
0	0	PVD	PVD + LSB
0	1	PVD + LSB	PVD
1	0	PVD + LSB	PVD
1	1	PVD	PVD + LSB

On the embedding process, the difference d_i is calculated. If $d_i < div$, this area is smooth and three bits will be directly embedded into the least significant bits of each pixel. The difference d'_i will be calculated after embedding the secret data and will be compared with the division div . If $d'_i \geq div$ the readjusting equations will be used, otherwise pixels belong to an edge area and the PVD method will be used.

3 Related Work

Gutub et al. [13], proposed a pixel indicator technique for RGB images. This approach based on the randomization principle to indicate the intensity of Red, Green and Blue channels. They used one of the color channels to select one of the other two channels to hide data using the LSB method. However, their approach has a security problem as a result of using LSB alone. It does not also utilize the third channel in hiding secret data and hence the embedding capacity will be low.

To improve the capacity and security, Parvez and Gutub [16] proposed another method to store a variable number of bits in one of the channels based on the actual color values. They changed the selected channel based on a sequence from a shared secret key. Thus, one channel is acting as an indicator to select the channel that has the lowest value among the other two channels for hiding data.

To improve the capacity further of [13], Amirtharajan et al. [12] used LSB with an adapted version of the pixel indicator approach. However, this approach is still using LSB and the capacity can be improved further since one of the channels is not used to hide any secret data. To remove the security limitation of LSB, Amirtharajan et al. [11] proposed a combination of the pixel indicator method [13], zig-zag PVD [14] and OPAP (Optimal Pixel Adjustment Process) [4] to enhance the stego-image quality while making it indistinguishable from the cover image.

4 The Proposed Method

The proposed method for color images is based on merging two techniques: PVD and LSB. It aims at increasing the embedding capacity without great quality loss. It also provides another level of security by varying the method used for each color channel. The embedding procedure is as shown in Fig. 2. It starts by separating the image into three channels: red, green, and blue. Then, it hides part of the secret data in one of the channels, for example the red channel, using LSB substitution with 2 bits. After that, the selected channel serves as an indicator to choose which method to apply on each of the other two channels as indicated in Table 1. The recovery procedure can be drawn similarly but using reverse operations.

5 Experimental Work

We implemented the proposed method in MATLAB and tested it using three 512×512 color images as cover images (as shown in Fig. 3): Lena, Peppers, and Baboon, which are commonly used in image processing, compression and steganography. The ranges [8, 8, 16, 32, 64, 128] were used in PVD approach. The embedding capacity, signal-to-noise ratio (SNR), and peak signal-to-noise ratio (PSNR) were utilized as performance measures. The PSNR is defined as follows:

$$PSNR = 10 \frac{\log_{10} 255^2}{MSE} \text{ dB} \quad (4)$$

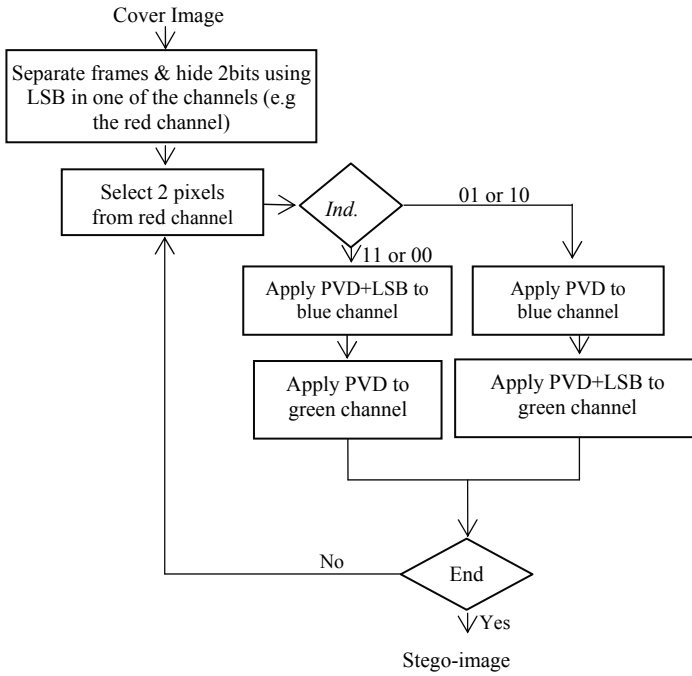


Fig. 2. Flowchart of the embedding procedure

where MSE is the mean square error and is defined as follows:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I - I_0)^2 \tag{5}$$

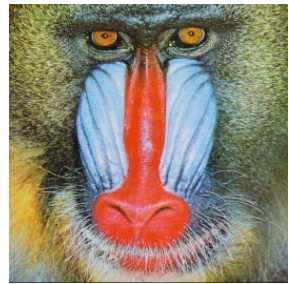
where I_0 is the pixel of the cover image (i.e. before embedding), I is the pixel of the stego-image (i.e. after embedding), and $m \times n$ represents the size of the image.



Lena



Peppers



Baboon

Fig. 3. The original cover images

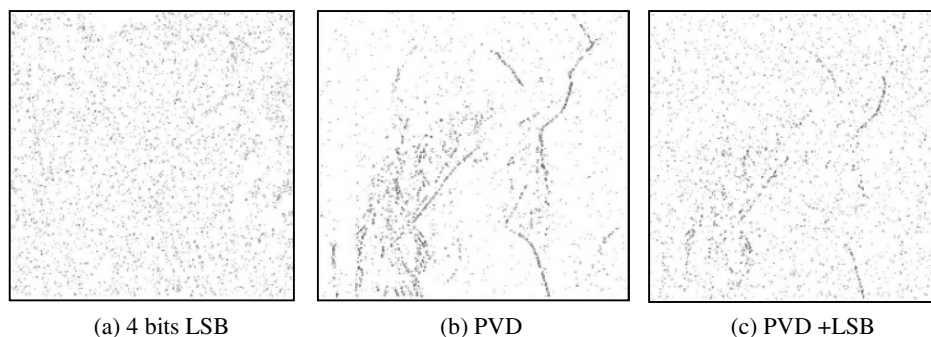


Fig. 4. Comparison between three methods using Lena image

Table 2. Embedding capacity and PSNR for Peppers

	Lena			Peppers			Baboon		
	Red	Green	Blue	Red	Green	Blue	Red	Green	Blue
Method	LSB	PVD or PVD+LSB		LSB	PVD or PVD+LSB		LSB	PVD or PVD+LSB	
Hidden bits	2	n (PVD), 3 (PVD+LSB)		2	n (PVD), 3 (PVD+LSB)		2	n (PVD), 3 (PVD+LSB)	
SNR	41.45	25.93	27.30	39.90	28.34	24.49	39.44	24.04	26.59
PSNR/channel	44.15	33.06	34.54	44.15	33.69	34.57	44.15	29.41	32.52
Capacity (bytes)	216830			212460			212427		
Avg SNR	31.56			30.91			30.02		
Avg PSNR	37.25			37.47			35.36		

In our experiments, a secret is a randomly generated and is used for all evaluations. We first compared LSB, PVD and PVD + LSB on grayscale images to understand the difference in distributing the data in the cover image. As shown in Fig. 4, LSB uniformly distributes the secret data over the whole image whereas PVD hides most of the secret data at the edges. On the other hand, PVD+LSB merges the advantages of both by utilizing smooth areas and edges to hide more data. After that, we tested our proposed method for color images and evaluated the performance measures for the three cover images as shown in Table 2. The PSNR of the stego-images for the proposed method is above 30 dB. At this level, the changes in the original cover image due to embedding secret data are hard to be recognized by human eyes. Fig. 5 shows the stego-images after embedding the secret data. Fig. 6 compares the histograms of the three color channels for Lena image before and after hiding the secret data. It is clear that the differences are not significant which makes it hard to be detected by steganoanalysis.

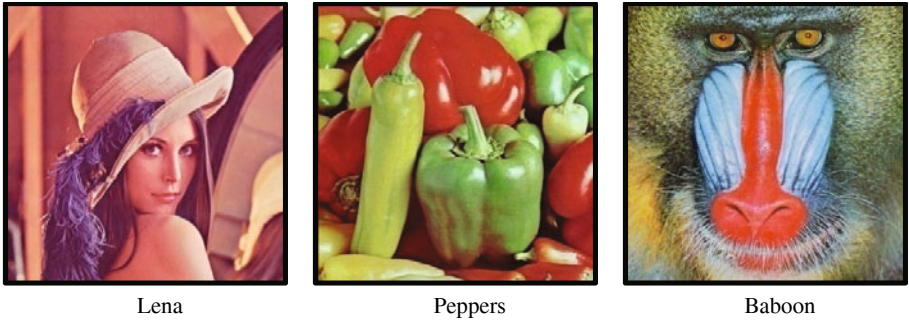


Fig. 5. Images after embedding data (stego-images)

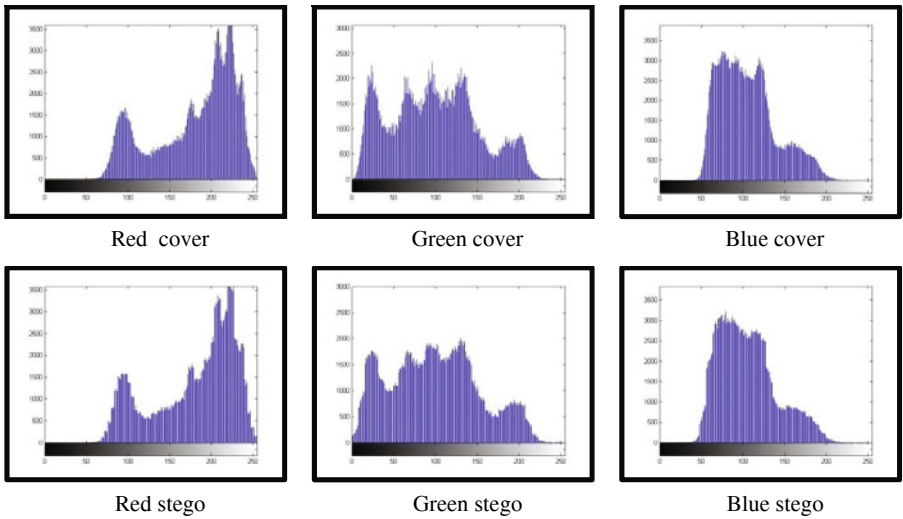


Fig. 6. Histogram comparison before and after embedding data in Lena image

We also implemented two other methods (the pixel indicator [13] and the color guided [12]) and compared the capacities with our method as shown in Fig. 7. Our method succeeded to increase the capacity by 3.29 times more than the pixel indicator method and by about 1.47 times more than the color guided method. The average PSNR for our method is still above 35 dB. This also makes the stego-image indistinguishable from the cover image by the human eye.

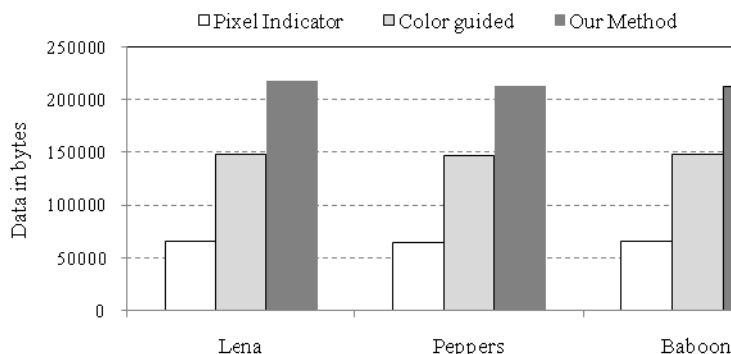


Fig. 7. Capacity comparison between the three methods

6 Conclusion

Invisibility, capacity and security are three aspects for a good steganographic approach. In this paper, we proposed a novel scheme using adaptive version of pixel indicator and variations of PVD and LSB. Our method increased the capacity of hidden data by utilizing the indicator channel to hide data in addition to the other channels. Furthermore, we randomize the secret data on the other two channels using two types of steganography which makes more secure. To increase the security further, we can encrypt the confidential message before hiding it.

Acknowledgments. The authors would like to thank King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia and the Hadhramout Est. for Human Development in Yemen for their support during this work.

References

1. Petitcolas, F.A.P., Anderson, R.J., Kuhn, M.G.: Information Hiding: A Survey. Proc. of the IEEE (special issue) 87, 1062–1078 (1999)
2. Rabah, K.: Steganography: The Art of Hiding Data. Information Tech. J. 3, 245–269 (2004)
3. Wang, R.-Z., Lin, C.-F., Lin, J.-C.: Image Hiding by Optimal LSB Substitution and Genetic Algorithm. Pattern Recognition 34, 671–683 (2001)
4. Chan, C.K., Cheng, L.M.: Hiding Data in Images by Simple LSB Substitution. Pattern Recognition 37, 469–474 (2004)
5. Wu, D.-C., Tsai, W.-H.: A Steganographic Method for Images by Pixel-Value Differencing. Pattern Recognition Letters 24, 1613–1626 (2003)
6. Yang, C.-H., Wang, S.-J., Weng, C.-Y.: Analyses of Pixel-Value-Differencing Schemes with LSB Replacement in Steganography. In: Proc. of the 3rd Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing, vol. 1, pp. 445–448 (2007)
7. Chang, C.C., Hsiao, J.Y., Chan, C.S.: Finding Optimal Least-Significant-Bit Substitution in Image Hiding by Dynamic Programming Strategy. Pattern Recognition 36, 1583–1595 (2003)

8. Yang, C.H., Weng, C.Y.: A Steganographic Method for Digital Images by Multi-pixel Differencing. In: Proc. Int. Comput. Symp (2006)
9. Wu, H.-C., Tsai, C.-S., Hwang, M.-S.: Image Steganographic Scheme Based on Pixel-Value Differencing and LSB Replacement Methods. In: IEE Proceedings Vision, Image and Signal Processing, vol. 152, pp. 611–615 (2005)
10. Ker, A.: Improved Detection of LSB Steganography in Grayscale Images. *Information Hiding*, 583–592 (2005)
11. Amirtharajan, R., Adharsh, D., Vignesh, V., Balaguru, R.J.B.: PVD Blend with Pixel Indicator-OPAP Composite for High Fidelity Steganography. *International Journal of Computer Applications* 9 (2010)
12. Amirtharajan, R., Behera, S.K., Swarup, M.A., Ashfaq, M., Rayappan, J.B.B.: Colour Guided Colour Image Steganography. *Universal Journal of Computer Science and Engineering Technology* 1, 16–23 (2010)
13. Gutub, A., Mahmoud, A., Muhammad, A., Abdulrahman, S., Aleem, A.: Pixel Indicator High Capacity Technique for RGB Image Based Steganography. In: *International Workshop on Signal Processing and its Applications* (2008)
14. Padmaa, M., Venkataramani, Y.: Zig-Zag PVD: A Nontraditional Approach. *International Journal of Computer Applications*, 5–10 (2010)
15. Cheddad, A., Condell, J., Curran, K., Mc Kevitt, P.: Digital Image Steganography: Survey and Analysis of Current Methods. *Signal Processing* 90, 727–752 (2010)
16. Parvez, M., Gutub, A.: RGB Intensity Based Variable-Bits Image Steganography. In: *Proceedings of IEEE Asia-Pacific Services Computing Conference, APSCC 2008* (2008)

Communication Aware Co-scheduling for Parallel Job Scheduling in Cluster Computing

A. Neela Madheswari¹ and R.S.D. Wahida Banu²

¹ Associate Professor & Head, Department of Information Technology,
KMEA Engineering College, Aluva
neela.madheswari@gmail.com

² Professor & Head, Department of Electronics and Communication Engineering,
Government College of Engineering, Salem
drwahidabanu@gmail.com

Abstract. Parallel job scheduling is gaining an important topic in the field of research. The main issue is how to allocate jobs to the multiprocessor systems and to improve the system performance. Co-scheduling is the concurrent scheduling of parallel jobs on individual nodes of multiprocessor systems. There are three main co-scheduling policies namely explicit co-scheduling, local scheduling and implicit co-scheduling. But no consideration is given for parallel jobs which are communication intensive. A new algorithm is devised for this issue named communication aware co-scheduling. The workload from Grid5000 is considered for performance evaluation.

Keywords: Parallel jobs, co-scheduling, time sharing, space sharing, context switch, workload, communication aware co-scheduling.

1 Introduction

Job scheduling plays an important role in various fields starting from mainframe computing to high performance computing. There are two general approaches for job scheduling in multiprocessor systems. They are time sharing and space sharing.

Time sharing allows more flexible scheduling but it introduces its own set of problems. An executing job may be slowed down by an arbitrary amount if it must share its resources with many other jobs. It is therefore proposed to limit the number of jobs that can share the processors so that each user is assured a minimum QoS. Another potential problem with time sharing is that the performance of a job may suffer if all of its components do not execute at the same time. Space sharing without time sharing can cause the system to be underutilized and make it hard to execute highly parallel jobs. Another problem of space sharing without time sharing is that too many processors may remain idle while jobs are awaiting execution [2].

In recent years, researchers have developed parallel scheduling algorithms that can be loosely organized into three main classes, according to the degree of coordination between processors: explicit co-scheduling, local scheduling and implicit or dynamic co-scheduling. Explicit co-scheduling approach is neither scalable nor reliable.

Explicit co-scheduling of parallel jobs also adversely affects performance on interactive and I/O based jobs. Local scheduling allows each processor to independently schedule its processes. Although local scheduling is attractive due to its ease of construction, the performance of fine grain communicating jobs degrade significantly because scheduling is not coordinated across processors. In dynamic co-scheduling, each local scheduler makes decisions that dynamically coordinate the scheduling actions cooperating processes across processors. These actions are based on local events that occur naturally within communicating applications [5].

Clusters are gaining acceptance not only in scientific applications that need super computing power, but also in domains such as databases, web service and multimedia which place diverse quality of service demands on the underlying system. Resource management and scheduling on workstation clusters is complicated by the fact that the number of idle workstations available to execute parallel applications is constantly fluctuating [9]. Most universities make heavy use of shared clusters for development and research. Industries too use such clusters for internal product development and testing. On any such parallel computing platform, optimally scheduling processes of a parallel job onto various nodes of the system has always been a challenging problem [10].

2 Motivation

Co-scheduling is the term first introduced by John K.Ousterhout. Close interactive parallel programs have a process working set that must be co-scheduled for the parallel program to make progress. Process thrashing occurs when the scheduling of each process whose services are awaited causes another process which in turn waited for the resource hold by another process, to be de-scheduled. The progress of parallel program is limited by the rate at which the scheduling decisions are made. If efficient inter-process communication primitives are to be used to their fullest, mechanisms must be provided to avoid process thrashing [1].

Many job scheduling studies regard parallel jobs as rectangles in processors \times time space: they use a fixed number of processors for certain interval of time. This is justifiable when the discussion is limited to the workings of the scheduling proper, and jobs are assumed not to interact with each other or with the hardware platform. In reality, this is not always the case. Running the same jobs on different architectures can lead to very different runtimes, changing the structure of the workload. Running applications side by side may lead to contention if their partitions share communication channels. Contention effects are especially bad for systems using time slicing, as they may also suffer from cache and memory interference [14].

Many scientific and high performance computing applications consists of multiple processes running on different processors that communicate frequently. Because of their synchronization needs, these applications can suffer severe performance penalties if their processes are not co-scheduled to run together. Two common approaches to co-scheduling jobs are batch scheduling, wherein nodes are dedicated for the duration of the run, and gang scheduling, wherein time slicing is coordinated across processors [13]. There are numerous solutions for gang scheduling [11], [4], [8], [6], [7], [12], [3], but lack of addressing the solution for communicating jobs.

An evaluator needs to be aware of the special role of interactive jobs and make a decision if they are to be incorporated in the evaluation or not. To incorporate them, one can use workload traces or models that include them. If choosing not to incorporate them, the decision should be stated and explained [14].

While considering the workload for parallel jobs, there are generally two kinds of jobs are considered. They are serial and parallel. The serial jobs can be executed independently without affecting the number of other jobs. So the scheduling of serial jobs cannot create much effort while scheduling. The parallel jobs are classified into batch and co-scheduled jobs. Batch jobs are specified by the group of jobs to be executed without interrupting the system after start its execution. But the co-scheduled jobs are the jobs in which the execution of a single job depends upon the execution of all its related jobs.

3 System Model

For experimentation, the adopted workload from Grid5000 [15] is considered. In this work, it is assumed that there are 124 nodes in the system and the communication between them takes place with the help of Message passing interface environment. One of the nodes will act as a scheduler and before scheduling a job, it retrieves the status of the nodes whether it is idle or the minimum time to complete its process using heartbeat messages. Assuming each job submission from the scheduler to the processors will take 2 seconds of latencies, one for submitting the job and another second for retrieving the results. A list of co-scheduled jobs is provided in Grid5000. Assuming all the co-scheduled jobs to be communication intensive and one job is initiating the remaining co-scheduled jobs list and the processor is scheduled such that all the list of jobs involved in the co-scheduled job list is scheduled together.

The main focus of this work is to avoid process thrashing, since the communication intensive jobs are considered for scheduling. At the same time, blockade situation [8] must be avoided when space sharing policy is followed. The idle time of the processors must be reduced and the wait time of jobs in the scheduler must also be minimized. A new algorithm is devised that address the solution for the above mentioned problems for scheduling the parallel jobs.

4 Communication Aware Co-scheduling

The Grid5000 workload contains serial and parallel jobs. For performance evaluation, both the serial and parallel jobs are taken for consideration. A new scheduling algorithm called communication aware co-scheduling is proposed in which the jobs of serial and parallel are subjected under test for job scheduling evaluation. There are three cases to be considered.

4.1 All Jobs Are Serial

Consider from the workload, first 300 serial jobs. The algorithm for job scheduling is given by algorithm1.

Algorithm1: For all serial jobs

1. For each job (j_1) to be scheduled
 - a. Calculate the start time, end time and actual time from runtime.
 - b. Select any node (n_1) from the given set of nodes which is not allotted with job by scheduler and the job j_1 is scheduled.
 - c. If no one node is available, the minimum completed node is calculated and is scheduled for the job. Calculate the wait time for the job $j_1 =$ current time - minimum time to be completed by the any of the nodes, which is to be scheduled with the job j_1 .
 - d. Continue the step (1) till the scheduler has more number of jobs to be scheduled.
2. Calculate the idle time of node $n_1 =$ wait time, till next job to be scheduled to n_1 .
3. Idle time of every processor is added further till the last jobs are scheduled and the same nodes are allotted for the jobs.

4.2 All Jobs Are Parallel

Consider from the workload, the first 300 parallel co-scheduled jobs. Since the number of jobs for each co-scheduled entry is known, i.e., for a co-scheduled job id, a set of jobs that are needed to execute at the same time is given. The algorithm for scheduling co-scheduled jobs is given by algorithm2.

Algorithm2: For all parallel co-scheduled jobs

1. For each co-scheduled job (c_1) to be scheduled
 - a. Calculate the total time needed for the entire job as $tr = tr_1 + tr_2 + \dots + tr_n$, where n is the total number of jobs that are to be co-scheduled and tr_1, tr_2, \dots, tr_n are their respective runtimes.
 - b. Calculate start time, end time and actual time from the run time tr .
 - c. Select any node n_1 from the given set of nodes which is available and the job c_1 is scheduled, implies all the jobs coming under c_1 i.e., $c_{j1}, c_{j2}, \dots, c_{jn}$ are scheduled together for runtime tr .
 - d. If no one node is available, the minimum completed node is calculated and is scheduled for the job. Calculate the waiting time for the job $c_1 =$ current time - minimum time to be completed by any of the nodes, which is to be scheduled with the job c_1 .
 - e. Continue the step (1) till the scheduler has more number of jobs to be scheduled.
2. Calculate the idle time of node $n_1 =$ wait time, till next job to be scheduled to n_1 .

3. Idle time of every processor is added further till the last jobs are scheduled and the same nodes are allotted for the jobs. Calculate the idle time of node n_1 = wait time, till next job to be scheduled to n_1 .
4. Idle time of every processor is added further till the last jobs are scheduled and the same nodes are allotted for the jobs.

4.3 Mixed Jobs

Consider from the workload, the first 25 parallel jobs and the next 25 serial jobs, likewise the total of 300 jobs are considered. The algorithm for scheduling mixed jobs is done by make use of alternative scheduling of 25 alternative serial and parallel jobs using the above algorithms namely algorithm1 for serial and algorithm2 for parallel jobs.

4.4 Run Time Calculation of Co-scheduled Job

Let us consider a co-scheduled job id c_1 . It contains a set of jobs to be scheduled together as given by: j_1, j_2, \dots, j_n . Here n is the number of jobs involved in the co-scheduled job id j_1 . Then the total run time for the execution is given by

$$rt = \sum_{i=0}^n rt(j_i) \quad (1)$$

Here rt is the total run time needed for running the job c_1 , $rt(j_i)$ represents the run time of a particular job j_i .

4.5 Idle Time Calculation of a Node

Initially the idle time of all nodes are set to zero. After the scheduler starts scheduling the jobs, the idle time is calculated for every node and is updated while scheduling the job at the same node further. Consider a node n_1 . Let the idle time initially is given by: $I(n_1) = 0$. Let the node n_1 is initially scheduled with the job j_1 and completes its execution. Further, the job scheduled again in the same node be j_m . Let ct be the current time while scheduling the job j_m to the node n_1 . Then the idle time is calculated as

$$I(n_1) = I(n_1) + (\text{End time of job } j_1 - \text{start time of job } j_m) \quad (2)$$

5 Experimental Results

An algorithm is proposed for communication intensive jobs. The performance of the communication aware co-scheduling algorithm is evaluated by deriving the idle time of the nodes involved in the system, total wait time at the end of scheduling all the jobs and wait time of all the jobs.

The algorithm proposed work for all serial jobs. The idle time of the nodes while scheduling the job is given in fig. 1 where the horizontal axis specifies the number of nodes and the vertical axis specifies the idle time in seconds. The wait time of all the jobs scheduled is given in fig. 2 where the horizontal axis specifies the number of jobs and the vertical axis specifies the wait time in seconds.

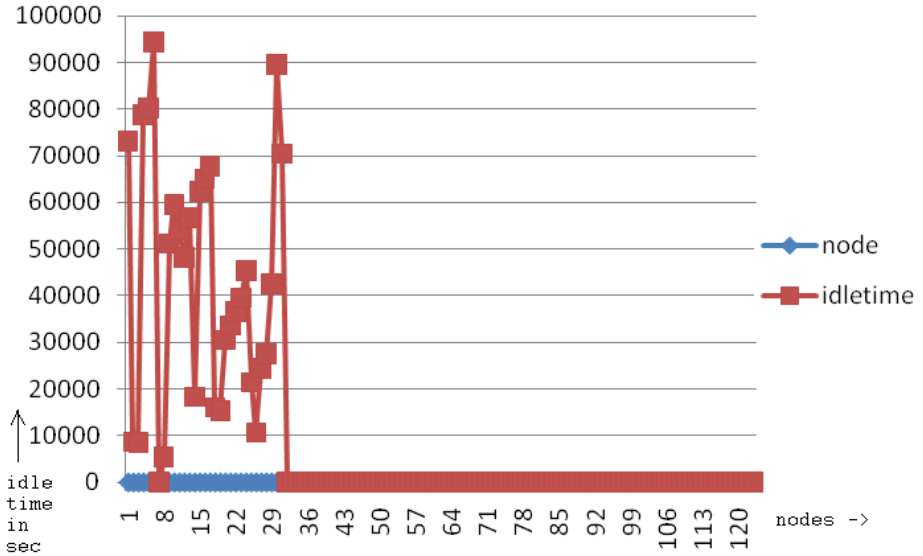


Fig. 1. The idle time of the nodes at the end of scheduling serial jobs

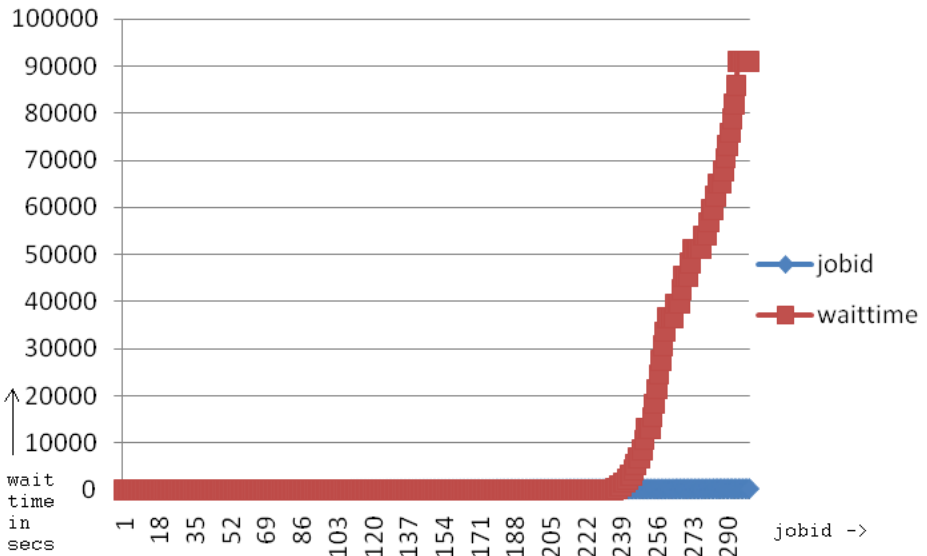


Fig. 2. The wait time for all the jobs in seconds at the end of scheduling serial jobs

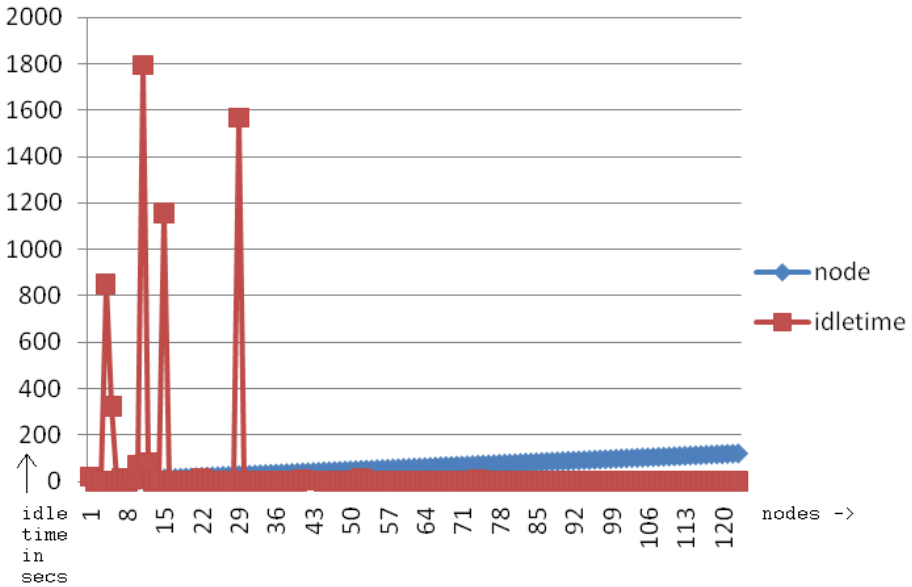


Fig. 3. The idle time of all the nodes in seconds at the end of scheduling parallel jobs

For all parallel jobs at the end of scheduling 300 jobs, the idle time of the nodes while scheduling the job is given by the fig. 3 where the horizontal axis specifies the number of nodes and the vertical axis specifies the idle time in seconds. The wait time of all the jobs scheduled is given in fig. 4 where the horizontal axis specifies the number of jobs and the vertical axis specifies the wait time in seconds. It is observed from the fig. 4 that at the end of the scheduling, few jobs have significant wait time.

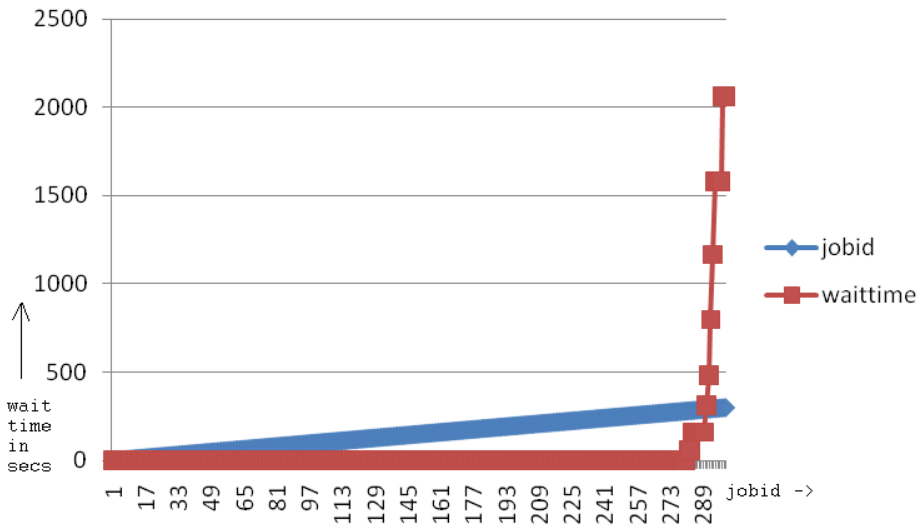


Fig. 4. The wait time of the all the jobs in seconds at the end of scheduling parallel jobs

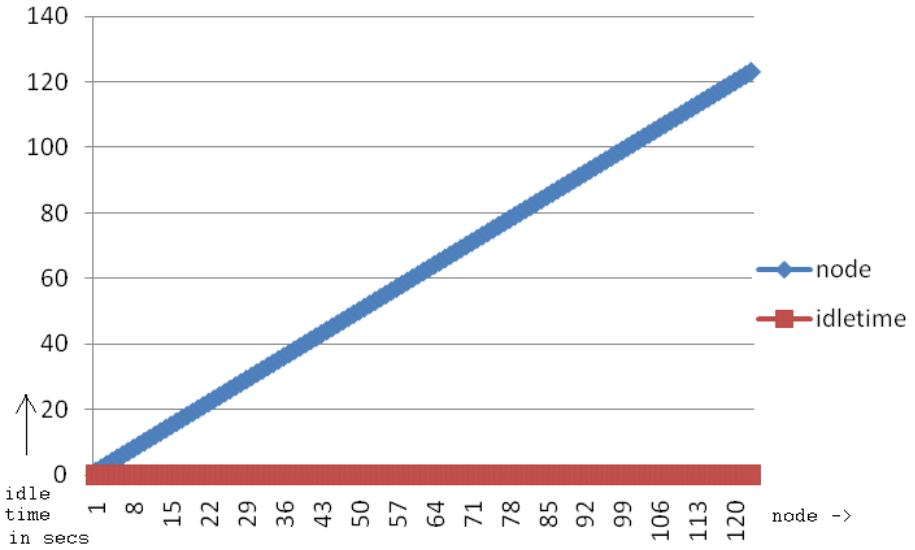


Fig. 5. The idle time of the nodes (here 0 second) at the end of the scheduling for mixed jobs

For the mixed jobs, i.e. alternative scheduling of parallel and serial jobs, at the end of scheduling 300 jobs, the idle time of the nodes while scheduling the job is given in fig. 5 where the horizontal axis specifies the number of nodes and the vertical axis specifies the idle time in seconds. The wait time of all the jobs scheduled is given in fig. 6 where the horizontal axis specifies the number of jobs and the vertical axis specifies the wait time in seconds.

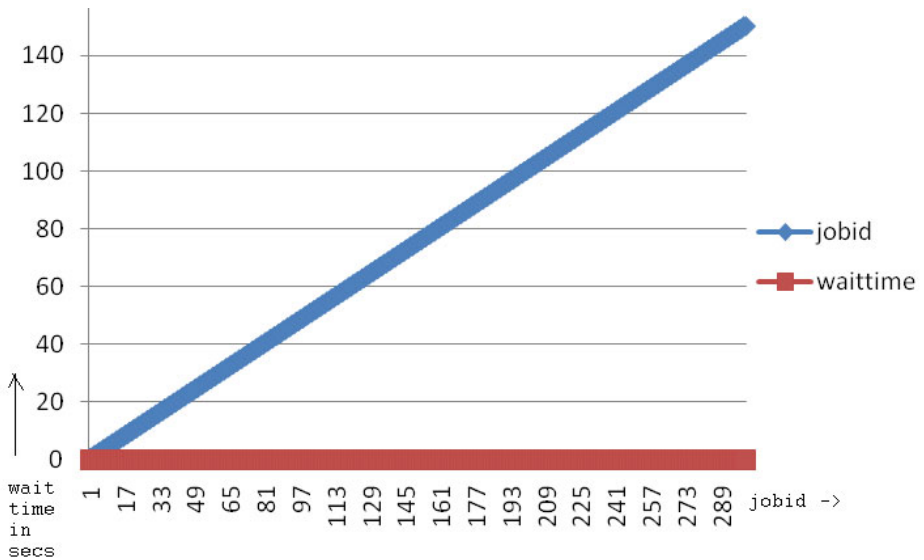


Fig. 6. The wait time of the jobs (here 0 second) at the end of scheduling mixed jobs

6 Conclusion

A new co-scheduling algorithm called communication aware co-scheduling is proposed and the communication intensive jobs are considered for parallel jobs. The workload from Grid5000 is considered and the information of co-scheduled jobs, serial jobs with runtime is used for scheduling. The experimental results prove that the algorithm work well for mixed jobs. But the total wait time for scheduling increases while scheduling all co-scheduled jobs. There is no wait time for mixed jobs. The real workload is the combination of serial and parallel jobs. For performance analysis, various combinations of workload are considered. It is concluded from the experimental results that the proposed scheduling algorithm is working alright for mixed jobs.

References

1. Ousterhout, J.K.: Scheduling Techniques for concurrent systems, Identification of Common Molecular Subsequences. In: Smith, T.F., Waterman, M.S. (eds.) International Conference on Distributed Computing Systems (1982)
2. Franke, H., Pattnaik, P., Rudolph, L.: Gang scheduling for highly efficient distributed multiprocessor systems. In: Sixth Symposium on the Frontiers of Massively Parallel Computation (1996)
3. Jette, M.A.: Performance characteristics of Gang scheduling in multiprogrammed environments. In: ACM/IEEE Conference on Super Computing (1997)
4. Silva, F.A.B., Scherson, I.D.: Improving throughput and Utilization in Parallel machines through concurrent gang. In: 14th International Parallel and Distributed Processing Symposium (2000)
5. Petrini, F., Feng, W.C.: Buffered Coscheduling: A new methodology for multitasking parallel jobs on distributed systems. In: International Parallel and distributed Processing Symposium (2000)
6. Batat, A., Feitelson, D.G.: Gang scheduling with memory considerations. In: 14th International Parallel and Distributed Processing Symposium (2000)
7. Zhang, Y., Franke, H., Moreira, J.E., Sivasubramaniam, A.: Improving parallel job scheduling by combining gang scheduling and backfilling techniques. In: 14th International Parallel and Distributed Processing Symposium (2000)
8. Zhou, B.B., Brent, R.P.: Gang scheduling with a queue for large jobs. In: 15th International Parallel and Distributed Processing Symposium (2001)
9. Karatza, H.D.: Gang scheduling performance on a cluster of non-dedicated workstations. In: 35th IEEE Annual Simulation Symposium (2002)
10. Agarwal, S., Choi, G.S., Das, C.R.: Co-ordinated coscheduling in time-sharing clusters through a generic framework. In: IEEE International Conference on Cluster Computing (2003)
11. Ryu, K.D., Pachapurkar, M., Fong, L.L.: Adaptive memory paging for efficient gang scheduling of parallel applications. In: 18th IEEE International Parallel and Distributed Processing Symposium (2004)
12. Strazdins, P., Uhlmann, J.: A comparison of local and gang scheduling on a Beowulf cluster. In: IEEE International Conference on Cluster Computing (2004)

13. Frachtenberg, E., Feitelson, D.G., Petrini, F., Fernandez, J.: Adaptive parallel job scheduling with flexible coscheduling. In: IEEE Transactions on Parallel and Distributed Systems (2005)
14. Frachtenberg, E., Feitelson, D.G.: Pitfalls in parallel job scheduling evaluation. In: Feitelson, D.G., Frachtenberg, E., Rudolph, L., Schwiegelshohn, U. (eds.) JSSPP 2005. LNCS, vol. 3834, pp. 257–282. Springer, Heidelberg (2005)
15. Grid'5000 team, Dr.Franck Cappello (June 2010), <http://gwa.ewi.tudelf.nl/pmwiki/>

Shared Resource Allocation Using Token Based Control Strategy in Augmented Ring Networks

Rajendra Prasath

Department of Computer Science and Engineering
Indian Institute of Technology

Kharagpur - 721 302, West Bengal, India

rajendra@cse.iitkgp.ernet.in, drrprasath@gmail.com

Abstract. In this paper, we consider the token distribution problem in logically structured augmented ring networks. We propose *request-message-based* token control strategy for a single shared resource allocation in augmented ring networks like touching rings, interconnected rings and for two shared resources allocation in a bidirectional ring network. This strategy assures that request messages as well as token do not carry any information about their source as well as destination processors and request messages are served within a finite time. It is observed that the proposed protocols permit only a bounded number of message exchanges per request and can be extended for various higher dimensional interconnection networks.

1 Introduction

Consider the problem of controlling the allocation of a shared resource among a set of n processors that are arranged in augmented ring networks like touching rings, interconnected rings and two shared resources in a bidirectional ring network. We propose the solution by means of a *control token*. A processor which needs the shared resource must first get a free token and then changes it into a busy token. The busy token allows to utilize the shared resource in the critical section. After the critical section, the busy token is freed and this free token is again circulated in the network enabling one of the competing processes to access the shared resource.

This type of circular token based communication model is used in various combinatorial problems such as election problem, mutual exclusion problem [14, 9, 10, 11, 19, 20]. The original concept of “PREVILEGE” (token) was coined by Suzuki and Kasami in 1985 [21]. All these problems use the same procedure: an entity holding the token will pass it along the circular communication model as soon as it no longer needs it [8, 16]. But the unnecessary movements of the token amount to an unbounded number of message exchanges even for a finite set of requests. To avoid the unnecessary movements, we adopt the *request-message-based* token control strategy introduced by Feuerstein *et al.* [7] in which token is circulated only if it is informed of a request message generated for the shared resource.

Although a vast amount of literature [4, 17] exists on the token based control strategy, but little is known about the use of circular token based control mechanism. In distributed computing, detection of token loss and on self stabilizing aspects have been

investigated with special focus [3,5,10]. They assumed that the request message carries routing information about the source-destination identities. Using this routing information, we can find the shortest path [2,6,12] on which routing from source to destination node is made. Feuerstein *et al.* [7] have proposed two protocols based on the *request-message-based* token control strategy for controlling the allocation of a shared resource in an unidirectional fault-free ring network in which both the token as well as request messages move in a single direction. Then it had been applied to linear arrays, bidirectional rings extensions and mesh networks [13,14,15,23].

In this paper, we propose three algorithms for the shared resource allocation problem. Section 3 presents an algorithm for touching rings network. In section 4, we present an algorithm for single shared resource allocation among processors in two interconnected rings. Section 5 describes an algorithm for the allocation of two shared resources in a bidirectional ring network. The last section concludes the paper.

2 Preliminaries

Consider a touching rings network in which n processors are arranged in two subrings that touch at a common processor through which routing of the requests as well as token from one subring to another subring takes place either in an unidirectional or in a bidirectional way. The processor indexing starts with the processor next to the central touching processor and continues over the processors in clockwise direction; crosses the central touching processor; then progresses over the processors in the next subring (in clockwise) and finally ends up the indexing with the processor before the central processor. Here the role of the central processor is significant as it can switch the token as well as request messages into the subrings appropriately according to movements of request messages and token as well.

In an interconnected rings network, n processors are arranged in two subrings of $\frac{n}{2}$ processors each. The two subrings are interconnected in such a way that processor P_i in the first subring is connected with processor $P_{i+\frac{n}{2}}$ by a bidirectional link in which both the token as well as request messages can flow in either directions. We assume unidirectional movements for requests and token, over the ring edges and bidirectional over the edges connecting two subrings. In an interconnected rings network, we assume that each processor in the second subring follows a mechanism to sensor the movements of the token so as to send the token from the processor in the second subring to the processor in the first subring.

Next we suppose that a bidirectional ring network with n processors in which token moves in anticlockwise direction and request messages move in clockwise direction. Let $P = \{P_i\}$, $i = 1, 2, \dots, n$ be a set of n processors. For $i = 1, 2, \dots, n$ processor P_i may directly communicate with processor P_{i+1} and processor P_n is connected to P_1 . It is assumed that request messages do not carry any information about their origins.

We have used two distinct measures to analyze the proposed algorithms [7,22].

- (i) *Average number of messages per request*: How many messages are necessary to satisfy a sequence of requests on an average / worst case?, i.e., the *worst case ratio* between the total number of (token and request) messages and the number of requests in the sequence.

- (ii) *Service traffic*: The worst case number of messages that are exchanged in the network between the time in which a processor sends a request message and the time in which it gets served.

The service traffic should be bounded to arrive at a feasible solution.

We start with the lower bound derived for a fault-free unidirectional ring network.

Theorem 1. (Feuerstein et al, 1998)

Any algorithm for the token distribution problem in an unidirectional ring with anonymous request messages requires atleast $(n - 1)$ messages per request and the service traffic is atleast $(n - 1)$. ■

3 Touching Rings Network

We consider unidirectional touching rings network with the central touching processor as $P_{\lceil \frac{n}{2} \rceil}$. Incoming links to the central processor are from $P_{\lceil \frac{n}{2} \rceil - 1}$ and P_n and outgoing links are to P_1 and $P_{\lceil \frac{n}{2} \rceil + 1}$. Each processor P_i [$i = 1, 2, 3, \dots, n$] has a local variable M_p , used to stop further movements of the following request messages. The value of M_p is set to 1, if P_i has a pending request message of its own or a request message has been passed through it and 0, otherwise. Now assume that the central processor $P_{\lceil \frac{n}{2} \rceil}$ has a token locator T_l to trace the token location in the subrings and a request recorder V_r that helps to record the pending requests of one subring when token has passed through the other subring. The value for T_l is assigned to 0, if token has passed to the processor in the first subring from the central processor and 1, otherwise.

V_r is assigned to 1, if a request is received from P_n to $P_{\lceil \frac{n}{2} \rceil}$ [similarly from $P_{\lceil \frac{n}{2} \rceil - 1}$ to $P_{\lceil \frac{n}{2} \rceil}$] when $T_l = 0(1)$ and 0, otherwise. Initially the values of M_p and V_r are assumed to be zero and token is placed at the central processor. It can be observed that whenever $T_l = 0(1)$ then V_r indicates the receipt of a request message from the second (respectively first) subring. Token maintains a flag bit with two cells and sets as and when it recognizes either $V_r=1$ at $P_{\lceil \frac{n}{2} \rceil}$ or the receipt of a request message. The flag bit cells are indexed as T_1 and T_2 which takes 1, if a request is received from first and second subrings respectively and 0, otherwise. The value of flag bit cells helps token to move through the processors whose M_p might be zero in the next subring. At the central processor, token moves according to the value of M_p and V_r . Token stops only if the values of M_p, T_1 and T_2 are zero.

ALGORITHM : UTRN

- (a) When P_i needs resource then
 - If P_i has token then it enters critical section.
 - If P_i has no token then
 - if P_i is the central processor then
 - if $T_l = 0(1)$ then set $M_p = 1$ and send the request to P_1 ($P_{\lceil \frac{n}{2} \rceil + 1}$)
 - else (if P_i is not the central processor then)

if $M_p = 0$ then set $M_p = 1$ and send the request to next processor
else no request message is sent.

(b1) When P_i receives a request message

If P_i has token then

if P_i is the central processor then

if request is from $P_{\lceil \frac{n}{2} \rceil - 1} (P_n)$ then set $T_l = 0(1)$; set $T_1(T_2) = 1$;

set $M_p = 0$ and send token to $P_1 (P_{\lceil \frac{n}{2} \rceil + 1})$ in the current subring

else(if P_i is not the central processor then)

if $1 \leq i < \lceil \frac{n}{2} \rceil$ then set $T_1 = 1$;

else (if $\lceil \frac{n}{2} \rceil < i \leq n$ then) set $T_2 = 1$

and send token to P_{i+1} in clockwise direction.

(b2) When P_i receives a request message and has no token then

If P_i is the central processor then

if request is from $P_{\lceil \frac{n}{2} \rceil - 1} (P_n)$ then

if $T_l = 0(1)$ then set $M_p = 1$;

if $V_r = 1$ then stop the request at P_i ;

else send the request to $P_1(P_{\lceil \frac{n}{2} \rceil + 1})$

else(if $T_l = 1(0)$ then) set $V_r = 1$

if $M_p = 1$ then stop the request at P_i ;

else send the request to $P_{\lceil \frac{n}{2} \rceil + 1}(P_1)$

else (if P_i is not the central processor then)

if $M_p = 0$ then set $M_p = 1$ and send the request to P_{i+1} ;

else stop the request at P_i .

(c1) When P_i receives token and has a pending request then

it enters critical section; set $M_p=0$ and at the end of critical section,

if P_i is the central processor then

if $V_r = 0$ then set $T_{1(2)} = 0$; $T_l = 0(1)$ and send token to $P_1 (P_{\lceil \frac{n}{2} \rceil + 1})$
in the current subring

else(if $V_r = 1$ then)

if token is from $P_{\lceil \frac{n}{2} \rceil - 1} (P_n)$ then set $T_{2(1)} = 1$; $T_l = 1(0)$; $V_r = 0$ and

send token to $P_{\lceil \frac{n}{2} \rceil + 1}(P_1)$

else (if P_i is not the central processor then)

if $1 \leq i < \lceil \frac{n}{2} \rceil$ then set $T_1 = 0$; else set $T_2 = 0$

and send token to P_{i+1} .

(c2) When P_i receives token and has no pending request then

if P_i is the central processor then

if $V_r = 1$ then

if token is from $P_{\lceil \frac{n}{2} \rceil - 1} (P_n)$ then set $T_l = 1(0)$; reset $V_r = 0$;

if $M_p = 1$ then set $T_1=T_2 = 1$; otherwise set $T_2(T_1) = 1$

and send token to $P_{\lceil \frac{n}{2} \rceil + 1} (P_1)$

else (if $V_r = 0$ then)

If $M_p = 1$ then set $M_p = 0$ and send token to the next processor in
the current subring

else(if $M_p = 0$ then)

if T_1 AND $T_2 = 0$ then stop token at P_i
 else send token to the next processor
 else (if P_i is not the central processor then)
 if M_p OR T_1 OR $T_2 = 1$ then reset $M_p = 0$ and pass token to P_{i+1}
 else if M_p AND T_1 AND $T_2 = 0$ then stop the token at P_i itself.

Theorem 2. The algorithm UTRN satisfies all request messages generated for the single shared resource.

Proof: Initially assume that token is placed at the central processor. Each request message will eventually reach the token and informs it to access the shared resource. Token, on receiving a request message, moves over processors in the appropriate subring by setting T_i value at the central processor accordingly. Meantime V_r and M_p are updated at the central processor to track the receipt of request messages from the respective subrings.

In another case, assume that token is present at any one of the processors in the subrings. Then the central processor when it receives a request message checks the values of T_i , V_r and M_p and sets them as in steps (b1) and (b2) and then the request message is sent to the corresponding subring based on the availability of the token. Whenever token reaches the central processor, it checks the values of V_r and M_p . If there is a pending request message of the next subring then the token is forwarded to the next subring; otherwise the token is forwarded to the next processor in the current subring. As and when token switches over subrings, V_r and M_p are dynamically maintained. Token moves through the processors until T_1 AND $T_2 = 0$ and $M_p = 0$. Thus request messages are not lost and token is aware of all request messages. Thus the algorithm UTRN satisfies all request messages in the network. ■

Theorem 3. The algorithm UTRN needs atmost $(2n - 1)$ messages per request and service traffic, in the worst case, amounts to $(\frac{5n-2}{2})$.

Proof: Assume that the token is at the central processor. Now suppose that a request message has been originated at some processor in a subring. In fact, regardless of n , the total number of processors, at most $(\frac{n}{2} - 1)$ movements for each subring are required in the worst case to inform the token. If token has passed through the first subring then the request coming from second subring moves to the first subring if $M_p = 0$ at the central processor, else the request message stops. Meantime, atmost $(n - 1)$ processors may generate request messages. Thus in addition, token may have to move over atmost n processors. Thus in the worst case, the total number of message exchanges per request amounts to $(2n - 1)$.

To prove the second measure, two types of request messages may be sent between the time in which P_i generates a request message and the time in which it is served. The first type includes the request messages that require atmost $(n-1)$ movements to inform the token. In response to this request, token needs to move atmost $\frac{n}{2}$ moves for switching it into the next subring; $\frac{n}{2}$ moves for the next subring and $\frac{n}{2}$ movements of either the current or the next subring. Thus in total $\frac{3n}{2}$ messages are needed for the token to reach the processor with a pending request. The last $\frac{n}{2}$ movements guarantee that token is aware of all pending request messages of one subring when token is passing through another subring. Thus the service traffic in the worst case amounts to $(\frac{5n-2}{2})$. ■

Here the total number of message exchanges per request in UTRN resembles with the complexity of an unidirectional ring network [7]. Also in the best case, the complexity of the algorithm UTRN converges to Theorem 1.

Instead of assuming unidirectional movement over the touching rings network, we can assume opposite directional movement for both the token as well as request messages. Such a touching ring network is termed as a *bidirectional touching rings network* in which token flows in clockwise and request messages flow in anticlockwise direction. The algorithm UTRN still can be extended to a bidirectional touching rings network by allowing the movement of request messages only in anticlockwise direction. The resulting algorithm with bidirectional movements, we call it as BTRN, ensures the fastest service to the pending request messages. An immediate result follows in the sequel.

Theorem 4. Algorithm BTRN for *bidirectional touching rings networks* serves all requests, requires atmost $(n-1)$ messages per request and service traffic is bounded by $(2n-1)$.

Proof: As the request messages move in anticlockwise direction and token in clockwise direction, no request message would be skipped by the token and hence all request messages are served. Now to prove the first measure, we consider the variables that are dynamically maintained according to the opposite movements of the token as well as the request messages. Here $(n - 1)$ movements are necessary to satisfy a sequence of requests and hence average number of messages per request amounts to $(n - 1)$. Next we prove the second measure. As the token and request messages move in opposite direction, token can easily sensor and serve all the pending requests in its way to the destination. Token stops moving only when it reaches a processor with no pending request message. Hence the service traffic amounts to $(2n-1)$. ■

The algorithm UTRN can further be generalized for *multi-touching rings network* with n number of processors in the central ring and m_i number of subrings each with k processors [excluding the touching processor] and $n, k \geq 2$. The multi-touching rings network is formed in such a way that each subring m_i with k processors is connected with only one processor in the central ring network (*see* Fig. 1). Thus there are totally $N = (n + \sum_{i=1}^n m_i * k)$ processors.

We assume that the multi-touching rings network with N processors is completely fault-free. Now the proposed UTRN algorithm can be extended to the multi-touching rings network by considering each subring with the remaining part of the network and for each processor in the central ring network. By this way, each processor in the central ring assumes all the variables as in UTRN and operates in parallel. Still the basic assumptions, like processors have neither a shared memory nor a common clock and their speeds are not related, are hold. Then we derive the following result.

Theorem 5. The generalized algorithm for multi-touching rings network amounts to $(N - 1)$ messages per request and $(2N + n - 1)$ service traffic in the worst case. ■

It can be guaranteed that this generalized algorithm works perfectly for variable k also. But as n and k grow, the transmission delay for serving the pending requests would be additional overhead.

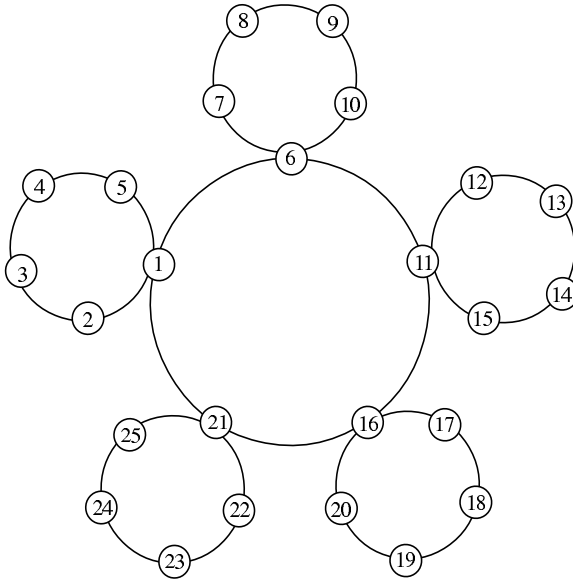


Fig. 1. A multi-touching rings network with $N = 25$ processors (Here $n = 5$, $m = 5$ and $k = 4$)

4 Interconnected Rings Network

Next we describe *request-message-based* token control mechanism for shared resource allocation in an interconnected rings network. In this architecture, there are two subrings with $\frac{n}{2}$ processors each, where n is the total number of processors (assumed to be even). The interconnections between the subrings are made in such a way that processor P_i ($i=1,2,3,\dots,\frac{n}{2}$) in the first subring is connected to the processor $P_{i+\frac{n}{2}}$ in the second subring by a bidirectional channel in which the token as well as the request messages can pass through in either directions.

We follow a few assumptions: Each processor has a local variable M_p that takes 1, if it has a pending request(may be of its own or received from the previous processor in the first subring) and 0, otherwise. Each processor in the first subring has a flag variable V_p whose value is 1, if a pending request has been received from $P_{i+\frac{n}{2}}$ by P_i and 0, if there is a non-receipt of the request messages from $P_{i+\frac{n}{2}}$. The value of V_p will be checked by the token to identify the request messages received from the second subring. Each processor P_i resets V_p value from 1 to 0 only on receipt of a request message and not by the movement of the token through it. Token has a counter to record the receipt of the request messages and token stops moving only when the values of the token counter and M_p are zero.

ALGORITHM : ICRN

- (a) When P_i needs resource then
 - If P_i has token then it enters the critical section

If P_i has no token then

If $1 \leq i \leq \frac{n}{2}$ then

if $M_p = 0$ then set $M_p = 1$;

if $V_p = 0$ then send the request to P_{i+1}

else ($V_p = 1$ and $M_p = 1$) stop the request at P_i

else (if $\frac{n}{2} + 1 \leq i \leq n$ then) send the request from P_i to $P_{i-\frac{n}{2}}$

(b1) When P_i receives a request message and has token then

increase the token counter by one; reset $M_p = 0$ and

if $V_p = 1$ then set $V_p = 0$ and send token to $P_{i+\frac{n}{2}}$

else send token to the next processor P_{i+1} .

(b2) When P_i receives a request message and has no token then

if the request is from $P_{i+\frac{n}{2}}$ then set $V_p = 1$

if $M_p = 1$ then stop the request at P_i

else send the request to P_{i+1}

else (if the request is from P_{i-1} then)

if $M_p = 1$ then stop the request at P_i

else set $M_p = 1$ and

if $V_p = 1$ then stop the request at P_i

else send the request to P_{i+1} .

(c1) When P_i receives token and has a pending request message then

it enters critical section; decrease token counter by 1;

reset $M_p = 0$; at the end of critical section,

if $1 \leq i \leq \frac{n}{2}$ then

if token is from P_{i-1} then

if $V_p = 1$ then increase counter by 1; set $V_p = 0$ & send token to $P_{i+\frac{n}{2}}$

else send token to P_{i+1}

else (if token is from $P_{i+\frac{n}{2}}$) reset $V_p = 0$;

if token counter is bigger than zero then, send token to P_{i+1} ;

else stop token at P_i

else (if $\frac{n}{2} + 1 \leq i \leq n$ then)

if token is from P_{i-1} then send token to $P_{i-\frac{n}{2}}$

else send token to P_{i+1} .

(c2) When P_i receives token and has no pending request message then

if $1 \leq i \leq \frac{n}{2}$ then

if token is from P_{i-1} then

if $V_p = 0$ then

if $M_p = 1$ then set $M_p = 0$ and send token to P_{i+1}

else if (token counter > 0) then send token to P_{i+1} ;

else stop token at P_i

else increase token counter by one; reset

$M_p = 0$; $V_p = 0$ and send token to $P_{i+\frac{n}{2}}$

else (if token is from $P_{i+\frac{n}{2}}$ then)

if $M_p = 1$ then reset $M_p = 0$; $V_p = 0$;

if (token counter > 0) then send token to P_{i+1} ; else stop token at P_i

else (if $M_p = 0$ then) send token to the next processor if token counter is bigger than zero
 else (if $\frac{n}{2} + 1 \leq i \leq n$ then)
 if token is from P_{i-1} then send token to $P_{i-\frac{n}{2}}$;
 else send token to the processor P_{i+1} .

Theorem 6. The algorithm ICRN serves all request messages generated for the single shared resource.

Proof: Assume an interconnected rings network with two subrings. If there is no request message then token resides at any one of the first subring processors. So each originated request message from the processor in the first subring will eventually find out the token and push it till the next processor with a pending request. Infact each request message generated at the inner ring processors needs atmost $(\frac{n}{2} - 1)$ moves to reach the token in the worst case. Meanwhile each processor in the second subring requires just one move from processor P_i to $P_{i-\frac{n}{2}}$ and then moves over the processors in the first subring. In the worst case, as $(n - 1)$ processors may originate request messages for the token, the zig-zag movements of the token between the subrings help to serve all pending request messages. Thus the algorithm ICRN satisfies all processors in the network. ■

Theorem 7. The algorithm ICRN requires $(n - 1)$ messages per request and service traffic is $(3n - 2)$.

Proof: To prove the first measure, let the token be initially at any one of the processors P_i in the first subring. Now each request at P_{i+1} has to traverse atmost $(\frac{n}{2} - 1)$ movements before it reaches the token or it has been stopped by the preceding request message. Meanwhile each processor in the second subring may originate and send requests for token. Thus in the worst case, request messages generated at the second subring require one move to reach their corresponding processors in the first subring and then moves $(\frac{n}{2} - 1)$ processors in the first subring. This overall results in $\frac{n}{2} + (\frac{n}{2} - 1) = (n - 1)$ moves required to inform the token. Hence, in the worst case, the number of messages per request amounts to $(n - 1)$.

In order to prove the second measure, service traffic, consider two types of movements (request as well as token) between the two subrings. The first type includes requests from the first subring and in the absence of the second subring, atmost $(\frac{n}{2} - 1)$ messages may be sent to inform the token. But by the interconnection of the second subring, each request originated from second subring processors requires $\frac{n}{2}$ moves in total and then moves over the processors in the first subring provided if $M_p = 0$. Thus totally $(n - 1)$ messages are needed to inform the token in the worst case. Token starts serving the processors in the first and second subrings according to the value of V_p . The zig-zag movements of the token guarantee that it serves all requests by visiting each processor atleast once. Thus n messages are required to visit each processor atleast once. So atmost $(2n - 1)$ messages are needed to service the processors with a pending request, regardless of the position of the processors in the subrings. As each processor is again eligible to generate a request message after its previous request message has been served, additionally $(n - 1)$ movements are needed. Hence the service traffic in the worst case amounts to $(3n - 2)$ and token skips no request messages. ■

5 Two Shared Resources

Next we consider the problem of token based controlling of two shared resources in a bidirectional ring network. We assume two tokens, T_1 and T_2 , that are permitted to pass through anticlockwise direction where as request messages are sent in clockwise direction. We define the variables as follows: M_p assumes 1, if P_i has a pending request of its own and 0, otherwise. Also we assume a request recorder V_n that takes the value 1, for a pending request received from P_{i-1} and 0, for non-receipt of a pending request message. Initially T_1 is placed at P_1 , T_2 is at $P_{\lfloor \frac{n}{2} \rfloor}$ and $M_p=0=V_n$ in all processors. We restrict the movement of the following request message according to the value of M_p . Also token moves until V_n is non-zero. The behavior of a processor that executes this algorithm is as follows.

Algorithm : R2 Token based control of two shared resources in bidirectional rings

- (1) When P_i needs resource
 - * If P_i has token then it enters the critical section;
 - * otherwise set $M_p=1$ and
if $V_n = 0$ then send the request to P_{i+1}
else stop the request at P_i itself.
- (2) When P_i receives request
 - * If P_i has token then send token to P_{i-1} by resetting $M_p=0$; $V_n=0$;
 - * otherwise set $V_n = 1$;
if $M_p=0$ then send the request to P_{i+1}
else($M_p=1$) stop the request at P_i itself.
- (3) When P_i receives token
 - * If P_i has a pending request then it enters critical section; set $M_p=0$ and
if $V_n=1$ then set $V_n=0$; exit the critical section and send token to P_{i-1}
else (if $V_n=0$ then) exit the critical section and stop the token at P_i itself
 - * If P_i has no pending request then set $V_n = 0$ and send token to P_{i-1} .

Theorem 8. The algorithm R2 satisfies all requirements of the processors. It needs atmost $2(n - 2)$ messages per request and its service traffic is $5(n - 2)$.

Proof: A request message is allowed to circulate in clockwise direction and tokens in anticlockwise direction. At first, T_1 is placed at P_1 and T_2 is placed at $P_{\lfloor \frac{n}{2} \rfloor}$. Each request originated between T_1 and T_2 in either side sets $M_p = 1$ and moves in clockwise direction searching the token. On the way it sets $V_n = 1$ if that processor has no pending request of its own. Finally it informs the token. Then token starts moving in anticlockwise direction by searching the pending requested processor. If $M_p = 1$ at that processor then token serves it and sets $M_p = 0$ and if $V_n = 1$ then token moves to the next processor; else stops at that processor itself. The value of V_n helps token to recognize the pending request of the next processor in clockwise direction and hence in the worst case, token $T_1(T_2)$ stops before the processor in which $T_2(T_1)$ is available. Thus the algorithm R2 serves all requirements of the processors.

To prove the first measure, we use amortized analysis techniques [7, 22]. First we assume that R is the set of total pending request messages spread over the processors in the network; M is the set of waiting request messages for the token and C is the current value of the local variable V_n of the processor at which the token is residing. It is clear that $R = M + C$.

Now for each pending request message $r_i \in R$, $d(r_i)$ denotes the anticlockwise distance from token T_1 (or T_2) to the processor P_i with a pending request r_i . Similarly for each waiting request $m_j \in M$, $d(m_j)$ denotes the clockwise distance from the processor currently possessing a waiting request m_j to the processor presently having the token T_1 (or T_2).

Next we consider a potential function Φ as:

$$\Phi = - \sum_{r_i \in R} d(r_i) + \sum_{m_j \in M} d(m_j)$$

Here the *minus* sign denotes the anticlockwise direction from T_1 (or T_2) to the pending request. The formerly generated request message only reaches the token and all other following request (waiting) messages sets $V_n = 1$ and stops if $M_p = 1$ or forwarded to the next processor in clockwise direction if $M_p = 0$.

Token, after initiation, is passed to the next processor P_{i-1} from P_i (P_n from P_i , if $i = 1$). Then by the algorithm R2, the variable V_n is exactly $1 > 0$ implies $C = 1$. In the worst case, as two tokens T_1 and T_2 are in adjacent processors, there would be $(n - 3)$ pending (waiting) requests. So the potential function increases for the first request received by token and decreased by one for each pending request as well as waiting request message. Since token moves in anticlockwise direction, it traces all request messages on its way to the destination due to M_p and V_n as well. The variation in Φ is $\Delta\Phi = +M - R = -C = -1$. Hence the amortized cost is $1 + \Delta\Phi = 0$.

In contrast, each new request in the worst case increments Φ exactly $2(n - 2)$ when $V_n > 0$ and $M_p > 0$ and for k pending (including waiting) request messages, the variation in Φ can not exceed $k * 2(n - 2)$. This implies an average of $2(n - 2)$ message exchanges per request.

The service traffic accounts the total number of message exchanges per request between the time in which a processor P_i sends a request and the time in which it gets served. Also we assume that all the request messages are originated only after P_i 's request. Here two different types of requests may be exchanged. The first type includes the requests from P_i , which requires atmost $(n - 2)$ movements to reach the token T_1 (T_2 is at the adjacent processor) in the clockwise direction. In order to serve the request of P_i , token T_1 (similarly T_2) may have to pass atmost $(n - 2)$ processors. As each token T_1 (or T_2) serves the pending requests along the path and moves, these processors can further ask for resource and cause movement of the second token. Even after getting served by the second token, the processors can further ask the resource and the number of messages exchanged in the worst case amounts to $5(n - 2)$. Hence the service traffic in the worst case is bounded by $5(n - 2)$. ■

It is to be noted that this strategy could further be extended to control any number of shared resources in the massively larger networks. It seems that the measure of service

traffic might be relative as the number of resources grows and in this case we can try to look for only how fast a single shared resource can serve request messages.

6 Conclusion

In this paper, we considered the token distribution problem for shared resource allocation using *request-message-based* strategy in various augmented ring networks. First algorithm UTRN allocates a single shared resource in an unidirectional touching rings network. This could be generalized to multi-touching rings network. Then we presented an algorithm ICRN for unidirectional interconnected rings network with two interconnected subrings. Also this algorithm could be extended to several wrap around interconnection networks. But the communication delay would be overhead as the total number of interconnected layers increases. Finally we presented an algorithm for the allocation of two shared resources in a bidirectional ring network. The next step is to develop a model using *request based token passing* strategy for various higher dimensional interconnection networks.

Acknowledgement

The partial support provided by the University of Madras during the early stages of this work is gratefully acknowledged.

References

1. Andrews, D., Schulz, G.: A token-ring architecture for local area network: an update. In: Proc. IEEE COMPCON (1982)
2. Arden, B.W., Lee, H.: Analysis of chordal ring network. IEEE Trans. on Computers C-30(4), 291–295 (1981)
3. Ben-Dor, A., Haveli, S., Schuster, A.: Potential function analysis of Greedy hot potato routing. Theory of Computing Systems 31, 41–61 (1998)
4. Bertsekas, D., Gallager, R.: Data Networks, 2nd edn. Prentice Hall, Englewood Cliffs (1999)
5. Burns, J.E., Pahl, J.: Uniform self-stabilizing rings. ACM Trans. on Prog. Languages and Systems 11(2), 330–334 (1989)
6. Chalamaiah, N., Ramamurthy, B.: Finding shortest paths in distributed loop networks. Information Processing Letters 67, 157–161 (1998)
7. Feuerstein, E., Leonardi, S., Marchetti-Spaccamela, A., Santoro, N.: Efficient token based control in rings. Information Processing Letters 66, 175–180 (1998)
8. Halsall, F.: Data Communication, Computer Networks and Open Systems, 3rd edn. Addison-Wesley, Reading (1992)
9. Wu, J.: Distributed System Design. CRC Press, Boca Raton (1999)
10. Lann, G.L.: Distributed Systems - Towards a formal approach. Information Processing Letters 77, 155–160 (1977)
11. Lodya, S., Kshemkalyani, A.: A fair distributed mutual exclusion algorithm. IEEE Trans. in Parallel and Distributed Systems 11(6), 537–549 (2000)
12. Mukhopadhyaya, K., Sinha, B.P.: Fault-Tolerant routing in distributed loop networks. IEEE Trans. on Computers 44(12), 1452–1456 (1995)

13. Rajendra Prasath, R., Thangavel, P.: Token based message passing in bidirectional ring extensions. *Journal of the Madras University* 52, 145–159 (2000)
14. Rajendra Prasath, R., Thangavel, P.: Shared resource allocation using token based control in ring extension topologies. In: *Proc. of Int. Conference on Recent Advances in Mathematical Sciences*, pp. 53–63 (December 2000)
15. Rajendra Prasath, R., Thangavel, P.: Shared resource allocation using token passing strategy in interconnected networks. *Information* 6(2), 197–206 (2003)
16. Raynal, M.: *Distributed Algorithms and Protocols*. John Wiley and Sons, Chichester (1988)
17. Rego, V., Ni, L.M.: Analytic models of cyclic service systems and their applications to token-passing local networks. *IEEE Trans. on Computers* C-37(10), 1224–1234 (1988)
18. Reisig, W.: *Elements of Distributed Algorithms*. Springer, Heidelberg (1998)
19. Ross, F.E.: FDDI- A tutorial. *IEEE Communication magazine* 24, 10–17 (1986)
20. Stallings, W.: *Data and Computer Networks*, 5th edn. Prentice-Hall, Englewood Cliffs (1997)
21. Suzuki, I., Kasami, T.: A distributed mutual exclusion algorithm. *ACM Transactions on Computer Systems* 3(4), 344–349 (1985)
22. Tarjan, R.E.: Amortized computational complexity. *SIAM J. Alg. Discrete Math.* 6(2), 306–318 (1985)
23. Thangavel, P., Rajendra Prasath, R.: A note on token based control in rings and linear arrays. *Journal of Comp. Soc. India* 32(3), 62–65 (2002)

An Algorithmic Approach to Minimize the Conflicts in an Optical Multistage Interconnection Network

Ved Prakash Bhardwaj¹, Nitin², and Vipin Tyagi³

¹ Department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology, Waknaghat, Solan-173234, Himachal Pradesh, India
ved.juit@gmail.com

² Department of Computer Science, College of Information Science and Technology, University of Nebraska at Omaha, Omaha-68182-0116, Nebraska, United States of America
fnunitin@mail.unomaha.edu

³ Department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology, Waknaghat, Solan-173234, Himachal Pradesh, India
dr.vipin.tyagi@gmail.com

Abstract. Multistage interconnection networks (MINs) consist of more than one stage of small interconnection elements called switching elements and links interconnecting them. A MIN connects N inputs to N outputs and is referred as an $N \times N$ MIN, having size N . An Optical MIN (OMIN) is an important class of Interconnection networks. The problem of crosstalk is caused by coupling two signals within switching elements. A number of techniques like Optical window, Heuristic, Genetic, and Zero have been proposed earlier in this regard. In this paper, we have proposed an Address Selection Algorithm (ASA) and we have applied to existing Omega network, having shuffle-exchange connection pattern. The aim of this algorithm is to minimize the number of switch conflicts in the network and to provide conflict free routes.

Keywords: Optical Multistage Interconnection Networks, Crosstalk, Omega Network and Time Domain Approach.

1 Introduction and Motivation

Advances in optical technologies have shown the interest for optical implementation in MINs [1] to achieve high bandwidth capacity, low error probability. OMINs [2] are the advance field of research for the communication network. This network consists of N inputs, N outputs, and n stages ($n=\log_2 N$). Each stage has $N/2$ Switching Elements (SEs) comprising of two inputs and two outputs connected in a particular pattern [3]. The three common approaches space, time or wavelength domain is used to reduce the effect of crosstalk. In the present paper, our interest is on the Time Domain Approach (TDA) for solving optical crosstalk in Optical Omega Network (OON) [4]. In this approach, the two communication signals will be passed at different times if they are using same switching element. In this paper, we have presented a new algorithm called Address Selection Algorithm. It minimizes the number of conflicts in the network and provides crosstalk free routes in the network.

The rest of the paper is as follows: Section 2 and Section 3 discusses the Problems arises due to Crosstalk and Preliminaries and Background. In Section 4, we have discussed our proposed algorithm and applied the same on OMIN by taking an example in Section 5 is followed by conclusion and references.

2 Problem of Crosstalk

The problem of crosstalk [5] may occur when two signals within a switch tries to interact with each other .There are various reasons of crosstalk. Some basic reasons of crosstalk within a switching element are shown in figure (1). Firstly, when two signal channels having exchange connection pattern. Secondly, when one signal channel is having upper straight route and second is having upper broadcast route, third when one signal is having lower broadcast route and second is having lower straight route. To avoid this problem the legal passes is shown in figure (2). First, both signal channels should have straight route. Second, only one signal should pass through the switching element either it has lower broadcast or upper broadcast route.

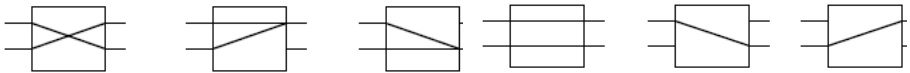


Fig. 1. Crosstalk

Fig. 2. Legal Passes

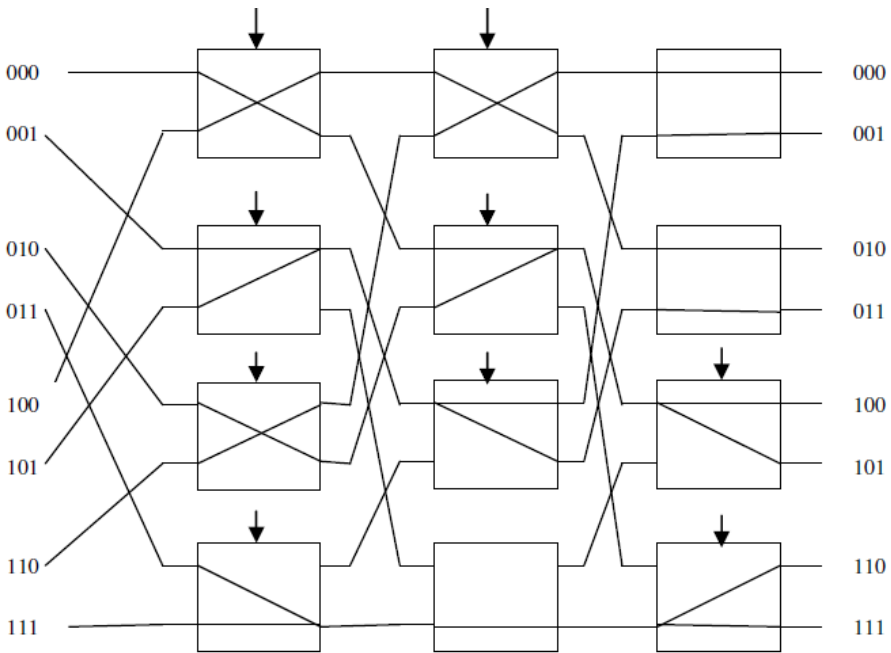


Fig. 3. An 8x8 Optical Omega Network With Conflicts

2.1 Crosstalk in Optical Omega Network (OON)

The Omega Network (ON) is an example of a banyan multistage interconnection network and it can be used as a switch fabric [2]. The Omega MIN uses the “Perfect Shuffle”. It has self-routing property. This network connects N input to N output nodes using n stages, where $n = \log_2 N$ with each stage containing 2^{n-1} SEs. In the Optical Omega Network [6], each communication signal must go through a number of switching stages. The routing path of each packet is set according to its destination. Figure (3) shows a 3-stage, 8×8 Optical Omega Network. The problem of crosstalk within a SE is shown in figure (3). The arrows in figure (3) indicate the conflicted switching element. There are nine conflicts in the network. All conflicted SEs having a common reason viz. the communication signals are trying to interact with each other [7, 8].

3 Preliminaries and Background

Time Domain Approach (TDA) is the basic approach of this paper [9]. In TDA, Crosstalk is considered as a conflict. This approach makes its important because of various reasons like most of the multiprocessors use electronic processor and Optical MINs. Therefore, there is a big mismatch between the processing speeds of the network carrying optical signals. In TDA, Permutation and Semi-permutation [10] is applied on the message groups so that each group is routed in a different time slot. The source and destination address is combined to build combination matrix [7]. On the basis of combination matrix message partitioning is performed so that some specific message should get their destination in the first pass and network remains crosstalk free [11]. Window Method [2], Heuristic Routing Algorithm [5] and many other algorithms are available for message partitioning. In this paper, the focus is to provide best message partitioning scheme so that a conflict free network can be obtained. Before describing our algorithm just have a look on the Window Method and Heuristic Routing Algorithm.

3.1 Window Method

This method [1, 2] eliminate the source addresses which having same bit pattern to avoid crosstalk problem and these eliminated source addresses will get their destination in next pass. For network size $N \times N$, there are N source and N destination address. Each source and its corresponding destination address are combined to produce a combination matrix. From this matrix, the optical window size is $M-1$, where $M = \log_2 N$ and N is the size of the network [2, 3]. The first and last columns are not considered of the combination matrix. We have to focus on the remaining columns for the further processing so that a crosstalk free network can be obtained.

3.2 Heuristic Routing Algorithm

There are four approaches of this algorithm to schedule the messages in different passes in order to avoid the path conflicts in the network [5]. In the first approach, messages are chosen sequentially in increasing order of the source addresses. The

same job is performed for the second approach but source addresses will be in decreasing order. In the third approach messages are chosen on the basis of increasing degree in the conflict graph. The fourth approach is same as the third one but here messages are chosen on the basis of decreasing degree of the message source address. Scheduling the message in decreasing degree of the message conflicts gave the good performance among these four approaches [5].

4 Proposed Approach

The aim of present algorithm is to select such particular source address in first pass, which do not create conflict in the network, and the remaining source address can be transmitted in second pass. This algorithm is applicable on 8×8 Optical Multistage Interconnection Network and its above series. In our approach first, we get the source and destination address sequentially. Second, we find the combination matrix of the source and corresponding destination address. Now transformation is applied on the combination matrix. The transposed matrix will have a particular set of rows. Now select the middle four rows and eliminate the remaining rows. In this way, two pair of rows can be obtained. In the next step, addition operation will be performed between corresponding bits in each pair. Therefore, two different sets will be obtained for the next step. We subtract the obtained result of first set with its corresponding result of second set.

Address Selection Algorithm (ASA)

1. Get the source and destination address sequentially.
2. Make combination matrix of the source and corresponding destination address.
3. Transform the matrix. A complete set of rows are thus formed $r_0, r_1 \dots r_n$, where $n =$ total number of bits in source and destination address.
4. Select the middle four rows.
5. Pair the obtained rows in set of two.
6. Add the corresponding bits of each pair.
7. Subtract the result of first set from second set.
8. If (result ≤ 0)
 - Then take corresponding address and transmit them in current pass and go to step 9.
 - Else
 - Store the address in remaining_address.
9. If (Conflict)
 - Then transmit the address with higher magnitude of the conflicting address pair and add the lower magnitude address to the remaining_address.
10. Transmit the remaining_address.
11. End.

If the result is a positive number then store it in a variable called the remaining_address. If conflict occurs in the current pass then we transmit only those addresses, which have higher magnitude and the address, which has lower magnitude, store it in the remaining_address. Now in the second pass, transmit the all remaining addresses, which are store in remaining_address. In this way, a conflict free network can be obtained.

5 Results

Example1: Let the source and destination address as follows and these are going to be inputs for the Algorithmic steps.

Source	Destination
000	100
001	011
010	101
011	110
100	010
101	001
110	000
111	111

Algorithmic Step1: Get the source and destination address sequentially.

Source	Destination
000	100
001	011
010	101
011	110
100	010
101	001
110	000
111	111

Algorithmic Step2: Make combination matrix of the source and destination address. Like source address is 000 and destination address is 100. Therefore, it will become 000100. Similarly we can get the other combination. It is clear from the combination matrix.

0	0	0	1	0	0
0	0	1	0	1	1
0	1	0	1	0	1
0	1	1	1	1	0
1	0	0	0	1	0
1	0	1	0	0	1
1	1	0	0	0	0
1	1	1	1	1	1

Algorithmic Step3: Transpose the matrix. Row 1 of the matrix represents r_0 , Similarly the other row of the matrix represents r_1 till r_5 .

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Algorithmic Step4: select the middle four rows, eliminate the first and last row, and get the remaining rows. Row 1 of the matrix represents r_1 , Similarly the other row of the matrix represents r_2 till r_4 .

$$\begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Algorithmic Step5: Pair the obtained rows in set of two. The first pair will have rows r_1 and r_2 . The second pair will have rows r_3 and r_4 .

r_1	r_2	r_3	r_4
0	0	1	0
0	1	0	1
1	0	1	0
1	1	1	1
0	0	0	1
0	1	0	0
1	0	0	0
1	1	1	1

Algorithmic Step6: Add the corresponding bits of each pair.

r_1+r_2	r_3+r_4
0	1
1	1
1	1
2	2
0	1
1	0
1	0
2	2

Algorithmic Step7: Subtract the result of the first set from the second set.

- $(r_1+r_2) - (r_3+r_4)$
- -1
- 0
- 0
- 0
- -1
- 1
- 1
- 0

Algorithmic Step8: Now the corresponding addresses, having their result either zero or negative number will be transmitted. These addresses are 000, 001, 010, 011, 100 and 111. The transmission of these selected addresses is clear from figure (4).

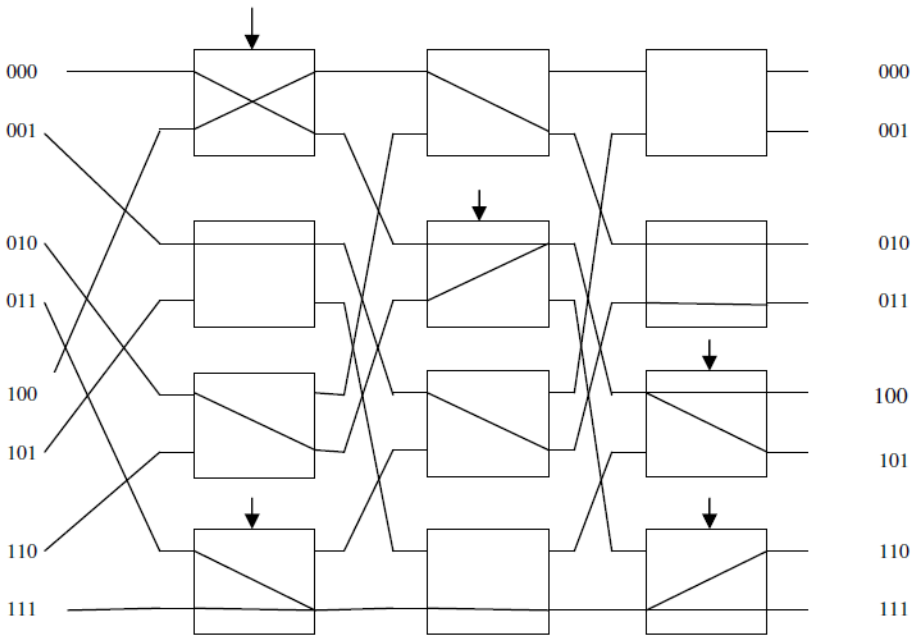


Fig. 4. Optical Omega Network with Reduced Conflict

Now the addresses, which are having results in positive numbers, will be stored in remaining_address variable. These addresses are 101 and 110.

Algorithmic Step9: Now finding the conflicted addresses. From figure (4), the arrows show the conflict in the network. So select the conflicting addresses and find which is having greater magnitude. Therefore, in the above network 000 and 100 is having conflict. The addresses 011 and 111 having conflicts. Similarly, the other conflicted address can be obtained as shown in figure (4). Finally, store the address 000 and 011 in the remaining_address variable.

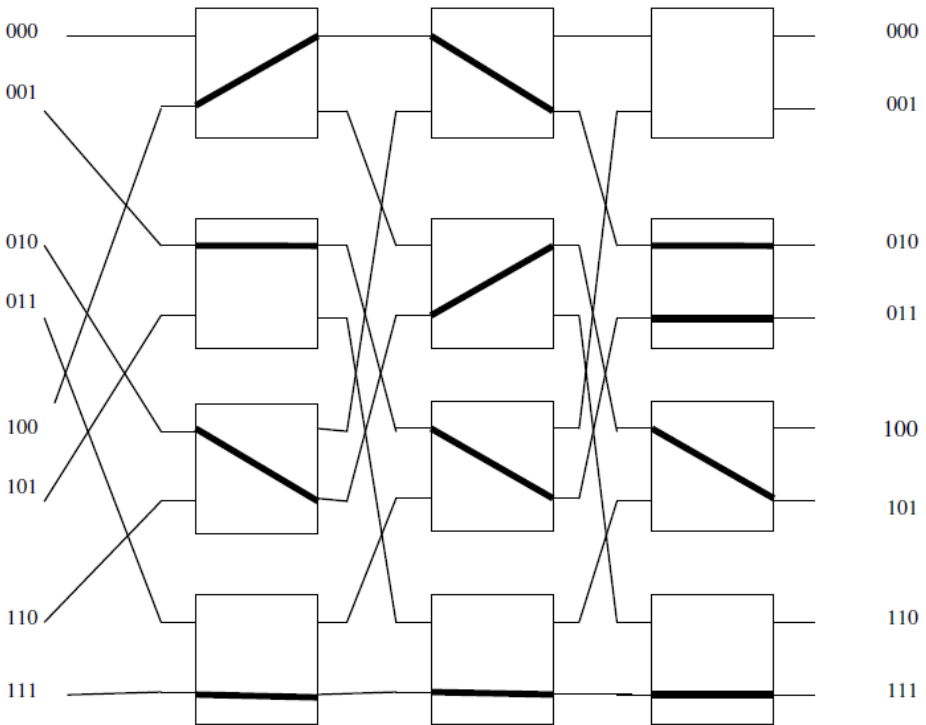


Fig. 5. Conflict free Optical Omega Network

In this way, the addresses 001, 010, 100 and 111 will be transmitted as shown in figure (5). It is clear from the figure that there is no conflict in the network.

Algorithmic Step10: remaining_address variable will store four source addresses for next pass i.e. 101, 110, 000 and 011. Similarly, the omega network for addresses 101, 110, 000 and 111 can be obtained in second pass. These addresses will get their destination without any conflict. So that the network remains conflict free.

Algorithmic Step11: Finally, goal of the algorithm is achieved.

6 Conclusion

The Crosstalk is the biggest problem in OMIN and because of this problem; the communication signals do not reach to their proper destination. In this paper, we have presented a new algorithm called Address Selection Algorithm. It selects some specific source addresses for first pass and these addresses do not create conflict in the network. In the second pass, it transmits the remaining source address to their destination. Therefore, this proposed algorithm can work as a solution to avoid crosstalk in OMIN. This approach can be applied to other Time Domain Algorithms.

References

1. Shahida, T.D., Othman, M., Abdullah, M.K.: Fast Zerox Algorithm for Routing in Optical Multistage Interconnection Networks. *IIUM Engineering Journal* 11(1) (2010)
2. Al-Shabi, M.A., Othman, M.: A New Algorithm for Routing and Scheduling in Optical Omega Network. *International Journal of the Computer, the Internet and Management* 16(1), 26–31 (2008)
3. Abed, F., Othman, M.: Fast Method to Find Conflicts in Optical Multistage Interconnection Networks. *International Journal of the Computer, the Internet and Management* 16(1), 18–25 (2008)
4. Shahida, D., Othman, M., Khazani, M.: Routing Algorithms in Optical Multistage Interconnection Networks: Revisited. In: *Proceedings of the 3rd World Engineering Congress*, pp. 63–70 (2007)
5. Abdullah, M., Othman, M., Johari, R.: An Efficient Approach for Message Routing in Optical Omega Network. *International Journal of the Computer, the Internet and Management* 14(1), 50–60 (2006)
6. Chau, S.-C., Xiao, T., Fu, A.W.-c.: Routing and scheduling for a novel optical multistage interconnection network. In: Cunha, J.C., Medeiros, P.D. (eds.) *Euro-Par 2005*. LNCS, vol. 3648, pp. 984–993. Springer, Heidelberg (2005)
7. Al-Shabi, M.A., Othman, M., Johari, R., Subramaniam, S.: New Algorithm to avoid Crosstalk in Optical Omega Network. In: *Proceedings of the IEEE International Conference on Network (MICC-ICON 2005)*, pp. 501–504 (2005)
8. Katangur, A.K., Akkaladevi, S., Pan, Y., Fraser, M.D.: Applying Ant Colony Optimization to Routing Optical Multistage Interconnection Networks with Limited Crosstalk. In: *Proceedings of the 18th International Parallel and Distributed Processing Symposium*, pp. 163–170 (2004)
9. Chau, S.C., Xiao, T.: A New Algorithm for Routing and Scheduling in Optical Multistage Interconnection Networks. In: *Proceedings of the 4th IASTED International Multi-Conference on Wireless and Optical Communications, Banff, Canada*, pp. 749–755 (2004)
10. Yang, Y., Wang, J., Pan, Y.: Routing Permutations with Link-Disjoint and Node Disjoint Paths in a Class of Self-Routable Interconnects. *IEEE Transactions on Parallel and Distributed Systems* 14(4), 383–393 (2003)
11. Lu, E., Zheng, S.Q.: High-Speed Crosstalk-Free Routing for Optical Multistage Interconnection Networks. In: *Proceedings of the 12th International Conference on Computer Communications and Networks*, pp. 249–254 (2003)

An Efficient Methodology for Realization of Parallel FFT for Large Data Set

Peter Joseph Basil Morris, Saikat Roy Chowdhury, and Debasish Deb

Electronics and Radar Development Establishment,
Defence Research and Development Organization
Bangalore, India

Abstract. The paper presents an efficient methodology for the parallel realization of FFT on multiprocessor architecture for larger data sets. FFT algorithm demands batch processing on data; large data sizes thus gives rise to performance related issues namely latency, execution time degradation triggered by cache misses, buffer management etc. The realization of FFT on multiprocessor platform presented here is based on two dimensional formulation, wherein an N point radix 2 FFT is re-ordered as a two dimensional matrix organization. The implementation employs data partitioning and parallelism so as to minimize the number of cache misses thereby ensuring cache friendliness. The implementation also employs parallel pipelined architecture. Parallel pipelined architecture and two dimensional formulation of FFT coupled with vectorization improves the performance figure of FFT realization of large data sets. A detailed analysis is presented in this paper.

1 Introduction

The impact of Fast Fourier Transforms in the field of digital signal processing has even been increasing since its recorded inception by J.W Cooley and Tuckey. The Fast Fourier Transform reduces the computational complexity of DFT from $O(N^2)$ to $O(N \log_2 N)$ which still remains to be high for larger values of N . Thus for real-time execution of FFT, large data sizes pose considerable challenge demanding the need for multiprocessor based systems.

Batch processing of large data sets for computation of FFT increases latency, degrades cache utilisation and makes buffer management difficult to handle. The presence of larger data sets poses a considerable challenge in the effective cache utilization thereby degrading the performance of traditional FFT implementation on even multiprocessor platforms operating with pipelined architectures. Hence the basic underlying strategy for realization of FFT for large data size on multiprocessor platform lies in effective data partitioning, optimal organization of processor architecture and selection of an effective and efficient parallel FFT algorithm as discussed in [1]. The present work employs the classic four step FFT algorithm chosen specifically to enhance the operational performance of the hardware.

The paper is organized as follows:- Section 2 describes the classic four step algorithm. Multiprocessor architecture and analysis of implementation is brought out in Section 3. Section 4 brings out the specification of multiprocessor platform used for the implementation and the performance achieved for different data sizes. Section 5 discusses the results of implementation while conclusion is drawn in Section 6.

2 Algorithm Description

The algorithm incorporates the divide and conquer approach. The four step radix two FFT algorithm is based on the decomposition of the N point FFT into two factors say L and M. The two dimensional formulation is derived from the multidimensional formulation [2] by setting the number of dimensions to two. The following summarises the derivation of the algorithm.

$$X_k = \sum_{n=0}^{N-1} (x_n W_n^{nk}) \tag{1}$$

where x_n is the input signal of length N, X_k is the complex FFT of x_n and $W_n = \exp(-j2\pi/N)$ is the twiddle factor.

If N has factors say L and M ($N = LM$) then the indices could be represented as:

$$n = n_1 + n_2L; \text{ where } 0 \leq n_1 \leq L - 1, 0 \leq n_2 \leq M - 1 \tag{2}$$

$$k = Mk_1 + k_2; \text{ where } 0 \leq k_1 \leq L - 1, 0 \leq k_2 \leq M - 1 \tag{3}$$

where X_k and x_n in (1) can be expressed as a two dimensional array as follows:

$$x_n = x(n_1, n_2); \tag{4}$$

$$X_k = X_k(k_2, k_1); \tag{5}$$

Substituting (2) (3) (4) (5) into (1) would yield

$$X_k(k_2, k_1) = \sum_{n_1=0}^{L-1} \sum_{n_2=0}^{M-1} x_n(n_1, n_2) W_M^{n_2k_2} W_n^{n_1k_2} W_L^{n_1k_1} \tag{6}$$

The above steps could be summarised as follows:

Step 1. Express N as a factor of two integers say L and M such that $N = LM$ where both L and M are radix two. Let $N = 2^r, L = 2^p, M = 2^q$ where r, p, q are integers. Let $p = q = r/2$ if r is even and $p = (r + 1)/2$ and $q = (r - 1)/2$; if r is odd. The N Point input vector x_n is reordered as a matrix $x(n_1, n_2)$ of dimension L x M. The mapping of the indices is governed by (2) and (3). The reordered matrix is represented as : $x(n_1, n_2) = x_n(n_1 + n_2L)$;

Step 2. Perform 'M' Point multi row FFT's for each of the L rows of 'x' and store the results in say Z.

Step 3. Multiply each entry of 'Z' with the pre-computed twiddle factors stored in the twiddle factor array $W_n(n_1, n_2)$ as follows : $A(n_1, n_2) = Z(n_1, n_2) W_n(n_1, n_2)$ where W_n is the pre-computed twiddle factor array

Step 4. Transpose the matrix A

Step 5. Perform 'L' Point multi row FFT's for each of the 'M' rows of A

Step 6. Reorder the data back into vector form from the rows of A

3 Algorithm Implementation

A hybrid of parallel and pipelined architecture is used for realization on multi-processor based platform.

3.1 Motivation for Hybrid Architecture

This paragraph dwells on the pure parallel and pure pipelined architecture and brings out the motivation for a hybrid architecture.

In a fully parallel architecture the data batch is partitioned among the processors functioning in parallel; e.g. if there are 'P' processors functioning in parallel each will get one $1/P^{th}$ of data. Though this arrangement reduces latency, it demands the algorithm to be directly parallelizable; which may not be the case always. The algorithm design should ensure the fair operation of the same on the partial data independently and in parallel to produce the intended result when the output is collected. Mostly the algorithms may not support fully parallel task with the present problem at hand.

The next candidate architecture is fully pipelined architecture which finds its application when a larger task is broken up into a number of sub tasks with each processor performing an atomic subtask. A major bottle-neck encountered in such an architecture is the pipeline stall (bubble) [3.3] which occurs due to the subtasks being of unequal size. A pure pipelined architecture may also perform the operation on the full data set leaving no scope for data partitioning. The pipelined architecture produces high latency and for larger data sizes, input buffering also presents a great challenge. Hence with a view to reduce the latency, ease the problem of data buffering and utilizing the partial parallelizability brought in by the four step FFT algorithm, we propose to use a hybrid architecture.

3.2 The Hybrid Architecture

The hybrid architecture shown in Figure 1 employs 'A' pipeline stages with 'B' processing elements(PE) per stage. Specific to this implementation, we have initially used two pipeline stages and two PE's per stage (i.e. a total of four processors). Data for a batch, as it arrives, gets divided in 'B' parts (such that each processor receives only L/B rows, each row containing M points). Each of the 'B'

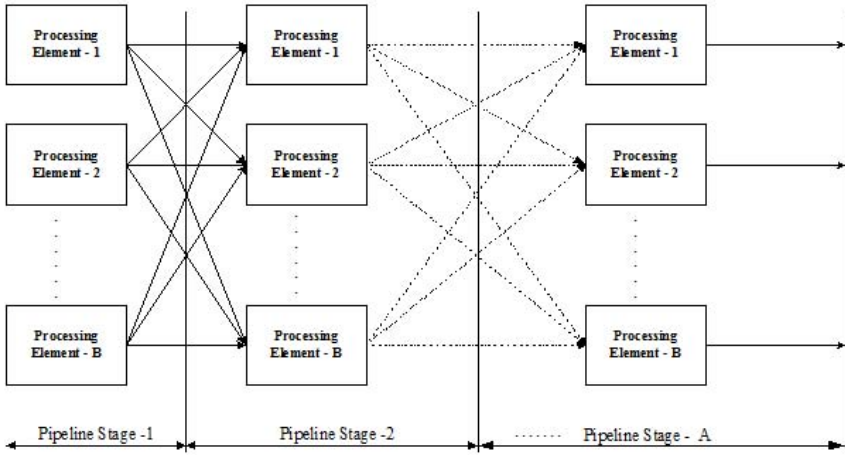


Fig. 1. Hybrid Architecture

processors perform 'L/B', 'M' point FFT's. The next step performs a distributed twiddle factor multiplication. This is followed by the transpose operation performed over the DMA channel. The final pipeline stage would then perform the FFT across the unoperated dimension.

3.3 Performance Issues

Stalling Due to Twiddle Factor Multiplication: Unequal distribution of twiddle factor multiplication among the processing elements presents a bottle neck in performance. With the first stage of processors performing both FFT as well as twiddle factor multiplication, while the second stage performs only FFT across other dimension, work distribution among the individual stages of the pipeline becomes unequal:- hence resulting in stalling. The current implementation solves this problem by distributing the number of points of twiddle factor multiplication equally among all the processing elements.

Load Balancing between the Processors: The implementation distributes multiplication operation so as to balance the load among the pool of processors. Thus, in the first stage of the pipeline, the twiddle factor multiplication for the first 'M/2' points is performed by the lower half processor set (indexing the N processors from 1 to N; the lower half set constitutes processors 1 to N/2 and upper half constitutes N/2+1 to N) in conjunction with the upper half processors performing the multiplication for the last 'M/2' points, as in Figure 2. The process is aped for the second stage of the pipeline where, the lower half processors perform the last 'L/2' points multiplication and the upper half processors perform the first 'L/2' points multiplication. Hence for a complete stage of processing, every processor performs FFT operation and multiplication

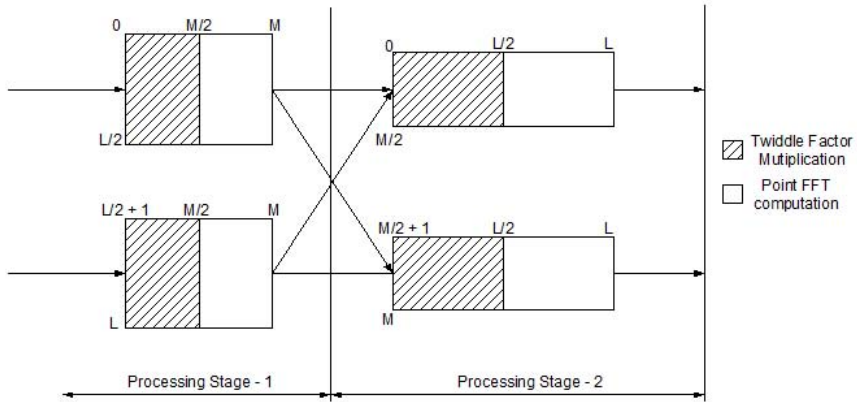


Fig. 2. Equal distribution of computation among PE’s eliminates stalling (bubble) in pipeline and ensures load balancing

operation leading to an improved load balancing capability. The algorithm is designed to ensure that L, M are always divisible by 2. Hence the processors can always be divided into two equal halves ensuring scalability for any L, M ($L \bmod 2 = 0, M \bmod 2 = 0$).

Data Transfer between Stages: Data transfer is accomplished through DMA, which happens in the background, while the processor continues execution. This provides two advantages:

As the first pipeline stage of processing completes, the DMA operation is initiated, which enables the set of processors to process the next batch of data, while the DMA transfer happens in the background. Thus when the pipe is full, two batches of data gets executed simultaneously at two stages, which leads to an increase in the throughput.

Secondly, a distributed transpose is performed over the DMA channel. Transposed data transfer takes additional time as compared to conventional data transfer, but this step is necessary so that data for the second stage processors is available in a contiguous fashion. Without the transpose operation, the second stage processors will have to operate across columns (requiring a stride greater than one between data points). This can lead to huge number of cache line misses and degrade the performance significantly. With the transpose done, processors operate across rows i.e., with unit strides. Moreover, this additional transpose time does not affect the efficiency of the algorithm, as it happens in the background, when the processors are performing operation on subsequent batch of data.

Ping- Pong Execution: The techniques employs two sets of processors, which perform their operations on alternate batches of data. As a result, the time available for execution of a current batch, effectively doubles.

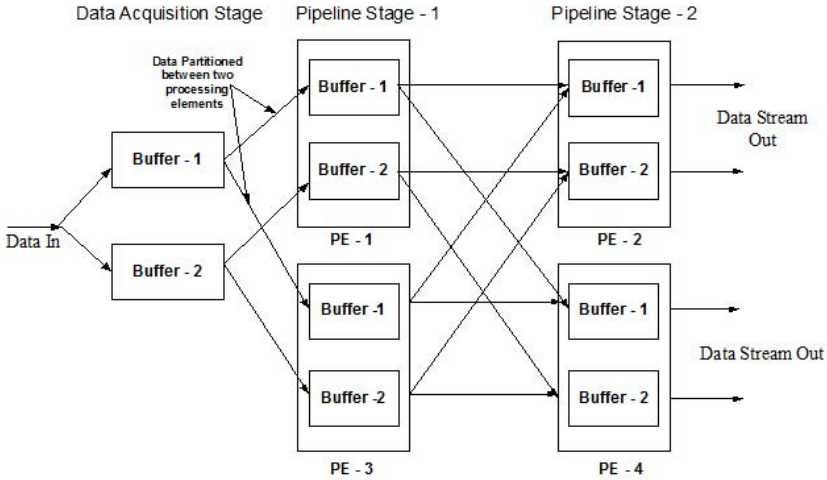


Fig. 3. Ping-Pong Arrangement of Processors

Table 1. Specification of Machine

Platform	8640D Based Multiprocessor Platform
Number of CPU's	2
Number of Cores	4
CPU Type	Freescale MPC8641D dual-core processor
L1 Cache	32kB
L2 Cache	1MB per core
Main memory	1GB DDR2 DRAM

4 Performance Analysis

4.1 Specification of the Machine

The specifications of the machine used for the implementation is enumerated in Table 1.

4.2 Performance Results

The performance results for a FreeScale MPC8640 multiprocessor is tabulated in Table 2. Performance results comparison is carried out between the execution times taken by the current implementation T_1 versus the execution time taken by a single PE for an off the shelf vector library FFT function T_2 . It is clear from the numbers that the current methodology outperforms the chosen vector library. The execution time taken by a single core for normal vector library function is found to be much higher than the execution time taken by all the

Table 2. Performance Results with four cores and two pipeline stages

FFT Size	T1(ms)	T2(ms)	Improvement(Percent)
4k	0.016	0.0195	17.9
8k	0.031	0.05	38
16k	0.050	0.11	56.8
32k	0.102	0.25	59.2
64k	0.197	0.75	73.7
128k	0.395	2.25	82.4
256k	0.796	5	84
512k	1.65	13.5	87

cores put together in the current implementation. The percentage is calculated using the following :

$Improvement = (T_1 - T_2) / T_2$ where T_1 is the execution time taken by the current architecture and T_2 is the time of execution taken by an off the shelf vector library FFT function for the same architecture.

4.3 Throughput and Latency

For a pure pipelined architecture with 'N' number of processing elements and 'N' pipeline stages, latency is given by the maximum execution time t_1 taken by a processing element. The latency L_1 for the first output could be expressed as $N * t_1$ and throughput T_1 is inversely proportional to t_1 (i.e one output every t_1 units of time).

The Hybrid Architecture used in the current implementation using the same set of 'N' processing elements and two pipeline stages with each stage having 'P' processing elements would yield a latency $L_2 = 2 * t_2$ and the throughput T_2 is inversely proportional to t_2 . where t_1 is the time to perform operation on full data size, where t_2 is the time to perform same operation on $1/P$ th of data size. Since $t_1 > t_2$, $T_1 < T_2$. Hence an increase in both latency and throughput is observed for the hybrid architecture for the same data size and number of processors.

4.4 Scalability

All the operations of the algorithm is performed on a basic matrix of size LM (where $L \geq M$). So the maximum number of processors per stage can be 'M'. However as the number of processing elements increases the cost and complexity increases rapidly. So we need to draw a line at some point where the desired peak performance is fairly achieved.

The number of pipeline stages is actually dependent on the number of atomic subtasks. In our implementation we have distributed the twiddle factor multiplication task among other stages; and the transpose operation happens in the background over the DMA channel. Hence a two stage pipeline would suffice for the algorithm, thereby, reducing the complexity of implementation.

5 Results

The algorithm was implemented on a multiprocessor platform with four nodes. The size of the input vector was varied from 4k to 512k data points with each entry being a complex number. The data was partitioned and communicated to each of the compute elements through the ping pong arrangement of processors. The ping pong arrangement coupled with the prudent usage of twiddle factor multiplication has resulted in considerable performance improvement. The results obtained are tabulated in Table 2. As expected the performance was found to increase for larger data sizes.

6 Conclusion

It is shown in this paper that by the use of classic four step algorithm coupled with judiciously chosen multiprocessor system architecture, a significant improvement in performance figure of FFT is observed. The paper focusses mainly on the issues of multiprocessor architecture and algorithm parallelization discussed in [3], but in addition to this efficient multiprocessor architecture, awareness of platform specific features are also explored coupled with effective programming strategies e.g cache utilization through stripmining, aligned memory access that are discussed in standard literature [4]. The scalability of the methodology guarantees the applicability for any arbitrarily large data sets.

References

- [1] Deb, D., Radhakrishna, P.: An Analysis Of Design Methodology Of Signal Processing Application On Multiprocessor Platform With Focus On Deterministic Execution Time And Algorithm Parallelizing. In: International Radar Symposium India (IRSI) Proceedings (December 2009)
- [2] Agarwal, R.C., Gustavson, F.G., Zubair, M.: A High Performance Parallel Algorithm for 1-D FFT. In: IEEE Proceedings. on SuperComputing, November 14-18, pp. 34-40 (1994)
- [3] Bhuyan, L.N., Agarwal, D.P.: Performance Analysis of FFT Algorithms on Multiprocessor Systems. IEEE Transactions on Software Engineering SE-9(4) (July 1983)
- [4] Abd-El-Barr, M., El-Rewini, H.: Fundamental of Computer Organization and Architecture. John Wiley & Sons Inc., Chichester (2005)

A Novel Approach for Adaptive Data Gathering in Sensor Networks by Dynamic Spanning Tree Switching

Suchetana Chakraborty and Sushanta Karmakar

Department of Computer Science and Engineering
Indian Institute of Technology, Guwahati
Assam, India

{suchetana, sushantak}@iitg.ernet.in

Abstract. A convergecast is a process in which all the sensor nodes sense data and fuse them to forward to a base station. A correct data gathering requires that there is no data loss or delivery of redundant data. Nodes can form a spanning tree rooted at the sink to perform the convergecast. Leaves of the tree can sense and forward data independently. But for any internal node data can be forwarded to its parent only after receiving data from all its children. It has been observed that a Breadth-First-Search (BFS) tree is a better choice for convergecast under low system load because the depth of any node from the root is always minimum; whereas under higher system load condition a Depth-First-Search (DFS) tree may be a better option as the degree of any node in a DFS tree is lower than that in a BFS tree. Hence per node load is lower in case of a DFS tree than that of a BFS tree. So to meet the requirement of load based adaptation, a dynamic tree switching algorithm has been proposed in this paper. The convergecast application remains transparent of the switching assuring the availability of the system at any instance of time. Also each convergecast message is assured to be delivered correctly to the base station without any loss or redundancy.

1 Introduction

A sensor is a small device which is capable of sensing, processing and transmitting data to other sensor nodes. A group of such sensors forms a network which can be used for the efficient monitoring of environment, health, military or inaccessible critical resources. Small sized sensor nodes are limited in processing capacity, memory and power consumption. It involves innovative techniques to get the best performance out of these sensors. Many wireless sensor network applications are based on broadcast and convergecast. In broadcast, data is delivered to all the sensor nodes from a particular node, generally the base-station. However in convergecast all the nodes sense data and fuse them to forward to the base-station. In recent years convergecast has received increasing attention because of many practical applications. One way to achieve an efficient convergecast is to model a sensor network as a tree rooted at the sink. All the leaf nodes of the tree collect and forward data independently, but an intermediate node can forward data to its parent only after receiving the data from all its children. Thus data from leaves are forwarded towards the sink node following various paths of the tree. A correct convergecast process guarantees that all the data from different nodes must reach the sink without any data loss or data redundancy.

1.1 Motivation

The performance of any application is environment dependent. Load of a network may vary depending on the changes in the environment. A fixed data-gathering tree may not be suitable for convergecast under changing environments. An efficient convergecast should adapt to the changes in the environment and thus continue with giving the best performance. The idea is to have two trees, a BFS tree and a DFS tree, both rooted at the sink for data gathering. The convergecast application uses the desired tree based on the variation of load. If load is low, it uses a BFS tree as the distance from any node to the root is minimum in BFS. However for higher system load the application uses a DFS tree because the degree of any node is lower in DFS than that in BFS. If load changes from low to high, the system switches from BFS to DFS, and similarly vice-versa. However a random switch between these two data-gathering trees would result in an incorrect convergecast due to data redundancy or loss of data. Also if the system switches from one tree to another using two-phase commit protocol then a large amount of delay is introduced where the convergecast process is stalled due to the switching and then resumed back after formation of the new tree is completed. Particular situations for these problems are illustrated below in Figure 1 and Figure 2 where a dashed link indicates a communication link between two nodes and an arrow indicates an edge of the tree.

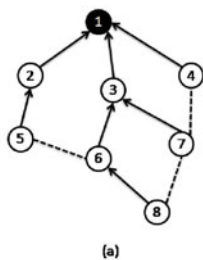
Let at time t convergecast was following the BFS tree at Figure 1a and at that moment leaf node 7 and 4 have sent data to their parents, node 3 and node 1 respectively. Now due to random switch data start flowing through the paths of a DFS tree in 1b. So at time $t + 1$, node 3 will forward the data to its current parent, node 7. Hence the same data will come back to node 7 creating data redundancy. To generalize the situation whenever the parent-child relation for any two nodes becomes reversed because of a random switch, the properties of correct convergecast will be compromised because of redundant data transmission in the network.

Again in Figure 2b, let at some time instant t node 3 has sent data to its parent. However, the data has not yet reached its destination because of the asynchronous property of channel. So data is in the channel between node 3 and node 5. Now a random switch occurs from DFS to BFS. At this moment node 5 has got two children i.e. node 6 and node 7. As soon as it receives data from both of its children it will fuse and forward to its parent which is node 2. So even if data from node 3 reaches to node 5 eventually, it will not consider it valid as at that moment node 3 is not its child. So the data that node 3 sent to node 5 will be dropped in this process and will never reach to the sink.

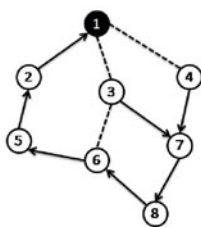
Thus we need to design a proper switching method that ensures correct convergecast in spite of switching from one tree to another.

1.2 The Problem Definition

A correct convergecast guarantees that all the data from different nodes must reach the sink without any data loss or data redundancy. Also each node can send data only once. Different topologies of the underlying data-gathering network will be useful in different scenarios. The main objective is to design a distributed system that adapts to the load of the network and still assures a correct application layer convergecast. Depending on the

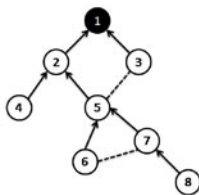


(a)

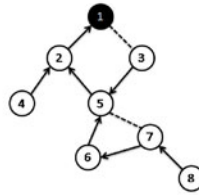


(b)

Fig. 1. Redundant Data Delivery



(a)



(b)

Fig. 2. Data Loss

system load the adaptive convergecast protocol uses either a BFS tree (at lower load) or a DFS tree (at higher load). But a random switching from a BFS tree to a DFS tree or vice-versa may cause the problem of *redundant data delivery*, *data loss* or an *indefinite stall of data flow*. Thus the problem is to design a distributed message-passing algorithm that performs this switching in such a way that the application layer convergecast remains unaffected. The proposed algorithm performs the switching between a BFS tree and a DFS tree while assuring that all the properties of a correct convergecast are maintained.

2 Related Works

There are many research works on convergecast in wireless sensor networks until recently. Annamalai et al. [1] proposed a heuristic algorithm for tree construction and TDMA based channel allocation for collision free convergecast to achieve greater efficiency in terms of latency and power consumption. The algorithm constructs a tree level by level and allocates a schedule for each node that specifies the time-slot(s) in which it can transmit data. Krishnamachari et al. [5] modeled a data-centric routing scheme and studied the energy-latency trade-off involved in data-aggregation that includes the effect of source-sink placements, communication network topology, and density of the network on convergecast. They have shown that the latency of convergecast is proportional to the number of hops between the sink and the furthest source and the formation of an optimal data-gathering tree is a NP-hard problem. So they proposed some data-aggregation tree generation heuristics which runs on polynomial time. Heinzelman et al. [3] proposed a family of adaptive protocols, called SPIN for data dissemination in wireless sensor network. To eliminate the transmission of redundant data flow in the network, the protocol uses high level data descriptors called meta-data. The performance of SPIN outperforms the traditional approaches like flooding or gossiping using its key features meta-data negotiation and resource adaptation. The proposed algorithm is energy-aware but useful only for highly available bandwidth. Also the adaptation incurs significant communication cost. Chen et al. [2] proposed an adaptive data-gathering scheme for clustered WSN. The objective here is to shift the burden of computation from ordinary sensor nodes to the resource-rich sink node through proper adjustment of

aggregation ratio and reporting frequency. The spatial and temporal aggregation degree is adaptive to the dynamic state of WSN via the interaction between the sink node and the cluster heads.

Liu et al. [6] worked on the adaptivity and the co-ordination among running protocols. They have built a hybrid protocol that can make smooth adaptation at run-time. The algorithm works in three steps and has little switching overhead. It is also scalable and delay-efficient compared to the traditional two-phase commit method of protocol switching. However the protocols used by the algorithm are abstract in nature, strictly homogeneous, and application specific. In another work, Liu et al. [7] designed a generic switching protocol which assures to preserve some of the communication properties like reliability, total order, integrity, no replay, confidentiality etc. The switching protocol runs below the application layer and is transparent. However the class of properties considered is specific and aggressive switching leads to the oscillating nature of the resulting protocol.

Karmakar et al. [4] designed a distributed protocol switching algorithm for broadcast applications that switches between a BFS tree and a DFS tree adapting to the system load. At low load a BFS tree is used and at higher load a DFS tree is used. In the proposed algorithm, the switching is done adapting to the network load as well as without affecting the application layer broadcast. The algorithm also ensures the correct delivery of each broadcast packet and guarantees that some spanning tree of the graph is always maintained even at the time of switching. In this paper we have proposed a distributed algorithm for load-adaptive convergecast using tree switching.

3 System Model

The random distribution of the sensor nodes are represented by a communication graph $G(V, E)$ where V is the set of sensor nodes and E is the set of communication links between the nodes. $N(v)$ denotes the neighbors of any node v . The switching algorithm is used to switch between a BFS tree and a DFS tree (or vice-versa) depending on the load of the system. The BFS tree and the DFS tree are assumed to be pre-computed for a given network topology using standard algorithm and are rooted at the sink of the convergecast application, which is essentially the root of the tree. The computation model is assumed to be asynchronous and the network is static. Also it is assumed that the channels are reliable and FIFO and there is no node or link failure in the network.

4 Algorithm Design

4.1 Tree Switching Algorithm

As root of the tree (sink) is the ultimate receiver of all the data, it can detect the change of environment and initiate tree switching to adapt to the change. In this work we skip the details of how to detect the change in environment and concentrate on how the switching works. The proposed algorithm is a distributed message-passing algorithm which works in two phases. Let T and T' denote the two trees.

<pre> On receiving <i>TOKEN</i> from <i>u</i> if $u \in C_{new}$ then $C_{curr} \leftarrow u$ $Counter1 \leftarrow Counter1 - 1$ if $Counter1 = 0$ then if $p_{curr} \neq p_{new}$ then Send(<i>CANCEL</i>, p_{curr}) $p_{curr} \leftarrow p_{new}$ endif endif Send(<i>TOKEN</i>, p_{new}) if $BLOCK = TRUE$ then $BLOCK \leftarrow FALSE$ endif else $p_{new} \leftarrow u$ Send(<i>TOKEN</i>, v) where $v \in C_{new}$ $Counter1 \leftarrow C_{new}$ if $C_{new} = \phi$ then if $C_{curr} = \phi$ then if $p_{curr} \neq p_{new}$ then Send(<i>CANCEL</i>, p_{curr}) endif else Send(<i>TOKEN</i>, p_{new}) endif else Send(<i>CANCEL</i>, v) where $v \in C_{curr}$ $Counter2 \leftarrow C_{curr}$ endif endif </pre>	<pre> On receiving <i>CANCEL</i> from <i>u</i> if $u = p_{curr}$ then $BLOCK \leftarrow TRUE$ Send(<i>ACKC</i>, u) endif if $u = C_{curr}$ then $C_{curr} \leftarrow \{C_{curr} - u\}$ endif On receiving <i>ACKC</i> $Counter2 \leftarrow Counter2 - 1$ if $Counter2 = 0$ then Send(<i>CANCEL</i>, p_{curr}) Send(<i>TOKEN</i>, p_{new}) endif On receiving <i>M</i> Buffer[<i>REAR</i>] $\leftarrow M$ $REAR \leftarrow REAR + 1$ if $BLOCK = FALSE$ then Send(<i>M</i>, p_{curr}) $FRONT \leftarrow FRONT + 1$ endif </pre>
--	--

Fig. 3. Actions on receiving *TOKEN*, *CANCEL*, *ACKC*, or application message *M*

- **First Phase:** For a switch from tree T to T' , root passes a *TOKEN* to its children of new tree T' . As the *TOKEN* moves downwards a virtual new tree T' is constructed from the root to the leaves. However the convergecast still follows the old tree T .
- **Second Phase:** Leaf nodes start passing back the *TOKEN* to their respective parents for the switched tree T' through the return paths. Each node, on receiving *TOKEN* back from all its children, makes the links from its children permanent, which is an edge of the new tree T' . Now the data packets start flowing through this newly built paths. In this way the switching from T to T' occurs gradually from the bottom of the tree towards the root.

Thus it is evident that at some point of time data gathering is following T' at the bottom levels of the tree and T at the the upper ones. However the algorithm assures that convergecast process is not affected in this process. Three types of messages are exchanged among the nodes to complete the switching process.

- Each node sends *TOKEN* message to all its children for the switched tree and thus making a virtual construct of the switched tree. Also each node receives back the *TOKEN* from all its children in second phase to make sure that the paths from the leaves to that particular node have already been permanent. The *TOKEN* is generated by the root to initiate the switching and finally consumed back by the root itself after switching completes.
- When a node wants to change its parent pointer due to switching, it first sends a *CANCEL* message to its old parent. Also when an intermediate node becomes a leaf node for the switched tree, it sends *CANCEL* to all its old children from whom it used to receive data packets. Upon receiving *CANCEL* from a child, a node removes that particular node from its child set. Also on receiving *CANCEL* from parent, a node starts buffering incoming application data.
- A node sends *ACKC* to its parent as an acknowledgment to the *CANCEL* message. Upon receiving *ACKC* from all the expected neighbors, a node can guarantee that it is not going to receive any more data packets from any of those neighbors and hence changes its parent pointer for switching the tree.

Each node has a set a of variables whose descriptions are as follows:

- p_{curr} : Parent variable for the current tree T .
- p_{new} : Parent variable for the switched tree T' .
- C_{curr} : Set of children for the current tree T .
- C_{new} : Set of children for the switched tree T' .
- *Counter1* : Integer variable, keeps the track of the cardinality of the set C_{new} .
- *Counter2* : Integer variable, keeps the track of the cardinality of the set C_{curr} .
- *BLOCK* : Boolean variable to control enqueue or dequeue the buffer.

Counter1 and *Counter2* are initialized to 0 and p_{new} is NULL. The variables p_{curr} and p_{new} are NULL for the root node. Root node initiates the algorithm by sending *TOKEN* to all v where $v \in C_{new}$ and when it receives back the *TOKEN* from all v where $v \in C_{new}$ then the algorithm terminates. The formal description of the algorithm for BFS to DFS switching or vice-versa is given in figure 3.

Let each convergecast message coming from the application layer is denoted by M . The tree switching algorithm runs as a middle-ware layer below the application layer. So upon receiving M from the application layer the switching algorithm either forwards it or starts buffering. When the *BLOCK* value is *FALSE*, it simply forwards to its current parent, otherwise it starts buffering. The buffer that every node keeps to store incoming data from application layer is implemented through queue. *FRONT* and *REAR* denote the first and last pointers of the queue respectively and both are initialized to 0. The formal algorithm for handling application messages is given in figure 3.

4.2 Correct Delivery of Convergecast Messages

The problem of application data loss due to switching can be easily avoided through duplicate transmission. But no reconvergecast is allowed as it increases both the latency and traffic overhead of the system. Only proper design of the switching algorithm can

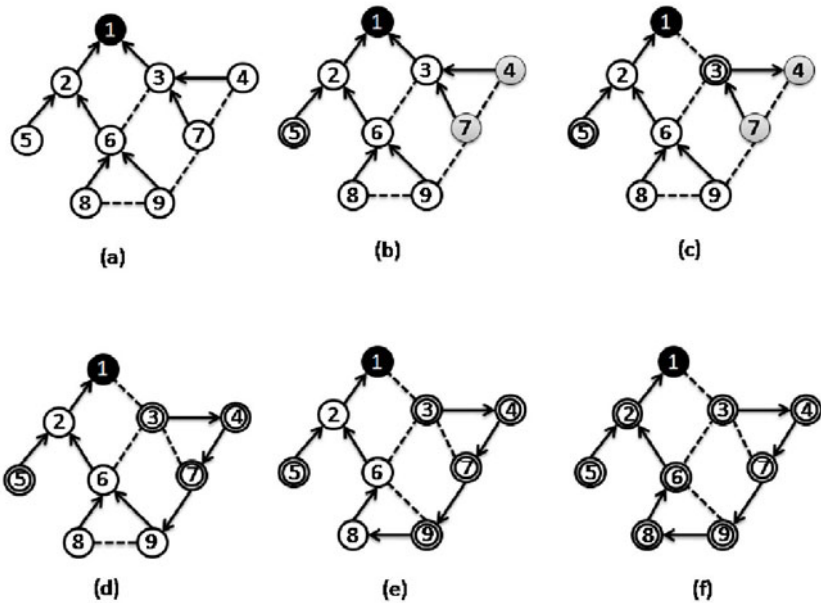


Fig. 4. Sample Run of the Switching Algorithm

handle this situation efficiently. The illustration in figure 4 is for the switching from a BFS tree to a DFS tree to show that the switching gets completed successfully as well as the properties of the convergecast are also satisfied.

- **No Redundant Data:** For T node 3 is the parent of node 4, but in T' this parent-child relationship gets reversed. So duplicate data may generate as a consequence. But on receiving *CANCEL* from node 3, node 4 starts buffering data. Receiving *ACKC* confirms node 3 that no more data is going to come from node 4. Before receiving *ACKC*, node 3 forwards all data through its old parent node 1, and after receiving *ACKC*, to its new parent node 4. So no data sent from node 4 to node 3 can come back to node 4 as duplicate.
- **No Data Loss:** Node 3 had two children node 4 and node 7 for T , whereas it is a leaf node for T' . Data sent from a child could get lost if its parent does not exist any more due to a sudden switch. But node 3 first makes sure through *ACKC* that no more data is going to come from its current child and then only can remove the link to its current parent. All the data received before receiving *ACKC* thus forwarded through the old path.
- **No Indefinite Stall:** A node starts buffering data after receiving either a *TOKEN* from its C_{new} or a *CANCEL* from its p_{curr} . Also a node before changing parent variable informs its p_{curr} through *CANCEL* message. So a node does not wait indefinite time expecting data from its child creating a stall in the network.

4.3 Proof of Correctness

The following Lemmas can be easily proved.

Lemma 1. *Each node receives $TOKEN$ exactly $|C_{new}| + 1$ times, one from its p_{new} and each one from u where $u \in C_{new}$.*

Lemma 2. *The $COUNTER1$ value at each node eventually becomes 0 and the node sends back $TOKEN$ to its p_{new} .*

Lemma 3. *If an intermediate node becomes a leaf node for the switched tree, then eventually its C_{curr} becomes ϕ .*

Theorem 1. *The switching algorithm will eventually terminate.*

Proof. Depending on the network load, the root node takes the decision of switching and starts the switching procedure by sending $TOKEN$ to each node $v \in C_{new}$. Each node u on receiving $TOKEN$ from its p_{new} , forwards it to all $v \in C_{new}$. The variable $Counter1$ at u tracks the cardinality of the set C_{new} . This phase of the algorithm terminates if $C_{new} = \phi$ for the node u ; or in other words u is a leaf node. At the second phase each leaf node u sends back the $TOKEN$ message to a node v where $v = p_{new}$. Thus the link between u and v becomes permanent according to Lemma 2. Let $Ncount(l)$ be the nodes at level l of the tree. Further, let all the links between $Ncount(l+1)$ and $Ncount(l)$ be already permanent. Then according to Lemma 2, the $COUNTER1$ value at each u where $u \in Ncount(l)$ becomes 0 eventually. They sent back the $TOKEN$ to v where $v \in Ncount(l-1)$ and $v = p_{new}(u)$. Then all the links between $Ncount(l)$ and $Ncount(l-1)$ become permanent. Thus a node u at level l does not transmit any more messages to a node at level $(l+1)$, or at level $(l-1)$ once all the edges incident on u have become the part of the new tree. This phase of the algorithm terminates producing a switched tree correctly when the root at level 0 receives back $TOKEN$ from all its children. The variable $Counter1$ at the root becomes 0 as a consequence. Thus there exists no more control messages in the system and all the nodes and edges become stable at this point. Hence the algorithm terminates successfully. \square

Lemma 4. *There will always be a path from any node to the root node at the time of switching.*

Proof. When a node u needs to change its parent due to switching, it has to be assured that there exists a path from $p_{new}(u)$ to the root. Now a node u sends $TOKEN$ to a node v where $v = p_{new}(u)$ to make a permanent link only when it has already received $TOKEN$ from v through forward $TOKEN$ traversal. That means node v had a path to the root. v could change its path to the root only after receiving $TOKEN$ from all $w \in C_{new}(v)$. Node u sending $TOKEN$ back to its p_{new}, v implies that the path from v to the root has not been changed yet. This implies there exists a path from any node to the root always, irrespective of the switching. \square

Lemma 5. *No cycle will be created in the system due to switching.*

Proof. All the leaf nodes u , whether it is newly created or already existing one, independently send back *TOKEN* to its p_{new} to build up a permanent link. According to Theorem 4 that node v , where $v = p_{new}$ is a part of the new tree and hence has a path to the root. A cycle can only form if $v \in C_{curr}(u)$ or there was a path $P = \{v, v + 1, \dots, u - 1, u\}$ from v to node u according to old tree construct. From Lemma 3 $C_{curr}(u)$ will become empty eventually breaking the cycle. Now suppose node v made a link to a node x , where $x \in P$ and also $x = p_{new}(v)$ making a cycle $C = \{v, v + 1, \dots, x, \dots, u - 1, u, v\}$. But there must be at least one node y where $y = x$ or x is a descendant of y such that there exists a path from node y to the root. This is because the *TOKEN* has already traversed from the root to the node v through node y along the forward path. So node y will eventually make a permanent link to its parent along the path to the root and thus breaking the cycle C . Thus no cycle will be formed due to the switching. \square

Theorem 2. *Each convergecast message will be eventually delivered to the root node.*

Proof. When an intermediate node u becomes a leaf of the switched tree, first it assures that all v where $v \in C_{curr}$ is blocked by sending *CANCEL* to them and changes its parent after receiving *ACKC* as response. So all the data received from $v \in C_{curr}$ are forwarded to its p_{curr} using the old link and no more data is received after receiving *ACKC* from all $v \in C_{curr}$. Thus no data is lost in this case.

When a node u needs to change its parent due to switching, first it informs the node v where $v = p_{curr}(u)$ by sending *CANCEL* message. No more data is sent to v after that. All the future data are forwarded using the link to the node w where $w = p_{new}(u)$. By Lemma 4 there exists a path from w to the root. So no data will be lost during the switching process. From the above discussion and Lemma 4 and Lemma 5 it can be proved that each convergecast message will be delivered to the root eventually. \square

The following Lemmas can be easily proved.

Lemma 6. *Each convergecast message will be delivered exactly once.*

Lemma 7. *There will be no indefinite stall in data gathering process.*

Theorem 3. *The message complexity of the switching algorithm is $O(|E|)$.*

Proof. Let n and E be the total number of nodes and edges in the network and d_v be the degree of node v . So v sends d_v number of *TOKEN* messages and receives the same in reverse direction. Hence the total number of *TOKEN* messages is, $M_T = 2 \times \sum_{v \in V} d_v = 4|E|$. In the worst case scenario there will be $(n - 2)$ number of nodes whose parents can be changed and there will be always less than n number of nodes which have been converted to leaf nodes from intermediate nodes due to the switching process. So there will be $O(n)$ numbers of *CANCEL* messages in worst case. Similarly there will be $O(n)$ number of *ACKC* messages. Hence the worst case message complexity for the switching between a DFS and a BFS tree = $4|E| + O(n) + O(n) = O(|E|)$. \square

The following theorems can be proved.

Theorem 4. *The algorithm runs in $O(D)$ steps, where D denotes the diameter of the network.*

Theorem 5. *The per node space complexity of the algorithm is $O(\delta)$ where δ is the maximum degree among all the nodes in the network.*

5 Conclusion

In this paper, we have proposed a distributed tree switching algorithm for load adaptive convergecast. The algorithm guarantees that even if a switching occurs between the underlying topologies, the application layer convergecast remains unaffected. We have shown that each convergecast packet is eventually delivered to the sink without any data loss or redundancy even during the switching. The switching algorithm eventually terminates. The algorithm runs within $O(D)$ steps, where D is the diameter of the network and the message complexity is $O(|E|)$. It will be interesting to investigate the problem when the trees used for the switching are not pre-computed.

References

1. Annamalai, V., Gupta, S.K.S., Schwiebert, L.: On tree-based convergecasting in wireless sensor networks. In: Proceedings of IEEE Wireless Communication and Networking Conference, pp. 1942–1947 (2003)
2. Chen, H., Mineno, H., Mizuno, T.: Adaptive data aggregation scheme in clustered wireless sensor networks. *Comput. Commun.* 31, 3579–3585 (2008)
3. Heinzelman, W.R., Kulik, J., Balakrishnan, H.: Adaptive protocols for information dissemination in wireless sensor networks. In: Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking, MobiCom 1999, pp. 174–185. ACM, New York (1999)
4. Karmakar, S., Gupta, A.: Adaptive broadcast by distributed protocol switching. In: Proceedings of the 2007 ACM Symposium on Applied Computing, pp. 588–589 (2007)
5. Krishnamachari, B., Estrin, D., Wicker, S.B.: The impact of data aggregation in wireless sensor networks. In: Proceedings of the 22nd International Conference on Distributed Computing Systems, ICDCSW 2002, pp. 575–578. IEEE Computer Society Press, Washington, DC, USA (2002)
6. Liu, X., van Renesse, R.: Fast protocol transition in a distributed environment (brief announcement). In: PODC 2000: Proceedings of the Nineteenth Annual ACM Symposium on Principles of Distributed Computing, p. 341. ACM, New York (2000)
7. Liu, X., van Renesse, R., Bickford, M., Kreitz, C., Constable, R.: Protocol switching: Exploiting meta-properties. In: Proceedings 21st International Conference on Distributed Computing Systems Workshops, Mesa, AZ, USA, pp. 37–42 (2001)

Hardware Efficient Root-Raised-Cosine Pulse Shaping Filter for DVB-S2 Receivers

Vikas Agarwal, Pansoo Kim, Deock-Gil Oh, and Do-Seob Ahn

Satellite & Wireless Convergence Research Department
Electronics and Telecommunications Research Institute (ETRI) Daejeon, S. Korea
vikasvikas85@gmail.com

Abstract. Hardware complexity of a root-raised-cosine (RRC) pulse shaping filter is dominated by the number of binary multiplications. In this paper an efficient multiplication scheme has been used to implement an RRC matched filter for Digital Video Broadcasting-Second Generation (DVB-S2) specifications. Nibble processing of the filter coefficients is performed to achieve computational efficiency. Quantization of the filter coefficient is used as a trade-off for the filter design. Twelve-bit quantized coefficients are used to obtain low hardware cost with satisfactory filter performance. Matlab simulation is used to compare and contrast the performance of an RRC filter with that of a standard floating point design. Hardware of the filter is simulated and synthesized using Xilinx ISE10.1. Implementation results show a 25% savings in LUT count. The use of nibble processing has the potential to reduce the hardware complexity of an RRC filter and is also applicable to other DVB-S2 receiver schemes.

Keywords: Digital Video Broadcasting-Second Generation; Root Raised Cosine filter; Quantization Noise.

1 Introduction

Digital Video Broadcasting - Second Generation (DVB-S2) via satellite [1] has proven to be a milestone in satellite communications. It has the ability to effectively utilize the channel bandwidth and provide a high data rate. However, due to an increase in demand for digital signal processing, high-speed and higher-order filters are frequently used to perform pulse-shaping. Digital filters are responsible for extracting a useful signal and suppressing unwanted interference. Basically, digital filters are used to modify the characteristics of a signal in the time and frequency domain. To meet the demand of high-speed, low- cost hardware, an efficient filter processing solution is needed.

Root-raised-cosine (RRC) filters are implemented differently depending on the applications in which they are applied. Different architectures have been studied for RRC filter implementation. An algorithm has been proposed to represent the filter coefficient [2] in such a way that the filter response can be represented with a minimum number of non-zero coefficients, therefore reducing the arithmetic

complexity needed to obtain the filter output; however, this results in a compromise in filter performance. A logarithmic approach is also [3] used for multiplierless filtering, but it has a drawback in terms of a high on-chip memory requirement. The higher the order of the filter is, the higher the memory usage. Another work on FIR filters basically constitutes a static filter implementation. Pre-processing of the filter coefficients has demonstrated that the multiplication block [4] can be realized using primitive operator graph synthesis techniques. As the word length increases, such techniques require high preprocessing time, and hence the complexity of finding the optimal graph increases. The sum-of-powers-of-two (SOPOT) [5] is another static technique for matched filter implementation. Here, the coefficients are designed as a tradeoff between system performance and hardware complexity.

The main task is to minimize the system cost by minimizing the number of multiplication operations involved in filter design and achieve a reliable performance. In this paper, a hardware-efficient multiplier scheme has been utilized to implement the RRC filter for DVB-S2 specifications. The approach relies on redundancy created by processing filter coefficients in nibbles. The output of a nibble multiplication with a random input signal can be used by other filter coefficients in a transposed form RRC filter. Hence, it reduces the number of additions involved in the multiplication operation of the RRC filter. This is referred to as a nibble processing scheme (NPS) in the rest of this paper. The performance of NPS is similar to that found in a floating point design. Hardware implementation of the filter is done on a Xilinx Virtex4-LX200 board. The use of NPS is shown to produce computationally-efficient, low-cost hardware. NPS can be used for other DVB-S2 receiver algorithms such as synchronization.

This paper is divided into seven sections. Section 2 explains the basic RRC filter. Section 3 focuses on the nibble processing scheme. In section 4, hardware implementation of the RRC filter using NPS is explained. Section 5 provides a performance comparison of the design with a floating point design. Implementation results of the RRC filter design are given in section 6, and finally, in section 7, some conclusions are drawn on the basis of the obtained results.

2 Root Raised Cosine Matched Filtering

RRCs filters are an area of great interest for communication systems. Unlike a rectangular pulse, a raised-cosine pulse takes on the shape of a sinc pulse, which can be implemented as a digital filter. It is used to pulse shape the signal such that the inter-symbol interference (ISI) is reduced. It also reduces the bandwidth needed for transmission. An RRC filter can be defined by its frequency response, $H(f)$, where T is the symbol duration and $\alpha = 0.2$ to 0.35 is excess bandwidth for DVB-S2. Eq. (1) is the frequency response of an RRC filter.

$$H(f) = \begin{cases} 1 & 0 \leq f \leq \frac{1-\alpha}{2T} \\ \cos \left[\frac{\pi T}{2\alpha} \left(1 - \left| f - \frac{1-\alpha}{2T} \right| \right) \right] & \frac{1-\alpha}{2T} \leq f \leq \frac{1+\alpha}{2T} \\ 0 & |f| \geq \frac{1+\alpha}{2T} \end{cases} \quad (1)$$

The impulse response of the RRC filter can be found by applying an inverse Fourier transform to $H(f)$, as Eq.(2)

$$h(t) = \frac{\sin\left(\pi \cdot \frac{t}{T}\right) \cdot \cos\left(\pi \cdot \alpha \cdot \frac{t}{T}\right)}{\alpha \cdot \left(\frac{t}{T}\right) \left[1 - 4 \cdot \alpha^2 \left(\frac{t}{T}\right)^2\right]} \tag{2}$$

The matched filter has a symmetric filter response and can be effectively implemented using a transposed direct form structure as shown in Fig. 1, where $x(n)$ is a complex random input and $y(n)$ is the filter output with filter order $L=25$. This type of implementation requires $(L+1)$ multipliers and $(L-1)$ adders, where L is the filter order.

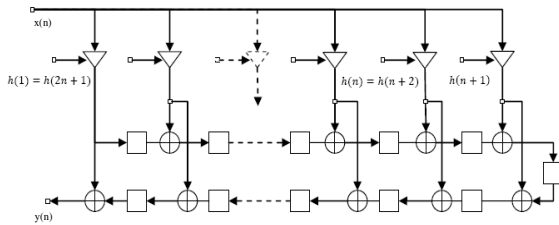


Fig. 1. Standard transposed form FIR filter with symmetric response

3 Nibble Processing Scheme

The dominance of binary multipliers over FIR filtering is a major hurdle in achieving low complexity architectures. NPS [6] is used for the implementation of an RRC filter for DVB-S2 receiver specifications. The scheme works on quantized coefficients and utilizes the redundancy involved in processing coefficients as nibbles. Each coefficient is paired in a set of nibbles. The

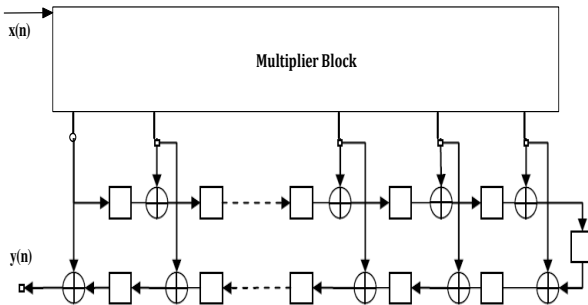


Fig. 2. RRC Matched Filter Block Diagram

multiplication output of $x(n)$ with a nibble (integer range 0 to 15) can be utilized for other filter coefficients. This approach reduces the computational complexity of the multiplication operation. NPS depends on the number of bits used for quantization of the filter coefficients. Fig. 2 shows a block diagram of the RRC matched filter. The multiplier block represents $(n+1)$ multipliers, and the output of the multiplier block is fed to the adder chain.

3.1 Mathematical Formulation

In a transposed form of FIR filtering, each input sample is multiplied with all filter coefficients. Eq. (3) represents the k^{th} coefficient multiplication at the n^{th} instant, where $k \in \{0 \dots L-1\}$.

$$z_k(n) = x(n) * h_k(n) \tag{3}$$

The k^{th} filter coefficient can also be written as in (4), where N is the coefficient bit length and $a_{i,k} \in \{0,1\}$.

$$h(k) = \sum_{i=0}^{N-1} a_{i,k} * 2^i \tag{4}$$

Since all coefficients are multiplied by the same random input data stream, if each coefficient can be grouped into m -nibbles, where $m \in \text{ceil}(N/4)$, the k^{th} coefficient can be expressed as (5).

$$h(k) = 2^{N-4(1+m)} \sum_{i=0}^3 a_{i,k} * 2^i \tag{5}$$

$$z_k(n) = \sum_{m=0}^{\text{ceil}(N/4)} 2^{N-4(1+m)} * x(n) * \sum_{i=0}^3 a_{i,k} * 2^i \tag{6}$$

(---Part II---) (----- Part I-----)

The output of the multiplier block given by Eq. (6) can be grouped in two parts: part I reflects the preprocessing of a nibble with a random input, and part II represents the desired binary shift. Finally, m -nibble outputs of the k^{th} coefficients are added. Thus, the multiplication is achieved by the nibble pre-processing, a few hardwired binary shifts, and $\text{ceil}(N/4)$ additions. Thus, for an L -Tap RRC filter with complex input, $(L+1) * \text{ceil}(N/4)$ adder units are required instead of $(L+1)$ binary multiplications. Table 1 shows the hardware cost in terms of adders required for an RRC filter design using NPS. The higher the bit length is, the higher the amount of hardware required. With an increase in filter coefficient quantization by nibble, 33% extra adders are required.

Table 1. Hardware cost for different quantization bit lengths

Quantization bit length of Coefficients	Adder Cost for NPS
8	$(L+1)$
12	$(L+1)*2$
16	$(L+1)*3$

3.2 Effect of Quantization on RRC Filter Coefficients

The response of the filter is improved with an increase in quantization bit length. Fig. 3 shows the frequency response of a filter for different quantization bit lengths.

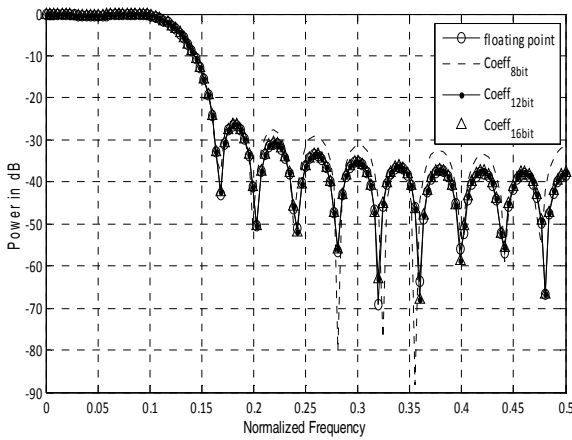


Fig. 3. RRC filter response

For 16- and 12-bit quantization, the filter response is much closer to the floating point. Table 2 shows that 12-bit quantization requires 33% less hardware compared to 16-bit quantization. Hence, a tradeoff is made between the RRC filter performance and hardware cost. In this paper, 12-bit quantization of the filter coefficient is used for hardware implementation.

4 Hardware Implementation

4.1 RRC Filter Architecture for DVB-S2

A block diagram of the RRC filtering scheme is shown in Fig. 4. It comprises a common processing unit and multiplexer units. The hardware requires adders and hardwired shift operations to do the filter processing.

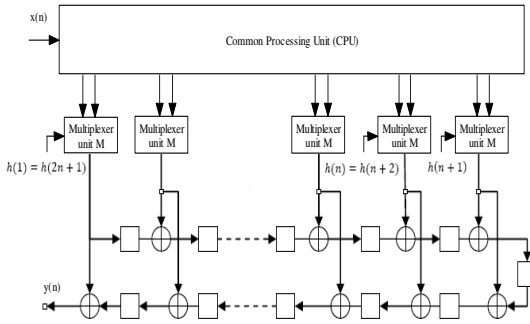


Fig. 4. Modified block diagram of the RRC filter

4.2 Common Processing Unit (CPU)

A graphical representation of the common processing unit is shown in Fig. 5, where each node represents an adder, and the base represents a shift in power of 2. A CPU is used to provide the fixed directed graph structure in order to compute small integer multiplication, $x(n) * [0 \ 1 \ \dots \ 15]$. For example, suppose $x(n) = X$ and 21 is a filter coefficient. Thus, instead of direct multiplication, we can compute $21X$ as $21 * X = (10101)_2 * X = X + X \ll 2 + X \ll 4$. Hence, multiplication can be implemented by a sequence of additions and shifts. Therefore, each input is pre-computed, and the outputs are fed to the multiplexer units to compute the final multiplication output.

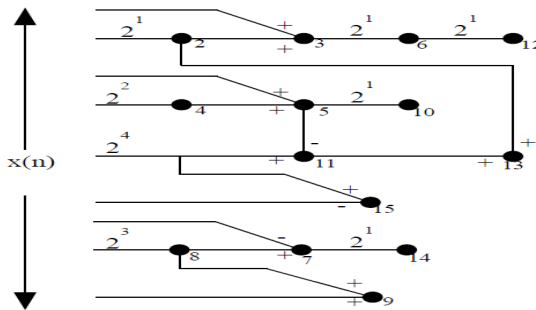


Fig. 5. Common Processing Unit

4.3 Architecture of Multiplexer Unit

The outputs obtained from the CPU are fed to all the coefficients of the RRC filter. For an L-tap RRC filter with complex input data, $2L$ multiplexer units are required. Each multiplexer unit consists of three 16:1 multiplexers. Each multiplexer unit uses a coefficient nibble as a control line as shown in Fig. 6. The outputs of the multiplexer are shifted by $\{2^0, 2^4, 2^8\}$, respectively, and after proper shifting, two binary adders are used to obtain the multiplication result. The above mechanism works fine for

unsigned numbers, but for signed representation, a small two’s-complement circuit has to be incorporated. The output of the multiplexer unit (M) is a two’s complement if the coefficient is negative in order to obtain the proper signed multiplication output.

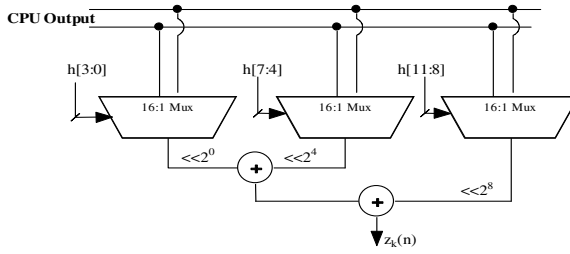


Fig. 6. Multiplexer unit (M)

5 Performance of RRC Filter

An RRC filter has been designed and implemented for DVB-S2 specifications. A performance evaluation of the RRC filter is conducted using Matlab simulation based on the DVB-S2 specifications as given in Table 2.

Table 2. RRC Filter Design Parameters

Filter Order	25
Oversampling	4
Symbol Rate	27.5Mbaud
Excess Bandwidth	0.25
Coefficient Quantization	8, 12 bits
Input Bit Length	16

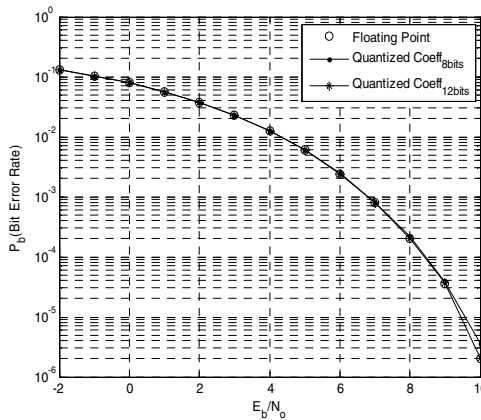


Fig. 7. BER plot for different SNQR

The performance of the RRC filter is verified in an AWGN channel for different quantization coefficient bit lengths. Figure 7 shows a BER plot for fixed- and floating-point implementations. The filter performance is acceptable with the use of the aforementioned parameters, and there is a savings in terms of the number of adders required for implementation.

6 Implementation Results

Implementation of an RRC filter was done using a Xilinx Vitex-4 LX200 kit. Digital simulation and synthesis of the designed RRC filter was conducted and the architecture verified on board. Hardware results are given in terms of LUT count. NPS is implemented using two adders per coefficient. The LUT count for NPS is 25% less than those of standard binary multiplier schemes. The number of slices required for NPS implementation is 8390. Hence the nibble processing architecture for pulse shaping is suitable for a DVB-S2 receiver.

Table 3. Comparison in terms of adder cost

Algorithms	Standard RRC Filter Architecture	Nibble Processing Architecture
Hardware Cost	26 Binary Multipliers	52 adders
LUTs count	13379(100%)	9950(74.4%)
Slices occupied	8390	6407

7 Conclusions

Digital implementation of a low-cost and computationally-efficient root-raised-cosine filter was performed using a nibble processing scheme. The performance of the RRC filter has been tested and verified on a Xilinx Virtex-4 LX200 board. Coefficient quantization is a critical parameter, and can be used to obtain a high filter performance or low hardware complexity. The proposed scheme reduces hardware complexity of the matched filter, and utilizes 75% of LUT compared to a conventional floating point design, validating the use of NPS in the RRC filter design for DVB-S2 receiver specifications.

References

1. Digital video broadcasting (DVB); Framing structure, channel coding and modulation for 11/12 GHz satellite services, EN300 421 (V1.1.2), European Telecommunications Standards Institute, ETSI (1997)
2. Eshtawie, M.A.M., Othman, M.B.: An Algorithm Proposed for FIR Filter Coefficients Representation. *International Journal of Applied Mathematics and Computer Sciences* 4(1), 24–30 (2008)
3. Lee, P.: An FPGA Prototype for a Multiplierless FIR Filter Built using the Logarithmic Number System. In: *Proc. 5th International Workshop on Field-Programmable Logic and Applications FPL*, pp. 303–310 (1995)

4. Dempster, G., Macleod, M.D.: Use of Minimum-Adder Multiplier Blocks in FIR Digital Filters. *IEEE Trans. Circuits Syst.* 42(9), 569–577 (1995)
5. Lim, Y.C., Parker, S.R.: FIR Filter Design Over Discrete Powers-of-Two. *IEEE Transactions on Acoustics, Speech, and Signal Proc.*, assp-31(3) (June 1983)
6. Park, J., Jeong, W., Mahmoodi-Meimand, H., et al.: Computation Sharing Programmable FIR Filter for Low-Power & high-Performance Applications. *IEEE Journal of solid-state Circuits* 39(2), 348–357 (2004)

Security Analysis of Multimodal Biometric Systems against Spoof Attacks

Zahid Akhtar¹ and Sandeep Kale²

¹ Dept. of Electrical and Electronic Engineering, University of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy

`z.momin@diee.unica.it`

² Department of Electronic Science, University of Pune
Ganeshkhind, 411007 Pune, India

Abstract. Biometrics, referred as the science of recognizing a person based on his or her physical or behavioral traits, has been widely accepted and deployed in various applications. However, recent researches show that many biometric traits are susceptible to spoof attacks. Moreover, a recent work showed that, contrary to a common claim, multimodal systems can be broken even if *only one* trait is spoofed. The result was obtained, using simulated spoofed samples, under the assumption that the spoofed and genuine samples are indistinguishable, which is not true for all biometric traits. We further investigate this security issue, focusing on behavior of fixed and trained score fusion rules, using real spoof attack samples. Preliminary empirical results on real biometric systems made up of face, fingerprint and iris confirm that multimodal biometric systems are not intrinsically robust against spoof attacks as believed so far. In particular, most used fixed rules can be even less robust than trained one. We found that trained rules are not only more flexible and accurate but more robust, also, against spoof attacks as compare to fixed rules. We also empirically observed that multimodal systems are more robust, under spoof attacks, than unimodal biometric systems, robustness increases as the number of matchers being fused increases.

Keywords: Biometrics, Multimodal biometric system, Score fusion rules, Spoof attacks.

1 Introduction

Biometrics, an automated security system that uses physiological or behavioral characteristics of an individual to verify the identity, has been greatly employed in several practical applications. The expected ideal characteristics of biometric traits are, it should be unique and hard to be duplicated. Unfortunately, recent researches have shown that biometric traits can be stolen, replicated, and reproduced fake trait can be used to attack biometric systems [1] [2]. Multimodal biometric systems have been proposed to increase the recognition rates as compared to the unimodal biometric systems that make them up. Extensive empirical evidences have shown that they are effective to accuracy improvement

but there is no empirical evidence to support the security improvement assumption. It has also been claimed that multimodal systems are more robust against spoof attacks, since evading several systems is more difficult than evading just one [3]. This claim implies that to evade a multimodal system it is necessary to evade *all*, or at least more than *one* of the individual systems. However, no one has investigated so far whether this is true or not with the exception of [4], where some evidence was provided by simulated spoof attacks that a multimodal biometric system made up of a face and a fingerprint matcher can be fooled by spoofing *only one* of the biometrics. In [4] the analysis was carried out under the assumption that spoofed and genuine samples are identical while this may be true for some biometric trait such as face but not for other traits like fingerprint, iris and so on. Hence, it is of great interest to investigate the robustness of multimodal biometric systems against real spoof attacks under more realistic scenarios where an attacker is not able to fabricate a perfect replica of biometric trait. In this work we further contribute to this goal using real spoof attack samples unlike [4]. We analyze the security of multimodal biometric systems by focusing on spoof attacks at the sensor level, which are so far the one raising the most of interest in the biometric community [5]. They consist in faking one or more biometrics by submitting to the system a biometric replica, like a fake fingerprint [5].

In this work we further empirically investigate the possibility of evading a multimodal system by real spoofing of any of the individual biometrics, by considering four multimodal biometric systems made up of face, fingerprint and iris matchers. We also analyze the behaviour of different fixed and three trained score fusion rules under spoof attacks. Our results show that multimodal biometric systems may be more vulnerable to spoof attacks when the most accurate biometric trait is spoofed. It is known that trained rules are more flexible and in principle more accurate than fixed one [3]. We provide empirical evidence that trained rules are more robust, as well, against spoof attacks as compare to fixed rules. We focus lastly on the robustness improvement, and investigate whether it is possible to improve robustness of multimodal biometric systems against spoof attacks through the selection of appropriate number of matchers to be fused.

2 State-of-the-Art and Goals

With the great acceptance and deployment of biometric systems, their issues of security and robustness against attacks are also increasing. Several researchers are studying the vulnerabilities of biometric systems, the possible attacks with their countermeasures. Eight possible different points where security of biometric systems can be compromised have been identified in [6]. Among the others, faking biometric input to the system is a growing concern. This kind of spoof attack is related to the sensor, and is also called direct attack. An attacker can stealthy procure the biometric trait of a legitimate user. The stolen biometric trait can be used to produce synthetic biometric samples. For instance, in [1] artificial fingers were created and 60% impostor acceptance rate, when submitting fake samples

to a fingerprint sensor, was reported. Several counteraction such as "liveness" detection at sensor level is a possible solution suggested by most researchers [2], but no effective one exist yet. The most notable point about biometric spoof attack is that it does not require any developed technical skill and information about the system operation mechanism.

The common claim about multimodal biometric systems is that they are intrinsically more robust against spoof attacks, because their evasion would require to spoof *all* biometric traits simultaneously [3]. However such claim is not based on theoretical or empirical evidences. In fact, in [4] multimodal systems made up by a face and a fingerprint matcher were considered, and it was shown using simulated spoof attack that they can be cracked by faking just *one* of the biometric traits. Such results are very interesting and motivate the need to further study the problem of spoof attacks, by assessing robustness of multimodal fusion rules under different aspects than [4], which is the goal of this paper.

1. Can we crack multimodal biometric system by spoofing *only one* biometric trait? Results presented in [4] on two data sets with simulated attacks and using two fusion rules showed that this is possible. We further investigate this issue by considering four multimodal systems with real spoof attack samples and six score fusion rules (three fixed and three trained rules).
2. How different fixed and trained score fusion rules behave when a multimodal system is subject to a spoof attack against one of the individual matchers?. We focus on the fact that trained rules are considered to be more flexible and yielded higher accuracy than fixed ones [3] and found that trained rules are more robust, also, against spoof attacks as compare to fixed rules. We also argue that spoofing the most accurate biometric matcher can create serious security breaches.
3. Is it possible to improve the robustness of a multimodal biometric systems against spoof attacks, by fusing appropriate number of matchers? The point here is that information fusion improves accuracy as the number of sources increases. Likewise, whether does it also improve robustness against spoof attacks?

3 Experimental Setup

We used Essex face [7], PolyU HRF DBII fingerprint [8] and IIT Delhi iris [9] databases. Since no biometric data sets including spoof attack samples are available publicly. Therefore, we used the above data sets and created spoof attacks. The data sets used in the experiments contain face, fingerprint and iris images of 100 individuals, with 10 genuine samples and 10 fake samples (spoof attacks) per individual.

Spoofed face and spoofed iris images were created with a "photo attack" method. We put in front of the camera the photo of each individual, displayed on a laptop screen. Fake fingerprint samples were created by the same procedure carried out in [10] to create fake palmprint images. For each individual, we created 10 spoofed face, iris and fingerprint images.

The face, iris and fingerprint recognition systems used for the experiments were implemented using the *Principle Component Analysis* (PCA) [11], the *Iris Code* [12] and the *minutiae*-based [13] methods, respectively. All the scores were normalized using the hyperbolic tangent method [3].

Using the face, fingerprint and iris individual systems, we built four different multimodal systems by pairing in all possible ways the face matcher with the fingerprint and the iris matchers. The resulting systems are therefore three multimodal systems using any of two individual systems and one multimodal system using all three individual systems, named as Face-Fingerprint, Iris-Face, Iris-Fingerprint and Iris-Fingerprint-Face system, respectively.

In our experiments, in order to focus on the effect of spoof attacks under the optimal configuration of the fusion rules, the decision thresholds on the fused scores for the considered operational points and the parameters of trained fusion rules were evaluated on the whole data sets. In particular, the decision thresholds and the parameters of trained fusion rules were evaluated on the original data sets (without spoofed samples), while the performance of the systems under attack was evaluated by replacing impostor score with respective spoofed score.

3.1 Fusion Rules

Three fixed and three trained score fusion rules were studied. Let s_1, s_2, s_3 and S be the scores of face, fingerprint, iris individual system and the fused score, respectively.

Fixed Rules

1. **Sum:** The sum rule is a direct summation of the matching scores produced by the set of N matchers, and the fused score is computed as $S = \sum_{i=1}^N s_i$
2. **Product:** The product rule is also applied directly to the matching scores produced by the set of N matchers as: $S = \prod_{i=1}^N s_i$
3. **Bayesian:** The fused score by bayesian rule [14] is computed as follows:

$$S = \frac{\prod_{i=1}^N s_i}{\prod_{i=1}^N s_i + \prod_{i=1}^N (1 - s_i)} \tag{1}$$

Trained Rules

1. **Weighted Sum:** The weights for the weighted sum rule can be computed using linear discriminant analysis (LDA) [15]. The aim of using LDA based fusion rule is to obtain a fused score which provides minimum within-class variations and maximum between-class variations. The fused score are computed as $S = \sum_{i=1}^N w_i s_i$, where w_i is the weight of biometric system i .
2. **Perceptron-based rule:** The perceptron-based fusion rule for N individual matchers can be implemented as follows [16]:

$$S = \frac{1}{1 + \exp[-(w_0 + \sum_{i=1}^N w_i s_i)]} \tag{2}$$

where $w_i (i = 0, \dots, N)$ are the weights of perceptron. Such weights are usually computed by a gradient descent algorithm with a least-squares loss function.

3. **Likelihood Ratio Rule (LLR):** This rule computes the matching scores as follows:

$$S = \frac{\prod_{i=1}^N p(s_i|G_i)}{\prod_{i=1}^N p(s_i|I_i)} \quad (3)$$

where $p(\cdot|G)$ and $p(\cdot|I)$ are the matching scores probability density function (PDF) of genuine and impostor users, respectively. LLR is referred as the optimal fusion rule, when all the PDFs are estimated accurately [17]. We used Gaussian to model the genuine and impostor score distributions.

The false acceptance rate (FAR) is the fraction of impostors being accepted as genuine users, to provide high security biometric systems operate at a low FAR operational point. To investigate issues 1, 2 and 3 of Section 2, we evaluated the increase of the FAR due to spoof attacks at 0%, 0.01% and 0.1% FAR operational points, known as the lowest threshold values that result into FAR on training data equal, respectively, to the operational points. We remind the reader that the corresponding decision thresholds and the parameters of the trained rules were estimated on the whole original data sets (without spoof attacks).

4 Experimental Results

We report results obtained, on unimodal (individual) systems in table 1, and on Iris-Fingerprint, Face-Fingerprint and Iris-Fingerprint-Face multimodal systems in table 2, 3 and 4, respectively, using sum, weighted sum (with LDA) and LLR score fusion rules. The results obtained on Iris-Face multimodal system and with other fixed and trained rules were qualitatively very similar, and hence are not reported for the lack of space.

Table 1. FAR(%) of the Iris, Fingerprint and Face individual recognition systems when the respective trait is spoofed, at three operational points

Operational Point	Iris System	Fingerprint System	Face System
0% FAR	39.60	44.60	7.40
0.01% FAR	43.70	59.70	16.30
0.1% FAR	51.00	67.50	26.30

Consider first issue 1 of Sect. 2, our results on all four multimodal systems with real spoof attacks clearly show that the answer is "yes", which is further support to the results obtained with simulation in 4. For instance, from the table 3, it can be seen that even at 0.1% FAR operational point the FAR under attack attained values up to 63.91%.

Table 2. FAR (%) of the Iris-Fingerprint System with the Sum, W. Sum and LLR rules, when either the Iris or the fingerprint is spoofed, at three operational points

Iris-Fingerprint System						
Operational Point	Sum		Weighted Sum		LLR	
	Iris Sp.	Fing. Sp.	Iris Sp.	Fing. Sp.	Iris Sp.	Fing. Sp.
0% FAR	34.12	1.89	12.58	1.01	32.89	1.08
0.01% FAR	40.91	3.67	21.85	2.78	38.02	2.40
0.1% FAR	48.74	6.03	37.81	3.90	45.51	3.71

Consider now the issue 2 of Sect. 2. We found that the two considered fixed rules perform similar. The increase of FAR was considerably high, when the most accurate matcher is spoofed. The less accurate matcher spoofing leads rather to low increase in FAR. For instance, on Iris-Fingerprint multimodal system, table 2, when most accurate matcher (iris matcher here) is spoofed the FAR attained under attack using sum rule is 48.74% while in less accurate (fingerprint in this case) matcher spoofing the FAR under attack is 6.03%. Since the operational points were computed using whole data set, this means when thresholds are computed in real systems on training data set even a small deviation in estimated operational point, caused by practically observed problem of concept drift in biometrics, can highly subvert the robustness of the system.

Table 3. FAR (%) of the Fingerprint-Face System with the Sum, W. Sum and LLR rules, when either the fingerprint or the face is spoofed, at three operational points

Face-Fingerprint System						
Operational Point	Sum		Weighted Sum		LLR	
	Fing. Sp.	Face Sp.	Fing. Sp.	Face Sp.	Fing Sp.	Face Sp.
0% FAR	41.39	1.69	39.87	1.00	2.79	1.01
0.01% FAR	52.14	3.09	50.75	2.30	3.51	2.23
0.1% FAR	63.91	5.05	59.84	3.66	4.59	2.96

With regard to the trained rules, weighted sum, perceptron and LLR rules, we observed that spoofing accurate matcher leads to increase in FAR more than spoofing less accurate matcher, since in training accurate matcher always gets higher contribution.

In general, trained score fusion rules are considered to be more flexible and accurate than fixed rules. Our results are evident that they are more robust against spoof attacks, as well, as compare to fixed rules, because of their peer ability to be tuned to the domain and data. It is possible to further enhance the robustness of trained rules by training them on data set with real spoof attack samples which is similar to noise injection methods used in neural network classifiers to improve its generalization capability [18].

Consider issue 3 of Sect. 2. We observed that robustness of the systems against spoof attacks increases with the increase of number of matchers being fused. For

Table 4. FAR (%) of the Iris-Fingerprint-Face System with the Sum, W. Sum and LLR rules, when either the Iris, the fingerprint or the face is spoofed, at three operational points

Iris-Fingerprint-Face System									
Operational Point	Sum			Weighted Sum			LLR		
	Iris Sp.	Fing. Sp.	Face Sp.	Iris Sp.	Fing. Sp.	Face Sp.	Iris Sp.	Fing Sp.	Face Sp.
0% FAR	29.04	0.65	0.51	8.79	0.30	0.01	3.96	0.23	0.01
0.01% FAR	36.32	2.15	1.70	17.18	1.05	0.80	4.96	0.98	0.61
0.1% FAR	43.97	3.63	2.18	31.87	1.76	0.93	10.03	1.31	0.84

example, at 0% FAR operational point when iris is spoofed on individual system (table 1), the FAR is 39.60% while on Iris-Fingerprint (two matchers based system, table 2) and Iris-Fingerprint-Face (three matchers based system, table 4) using sum rule, the FAR attained values up to 34.12% and 29.04%, respectively. Similar effect can be observed in other fusion rules. This means that information fusion not only improves accuracy but improves robustness also against spoof attacks, as the number of sources being fused increases. This phenomenon is materialized due to the effective variance reduction of the final fused score with respect to the variance of the original scores [18] [19]. Our empirical finding on relationship between the increase of robustness against spoof attacks and increase of number of sources being fused is an interesting point to be investigated more thoroughly and systematically which is subject of our on going work.

5 Conclusion

Our empirical investigation on *real* spoof attack samples verify that multimodal biometric systems are not intrinsically robust against spoof attacks as believed so far, providing the further confirmation to the results obtained with simulation in [4]. They can be cracked by spoofing *only one* biometric trait. In particular, the most used fixed score fusion rules determined to be very vulnerable to spoof attacks while the trained score fusion rules are less vulnerable. Spoofing most accurate matcher creates serious security breaches. We also experimentally observed that robustness of systems under spoof attacks increases with the increase of information sources being fused.

A possible way to improve the security against spoof attacks is to devise *ad-hoc* score fusion rules, which also undertake possibility of attacks. Such two rules were presented in [4], even though the parameters, the probability of particular trait is attempted to be spoofed and that attempt succeeds, could be relatively difficult to tune in practice, as also pointed out in [4].

Acknowledgment. The authors would like to thank Ajay Kumar and Lei Zhang of Department of Computing, The Hong Kong Polytechnic University for providing IITD Delhi Iris and PolyU HRF Fingerprint Databases, respectively.

References

1. Matsumoto, T., Matsumoto, H., Yamada, K., Hoshino, S.: Impact of artificial "gummy" fingers on fingerprint systems. In: Proc. of SPIE Optical Security and Counterfeit Deterrence Techniques IV, vol. 4677, pp. 275–289 (2002)
2. He, X., Lu, Y., Shi, P.: A Fake Iris Detection Method Based on FFT and Quality Assessment. In: Proc. Chinese Conf. on Pattern Recognition, pp. 316–319 (2008)
3. Ross, A., Nandakumar, K., Jain, A.K.: Handbook of Multibiometrics. Springer, Heidelberg (2006)
4. Rodrigues, R.N., Ling, L.L., Govindaraju, V.: Robustness of multimodal biometric methods against spoof attacks. J. of Visual Languages and Computing 20, 169–179 (2009)
5. Jain, A.K., Nandakumar, K., Nagar, A.: Biometric template security. EURASIP J. on Adv. in Sig. Proc., 1–17 (2008)
6. Ratha, N.K., Connell, J.H., Bolle, R.M.: An analysis of minutiae matching strength. In: Proceedings of third AVBPA, pp. 223–228 (2001)
7. Libor Spacek: University of Essex Face Database, <http://dces.essex.ac.uk/mv/allfaces/index.html>
8. PolyU HRF Database, <http://www.comp.polyu.edu.hk/~biometrics/HRF/HRF.htm>
9. IIT Delhi Iris Database version 1.0, http://web.iitd.ac.in/~biometrics/Database_Iris.htm
10. Zhang, D., Kanhangad, V., Luo, N., Kumar, A.: Robust Palmprint Verification Using 2D and 3D Features. Pattern Recognition 43(1), 358–368 (2010)
11. Kim, K.: Face Recognition using Principal Component Analysis. Department of Computer Science. University of Maryland, College Park (2000)
12. Daugman, J.G.: How iris recognition works. IEEE Trans. Circuits Syst. Video Technol. 14(1), 21–30 (2004)
13. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer, Heidelberg (2003)
14. Suen, C.Y., Lam, L.: Multiple Classifier Combination Methodologies for Different Output Levels. In: Proc. Int. Workshop on Multiple Classifier Systems, pp. 52–66 (2000)
15. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley Inc., Chichester (2001)
16. Jain, A.K., Prabhakar, S., Chen, S.: Combining Multiple Matchers for a High Security Fingerprint Verification System. Pattern Recognition Letters 20(11–13), 1371–1379 (1999)
17. Nandakumar, K., Chen, Y., Dass, S.C., Jain, A.K.: Likelihood Ratio Based Biometric Score Fusion. IEEE Trans. on PAMI 30(2), 342–347 (2008)
18. Bishop, C.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1999)
19. Poh, N., Bengio, S.: Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks? In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 893–896 (2004)

A Novel Copyright Protection Scheme Using Visual Cryptography

Amitava Nag¹, Jyoti Prakash Singh¹, Sushanta Biswas², D. Sarkar²,
and Partha Pratim Sarkar²

¹ Dept. of Information Technology

Academy of Technology, West Bengal, India

² Department of Engineering and Technological Studies

University of Kalyani, West Bengal, India

Abstract. Visual cryptography scheme is a cryptographic technique which allows visual information to be encrypted in such a way that the decryption can be performed by the human visual system, without the aid of computers. In this article, we propose a novel scheme for copyright protection for digital images based on visual cryptography. The proposed method employs encoding of Most Significant bits of host image which is to be copyrighted, to form a master image share. The master share is encoded with a copyright image to form another share called ownership share. The master share is kept with a central authority and ownership share is kept by the copyright owner. In case of any dispute, the master shares and ownership shares can be stacked together to give the copyright image verifying the ownership about the host image. The important feature of our procedure is that we do not disturb the host image either during copyright generation nor during copyright verification. The proposed method is also independent of the size of secret image.

Keywords: watermarking, secret sharing, visual cryptography, copyright protection, image processing.

1 Introduction

The Internet has proved to an excellent distribution system for the digital media because of its inexpensiveness and efficiency. Due to widespread use of Internet, the sharing and transmission of images has images in digital form has become quite easy and popular. However, the cyberspace is overloaded with duplication methods which violates the intellectual property rights of digital data, such as document, image, audio, and video. Therefore, the protection of the rightful ownership of digital data has become an important issue recently. Many researchers have proposed several techniques [12] to protect the intellectual property rights for digital images. Digital watermarking is a one such popular technique in which one embeds additional information called digital signature or watermark into the original digital content for copyright protection and image authentication. The digital signature or hidden watermark can be extracted to verify the rightful

ownership whenever there is a need to do so. Some watermark detection process uses the original image during the watermark detection process whereas some other watermark does not use it. However, in many cases such as image monitoring, the original image is usually unavailable, thus those techniques that can reveal watermarks without the aid of the original image become better solutions. The digital image watermarking scheme can be divided into two categories based on the visibility of the watermark. They are visible digital image watermarking and invisible image watermarking techniques. In visible watermarking [1], the information is visible in the picture or video. Typically, the information is text or a logo which identifies the owner of the original document. In invisible watermarking [4,8,9], information is added as digital data to audio, picture or video, but it cannot be perceived as such. The requirements of any effective digital watermarking scheme are imperceptibility, robustness, unambiguousness, security, capacity, and low computational complexity. Some of these requirements conflicts with each other and hence they introduce many technical challenges. For example, imperceptibility and capacity may conflict with robustness. Therefore, a reasonable compromise between some requirements is required to achieve better performance for the intended applications. Based on the way watermarking is done, we can group watermarking techniques into two categories: one is the spatial-domain approach and the other is the transform-domain approach. The most popular technique in spatial domain is Least Significant Bit (LSB). LSB [3] method is the simplest technique which directly modifies the intensities of some selected pixels in the spatial domain. The transform domain technique transforms an image into a set of transform domain coefficients [11]. The transformation adopted may be discrete cosine transform (DCT), discrete Fourier transforms (DFT) and discrete wavelet transforms (DWT) etc. After applying transformation, watermark is embedded in the transformed coefficients of the image such that watermark is not visible. Finally, the watermarked image is obtained by acquiring inverse transformation of the coefficients. In 1995, Naor and Shamir [10] introduced a perfectly secure way called visual cryptography (VC) for the protection of secret images. In addition to the property of perfect secrecy, the prominent feature provided by VC is the decryption method which is done by human eyes without use of any computers. In this article, we introduce a copyright protection scheme based on Most Significant Bit (MSB) and Visual Cryptography (VC). Our scheme follows a two step process. The first phase is the ownership share construction phase and next phase is the ownership identification phases. In the ownership share construction phase, the master share M will be generated from the host image by using the Most Significant Bit (MSB) of the host images. The master share and copyright image (secret image) is then encrypted according to the rules of visual cryptography to form another share called ownership share. The master share is kept with a central authority and ownership share is kept by the copyright owner. During the identification phase, the master shares and ownership shares are stacked together which gives the copyright image verifying the ownership about the host image.

The rest of this paper is organized as follows. In section 2, we briefly discuss the related work done in the area of watermarking using visual cryptography. We present our proposed algorithm in section 3. Section 4 presents our the experimental results. We conclude the paper in Section 5.

2 Related Works

Recently, many VC based copyright protection schemes were proposed, such as those in References [5,7]. Hou and Chen [5] use a modified two-out-of-two visual cryptography scheme to split the watermark into two meaningless shares, and the first share is embedded into the host image by decreasing the gray levels of some specific pixels. When the rightful ownership must be identified, the second share and the watermarked image are superimposed to reveal the watermark. The drawbacks of their method are that the host image should be altered and that the robustness to some attacks, such as jitter, geometric distortion, cropping, and rotation attacks, is rather weak. Chang et al. [2] proposed a copyright protection scheme which uses visual cryptography and discrete cosine transform (DCT) and enables embedding multiple watermarks without destroying other earlier hidden watermark. Their method comprises the ownership share construction and the watermark revelation phases. During the ownership share construction phase, the DC coefficients of all DCT blocks are extracted from the host image to form a master share. The ownership share is generated by combining the master share and the watermark. The ownership share works as a key to reveal the watermark whenever a dispute arises without using the original image. Since their method does not actually embed the watermark into the image, the host image will not be altered. However, their method does not really provide the key advantage of visual cryptography that uses human eyes to decrypt the secret without the aid of computers. In addition, their method requires the size of the watermark to be much smaller than that of the host image. Hwang [7] used the most significant bits of the host image to generate the first share so as to satisfy the requirement of robustness. Then, the first share is used together with the watermark to construct the second share according to the two-out-of-two visual cryptography scheme. The method has the advantages that the watermark can be of any size, and that the host image is not altered. Hsu and Hou [6] proposed a copyright protection scheme based on Visual Cryptography (VC) and Sampling Distribution of Means (SDM). In first phase, They generated the master share M from the host image by SDM. Then, the master share M is used together with the secret image S to generate the ownership share O according to some predefined encryption rules of visual cryptography. A private key K is used in both phases during sampling. The private key K is kept in secret by the copyright owner, and the ownership share O is registered with a trusted third party for further authentication. The private key K and the ownership share O are used to reveal the hidden secret image for settling the dispute wherever it arises. Wang and Chen [13] proposed copyright protection scheme based on Singular Value Decomposition (SVD) and Visual Cryptography (VC). During the ownership

share construction phase, they performed SVD on a small window centered at each randomly selected pixel position. The largest Singular Value (SV) of each window is utilized to construct a master share. Finally, an ownership share is constructed by using the master share and a secret image according to the visual cryptography technique. In this article, we propose a copyright protection scheme for digital images to using visual cryptography and most significant bits of an image. Our method uses the most significant bit of a host image to generate a binary master share from a gray-level image. Then, the master share and the secret image are used to construct the ownership share according to predefined rules of visual cryptography. When the rightful ownership must be identified, the master share, generated from the image to be identified, and the ownership share are superimposed to reveal the secret image without the aid of computers. The key advantages of our method are : (i) the host image will not be altered, (ii) the rightful ownership can be identified without the aid of the original image, (iii) the secret image can be of any size, and (iv) the advantage of visual cryptography, which uses human eyes to recover secret images without the aid of computers, can be fully utilized. In addition, our method can attain the requirement of robustness because the most significant bit an image can not be easily changed by many attacks. Finally, the security of the scheme is ensured by the properties of visual cryptography.

3 Proposed Method

In this section, we introduce our technique for copyright protection of a host image using visual cryptography. Our scheme follows a two step process. The first phase is the share construction phase and next phase is the ownership identification phases. We generate two shares based on the host image (image to be copyrighted) and the secret image (used as copyright image). The master share is generated by dividing the host image into 2×2 blocks of $\frac{M}{2} \times \frac{N}{2}$ number of images. We add the MSB of each block. We create a master share by putting 1 in $R_{i,j}$ if the sum of MSB of $(i, j)^{th}$ block is odd otherwise we store 0 in $R_{i,j}$. The collection of $R_{i,j}$ is the master share R. The ownership share is generated by comparing the value of $R_{i,j}$ and $P_{i,j}$ of secret image. The encryption rules are given in algorithm 1. The resultant ownership share O is registered to a certified authority (CA) for further authentication. When a dispute over the rightful ownership of the host image arises, the ownership identification procedure is performed to protect the owner's intellectual property. The hidden secret image can be revealed by stacking the generated master share and the ownership share by human eyes only without the aid of any computers. However, the secret image can also be revealed by computers as outlined in algorithm 2.

Algorithm 1: Share_generation_Procedure

Input: A gray-level host image H of size $M \times N$, a binary secret image I_s of size $\frac{M}{2} \times \frac{N}{2}$.

Output: An ownership share O of size $M \times N$

Steps

1. Decompose H into $\frac{M}{2} \times \frac{N}{2}$ number of 2×2 blocks
2. For each sub-blocks $B_{i,j}$
 - (a) Add the MSB of each pixel and stores the summation in $S_{i,j}$
 - (b) If $S_{i,j} \bmod 2$ is 0, then set $R_{i,j} = 0$, otherwise $R_{i,j} = 1$
3. For each pixel $P_{i,j}$ of the secret image I_s , calculate $O_{i,j}$ (with 4 sub-pixels) in the ownership share 0 according the following encryption rules:
 - (a) If $P_{i,j} = 0$ and $R_{i,j} = 0$ then $O_{i,j} = \begin{matrix} \blacksquare & \blacksquare \\ \blacksquare & \blacksquare \end{matrix}$
 - (b) If $P_{i,j} = 1$ and $R_{i,j} = 1$ then $O_{i,j} = \begin{matrix} \blacksquare & \blacksquare \\ \blacksquare & \blacksquare \end{matrix}$
 - (c) If $P_{i,j} = 0$ and $R_{i,j} = 1$ then $O_{i,j} = \begin{matrix} \blacksquare & \blacksquare \\ \blacksquare & \blacksquare \end{matrix}$
 - (d) Otherwise $O_{i,j} = \begin{matrix} \blacksquare & \blacksquare \\ \blacksquare & \blacksquare \end{matrix}$
4. Repeat step 3 until all pixels of secret image I_s are processed.
5. END.

Algorithm for share generation by computers.

Algorithm 2: Secret_Revelation_procedure

Input: A gray-level host image H' of size $M \times N$, a binary ownership share O of size $M \times N$ (each of which is composed of 4 sub pixels).

Output: A recovered binary secret image I'_s of size $\frac{M}{2} \times \frac{N}{2}$

Steps

1. Decompose H' into $\frac{M}{2} \times \frac{N}{2}$ number of 2×2 blocks
2. For each sub-blocks $B'_{i,j}$
 - (a) Add the MSB of each pixel and stores the summation in $S'_{i,j}$
 - (b) If $S'_{i,j} \bmod 2$ is 0, then set $R'_{i,j} = 0$, otherwise $R'_{i,j} = 1$
3. For each pixel $P_{i,j}$ of the ownership share O , calculate $P'_{i,j}$ of the secret image I'_s according the following encryption rules:
 - (a) If $R'_{i,j} = 0$ and $O_{i,j} = \begin{matrix} \blacksquare & \blacksquare \\ \blacksquare & \blacksquare \end{matrix}$ then $P'_{i,j} = 0$
 - (b) If $R'_{i,j} = 1$ and $O_{i,j} = \begin{matrix} \blacksquare & \blacksquare \\ \blacksquare & \blacksquare \end{matrix}$ then $P'_{i,j} = 1$
 - (c) If $R'_{i,j} = 0$ and $O_{i,j} = \begin{matrix} \blacksquare & \blacksquare \\ \blacksquare & \blacksquare \end{matrix}$ then $P'_{i,j} = 1$
 - (d) Otherwise $P'_{i,j} = 0$
4. Repeat step 3 until all pixels of ownership share O' are processed.
5. END.

Algorithm for secret revelation by computers

4 Results

In this section, the performance of the proposed copyright protection scheme is demonstrated. We have selected 512×512 gray level Lena image as the host image. The host image is shown in Fig. 1(a). The secret image is a visually recognizable binary image of "Rhino" of size 256×256 is shown in Fig. 1(b). The master share generated from the host image is shown in Fig. 2(a). The ownership share generated by the master share and the secret image "Rhino" is shown in Fig. 2(b). The revealed secret image by stacking the master share Fig. 2(a) and the ownership share Fig. 2(b) is shown in Fig. 3. The stacking operation is performed by simply "ORing" the two images in computers. As evident from Fig. 3, we can obtain the secret by using the techniques of visual

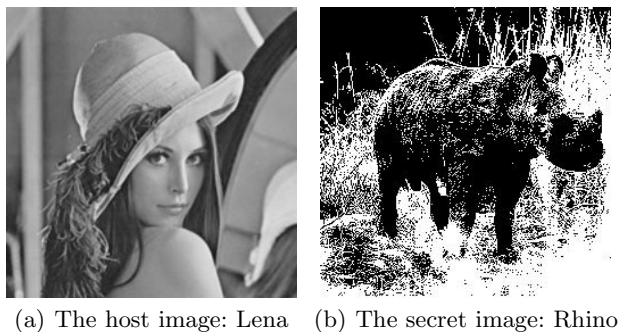


Fig. 1. The host image and secret image

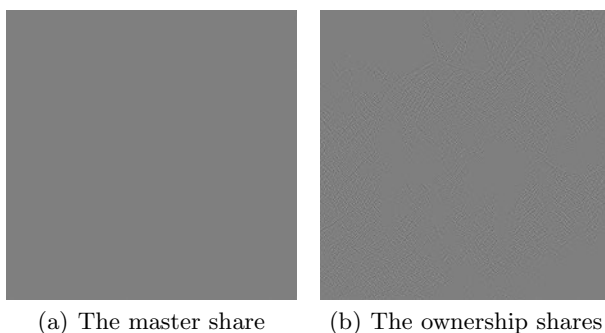


Fig. 2. The master share and the ownership shares generated by our algorithm

cryptography without using too much of computational power. We also tried to prove the resistance of the proposed scheme to various distortions like adding Gaussian noise, salt and pepper noise, median filtering, histogram equalization and rotation by distorting the host image. The imperceptibility of watermark in the proposed method has been evaluated against incidental attacks by using the metric Peak Signal to Noise Ratio (PSNR). The performance metric PSNR is defined as follows:

$$PSNR = 10 \times \log \frac{255^2}{MSE} \quad (1)$$

where

$$MSE = \frac{1}{M \times N} \sum_{i=1}^M \sum_{j=1}^N (h_{i,j} - h'_{i,j})^2, \quad (2)$$

where $h_{i,j}$ is the pixel value of the original image and the $h'_{i,j}$ is the pixel value of the recovered image. MSE is the Mean Squared Error.

To evaluate the robustness of our scheme, we added a Gaussian noise with variance 0 and mean 0.001. The PSNR of recovered image under attack to original image is found to 21.92. The salt and pepper noise with 0.002, the PSNR



Fig. 3. The recovered image after stacking master share and ownership share

Table 1. PSNR under different attacks

Attacks	parameters	PSNR
Gaussian Noise	0, 0.001	21.92
Salt and Pepper	0.002	30.31
Median Filtering	3 X 3	22.18
Histogram Equal		18.01

obtained was 30.31. The 3×3 median filtering of the host image give a PSNR value of 22.18. We also tested our procedure under histogram equalization, the result was a PSNR value of 18.01. The different attacks and along with their parameters and the obtained PSNR values are tabulated in Table 1.

5 Conclusion

A novel copyright protection scheme for digital images based on visual cryptography with using the most significant bit of the host image is proposed in this article. The requirements of robustness and unambiguousness were satisfied by the use of most significant bit (MSB), since a small change on most significant bit (MSB) will distort the complete image so most significant bit (MSB) of an image can not be easily changed by many common attacks. The experimental results proved that the proposed scheme can resist several common attacks. Additionally, the proposed scheme does not alter the host image, and can identify the ownership without resorting to the original image. Hence, it is very suitable to protect those digital images that can not be altered, such as medical images. Our scheme allows the secret image to be of any size regardless of the size of the host image. In our method, we fully utilized the advantages of visual cryptography which can recover the secret image with human eyes without the aid of computers. Although the present version of the proposed scheme deals only with bi-level secret images, it is possible to extend the method to gray-level or color secret images. The authors are currently trying to extend this idea to color and gray level images.

References

1. Braudaway, G.W.: Protecting publicly-available images with an invisible image watermark. In: ICIP (1), pp. 524–527 (1997)
2. Chang, C.C., Hsiao, J.Y., Yeh, J.C.: A colour image copyright protection scheme based on visual cryptography and discrete cosine transform. *Imaging Sciences* 50, 133–140 (2002)
3. Podilchuk, C.I., Delp, E.J.: Digital watermarking: algorithms and applications. In: *IEEE Signal Processing Magazine*, pp. 33–46 (2001)
4. Cox, I.J., Kilian, J., Thomson Leighton, F., Shamoon, T.: Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing* 6(12), 1673–1687 (1997)
5. Hou, Y.-C., Chen, P.-M.: An asymmetric watermarking scheme based on visual cryptography. In: 5th International Conference on Signal Processing, vol. 2, pp. 992–995 (2000)
6. Hsu, C.S., Hou, Y.C.: Copyright protection scheme for digital images using visual cryptography and sampling methods. *Optical Engineering* 44(7), 1–9 (2005)
7. Hwang, R.J.: A digital image copyright protection scheme based on visual cryptography. *Tambang Journal of science and Engineering* 3, 97–106 (2000)
8. Low, S.H., Maxemchuk, N.F.: Performance comparison of two text marking methods. *IEEE Journal on Selected Areas in Communications* 16, 561–572 (1998)
9. Matsui, K., Ohnishi, J., Nakamura, Y.: Embedding a signature to pictures under wavelet transform. *IEICE Trans. J79-D-II(6)*, 1017–1024 (1996)
10. Naor, M., Shamir, A.: Visual cryptography. In: De Santis, A. (ed.) *EUROCRYPT 1994*. LNCS, vol. 950, pp. 1–12. Springer, Heidelberg (1995)
11. Parthasarathy, A., Kak, S.: An improved method of content based image watermarking. *IEEE Transaction on broadcasting* 53(2), 468–479 (2007)
12. Petitcolas, F.A.P., Anderson, R.J., Kuhn, M.G.: Information hiding: A survey. *Proceedings of IEEE* 87(7), 1062–1078 (1999)
13. Wang, M.-S., Chen, W.-C.: Digital image copyright protection scheme based on visual cryptography and singular value decomposition. *Optical Engineering* 46, 67006 (2007)

A Weighted Location Based LSB Image Steganography Technique

Amitava Nag¹, Jyoti Prakash Singh¹, Srabani Khan¹, Saswati Ghosh¹,
Sushanta Biswas², D. Sarkar², and Partha Pratim Sarkar²

¹ Department of Information Technology
Academy of Technology, West Bengal, India

² Department of Engineering and Technological Studies
University of Kalyani, West Bengal, India

Abstract. Steganography is the art of hiding the presence of communication by embedding secret messages into innocent, innocuous looking cover documents, such as digital images, videos, sound files. We present here a novel steganographic method based on affine cipher encryption algorithm and the least significant bit (LSB) substitution in order to provide a strong security and imperceptible visual quality to secret message. We encrypt the 8 bit secret image by changing pixel values using affine cipher. After that each 8 bit pixel of encrypted secret image is divided into 4 groups of 2 bit each. Each part which have a decimal value between 0 to 3 determines the location in each pixel of cover image where to embed the message. We do not store the actual secret message instead we encode the secret message into cover image using the value of each group of secret message. Since, we have two layers of encoding: one using private keys of affine cipher and other for steganography, our methods proves to be more secure than others. Our experimental results also proves that the proposed method has got an acceptable image quality as supported by PSNR values.

Keywords: Steganography, Affine Cipher, LSB Technique, Information Hiding, image processing.

1 Introduction

The Internet has proved to be an excellent distribution system for the digital media because of its inexpensiveness and efficiency. Due to widespread use of Internet, the sharing and transmission of images in digital form has become quite easy. However, the transmitted data can be very easily copied or modified by unauthorized persons in cyberspace. Therefore, finding ways to transmit data secretly through Internet has become an important issue. Encryption is a one of the ancient way to provide a safe way by transforming data into a cipher text via cipher algorithms [10]. Encryption techniques scrambles the message so that it cannot be understood by unauthorized users. However, this can naturally raise the curiosity level of an eavesdropper. It would be rather more prudence if

the secret message is cleverly embedded in another media such that the secret message is concealed to everyone. This idea forms the basis for steganography [4], which is a branch of information hiding by camouflaging secret information within other information. The word steganography in Greek means "covered writing" [6]. Steganography is the art of hiding the presence of communication by embedding secret messages into innocent, innocuous looking cover documents, such as digital images, videos, sound files [6]. In context of steganography, a message represents the information that can be embedded into a bit stream. The cover medium is an image, video, or audio signal that conceals the message. The stego-medium is the result of embedding the message in cover-medium. A possible formula of the steganographic process may be represented as

$$Cover_medium + Embedded_message + Stego_key = Stego_medium \quad (1)$$

Images provide excellent carriers for hidden information. Many different techniques have been introduced to embed messages in images [6]. The most common approaches for message hiding in images are Least Significant Bit (LSB) modification, frequency domain techniques [1][2] and spread spectrum techniques. A recent survey of these techniques is given in [6]. One of the main objective of steganography is to hide a secret message inside cover media in such a way that the secret message is not visible to the observer. The unwanted parties should not be able to distinguish between the cover-image and stego-image to prevent any sense of message inside cover. Thus the stego-image should not deviate much from original cover-image. In this article, we have proposed a novel steganographic procedure based on affine cipher encryption and location based modified Least Significant Bit (LSB) replacement. Instead of substituting the exact message in the cover image, we encode the secret message based on their weight and put on and off some bits of pixel of the cover image. We also encrypt the message prior to embedding so that even some one senses the message in cover and decodes it from cover, he or she will get a encrypted version of the message. The decryption can be done by the parties which hold the correct keys. The rest of the article is organized as follows. In section 2, we briefly discuss the related work done in the area of least significant bit (LSB) substitution steganography. We present our proposed algorithm of encryption and steganography in section 3. Section 4 presents our the experimental results and security analysis. We conclude the paper in Section 5 pointing to some future directions.

2 Related Works

By far the most popular and frequently used steganographic method is the Least Significant Bit embedding (LSB). It works by embedding message bits in the LSBs of sequentially or randomly selected pixels. The selection of pixels depends upon the secret stego key shared by the communicating parties. The popularity of the LSB embedding is due to its simplicity. The least significant bit (LSB) substitution embeds secret data by replacing k LSBs of a pixel with k secret bits directly [3]. Mathematically, the pixel value $c_{i,j}$ of the chosen pixel of cover

image for storing the k-bit message $m_{i,j}$ is modified to form the stego-pixel $s_{i,j}$ as follows:

$$s_{i,j} = c_{i,j} - c_{i,j} \% 2^k + m_{i,j} \quad (2)$$

where % represent modulus operation. Many optimized LSB methods have been proposed to improve this work [12,4,8]. The human perceptibility has a property that it is sensitive to some changes in the pixels of the smooth areas, while it is not sensitive to changes in the edge areas. Not all pixels in a cover image can tolerate equal amount of changes without causing noticeable distortion. Hence, to improve the quality of stego images, several adaptive methods have been proposed in which the amount of bits to be embedded in each pixel is variable [14,13,5,9]. In 2003, Wu and Tsai proposed a novel steganographic method that uses the difference value between two neighboring pixels to determine how many secret bits should be embedded [13]. Chang and Tseng proposed a side match approach to embed secret data, where the number of bits to be embedded in a pixel is decided by the difference between the pixel and its upper and left side pixels [5]. In 2005, Wu et al. presented a novel steganographic method, which combined pixel-value differencing and LSB substitution [14]. Park et al. proposed a new method based on the difference value between two pixels adjacent to the target pixel [9]. In 2008, Wang et al. presented a steganographic method that utilizes the remainder of two consecutive pixels to record the information of secret data [11]. Yang et al. proposed an adaptive LSB steganographic method using the difference value of two consecutive pixels to distinguish between edge areas and smooth areas [15]. All pixels are embedded by the k-bit modified LSB substitution method, where k is decided by the range which the difference value belongs to [15]. Liao et. al. [7] proposed a steganographic method based on four-pixel differencing and modified least significant bit (LSB) substitution to improve the embedding capacity. A Nag et al. used transform domain technique along with Huffman coding for image steganography. In [1], they used discrete cosine transform and in [2], discrete wavelet transform to achieve quite better results in terms of security and visual quality.

3 Proposed Steganography Algorithm

Through out the article, the following notations are used.

- ◇ C represents a cover image with $c_{i,j}$ representing the value at location (i,j) of that image.
- ◇ B represents a block of the cover image with $b_{l,k}$ representing the block number (l,k)
- ◇ M represent the message with $m_{i,j}$ representing the value at location (i,j) of that message.
- ◇ E represent the encrypted message with $e_{i,j}$ representing the value at location (i,j) of that encrypted message.
- ◇ S represent the stego image with $s_{i,j}$ representing the value at location (i,j) of that stego image.

- ◇ E' represent the encrypted recovered message with $e'_{i,j}$ representing the value at location (i,j) of that message.
- ◇ M' represent the recovered message with $m'_{i,j}$ representing the value at location (i,j) of that message.
- ◇ B' represents a block of the stego image with $b'_{l,k}$ representing the block number (l,k)

Our steganography procedure consists of two phases. In first phase, we encrypt the message using affine cipher encryption method. The affine cipher encryption process changes the pixel value of location (i,j) according to the following equation

$$e_{i,j} = ((m_{i,j} \times K_1) + K_2)\%256; \tag{3}$$

where $e_{i,j}$ and $m_{i,j}$ represent encrypted value and original pixel value respectively of secret message, K_1 and K_2 are two private keys and % represent modulus operation. We replace the 4 least significant bits of the cover image with 0 by performing the operation.

$$c_{i,j} = c_{i,j} - c_{i,j}\%2^4 \tag{4}$$

where $c_{i,j}$ represents the $(i, j)^{th}$ pixel of the cover image and % represent modulus operation. The encrypted secret message is then divided into 4 groups of 2 bit each. The decimal value of each group decides the location where to put the message in the cover image. To embed the message, we also divide the cover image into blocks of 4 pixels. Depending on the value of each group of pixel in encrypted secret message, we put a 1 in each pixel of the block of the cover image. The mapping of 4 groups of secret message to 4 blocks of cover image is given in detail in algorithm 1.

3.1 The Embedding Algorithm

Algorithm 1: The embedding algorithm

Input: A gray-level cover image C of size $h \times w$, a 8 bit gray-level secret image M of size $\frac{h}{2} \times \frac{w}{2}$ and two private keys K_1 and K_2 .

Output: An stego image of size $h \times w$

Steps

1. for each pixel $c_{i,j}$ of cover image C
 - (a) perform $c_{i,j} = c_{i,j} - c_{i,j}\%2^4$
2. for each pixel $m_{i,j}$ of cover image M
 - (a) perform $e_{i,j} = ((m_{i,j} \times K_1) + K_2)\%256$
3. Decompose C into $\frac{h}{2} \times \frac{w}{2}$ number of 2×2 blocks
4. For each pixel $e_{i,j}$ of E and block $B_{i,j}$ of C do
 - (a) Divide the pixel value $e_{i,j}$ as follows
 $(b_7b_6)_3(b_5b_4)_2(b_3b_2)_1(b_1b_0)_0$.
 - (b) place a 1 at $(b_7b_6)_3^{th}$ location in pixel (1,1) of $B_{i,j}$ block
 - (c) place a 1 at $(b_5b_4)_2^{th}$ location in pixel (1,2) of $B_{i,j}$ block

- (d) place a 1 at $(b_3b_2)_1^{th}$ location in pixel (2,1) of $B_{i,j}$ block
 - (e) place a 1 at $(b_1b_0)_0^{th}$ location in pixel (2,2) of $B_{i,j}$ block
5. END.

The extraction algorithm works in the reverse way by first making the recovered secret image and then decrypting the recovered secret image to get the actual secret image. The extraction process divides the stego image into blocks of 2×2 . All 4 pixels of a block is read to get a pixel for recovered secret image. The complete process of recovering the secret image is given in algorithm 2. The recovered secret image is then decrypted by the following equation

$$m'_{i,j} = ((e'_{i,j} - K_2)/K_1)\%256 \quad (5)$$

where % represent modulus operation and K_1 and K_2 are the same private keys which were used during encryption.

3.2 The Extracting Algorithm

Algorithm 2: The extracting algorithm

Input: An stego image S of size $h \times w$ and two private keys K_1 and K_2 which were used for encryption

Output: The recovered image M' of size $\frac{h}{2} \times \frac{w}{2}$

Steps

1. Decompose stego image S into $\frac{h}{2} \times \frac{w}{2}$ number of 2×2 blocks
2. For each block $B'_{i,j}$ of the stego image S do
 - (a) Read the 4 LSB of each pixel of block $B'_{i,j}$
 - (b) For 1 in location L_i of pixel (1,1), write $(b_7b_6) =$ which is binary equivalent of location L_i .
 - (c) For 1 in location L_i of pixel (1,2), write (b_5b_4) which is binary equivalent of location L_i .
 - (d) For 1 in location L_i of pixel (2,1), write (b_3b_2) which is binary equivalent of location L_i .
 - (e) For 1 in location L_i of pixel (2,2), write (b_1b_0) which is binary equivalent of location L_i .
 - (f) write $(b_7b_6)(b_5b_4)(b_3b_2)(b_1b_0)$ into pixel $e'_{i,j}$ of recovered encrypted secret image E' .
3. for each pixel $e'_{i,j}$ of recovered encrypted image E'
 - (a) perform $m'_{i,j} = ((e'_{i,j} - K_2)/K_1)\%256$
4. END

Let us consider that one pixel of the encrypted secret image is 10010011 with a block of the cover image with 4-least significant bits replaced by 0's is given in Table 1. The bits of the secret image is divided into 4 groups as 10, 01, 00 and 11. For bit group 10, the decimal value is 2. We put a 1 in 2^{nd} position of pixel (1,1) of given block. Similarly, for group 2, 3 and 4, we put 1 in 1^{st} position of pixel (1,2), 1 in 0^{th} position of pixel (2,1) and 1 in 3^{rd} position of pixel (2,2)

Table 1. A block of the cover image with 4-LSB replaced by 0's

10100000	11100000
00110000	10110000

Table 2. A block of the cover image after insertion of secret image

10100 <u>1</u> 00	111000 <u>1</u> 0
0011000 <u>1</u>	1011 <u>1</u> 000

respectively. The block after the changed pixels values are shown in Table 2 with changed bit are underlined.

During extraction, 1 in 2nd location of pixel (1,1) gives 10, 1 in 1st location of pixel (1,2) gives 01, 1 in 0th location of pixel (2,2) gives 00 and 1 in 3rd location of pixel (2,2) gives 11. When these values are put in order, they form the bit patterns 10010011 of recovered image.

4 Experimental Results

In this section, the performance of the proposed copyright protection scheme is evaluated and discussed. We have selected 256 × 256 sized 8-bit gray level image of Lena as the host image as shown in Fig. 3. The secret image is a visually recognizable 8-bit gray level image of "Ship" of size 64 × 64 is shown in Fig. 1. The encrypted secret generated by applying the affine cipher encryption is shown in Fig. 2. The stego image generated by embedding the encrypted secret image shown in Fig. 4. We have used the PSNR values to measure how close our stego images are to the original cover images.



Fig. 1. The secret image

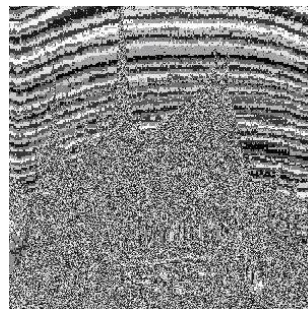


Fig. 2. The affine transformed secret image

The larger PSNR indicates that the difference between the cover-image and the stego-image is very small and this is what is desirable by any steganographic algorithm. We have got PSNR value of around 30 for all test images. The PSNR values for different images are shown in Table 3. The PSNR values as well as the visual appearance of the stego image shown in Fig. 4 suggests that the distortion level in our stego image is very less and insensitivity to human eye. Our method



Fig. 3. The cover image: Lena

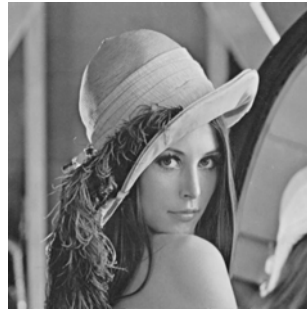


Fig. 4. The Stego image: Lena

Table 3. Capacity and PSNR for different Images

Images	Size (pixel)	Capacity (pixel)	PSNR
Lena	256	64	30.48
Baboon	256	64	30.28
Airplane	256	64	30.91
Boat	256	64	30.36

does not only improves the visual quality, but also provides strong security to the message. We do not store the actual data from the secret message which makes it robust to steganalysis. Even if attackers are able to know that LSB techniques are used, normal LSB steganalysis will not suffice to get the secret image from cover image. If by some improved steganalysis, some attackers recovers the secret image. They will get an encrypted message which will be meaningless to them. They needs to find out the keys of the Affine ciphers to decrypt that message and get the actual content.

5 Conclusion

In this paper, we have proposed a novel steganographic method based on Affine cipher and location based LSB substitution. Secret data are encoded into each pixel of cover image by 4-bit LSB modification method. We do not embed the actual data instead we change the bit values of certain position in LSB of the cover image. Along with this encoding, we also perform encryption using affine cipher to encrypt the message which makes it more secure than existing steganographic techniques. Experiments show that the stego-image of our method are almost identical to the cover image. The stego image generated by our method has got just one 1 in 4 LSB of the stego image which can be a point of attack by steganalyzers. The authors are currently engaged into finding ways to mitigate this limitation and make a more robust signature free stego image.

References

1. Nag, A., Biswas, S., Sarkar, D., Sarkar, P.P.: A novel technique for image steganography based on block-dct and huffman encoding. *International Journal of Computer Science and Information Technology* 2(3), 103–111 (2010)
2. Nag, A., Biswas, S., Sarkar, D., Sarkar, P.P.: A novel technique for image steganography based on dwt and huffman encoding. *International Journal of Computer Science and Security* 4(5), 561–570 (2010)
3. Bender, D.W., Gruhl, N.M., Lu, A.: Techniques for data hiding. *IBM Systems Journal* 35, 313–316 (1996)
4. Chan, C.K., Cheng, L.M.: Hiding data in images by simple lsb substitution. *Pattern Recognition* 37(3), 469–474 (2004)
5. Chang, C.C., Tseng, H.W.: A steganographic method for digital images using side match. *Pattern Recognition Letters* 25(12), 1431–1437 (2004)
6. Cheddad, A., Condell, J., Curran, K., McKevitt, P.: Digital image steganography: Survey and analysis of current methods. *Signal Processing* 90, 727–752 (2010)
7. Liao, X., Wen, Q.-Y., Zhang, J.: A steganographic method for digital images with four-pixel differencing and modified lsb substitution. *Journal Visual Communication and Image Representation* 22, 1–8 (2011)
8. Lin, I.-C., Lin, Y.-B., Wang, C.-M.: Hiding data in spatial domain images with distortion tolerance. *Comput. Stand. Interfaces* 31, 458–464 (2009)
9. Park, Y.-R., Kang, H.-H., Shin, S.-U., Kwon, K.-R.: A steganographic scheme in digital images using information of neighboring pixels. In: Wang, L., Chen, K., S. Ong, Y. (eds.) *ICNC 2005*. LNCS, vol. 3612, pp. 962–967. Springer, Heidelberg (2005)
10. Stallings, W.: *Cryptography and Network Security: Principles and Practices*, 4th edn. Pearson Education Pvt. Ltd, India (2004)
11. Wang, C.-M., Wu, N.-I., Tsai, C.-S., Hwang, M.-S.: A high quality steganographic method with pixel-value differencing and modulus function. *Journal of System Software* 81, 150–158 (2008)
12. Wang, R.Z., Lin, C.F., Lin, J.C.: Image hiding by optimal lsb substitution and genetic algorithm. *Pattern Recognition* 34(3), 671–683 (2001)
13. Wu, D.C., Tsai, W.H.: A steganographic method for images by pixel-value differencing. *Pattern Recognition Letters* 24(9-10), 1613–1626 (2003)
14. Wu, H.C., Wu, N.I., Tsai, C.S., Hwang, M.S.: Image steganographic scheme based on pixel-value differencing and lsb replacement methods. *Images Signal Processing* 152(5), 611–615 (2005)
15. Yang, C.-H., Weng, C.-Y., Wang, S.-J., Sun, H.-M.: Adaptive data hiding in edge areas of images with spatial lsb domain systems. *IEEE Transactions on Information Forensics and Security* 3(3), 488–497 (2008)

Comments on ID-Based Client Authentication with Key Agreement Protocol on ECC for Mobile Client-Server Environment

SK Hafizul Islam and G.P. Biswas

Department of Computer Science and Engineering
Indian School of Mines, Dhanbad-826004, Jharkhand, India
hafi786@gmail.com, gpbiswas@gmail.com

Abstract. In 2011, Debiao et al. proposed an ID-based remote mutual authentication with key agreement scheme on ECC for mobile client-server environment [H. Debiao, C. Jianhua, H. Jin: An ID-based client authentication with key agreement protocol for mobile client-server environment on ECC with provable security, Information Fusion, 2011]. They claimed their scheme provides remote mutual authentication with key agreement and is secured against various known attacks. In this paper, we show that their proposed scheme has some other security flaws.

Keywords: Mutual authentication, elliptic curve, user anonymity, identity based cryptosystem.

1 Introduction

Remote user authentication means, a remote server and a user mutually authenticates the legitimacy of each other over unreliable networks. With the rapid growth of Internet and wireless network technologies and due to the portability of mobile devices (i.e. cell phone, PDA, notebook PC etc.), online transaction by mobile devices can be realized easily. Remote user authentication based on public key infrastructure (PKI) [1], [2] is the traditional approach. However PKI-based remote user authentication schemes suffer from heavy management, revocation, delivery and verification of public key certificate. In addition to that, PKI-based remote authentication scheme works correctly on application of modular exponentiation, which is a very time consuming operations. Mobile devices have very limited battery capacity, storage space and computation ability, so the PKI-based remote authentication cannot be applied for mobile devices. An alternative solution to traditional PKI-based remote authentication system is elliptic curve cryptosystem (ECC) [3], [4] based remote user authentication. Recently ECC-based mutual authentication scheme [5], [6], [7], [8], [9] has been extensively deployed in the Internet or wireless networks for mobile users. Because 160 bit ECC key provides same level of security as of 1024 bit RSA key. So the computation cost and storage cost can be reduced tremendously by means of ECC in any remote login scheme for mobile devices. But ECC-based remote login scheme has some problems. Like other

PKI cryptosystem, ECC-based system needs additional storage space to store users' public keys and certificates. Moreover, user needs additional computation capability to verify the other's public key certificate. Above mentioned problems can be overcome with the help of identity-based cryptosystem (IBC), which was initially proposed by Adi Shamir [10] in 1984. In 2001, Boneh and Franklin [11] gave full functional solution for IBC using bilinear pairing over elliptic curve. IBC reduce the overheads of public key certificate, because public key of an entity, is an easily computable function of his identity such as email address, physical IP address, etc., rather than a random number and the corresponding private key is generated by binding the identity of the entity with a master secret key of a trusted authority, called private key generator (PKG). The private key of the entity, is given through a secure channel and known to only particular entity and PKG, but its legitimacy can be verified publicly.

In 2006, Das et al. [6] proposed a remote user authentication scheme using bilinear pairings with smart cards. But Chou et al. [7] identified that the Das et al.'s scheme is breakable against replay attack and then made an improvement over Das et al. Later on, Goriparthi et al. [8] shows that Chou et al. still suffer from the replay attack. In 2008, Tseng et al. [9] presented a provably secure and efficient pairing-based client authentication protocol for wireless clients with smart cards. In 2009, Wang et al. [12] proposed a dynamic ID-based remote user authentication scheme and claimed that their scheme is more efficient and secure than Das et al.'s scheme. In 2011, Khan et al. [13] identify that Wang's scheme has the following security flaws: no provision of user's anonymity, inability to offer user free choice in choosing his password, vulnerability to insider attack, no provision for revocation of lost or stolen smart card, and does not provide session key agreement. In 2009, Yang and Chang [14] proposed an ID-based remote mutual authentication with key agreement protocol on ECC. Yoon and Yoo [15] found Yang and Chang's protocol is vulnerable to an impersonation attack and does not provide perfect forward secrecy. In 2010, Chen et al. [19] independently showed that Yang and Chang's scheme vulnerable to impersonation attack and insider attack.

Recently, Debiao et al. [17] proposed an ID-based remote mutual authentication with key agreement scheme on ECC for mobile client-server environments. They claimed their scheme provides remote mutual authentication and session key agreement with low computation cost and is secured against various known attacks. In this paper, we show that Debiao's scheme cannot withstands the clock synchronization problem, many logged-in users' problem, known key session-specific temporary attack, impersonation attack, privilege-insider attack, inability to protect user's anonymity and no provision for changing/updating private key with same identity.

The rest of this paper is organized as follows. Section 2 briefly reviews Debiao's scheme. Weaknesses of Debiao et al.'s scheme is presented in section 3. Section 4 concludes the paper.

2 Review of Debiao et al.'s Scheme

This section describes the Debiao's ID-based remote mutual authentication scheme using ECC for client-server environment. Description about elliptic curve can be

found in [3], [4]. The proposed protocol is divided into three phases: system initialization phase, client registration phase, and mutual authentication with key agreement phase. Following notations are used throughout the scheme.

Table 1. Notations used in Debiao’s scheme

Notations	Description
C_i	A client.
S	The remote server.
ID_{C_i}	Identity of the client C_i .
p, n	Two large prime numbers.
F_p	A finite field.
$E_p(a,b)$	An elliptic curve defined on finite field F_p with prime order n .
$G_p(a,b)$	An additive cyclic group of elliptic curve points on $E_p(a,b)$.
P	A base point on elliptic curve $E_p(a,b)$ with order n .
$H_1:\{0,1\}^* \rightarrow Z_n^*$	Three secure one-way hash function (i.e. SHA-1).
$H_2:\{0,1\}^* \rightarrow Z_p^*$	
$H_3:\{0,1\}^* \rightarrow Z_p^*$	
$MAC_K(m)$	Secure message authentication code of message m under the key K .
(x,P_S)	Private/public key pair of the server S , where $P_S = x \cdot P$.

2.1 System Initialization Phase

In this phase, S generates the following system parameters.

- (1) S chooses an elliptic curve equation $E_p(a,b)$.
- (2) S selects a base point P with the order n over $E_p(a,b)$.
- (3) S selects its master key $x \in Z_n^*$ and computes public key $P_S = x \cdot P$.
- (4) S chooses three secure one-way hash functions $H_1(\cdot), H_2(\cdot), H_3(\cdot)$ and a message authentication code $MAC_K(m)$.

The server S keeps x in private and publishes $\{F_p, E, n, P, P_S, H_1, H_2, H_3, MAC_K(m)\}$.

2.2 Client Registration Phase

In the registration phase, the client C_i submits his identity ID_{C_i} to the remote server S for registration. The server S computes $h_{C_i} = H_1(ID_{C_i})$ and clients private key/public key pair $D_{C_i} = (x + h_{C_i})^{-1}P$ and $P_{C_i} = (x + h_{C_i})P$. The server S returns the private key D_{C_i} with identity ID_{C_i} to the client C_i through secured channel. After receiving (ID_{C_i}, D_{C_i}) , the client C_i validates the public P_{C_i} by checking the equation $P_{C_i} = D_{C_i} \cdot P = x \cdot P + P_S$.

2.3 Mutual Authentication with Key Agreement Phase

In this phase, both the client C_i and the server S mutually authenticate each other and then generate a common session key. To access resources stored on server’s database,

first the client C_i sends a login request to the server S . The server S verifies the authenticity of the client's login request. If the login request is valid, then the server S returns another authentication message to the client C_i , which helps the client to prove the authenticity of the remote server S . Subsequently, both the client and the server generate a common session key. A brief description about this scheme is given below.

Step 1. The client C_i chooses a random number $r_{C_i} \in Z_n^*$ and computes $M = r_{C_i} P$,

$M' = r_{C_i} D_{C_i}$ and $K = H_2(ID_{C_i}, T_{C_i}, M, M')$, where T_{C_i} is the current timestamp of the client's system. Now the client C_i sends login request message $M_1 = \{ID_{C_i}, T_{C_i}, M, MAC_K(ID_{C_i}, T_{C_i}, M)\}$ to the server S .

Step 2. Upon receiving the login message $M_1 = \{ID_{C_i}, T_{C_i}, M, MAC_K(ID_{C_i}, T_{C_i}, M)\}$ from C_i , then validity of ID_{C_i} and the timestamp T_{C_i} is checked by the server S using the condition $T' - T \leq \Delta T$, where T' indicate received time of M_1 and ΔT is valid time interval. Server S rejects the login request if any condition is violated. Otherwise the server S computes $h_{C_i} = H_1(ID_{C_i})$, $M' = (x + h_{C_i})^{-1} M$ and $K = H_2(ID_{C_i}, T_{C_i}, M, M')$. Integrity of (ID_{C_i}, T_{C_i}, M) is checked with $MAC_K(ID_{C_i}, T_{C_i}, M)$ using the key K , and the server S rejects the login request if integrity check fails. Otherwise, S chooses a random number $r_S \in Z_n^*$, then computes $W = r_S P$, $K_S = r_S M$ and the session key $SK = H_3(ID_{C_i}, T_{C_i}, T_S, M, W, K_S)$. Now the server S sends the message $M_2 = \{ID_{C_i}, T_S, W, MAC_K(ID_{C_i}, T_S, W)\}$ to the client C_i , where T_S is current system time of the server.

Step 3. When the client C_i receives the message M_2 from S , then C_i validates the timestamp T_S and the integrity of $\{ID_{C_i}, T_S, W\}$ to check the authenticity of the server S . The client C_i computes $K_{C_i} = r_{C_i} W$ and the session key $SK = H_3(ID_{C_i}, T_{C_i}, T_S, M, W, K_{C_i})$ if all provisions are met.

3 Weaknesses of Debiao et al.'s Scheme

Debiao et al.'s remote user authentication scheme provides mutual authentication and a session key agreement between the client and the server, which is based on ID-based public key cryptosystem. But the conventional remote login scheme follows password based protocol. Various password-based remote login schemes [5]-[8], [12]-[13] have been proposed by researchers. Client can remember short-length password easily, however in the proposed scheme, the remote server generates a long private key (160 bit ECC key) for each client which is difficult to remember provided the private key is frequently used. Except this problem, Debiao et al.'s scheme has some other disadvantages as explained below.

3.1 Many Logged-in Users' Problem

Consider the scenario, if the private key D_{C_i} of C_i , is leaked to more than one person then all who know the private key D_{C_i} and the identity ID_{C_i} of C_i , may attempt to

login the same server S at the same time by originating valid login requests. Suppose that more than one adversary knows ID_{C_i} and D_{C_i} of C_i , they can login the server at the same time by selecting the random number $r_{C_i} \in \mathbb{Z}_n^*$ and executing the following steps of Debiao et al.'s mutual authentication with session key agreement scheme, because all of them employs the same authentication process using C_i 's private key D_{C_i} . The server S cannot detect this scenario.

3.2 Privileged-Insider Attack

Debiao's scheme is based on the identity based cryptosystem and all of the IBC systems suffer from the well known key escrow problem as the remote server knows private key of each client. If the privileged insider of the remote server has the knowledge of C_i 's private key D_{C_i} , he may try to impersonate C_i to access the remote server S . Assuming a privileged-insider obtains the private key D_{C_i} of C_i , now he acts as valid client and can access the remote server without facing any problem. Since the privileged insider knows the private key D_{C_i} , following the Debiao's scheme he is easily capable of login the server S and generates a valid common session key.

3.3 Impersonation Attack

Generally the server stores information about each client in the database which helps to verify the legitimacy of clients during mutual authentication phase, but Debiao et al.'s scheme does not store any information about clients. Suppose the adversary steals the identity ID_{C_i} of a legitimate client C_i , and then re-registers to the server S with ID_{C_i} , then server returns the private key D_{C_i} to the adversary, which is nothing but the private key of C_i . Therefore the adversary can impersonate C_i without any difficulty and can access the remote server S .

3.4 Known Session-Specific Temporary Information Attack

In 2001, Canetti and Krawczyk [18] investigated the known session-specific temporary information attack. Later on, Cheng et al. [19] pointed out that if the session ephemeral secrets are leaked to an adversary accidentally, then from this disclosure, secrecy of the generated session key should not be affected. In Debiao's remote login scheme, the client C_i and the server S generates the common session key $SK = H_3(ID_{C_i}, T_{C_i}, T_S, M, W, K_S)$, where all of $(ID_{C_i}, T_{C_i}, T_S, M, W)$ are public information except K_S and the security of the session key SK depends on the secrecy of $K_S = r_S \cdot M = r_{C_i} \cdot W = r_s \cdot r_{C_i} \cdot P$. According to [18], [19], if the session ephemeral secrets r_{C_i} and r_S are exposed accidentally to an outsiders by somehow, then he can compute K_S and the resulting session key $SK = H_3(ID_{C_i}, T_{C_i}, T_S, M, W, K_S)$ is compromised. Thus the Debiao et al.'s scheme does not prevent this attack.

3.5 No Provision for Changing/Updating Private Key

In real life environments, it is a common practice that conventional remote login scheme supports password change/update scheme to provide the adequate security.

However Debiao et al.'s remote login scheme does not provide private key change/update scheme. Suppose the private key D_{C_i} of C_i is exposed to an adversary in some way and the client C_i detects it. Therefore without any loss, the client C_i needs another fresh private key with the same identity ID_{C_i} . But the Debiao et al.'s scheme does not have such a scheme. We may assume that, following the Debiao et al.'s scheme, client C_i can re-registered to the server S many times, but the client C_i cannot get the fresh private key with the same identity ID_{C_i} , because the private $D_{C_i}=(x+H_1(ID_{C_i}))^{-1}P$ is computed by the identity ID_{C_i} and the server's secret x . Therefore the client C_i has to choose new identity each and every time for fresh private key. Therefore, Debiao's scheme does not provide flexibility for changing/updating the private key.

3.6 Inability to Protect User's Anonymity

User anonymity is one of the security aspects of remote login system. Debiao et al.'s scheme does not preserve the user anonymity. In mutual authentication phase, the identity ID_{C_i} is openly transmitted with the message $M_I=\{ID_{C_i},T_{C_i},M,MAC_K(ID_{C_i},T_{C_i},M)\}$ through open channel. In some environments, i.e. e-voting or secret online-order placement, etc, it is very important to maintain the user secrecy, because from the identity ID_{C_i} , some personal secret information may be leaked about the client C_i . In other words, without employing any effort an adversary recognizes the particular transaction being performed by the client C_i . Therefore a well sound remote login scheme should preserve the user anonymity.

3.7 Clock Synchronization Problem

The mutual authentication phase of Debiao's scheme employs the time stamp to prevent the replay attack and man-in-the-middle attack. But the timestamp raises the clock synchronization problem in large networks, such as wide area networks, mobile communication networks, and satellite communication networks. All the schemes based on the concept of time-stamp can withstand the replay attack using systems' time-stamp provided the system clock must be synchronized; otherwise the scheme will not work properly. Since network environment and transmission delay is unpredictable [20], a potential replay attack exists in all schemes that employ the time-stamp.

4 Conclusion

To achieve better efficiency and security, both the elliptic curve cryptosystem and identity based cryptosystems are used in many remote mutual authentication scheme. The scheme proposed by Debiao et al. is one of such scheme. Debiao et al. claimed that their scheme provides resilience against many cryptographic attacks. In this paper, we have shown that their scheme fails to protect user's anonymity, known session-specific temporary attack, impersonation attack, privileged-insider attack, and many logged-in users' attack and has no provision for leaked key revocation phase.

Acknowledgement

SK Hafizul Islam is working as a fulltime research scholar in the Department of Computer Science and Engineering, Indian School of Mines Dhanbad, under INSPIRE fellowship (Reg. No. IF10247), funded by Department of Science and Technology, Ministry of Science and Technology, Govt. of India.

References

1. ElGamal, T.: A public key cryptosystem and a signature protocol based on discrete logarithms. *IEEE Trans. on Info.* 31, 469–472 (1985)
2. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public key cryptosystems. *Comm. of the ACM* 21(2), 120–126 (1978)
3. Miller, V.S.: Use of elliptic curves in cryptography. In: Williams, H.C. (ed.) *CRYPTO 1985*. LNCS, vol. 218, pp. 417–426. Springer, Heidelberg (1986)
4. Koblitz, N.: Elliptic curve cryptosystem. *J. of Math. of Comp.* 48(177), 203–209 (1987)
5. Das, M.L., Saxena, A., Gulati, V.P., Phatak, D.B.: A novel remote client authentication protocol using bilinear pairings. *Comp. & Secu.* 25(3), 184–189 (2006)
6. Das, M.L., Saxena, A., Gulati, V.P.: A dynamic ID-based remote user authentication scheme. *IEEE Trans. on Cons. Elec.* 50(2), 629–631 (2004)
7. Chou, J.S., Chen, Y., Lin, J.Y.: Improvement of Das et al.'s remote user authentication scheme (2005), <http://eprint.iacr.org/2005/450.pdf>
8. Goriparthi, T., Das, M.L., Saxena, A.: An improved bilinear pairing based remote user authentication scheme. *Comp. Stan. & Inte.* 31, 181–185 (2009)
9. Tseng, Y.M., Wu, T.Y., Wu, J.D.: A pairing-based client authentication protocol for wireless clients with smart cards. *Informatica* 19(2), 285–302 (2008)
10. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Blakely, G.R., Chaum, D. (eds.) *CRYPTO 1984*. LNCS, vol. 196, pp. 47–53. Springer, Heidelberg (1985)
11. Boneh, D., Franklin, M.: Identity-Based Encryption from the Weil Pairing. In: Kilian, J. (ed.) *CRYPTO 2001*. LNCS, vol. 2139, pp. 213–229. Springer, Heidelberg (2001)
12. Wang, Y.Y., Kiu, J.Y., Xiao, F.X., Dan, J.: A more efficient and secure dynamic ID-based remote user authentication scheme. *Comp. Comm.* 32, 583–585 (2009)
13. Khan, M.K.: Cryptanalysis and security enhancement of a 'more efficient & secure dynamic ID-based remote user authentication scheme'. *Comp. Comm.* 34(3), 305–309 (2011)
14. Yang, J.H., Chang, C.C.: An ID-based remote mutual authentication with key agreement scheme for mobile devices on elliptic curve cryptosystem. *Comp. & Secu.* 28(3), 138–143 (2011)
15. Yoon, E., Yoo, K.: Robust ID-based remote mutual authentication with key agreement protocol for mobile devices on ECC. In: 2009 International Conference on Computational Science and Engineering, Vancouver, Canada, pp. 633–640 (2009)
16. Chen, T.H., Chen, Y.C., Shih, W.K.: An Advanced ECC ID-Based remote mutual authentication scheme for mobile devices. In: *Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing*, pp. 116–120 (2010)
17. Debiao, H., Jianhua, C., Jin, H.: An ID-based client authentication with key agreement protocol for mobile client-server environment on ECC with provable security. *Info. Fusi* (2011), doi:10.1016/j.inffus.2011.01.001

18. Canetti, R., Krawczyk, H.: Analysis of key-exchange protocols and their use for building secure channels. In: Pfitzmann, B. (ed.) EUROCRYPT 2001. LNCS, vol. 2045, pp. 453–472. Springer, Heidelberg (2001)
19. Cheng, Z., Nistazakis, M., Comley, R., Vaslu, L.: On the indistinguishability-based security model of key agreement protocols-simple cases. Cryptology ePrint Archive, Report 2005/129 (2005)
20. Gong, L.: A security risk of depending on synchronized clocks. ACM SIGOPS Operating System Review 26(1), 49–53 (1992)

Covariance Based Steganography Using DCT

N. Sathisha¹, K. Suresh Babu², K.B. Raja², K.R. Venugopal², and L.M. Patnaik³

¹ Department of Electronics and Communication Engineering,
R L Jalappa Institute of Technology, Doddaballapura, Bangalore Rural Dist. 561 203, India
nsathisha@gmail.com

² Department of Computer Science and Engineering,
University Visvesvaraya College of Engineering, Bangalore University,
Bangalore 560 001, India
raja_kb@yahoo.com

³ Defence Institute of Advanced Technologies, Pune, India

Abstract. The confidential information is communicated through the open channel in a covert way by using steganography. In this paper we propose the Covariance based Steganography using Discrete Cosine Transform (CSDCT) algorithm. The Average Covariance of the Cover Image (ACCI) is computed and threshold ACCI value is fixed at 0.15. The cover image is segmented into 8*8 cells and the Least Significant Bit (LSBs) are replaced by Most Significant Bits (MSBs) of payload based on ACCI values. It is observed that the capacity, Peak Signal to Noise Ratio (PSNR) and security is better compared to the existing algorithm.

Keywords: Covariance, Cover Image, DCT, Payload, Steganography.

1 Introduction

The rapid developments for exchanging information over the internet require security for confidential information. The important methods employed to protect the confidential information are cryptography, digital water marking and steganography. Cryptography is a method of protecting confidential information by scrambling and mapping pieces of data into cipher text with a key. Digital water marking is the process of embedding information (watermark) into digital multimedia contents such that the information can later be extracted to ascertain the authenticity of the object. Steganography is a word originated from Greek which literally means *covered or hidden writing*. Steganography is an art and science of hiding confidential information into a cover media to ensure the security of information over the communication channel. The cover media can be text, audio, video and image. The text steganography is difficult technique due to lack of redundant information in a text file compared to an image or a sound file. Audio steganography embeds the secret message into digitized audio signal which results in slight altering of binary sequence of the corresponding audio file. Some of the available audio steganography methods are Least Significant Bit (LSB) coding, phase coding and echo hiding. In video steganography secret information to be communicated is embedded into video files. The payload image is hidden in the cover image for image steganography.

Two major steganography methods are spatial and transform domain based. The two approaches of spatial domain steganography are LSB and Bit Plane Complexity Steganography (BPCS). In LSB method the LSB of cover image are replaced by the payload bits. In BPCS the cover image is divided into blocks and categorized into information and noisy blocks, the payload is embedded into noisy blocks. The transform domain steganography methods are Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT) and Fast Fourier Transform (FFT). In DCT the secret information is embedded into the DCT coefficients of cover image [1]. In DWT the secret information is embedded into the higher frequency coefficients of the wavelet transform without altering the low frequency sub bands [2].

Steganographic techniques are (1) *physical steganography* is old manual type of technique for transmitting confidential information from one end to an other end in ancient days and examples of physical steganography are (i) writing messages on the wood and then covered with wax. (ii) message tattooed on a shaved head and is hidden by the hair growth and (iii) writing messages using invisible ink between the lines. (2) *Digital steganography* evolved with advent of the personnel computer being applied to classical steganography problems. Digital steganography techniques are (i) hiding information in the least significant bits of images or sound files. (ii) Pictures embedded in video files played at slower or faster speed. (3) *Network steganography* involve modification of the properties of a single network protocol applied to the protocol data unit. Network steganography covers a broad spectrum of techniques like concealing messages in voice over internet protocol conversations and wireless local area network steganography. (4) *Printed steganography* is the form of printed document. The printed document may be produced by inserting the confidential message into the cipher cover text by changing letter size, spacing or other characteristics of a cipher cover text is manipulated to carry the confidential message. Only a recipient who knows the technique can recover the confidential message by decrypt. Applications of steganography are confidential communication and secret data storing, copyright protection of electronic products, Bank Transactions, Healthcare information, Internet security, Authentication and Information assurance.

Contribution: In this paper we presented CSDCT that hides secret information in the frequency domain using DCT by calculating the average covariance of the cover image which results in increase of the security and capacity.

Organization: This paper is organized into following sections. Section 2 is an overview of related work. The embedding and retrieval model is described in section 3. Section 4 discusses the algorithms used for embedding process. Performance analysis is discussed in section 5 and conclusion is discussed in section 6.

2 Related Work

WeiQi Luo et al., [3] proposed a method which embeds the secret message into sharper edge regions of cover image adaptively according to size of the message and

the gradients of the content edges of cover image. Gyankamal and Shinde [4] developed a method of embedding the encrypted secret message into the black and white cover picture images. This method aims to utilize the cover image as much as possible. The encrypted secret message bits are compared with the blocks of cover image and the maximum matching block is selected for embedding the secret information. Cheng-hsing yang et al., [5] proposed a technique to embed the secret information by Pixel Value Differing (PVD) method. The number of secret bits embedded depends on the difference between two consecutive pixels.

Vijay kumar and Dinesh kumar [6] has presented a performance evaluation of image steganography using DWT applied on cover image and payload to derive four sub bands such as Approximation Coefficients (CA) Vertical Detail Coefficients (CV) Horizontal Detail Coefficients (CH) Diagonal Detail Coefficients (CD). The error blocks are calculated by subtracting the approximation coefficients of cover image from approximation coefficients of secret image. These blocks are replaced with the best matched CH blocks. They made use of CV and CD blocks also to embed the secret images. Aos. et al., [7] implemented a new means of hiding the secret information in the Executable (.EXE) file, such that it is unrevealed to any anti-virus software, since anti-virus software secretly read the furtive data embedded inside the cover file.

Nan-I Wu and Min-Shiang Hwang [8] developed steganographic techniques for gray scale images and introduced schemes such as high hiding capacity schemes and high stego-image degradation imperceptibility schemes. These schemes provide high imperceptibility and data hiding capabilities. Bo-Luen Lai and Long-Wen Chang [9] proposed a transform domain based adaptive data hiding method using haar discrete wavelet transform. The image was divided into sub-bands (LL1, HL1, LH1 and HH1) and most of the data is hidden in the edge region as it is insensitive to the human eye. If these sub-bands were complex, then further division of the bands were performed, so that more number of data bits could be embedded.

Raja et al., [10] proposed a high capacity, secure steganographic algorithm in which the payload bits are encrypted and embedded in the wavelet coefficients of the cover image. This method utilizes the approximation band of the wavelet domain to improve robustness. Han-ling zhang et al., [11] presented a superior embedding scheme in which out of 4 pixels, one pixel is considered as the target pixel. The largest difference values of the other three pixels are used to determine the number of payload pixels to be embedded inside the cover image. Guillermo et al., [12] presented the current techniques of data hiding wherein only the most important portion of the data is encrypted before embedding in the cover media instead of considering the entire bit stream. Moazzin Hossain et al., [13] employed three methods in their scheme, utilizing the pixel's dependency on its neighborhood and the visual redundancy to establish smooth and edge areas. In smooth area, only 3 bits of the secret data is embedded whereas in the edge area, variable rate of bits are embedded. Nazanin Zaker et al., [14] proposed some modifications to the existing pixel value differencing technique, there by attaining stability against histogram quantization at the cost of less capacity. Xiaoming Yao et al., [15] proposed a robust

exploiting modification direction scheme by replacing the gray scale cover image values with statistically stable quantities. This technique provides robustness in Gaussian and salt and pepper noise conditions. Souvik Bhattacharyya et al., [16] developed a technique to improve the security level of secret information by permuted and encoding through integer wavelet transformation. The cover image is segmented in different objects through normalized cut and each part of secret information is embedded through modified LSB embedding method on different cuts of the cover image to form stegoimage. Anjali and Kulkarni [17] proposed a method of embedding secret information in only skin region of the cover image. Secret information is embedded using DWT in one of the high frequency sub band. Cropping the cover image enhances the security because no one can extract information without having value of cropped region. Mamta Juneja and Parvinder Singh [18] introduced a image steganography technique based on LSB insertion and Rivest Shamir Algorithm (RSA) encryption. The secret information is encrypted using RSA public key algorithm and then embedded into least significant bits of cover image. Debnath Bhattacharyya et al., [19] presented a discrete fourier transform based image authentication technique. The cover image of spatial domain is converted into frequency domain using DFT. The bits of the secret information are then embedded at LSB within the real part of the transformed image. Carlos velsco et al., [20] proposed an adaptive data hiding method using convolutional codes and synchronization bits in DCT domain. The cover image is divided into suitable and ineligible blocks based on the DCT energy features from the horizontal, vertical and diagonal frequency information. The suitable blocks are used for embedding data using Quantization Index Modulation (QIM). The two synchronization bits are used for desynchronization problem and convolution codes are used for decoding errors.

3 Proposed CSDCT Model

The proposed embedding and retrieval models are discussed in this section.

1. CSDCT Embedding Model: The block diagram of proposed embedding model is shown in the Figure 1. The number of cover image bits are replaced by the MSBs of payload based on ACCI of cover image, which results in better stego image with reasonable PSNR for any kind of cover image. The payload is secure from intruder as the number of cover image DCT coefficient bits are replaced on the basis of ACCI and DCT coefficient values.

(i) *Cover Image:* The cover image is color or gray scale of any size and any format. If the cover image is color then convert into gray scale image. The gray scale image is converted into pixel intensity values.

(ii) *Cover Image Pixel Management:* The gray scale cover image pixel intensity values vary from lower zero to upper 255 values. During the payload embedding process the lower and higher intensity values of cover image may exceed and which results in difficulty to retrieve the payload at the destination. Hence the cover image pixel intensity values are limited to lower 15 and upper 240, instead of 0 and 255.

(iii) *Covariance and average covariance*: The covariance $\text{cov}(x, y)$ between two random variables x and y with expected values μ_x and μ_y is calculated using the equation

$$\text{cov}(x, y) = E [(x - \mu_x)(y - \mu_y)] \quad (1)$$

The correlation coefficients $\rho_{x, y}$ between two random variables x and y with standard deviations σ_x and σ_y is calculated using the equation

$$\rho_{x, y} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{E [(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y} \quad (2)$$

The average covariance is calculated by adding the correlation coefficients of the cover image and then dividing the sum by the size of the matrix. The threshold value of average covariance of cover image is fixed at 0.15 by trial and error method.

(iv) *Segmentation and DCT*: The cover image matrix is segmented into 8x8 matrices. The DCT is applied on each 8x8 block to get DCT coefficients which are used to hide the payload depending on the adaptive length L . The frequency domain improves the security and robustness during communication of payload.

(v) *Adaptive bit length L* : After converting the 8x8 matrices into the frequency domain, pixel values of the cover image are transformed to DCT coefficients (Co). The length L , which determines the number of LSBs of each coefficients of cover image to be replaced by the payload bits, is calculated according to the conditions given below.

If Average Covariance of Cover Image (ACCI) > 0.15

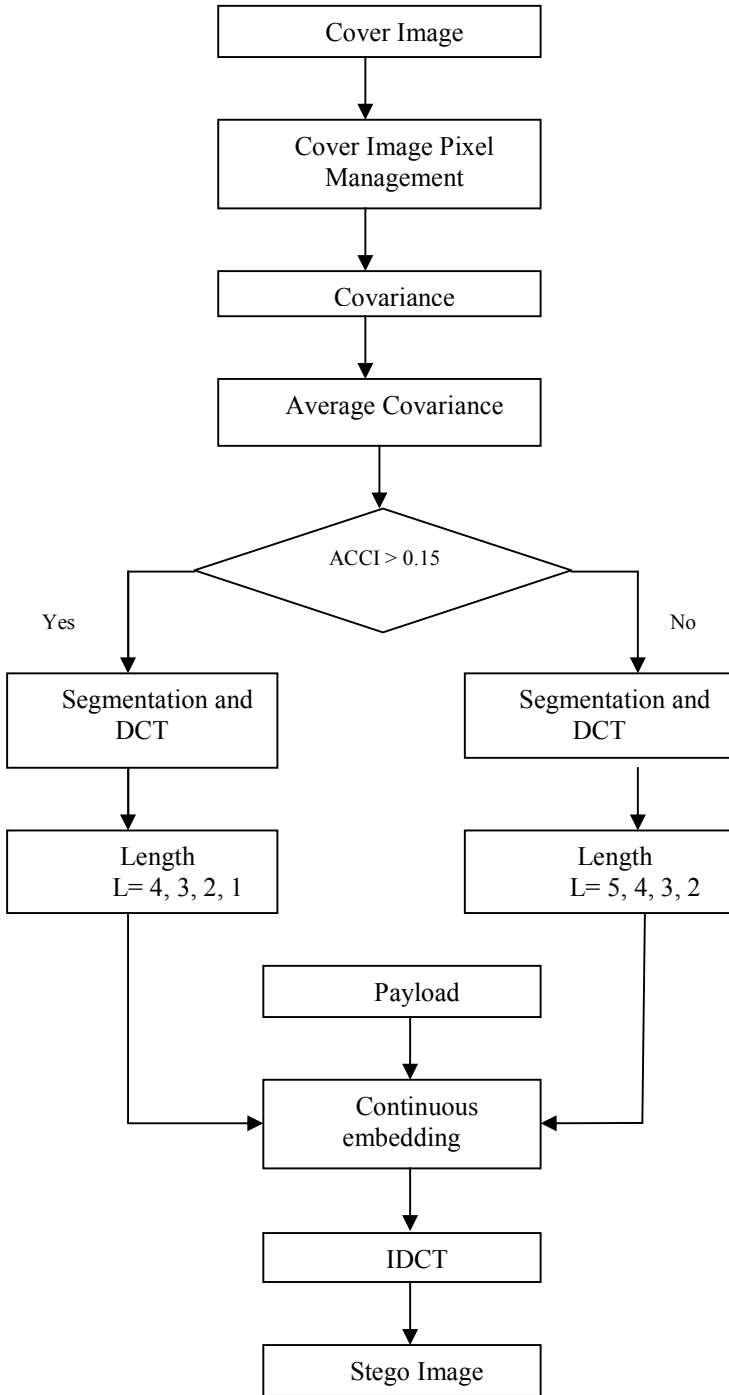
If $\text{Co} \geq 2^5$;	$L=4$,
If $2^5 \leq \text{Co} \leq 2^4$;	$L=3$
If $2^4 \leq \text{Co} \leq 2^3$;	$L=2$,
Else	$L=1$

Else

If $\text{Co} \geq 2^5$;	$L=5$,
If $2^5 \leq \text{Co} \leq 2^4$;	$L=4$
If $2^4 \leq \text{Co} \leq 2^3$;	$L=3$,
Else	$L=2$

(vi) *Embedding*: Four MSBs of each payload pixel are embedded into the segmented 8x8 cover image DCT coefficients in a continuous manner. After embedding the payload into each cover image block the 8x8 stego coefficient matrix is obtained.

(vii) *Inverse Discrete Cosine Transform (IDCT) and Stego image*: The 8x8 stego coefficient matrix is converted into the spatial domain by applying IDCT. The all 8x8 spatial domain matrix are arranged in a proper way to obtain stego image which is equivalent to the cover image.

**Fig. 1.** CSDCT Embedding Flowchart

2. CSDCT Retrieval Model: The payload is extracted from the stego image in the retrieval technique as shown in the Figure 2.

(i) *Stego image:* The stego image is received at the destination over the open channel. Any intruder interfering in the transmission process will only be able to read the stego image and cannot extract the secret image.

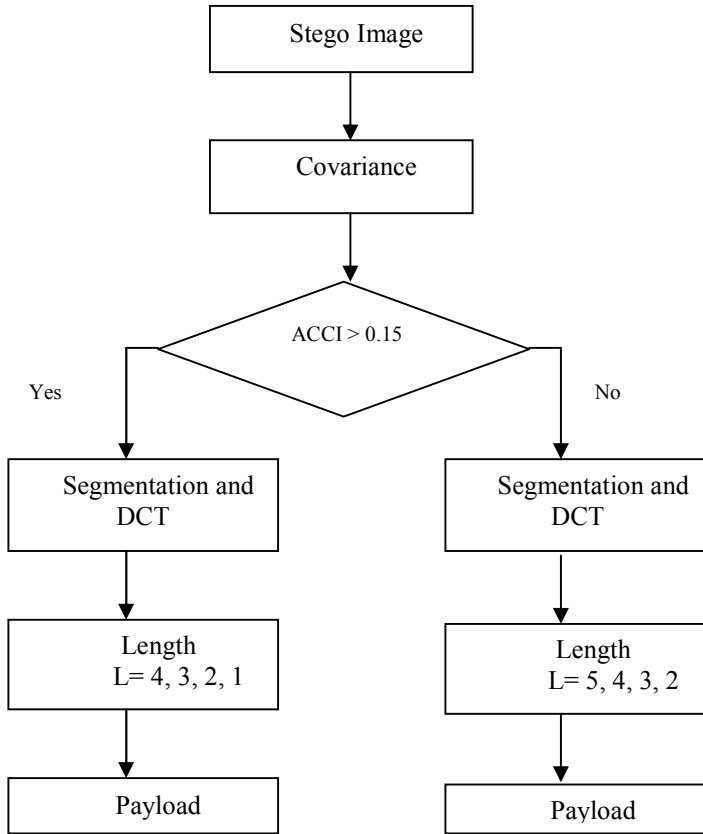


Fig. 2. CSDCT Retrieval Flowchart

(ii) *Covariance:* The covariance and average covariance of the stego image matrix is calculated. Based on the value of the average covariance the number bit length L is calculated at the receiver.

(iii) *Segmentation and DCT:* the stego image is converted into 8x8 blocks to ensure faster computation of DCT coefficients.

(iv) *Length L :* At the receiver L is calculated for the stego image using the procedure as adopted in the embedding process to extract payload.

(v) *Payload:* The extracted payload bits are rearranged in a proper way to get the payload.

4 Algorithm

Problem definition: Given a cover image and payload the objectives are (i) the payload is to be embedded into the cover image to derive stegoimage using average covariance, DCT and variable payload bit stream. (ii) The stego image with reasonable PSNR.

Assumptions: (i) The cover and payload objects are images with different dimensions and formats. (ii) The stego object is transmitted over an ideal channel.

The payload is embedded into the cover image DCT coefficients based on average covariance of cover image and the values of DCT coefficients is given in the Table 1.

Table 1. Algorithm of CSDCT

<p>Input: Cover Image and Payload.</p> <p>Output: Stego Image.</p> <p>Step 1) A cover image of any size and format is considered and if it is color image convert it into grayscale image.</p> <p>Step 2) Applying pixel management to the cover image to avoid overflow and underflow of the pixel values 0 and 255.</p> <p>Step 3) Covariance of cover image is determined and average is computed to get average covariance.</p> <p>Step 4) The average covariance of cover image value is fixed as 0.15, if ACCI > 0.15 go to step 5 else step 6</p> <p>Step 5)</p> <ul style="list-style-type: none"> (i) The cover image is segmented into 8*8 matrix and DCT is applied on each matrix. (ii) Embedding bit length L for each coefficient is calculated as following: <ul style="list-style-type: none"> $L=4$, if $Co \geq 2^5$; $L=3$, if $2^4 \leq Co \leq 2^5$; $L=2$, if $2^3 \leq Co \leq 2^4$; else $L=1$; (iii) Depending on the value of L the number of bits of cover image DCT coefficients is replaced by the MSB bits of payload. (iv) The stego image obtained in the DCT domain is converted back to the spatial domain using IDCT. <p>Step 6)</p> <ul style="list-style-type: none"> (i) The cover image is segmented into 8*8 matrix and DCT is applied on each matrix. (ii) Embedding bit length L for each coefficient is calculated as following: <ul style="list-style-type: none"> $L=5$, if $Co \geq 2^5$; $L=4$, if $2^4 \leq Co \leq 2^5$; $L=3$, if $2^3 \leq Co \leq 2^4$; else $L=2$; (iii) Depending on the value of L the number of bits of cover image DCT coefficients is replaced by the MSB bits of payload. (iv) The stego image obtained in the DCT domain is converted back into the spatial domain using DCT

5 Performance Analysis

For the performance analysis payload (PL) image of Boat and the cover images (CI) Lena, Old Image, Baboon, Barbara, Ranch house, Bridge and Casa are considered and shown in the Figure 3. The average covariance of cover image is calculated and the payload bits are embedded into the cover image depending on the ACCI.

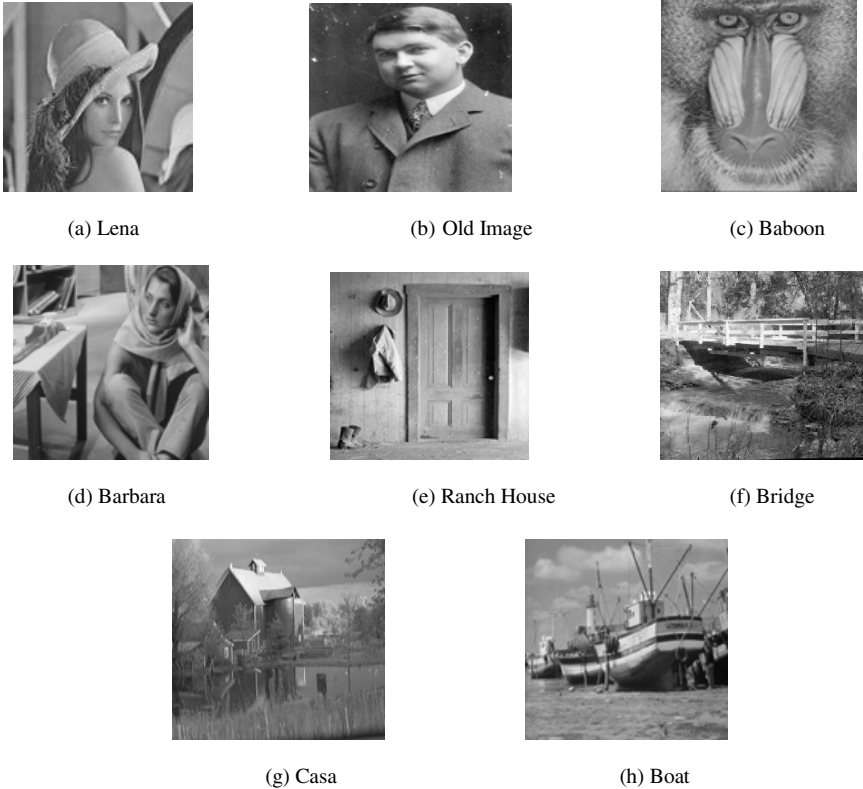


Fig. 3. (a) to (g) are Cover Images and (h) Payload

The payload is embedded into the DCT coefficients of cover image based on ACCI and length L . The performance parameter such as PSNR between cover image and stegoimage is computed and given in the Table 2. It is observed that the PSNR depends on the ACCI of the cover image and also the PSNR decreases as the Hiding Capacity (HC) increases. The PSNR value is maintained around 42 dB for the capacity of 34%.

Table 2. ACCI and PSNR for different cover images with payload boat

Cover Image	ACCI	HC 12.5%	HC 20.0%	HC 25.0%	HC 34.0%
		PSNR	PSNR	PSNR	PSNR
Lena	0.064	46.31	41.51	40.51	39.41
Old image	0.122	45.41	41.95	40.85	39.09
Baboon	0.154	47.25	42.64	41.88	40.71
Barbara	0.203	47.08	43.48	42.54	41.17
Ranch house	0.280	46.06	41.50	40.46	39.20
Bridge	0.358	46.13	41.78	40.96	39.69
Casa	0.504	47.56	43.33	42.57	41.35

The Maximum Hiding Capacity (MHC) and the PSNR between the cover image and stego image is tabulated for existing algorithm *An Adaptive Steganographic Technique Based on Integer Wavelet Transform (ASIWT)* [21] and the proposed algorithm ASCDCT is given in the Table 3. It is observed that the PSNR is improved in the proposed algorithm compared to the existing algorithm.

Table 3. PSNR of existing and proposed techniques for a MHC of 47%

Image	Existing Method (ASIWT)	Proposed Method (CSDCT)
	PSNR	PSNR
CI: Lena PL: Barbara	31.80	39.35
CI: Baboon PL: Cameraman	30.89	37.96

6 Conclusion and Future Work

The covert communication is achieved by steganography. In this paper covariance based steganography using DCT is proposed. The ACCI is computed and the payload bits are embedded into cover image DCT coefficients based on the ACCI values. It is observed that the capacity, security and the PSNR values are improved compared to the existing algorithm. In future the same technique can be extended by applying different transforms to both cover image as well as payload and thus the robustness of algorithm can be verified.

References

1. Sarreshtedari, S., Ghotbi, M., Ghaemmaghami, S.: On the Effect of Spatial to Compressed Domains Transformation in LSB based Image Steganography. In: International Conference on Computer Systems and Applications, pp. 260–264 (2009)
2. Chen, P.-Y., Lin, H.-J.: A DWT Based Approach for Image steganography. International Journal of Applied Science and Engineering, 275–290 (December 2006)
3. Luo, W., Huang, F., Huang, J.: Edge Adaptive Image Steganography Based on LSB Matching Revisited. IEEE Transactions on Information Forensics and Security 5(2), 173–178 (2010)
4. Chhajed, G.J., Shinde, S.A.: Efficient Embedding in B&W Picture Images. In: Second IEEE International Conference Information Management and Engineering, pp. 525–528 (2010)
5. Yang, C.-H., Weng, C.-Y., Wang, S.-J., Sun, H.-M.: Adaptive Data Hiding in Edge Areas of Images with Spatial LSB Domain Systems. IEEE Transactions on Information Forensics and Security 3, 488–497 (2008)
6. Kumar, V., Kumar, D.: Performance Evaluation of DWT based Image Steganography. In: IEEE Second International Conference on Advance Computing, pp. 223–228 (2010)
7. Aos, A.Z., Nazi, A.W., Hameed, S.A., Othman, F., Zaidan, B.B.: Approved Undetectable-Antivirus Steganography. In: International Conference on Computer and Information Technology, pp. 437–441 (2009)
8. Wu, N.-I., Hwang, M.-S.: Data Hiding: Current Status and Key Issues. International Journal of Network Security 4(1), 1–9 (2007)
9. Lai, B.-L., Chang, L.-W.: Adaptive Data Hiding for Images Based on Harr Discrete Wavelet Transform. In: Chang, L.-W., Lie, W.-N. (eds.) PSIVT 2006. LNCS, vol. 4319, pp. 1085–1093. Springer, Heidelberg (2006)
10. Raja, K.B., Vikas, Venugopal, K.R., Patnaik, L.M.: High Capacity Lossless Secure Image Steganography using Wavelets. In: International Conference on Advances Computing and Communications, pp. 230–235 (2006)
11. H.-l. Zhang, G.-Z., Geng, C.-Q., Xiong, C.-Q.: Image Steganography using Pixel Value Differencing. In: Second International Symposium on Electronic Commerce and Security, pp. 109–112 (2009)
12. Francia III, G.A., Yang, M., Trifas, M.: Applied Image Processing to Multimedia Information Security. In: International Conference on Image Analysis and Signal Processing, pp. 104–107 (2009)
13. Hossain, M., Haque, S.A., Sharmin, F.: Variable Rate Steganography in Gray Scale Digital Images using Neighborhood Pixel Information. In: International Conference on Computer and Information Technology, pp. 21–23 (2009)
14. Zaker, N., Hamzeh, A., Katebi, S.D., Samavi, S.: Improving Security of Pixel Value Differencing Steganographic Method. In: Third International Conference on New Technologies, Mobility and Security, pp. 1–4 (2009)
15. Yao, X., Du, W., Wu, W., Huang, M., Fu, J.: A Robust EMD-like Stegnographic Scheme. In: Third International Symposium on Intelligent Information Technology and Security Informatics, pp. 134–137 (2010)
16. Bhattacharyya, S., Kshitij, A.P., Sanyal, G.: A Novel Approach to Develop a Secure Image based Steganographic Model using Integer Wavelet Transform. In: International Conference on Recent Trends in Information, Telecommunication and Computing, pp. 173–178 (2010)

17. Shejul, A.A., Kulkarni, U.L.: A DWT Based Approach for Steganography using Biometrics. In: International Conference on Data Storage and Data Engineering, pp. 39–43 (2010)
18. Juneja, M., Sindhu, P.S.: Designing of Robust Image Steganography Technique Based on LSB Insertion and Encryption. In: International Conference on Advances in Recent Technologies in Communication and Computing, pp. 302–305 (2009)
19. Bhattacharyya, D., Dutta, J., Das, P., Bandyopadhyay, R., Bandyopadhyay, S.K.: Discrete Fourier Transformation based Image Authentication Technique. In: Eighth IEEE International Conference on Cognitive Informatics, pp. 196–200 (2009)
20. Velasco, C., Nano, M., Perez, H., Martinez, R., Yamaguchik.: Adaptive JPEG Steganography using Convolutional Codes and Synchronization Bits in DCT Domain. In: Fifty Second IEEE International Midwest Symposium on Circuits and Systems, pp. 842–847 (2010)
21. El Safy, R.O., Zayed, H.H., El Dessouki, A.: An Adaptive Steganographic Technique Based on Integer Wavelet Transform. In: IEEE Proceedings on International Conference on Networks and Media, pp. 111–117 (March 2009)

An Efficient Algorithm to Enable Login into Secure Systems Using Mouse Gestures

Usha Banerjee^{1,*} and A. Swaminathan²

¹ Department of Computer Science and Engineering,
College of Engineering Roorkee, Roorkee, India
ushaban@gmail.com

² Department of Mathematics, Indian Institute of Technology Roorkee, India
mathswami@gmail.com

Abstract. This paper proposes an algorithm to use mouse gestures as an optional login facility which can be integrated into existing systems that require an additional level of security. The aim of this facility is to eliminate the problem of login using the traditional keyboard in public systems by incorporating a real time user signature. The main advantage of this facility is that no additional hardware component or device is required apart from a standard mouse. This implies that this algorithm will make existing systems more secure at no additional cost. This system works by accepting mouse gestures as inputs from the user. The gestures are then interpreted as patterns of mouse motions and are converted to strings based on change of directions. These strings can be further improved in complexity and fed as password to the system authentication module via a pluggable authentication module (PAM).

1 Introduction

A Mouse gesture is a pattern of mouse movements constructed using different strokes produced by dragging mouse. Mouse Gesture quick login process may provide a low cost alternative to all existing login processes into low security requiring systems. It enables user to login through simple mouse gestures, substituting the need of any extra high cost hardware with conventional low cost hardware required for computers. This system even helps in eliminating the problem of key - logging associated with traditional login requiring input of passwords through keyboard. For high usability of gestures-based interface, three basic features must be preserved: accuracy, efficiency (of recognition), and adaptability to the possibilities and needs of the individual user.

1.1 Related Work

Concurrent login systems have several drawbacks. The traditional way of accepting passwords through keyboard [8] always suffers from the problem of being key

* Corresponding author.

logged. Thus, the world is now proceeding towards bio-metric authentication systems [1, 7, 9], which hold the promise of making the input difficult to tap. But to incorporate them in older machines may be in costly due to new hardware requirements. An alternative solution to key-logging problem may be the use of mouse, as in an on-screen virtual keyboard. However, it may be time consuming for several users to enter the password through it which in the meantime can be noticed by any person keeping a view of the screen. Fingerprint matching, face recognition and on-screen virtual keyboard are all successful existing login processes, but these either require either high cost external hardware or are insecure due to the noticeable way of taking the input. Moreover, some of them even require highly complex algorithms for processing input data which may be inefficient on some slower machines. But none of them accommodates the simplicity as of a gesture through a mouse. Using mouse gestures for login does not require any special hardware and can be quickly entered even by a novice, thus minimizing the possibility of being accurately seen by a person in the vicinity of the screen. The usability of the assumed notion of gesture was assessed during experiments described in Hofman [12], performed according to the standards, e.g. Newman and Lamming [13]. Devices such as finger print recognizer requires users to swipe their finger over a sensor. Such devices use highly reliable bio-metric information of the user for login processes and require additional expensive hardware and complex algorithms to process the input data thus effecting the overall efficiency of slower machines. Most of the previous works on gesture recognition [4, 5, 6] have either issued a command to a computer for a specific task or work as a character recognition application.

2 Our Approach

We devise a very simple and novel algorithm for the login process which can work even on slow machines and does not require any additional expensive hardware. The process is elaborated below. Our aim is to provide a system in which these gestures are stored in form of strings based on some algorithm and use the information for login purposes.

2.1 Method of Gesture Analysis

The main part of the system deals with the analysis of mouse strokes submitted by the user. Our algorithm provides a method to conveniently store and match the pattern of gestures entered by movement of the mouse. Figure-1 shows a typical gesture with corresponding X and Y values formed by dragging the mouse to form the gesture.

2.2 Pattern Recognition Approach

It is quite natural to suggest a pattern recognition [5] approach to the problem. In this approach, we analyze the curves formed by the sequences of x and

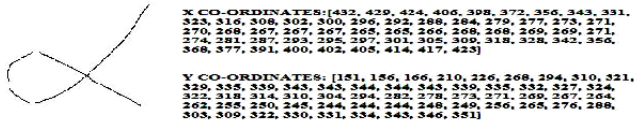


Fig. 1. A mouse gesture with X and Y coordinates

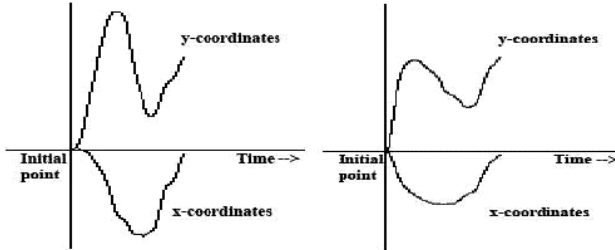


FIGURE 2

Fig. 2. Graphs formed by plotting x and y coordinates of the gesture similar to that shown in figure 1

y coordinates submitted by the user through mouse drags. Figure-2 shows two different graphs formed by plotting x and y coordinates of the gesture similar to that shown in figure-1. The vertical axis shows increase or decrease in the values of the respective coordinates. The differences in these graphs occur due to several reasons namely inability of the user to precisely control the mouse, difference of speed while entering the gesture (usually occurs while trying to redraw the same gesture), size of the gesture etc. Matching these will need pattern matching algorithms that either require large amount of training data or are computationally expensive.

2.3 Vector Chain Coding Approach

This approach is based on the change of directions in x and y plane [5] that can be analyzed by the sequences of x and y coordinates. In this approach, we sequentially process the lists of both x and y coordinates to get the direction vectors obtained by successive coordinates. If we enumerate the 8-directions as shown in figure-3, we will get long strings of these numerals as output which is known as the chain code representation of the curve.

For example, the list $x=[15,20,29]$ and $y=[25,22,24]$ will produce chain code as: 3 5 (Note: Here we have assumed right direction as positive x direction and downward direction as positive y direction). This approach has several problems like the variable length of output strings, repetition in chain code of directions,



Fig. 3. Numerals representing 8-directions for chain code representation of curves

large variation in code with slight change in curves etc. Thus, the improvement requires a method which can filter through minor variations and produce an output string of shorter length.

2.4 Change of Directions Approach

In this approach, rather than absolute chain code of the curve or stroke, we suggest the generation of relative coding. But before this coding, we would like to add a filter mechanism that reports only when an actual change of direction is there.

3 Algorithm

Input: Sequence of x and y coordinates.

Output: String of numerals representing change of directions.

Steps:

Let the original change in direction indicators, $dx=dy=0$. Now, for each successive pair of points in the sequences of x and y coordinates: Find new indicators of directions, i.e., nx and ny . The values of nx and ny will be either of +1 or -1 depending upon increase or decrease in values of x and y coordinates of successive coordinates respectively.

1. Compare nx and ny with dx and dy respectively. Now, change in these variables will add in the result string a value as per the following conditions:

Table 1. Conditions for value of string according to specified conditions

<i>Condition</i>	<i>Value : RelativeDirectionasinfigure2</i>
$nx > dx \text{ and } ny > dy$	5
$nx > dx \text{ and } ny < dy$	3
$nx > dx \text{ and } ny = dy$	4
$nx < dx \text{ and } ny > dy$	7
$nx < dx \text{ and } ny < dy$	1
$nx < dx \text{ and } ny = dy$	8
$nx = dx \text{ and } ny > dy$	6
$nx = dx \text{ and } ny < dy$	2

2. Return the result string.

3.1 Demonstration of the Algorithm

If the user makes a pattern that clearly follows the same sequence of direction changes, the output string will be the same. This must be noted that since this system notices only change in direction so there may be several similar gestures corresponding to a single output string. This is only there to support flexibility of user input. Figure-4 shows that for a M-like gesture the algorithm will produce an output string of 3626. Further, figure-5 shows the details of the process of conversion of a gesture into a pattern string. For the gesture given in figure-6(a) the above algorithm will produce an output string of 37. The gesture given in figure-6(b) consists of 2 (multiple) curves, which will require their sequential analysis. The algorithm will produce output string: 52#74, where, # is the separator separating the two output strings produced in same order as they were drawn.



Fig. 4. For the corresponding M-like gesture, the algorithm will produce output string:3626

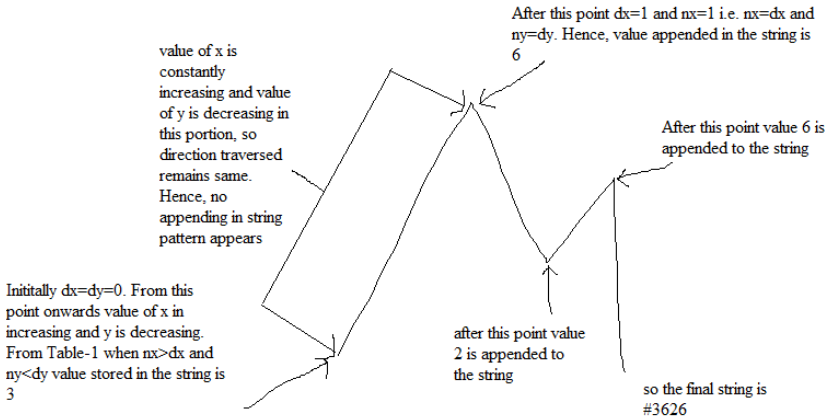


Fig. 5. Showing the process of converting gesture into pattern string

4 Operation

4.1 Initial Setup

This would require an application which will prompt the user to enter the gestures on a canvas using a mouse, which could be easily developed. These gestures

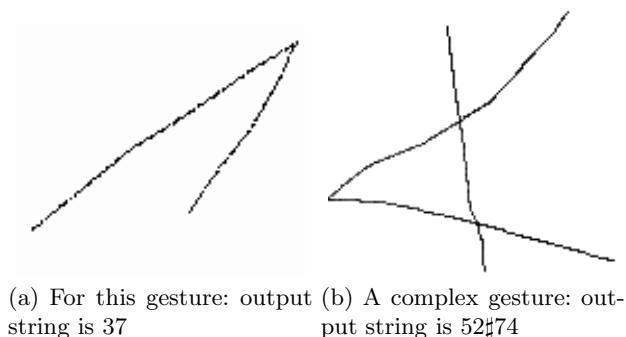


Fig. 6. Figure showing some typical output strings of gestures

must be easily reproducible by the user and must be hidden from other users of the system, since we are going to ask for the same gestures at each login time. The entry of gestures may involve several mouse strokes, which is perfectly valid from the view of our algorithm. The gestures drawn by a user will produce a sequence of mouse coordinates which will be analyzed by the algorithm described. This algorithm will produce a string that will help us to compare the gestures entered by user. To ensure that user will be able to produce them easily in future, we must ask for redrawing the gestures several times at this level and then compare the strings returned by analysis and check whether they are same or not. If they are same, this string must be used in some way to determine a password for the user and that password must be passed to the actual authentication system, otherwise user must be prompted to enter a new password.

4.2 Login System

We would also need an application which will ask for gestures at the time of login[9,10]. Upon entering the gestures, the generated mouse coordinates will be submitted to the algorithm (as explained) for analysis which will produce a string. This string must be similar to the one produced by the analysis of original gestures at the time of initial setup. Thus, it will be used in a similar manner to determine the password as it was done at setup time. Finally, the generated password is fed to the conventional login system, which may be accepted or rejected as the case may be. In case of rejection, the user will be prompted to re-draw the gesture. Due to obvious security reasons, the string produced by this algorithm should not be directly used as password. This string must be coded using any standard cryptographic algorithm available to generate another string that is more secure to be used as a password. An example could be to use this string as a key to a hash-table of random string passwords to get a more secure password for the authentication system.

5 Conclusion

Previous works in this area of research deviate from prevailing traditional text password systems. Later, efforts were made to move towards much more secure and reliable bio-metric authentication systems but these require additional sophisticated hardware to be incorporated into older machines. To overcome the limitations of current systems, we have proposed an algorithm that uses mouse gesture recognition for user authentication. The system uses relative coding techniques which consider any change in the direction of strokes relative to current direction. It eliminates the problem of pattern recognition approach using graphs which shows that large variations can occur due to small variations in gesture drawn by the user thus proving not of much use as pattern matching technique. It also overcomes complexity in vector chain coding approach which stores each and every variation in direction relative to every pixel (x and y co-ordinates) of gesture thus forming a highly complex and large string which forces very high accuracy level of gestures required for authentication. It stores the current and previous direction of co-ordinates of gesture (n_x , d_x and n_y , d_y), computes the current direction based on conditions given in table 1. If there is a change in either the direction along x-axis or y-axis (i.e. change in direction occurs) the pattern string is appended with the new value. This way each and every stroke of gesture is analyzed and corresponding pattern string is generated. This pattern string can be stored in any encrypted form. The application will prompt for gestures from users for user authentication. Users are required to draw the same gesture (or any other gesture producing same relative direction changes) they have previously stored. The entered gesture will be converted into pattern string and will be matched against pattern string which is already stored in system. If the two strings match the user is authenticated otherwise a message is sent to re-draw the gesture.

6 Future Work

The proposed method for quick login can be extended to cover many areas such as authentication system in archive files. This system can be used as authentication system in archive files in contrast to traditional password system thus providing user much secure and quick authentication method. This application can also be associated with archiving tools providing mouse gesture authentication option. Another major concern is the issue of integrating with all major operating systems. The current technique requires proper integration with user authentication modules of operating systems which poses a very major challenge as different operating system have different way of processing user authentication. An important area of use of this algorithm might be in using digital signatures for user authentication. Instead of using simple mouse gesture as user authentication passwords, much more complex digital signatures can be created using mouse gestures which require enhancing the current proposed method to cover minute details of signatures and can suitably process it and perform efficient comparisons.

Acknowledgments

This work is part of a WOS-A project(ref. no. : SR/WOS-A/ET-20/2008) funded by the Department of Science and Technology, Government of India.

References

1. Freeman, W.T., Roth, M.: Orientation histograms for hand gesture recognition. In: Intl. Workshop on Automatic Face- and Gesture- Recognition, pp. 296–301. IEEE Computer Society, Zurich (1995)
2. Hong, P., Turk, M., Huang, T.S.: Gesture Modelling and Recognition Using Finite State Machines. In: Proceedings of the IEEE conference on Face and Gesture Recognition (March 2000)
3. Lee, H.-K., Kim, J.-H.: A HMM Based threshold model approach for gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(10), 961–973 (1999)
4. Shilbayeh, N.F., Raho, G., Alkhateeb, M.: An efficient Structural Mouse Gesture approach for recognizing Hindi digits. Journal of Applied Sciences (2009)
5. Erickson, K.: Different methods of interpreting mouse gestures, www.cs.mtu.edu/~rpastel/Research/Code/KyleEricksen/gestures.pdf
6. Shilbayeh, N.F., Iskadaran, M.Z.: An intelligent multilingual mouse gesture recognition system. Journal of Computer Science 1(3), 346–350, doi:10.3844/jcssp.2005.346.350
7. Gafurov, D., Snekkenes, E.: Arm Swing as a Weak Biometric for Unobtrusive User Authentication. Intelligent Information Hiding and Multimedia Signal Processing, 1080–1087 (2008)
8. Conklin, A., Dietrich, G., Walz, D.: Password-based authentication: a system perspective. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences, p. 10 (2004)
9. Gandossi, A.J., Liu, W., Tjahyadi, R.: A Biometric Approach to Linux Login Access Control. In: 9th International Conference on Control, Automation, Robotics and Vision, ICARCV 2006, December 5-8 (2006)
10. Harn, L., Lin, H.-Y.: Integration of user authentication and access control” Computers and Digital Techniques. IEE Proceedings 139(2), 139–144 (1992)
11. Hofman, P.: Selected Issues of Artificial Intelligence in the Construction of User Interface to a CASE System. MSc Thesis, Wroclaw University of Technology (2005)
12. Newman, W., Lamming, M.: Interactive System Design. Addison-Wesley, Publisher (1995)

Intrusion Detection by Pipelined Approach Using Conditional Random Fields and Optimization Using SVM

R. Jayaprakash¹ and V. Uma²

¹ PG Scholar

² Assistant Professor, Department of Computer Science,
School of Engineering and Technology,
Pondicherry University, Puducherry, India
jpl6586@gmail.com, umabskr@gmail.com

Abstract. The rapid increase in network traffic and attacks made the Intrusion Detection Systems to fail in terms of accuracy and efficiency in many situations. In this paper we have proposed an approach for Intrusion Detection by Pipelined approach using Conditional Random Fields and Optimization using Support Vector Machine. The main goal of this approach in Intrusion Detection System is to achieve high accuracy and efficiency. The accuracy is maintained through the Pipelined approach and Conditional Random Fields and the efficiency is achieved through SVM. The proposed Intrusion Detection System can be used to build a network Intrusion Detection System which can detect a wide variety of attacks reliably and efficiently when compared to the traditional network intrusion detection systems. Another advantage of our system is that it is very general and is easily customizable depending upon the specific requirements of individual networks.

Keywords: Intrusion Detection System, IDS, Conditional Random Fields, CRF, SVM, Machine Learning.

1 Introduction

Intrusion of a computing system is an attempt to break into or misuse it. An intrusion is any kind of action that compromises the integrity, confidentiality and availability of some information or computer resource. Using the weakness or flaws in the system architecture, the intruder intrudes to circumvent the authentication or authorization process. With the tremendous growth of network based services and secured information on networks, network security is becoming more and more important than ever before. One solution to this is the use of Network Intrusion Detection System (NIDS) that detect attacks by observing various network activities. So it is more important that such systems should be more accurate in identifying attacks, quick to train and to generate as few false positives as possible. An Intrusion Detection System (IDS) identifies malicious anomalies and helps protect a network. Thus, IDS have become a necessary component of computer networks. Two

requirements for IDS are Responsiveness and Effectiveness [11]. Security is the sum of all measures taken to prevent any kind of loss. The important function of IDS is to provide a view of unusual activity and then raise an alarm/alert notifying the network administrators and/or block a suspected connection. In addition, IDS should also be capable of distinguishing between attacks produced internally (coming from own employees or customers or any other) inside the organization and external ones (attacks posted by hackers).

The common types of Intrusion Detection Systems (IDS) are Network based (Network IDS) and Host based (HIDS) [8]. In Network based IDS, it attempts to identify unauthorized, illicit and anomalous behaviour based solely on network traffic. Whereas, in HIDS, it attempts to identify unauthorized, illicit and anomalous behaviour on a specific device. In this paper we propose a new approach in Network based Intrusion Detection System. Most of the Intrusion Detection Systems use one of the two detection techniques: Anomaly based detection technique or Signature based detection technique. Anomaly based IDS identifies the anomalies from “normal” behavior and detects any deviation from it. It learns from the normal data collected when there are no anomalous activities. Whereas in Signature based IDS, it monitors the network traffic and it examines for preconfigured and predetermined attack patterns. *A Murali M Rao et. al.* [7], presented a survey on intrusion detection approaches, in which he lists the various techniques followed for implementing IDS. The techniques are: Data Mining techniques, Statistical Models, Neural Networks, State Transition Analysis, Genetic Algorithm, Immune System Approach, Expert Systems, File System Checking, Pattern Matching, Protocol Analysis and Keystroke Monitoring.

Our first approach is to develop hybrid IDS based on conditional random fields. The advantage of developing a hybrid IDS is that it can detect a wide variety of attacks with very few false alarms. The objectives of this paper are to develop an IDS which has broad attack detection coverage and which is not specific in detecting only the previous known attacks and to develop the system which can operate more efficiently even in high speed network traffic.

The rest of this paper is organized as follows: In Section 2, we discuss about the various methods and techniques followed in intrusion detection. In Section 3, we discuss about the proposed system for intrusion detection. In Section 4, we discuss about the experimental setup for IDS using conditional random fields.

2 Related Works

The concept of Intrusion detection is not a new one, it has existed for decades in the field of defence, research and early-warning systems. However Intrusion Detection System rose popularly in the public domain in 1980s. Since then many approaches, methods and frameworks have been proposed and implemented. Various classification methods used are Decision trees, neural networks, naïve Bayes method, Support Vector Machines, clustering (k-means, fuzzy c means and others), regression, statistical tests such as T^2 test and X^2 Multivariate test and many more. Various sequence labelling methods are Markov Chains, Hidden Markov Models, Bayesian Even Classification and many more.

Kapil Kumar Gupta et. al.[1] proposed a Layered based approach using Conditional Random fields (CRF) for Intrusion Detection. The approach addressed

two issues: Accuracy and Efficiency and was experimented using the 41 features represented in KDD Data set [10] in a Layered manner. The goal of using a layered model was to reduce computation time and the overall time required to detect anomalous activity. The system could detect most of the attacks by giving very few false alarms at each layer. Additionally it has been proved that their system is robust to noise and performs better even when the training data is noisy.

Jun Wang et. Al. [2] proposed an ABC-SVM model in a real time Intrusion Detection Systems. In this work it has been proved that Support Vector Machines (SVM) provides potential solutions for the IDSs problem. The authors [2] conducted a series of experiments on KDD Cup (1999) intrusion detection dataset [10] to examine the effectiveness of the proposed feature selection algorithm for building effective IDS.

You Chen et. al. [3] developed a lightweight Intrusion Detection System through modified Random Mutation Hill Climbing (RMHC) and Support Vector Machines (SVM). In this work it has been examined and proved that the feature selection on KDD'99 Intrusion Detection dataset [10] speeds up the process of Intrusion Detection and it also guarantees high detection rates.

Khaja Mohammad Shazzad et.al. [4] proposed an optimized Intrusion Detection through Fast hybrid feature selection technique. The authors [4] experimentally proved that all features are not relevant and some of them are redundant and useless. So the features were optimized using a fusion of correlation-based feature selection, Support Vector Machines and Genetic algorithms. The results of the developed model with optimized feature set had reduction of training and testing time with the good detection rate.

Charles Sutton et. al. [5] addressed on Dynamic Conditional Random Fields which are often used to represent complex interaction between labels. In this paper [5], the dynamic CRF was introduced which were conditionally trained undirected sequence model with repeated graphical structure and with tied parameters.

Kapil Kumar Gupta et. al. [6] introduced the conditional random fields technique in Intrusion Detection. It was experimentally proved that the CRF can be very effective in detecting intrusions when compared with all other known techniques. A Hybrid IDS using CRF technique was developed. Experimental results proved that CRF could be used for IDS which can outperform the existing techniques.

Motivated by the results of [1] [2] and [6], we proposed the approach for Intrusion Detection using CRF and optimization using SVM. The features common to n-layers are classified using SVM. Then a separate model is framed with Conditional Random Fields using the features classified by SVM and with the features selected specific to each layer. We can achieve better efficiency when compared to the layered approach [1] through this optimization.

3 Description of Our System

The figure 1 shows the architecture for Intrusion Detection System by pipelined approach using Conditional Random Fields and optimization using Support Vector Machine (SVM). The goal of using this approach is to reduce computation and the overall time required to detect anomalous events. Based on KDD dataset [10], attacks here are classified as 4 types such as Probe attack, DoS attack, R2L attack and U2R attack. A

separate layer is designed for each attack type. The layers essentially act as filters that block any anomalous connection. All the 41 features in the KDD dataset have not been considered for each of the four attack groups separately. In the training layer the features which are used commonly are identified and are classified using SVM. Each layer is specialised in detecting certain types of attacks. So accuracy is maintained and it reduces the false alarm rate. The classified features using SVM makes the system efficient too. The overall goal of the system is to maintain accuracy and efficiency.

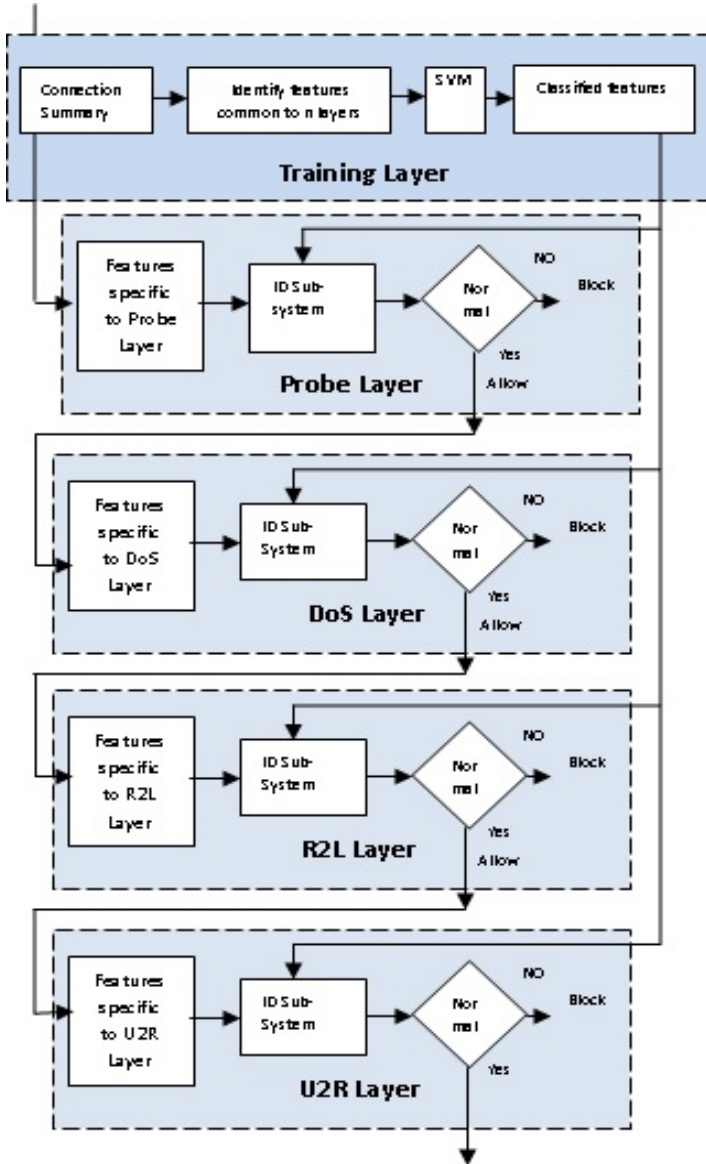


Fig. 1. Proposed IDS architecture by Pipelined approach using CRF and optimization using SVM

Algorithm for the Proposed IDS

The algorithm is based on CRF and Layered approach [1] with the extensions required for the proposed system.

Training:

Step 1: Select the 'n' layers needed for the whole IDS.

Step 2: Identify features common to n layers.

Step 3: Classify the common features using SVM

Step 4: Perform feature selection specific to each layer

Step 5: Train the each layer using the classified features selected from Step 3 and features specific to each layer from Step 4.

Step 6: Integrate the trained models sequentially so that only the connections which are labeled as normal are passed to the next layer.

Testing:

Step 7: For each (next) test instance perform Steps 8 to 11.

Step 8: Check whether the label is either labeled as normal or attack.

Step 9: If it is labeled as attack, block it and identify it as an attack represented by the layer name at which it is detected and go to Step 7. Else pass the sequence to the next layer.

Step 10: If the current layer is not the last layer in the system, test the instance and go to Step 9. Else go to Step 11.

Step 11: Test the instance and label it either as normal or as an attack. If the instance is labeled as an attack, block it and identify it as an attack corresponding to the layer name

4 Experiments

The KDD Data

The KDD Cup 99 Dataset [10] has been used to conduct the experiments. The KDD dataset is the data set used for the third International Knowledge Discovery and Data Mining Tools Competition. It has 41 features extracted from the DARPA Off-line Intrusion Detection Evaluation [10].

Attacks are grouped into four major categories. They are (1) Probe (2) Denial of Service (DoS) (3) Remote to Local (R2L) and (4) User to Root (U2R).

Probe: In this category, attacks are generated in the motive of collecting information for a possible intrusion

DoS: Denial of Service (DoS) attacks breaks the normal operation by making the target host or a server to crash.

R2L: Remote to Local (R2L), in this category, from remote host an attacker can execute commands by breaking the normal authentication which is known as Remote to Local.

U2R: User to Root (U2R), in this category of attacks, an attacker disguises himself as a root user in a network by obtaining authorized users login details which is known as User to Root.

Table 1. Features used in KDD Dataset

Feature Number	Feature Name	Description	Type
1	Duration	Duration of the connection (in seconds)	Continuous
2	Protocol type	Type of the connection protocol (e.g. tcp, udp)	Discrete
3	Service	Destination service(e.g. telnet, ftp)	Discrete
4	Flag	Status flag of the connection	Discrete
5	Source bytes	Number of bytes sent form source to destination	Continuous
6	Destination bytes	Number of bytes sent from destination to source	Continuous
7	Land	1 if connection is from/to the same host/port; 0 otherwise	Discrete
8	Wrong fragment	Number of wrong fragments	Continuous
9	Urgent	Number of urgent packets	Continuous
10	Hot	Number of "hot" indicators	Continuous
11	Failed logins	Failed logins	Continuous
12	Logged in	1 if successfully logged in; 0 otherwise	Discrete
13	Num Compromised	Number of "compromised" Conditions	Continuous
14	Root shell	1 if root shell is obtained; 0 otherwise	Continuous
15	Su attempted	1 if "su root" command attempted 0 otherwise	Continuous
16	Num Root	Number of "root" accesses	Continuous
17	Num File creations	Number of file creation operations	Continuous
18	Num Shells	Number of shell prompts	Continuous
19	Num Access files	Number of operation on access control files	Continuous
20	Num Outbound cmds	Number of outbound commands in an ftp session	Continuous
21	Is hot login	1 if the login belongs to the "hot" list; 0 otherwise	Discrete
22	Is guest login	1 if the login is a guest login 0 Otherwise	Discrete
23	Count	Number of connections to the same host as the current connection in the past two seconds	Continuous
24	Srv count	Number of connection to the same service as the current connection in past two seconds	Continuous
25	Serror rate	Percentage of connection that have "SYN" error	Continuous
26	Srv error rate	Percentage of connection that have "SYN" error	Continuous
27	Error rate	Percentage of connection that have "REJ" error	Continuous
28	Srv error rate	Percentage of connection that have "REJ" error	Continuous
29	Same srv rate	Percentage of connection to the same service	Continuous
30	Diff srv rate	Percentage of connection to different service	Continuous
31	Srv diff host rate	Percentage of connection to host	Continuous
32	Dst host count	Count of connection having same dest hot	Continuous
33	Dst host srv count	Count of connection having the same destination host and using same service	Continuous
34	Dst host same srv Rate	Percentage of connection having the same destination host and using same service	Continuous
35	Dst host diff srv Rate	Percentage of different service on the current host	Continuous
36	Dst host same src port rate	Percentage of connection to the current hot having same src port	Continuous
37	Dst host srv diff	Percentage of connection to the same service coming form different host	Continuous
38	Dst host serror rate	Percentage of connection to the current host that have an S0 error	Continuous
39	Dst host srv serror Rate	Percentage of connection to the current host and specified service that have an S0 error	Continuous

Table 1. (continued)

40	Dst host error rate	Percentage of connection to the current host that have an RST error	Continuous
41	Dst host srv error rate	Percentage of connection to the current host and specified service that have an RST error	Continuous

An important fact is that test data [10] is not from the same probability distribution as the training data, and it includes specific attack types not in the training data. This is somehow close to the real situation where the novel intrusion happens. The whole data set is so huge so that we use the 10% subsets has been used to create a smaller training and test set.

Table 2. Attack Types and Number of samples in 10% KDD Dataset

Types of Attacks	Number of Samples
Normal	Normal(97277)
Probe	Satan(1589), Ipsweep(1247), Portsweep(1040), Nmap(231)
DoS	Smurf(280790), Neptune(107201), Back(2203), Teardrop(979), Pod(264), Land(21)
R2L	Warezclient(1020), Guess_passwd(53), Warezmaster(20), Imap(12), ftp_write(8), Multihop(7), Phf(4), Spy(2)
U2R	Buffer_overflow(30), Rootkit(10), loadmodule(9), perl(3)

As represented in the figure 1, each layer is capable for detecting a specific class of attack groups. The table2 shows the features used in each layers of our proposed architecture.

Table 3. Features Used in Each Layer

Layer type	Features Selected (Feature number and name)
Common	1,2,3,4,5,23,33,34: duration, protocol_type, service, flag, src_bytes, Count, dst_host_srv_count, dst_host_same_srv_rate
Probe	1,2,3,4,5: duration, protocol_type, service, flag, src_bytes
DoS	1,2,4,5,23,34,38,39: duration, protocol_type, flag, src_bytes, count, dst_host_same_srv_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_error_rate
R2L	1,2,3,4,5,10,11,12,13,17,18,19,21,22: duration, protocol_type, service, flag, src_bytes, hot, num_failed_logins, logged_in, num_compromised, num_file_creations, num_shells, num_access_files, is_host_login, is_gust_login
U2R	10,13,14,16,17,18,19,21: hot, num_compromised, root_shell, num_root, num_file_creations, num_shells, num_access_files, is_host_login

Table 4. Parameters Setup for the experiments

Components Used	Configuration/Details
Dataset	10%KDD'99
Programming Tools	LibSVM, JAVA, Weka, Tanagra 1.4.38, MS-Excel 2007
Computer	Intel Core i3 Processor, 4GB RAM, 250 GB Hard Disk
Operating System	Microsoft Windows 7

Evaluation Index

To evaluate the proposed method we use the following evaluation measures as the test standard:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F-Value} = \frac{(1 + \beta^2) * \text{Recall} * \text{Precision}}{\beta^2 * (\text{Recall} + \text{Precision})}$$

Where β is Precision versus Recall and it is usually set to 1.

Here we evaluated the features which are used in the first layer (Training Layer) of the architecture. We used 20,000 records from KDD10% Dataset and evaluated it using Tanagra 1.4.38. The results are as follows:

Table 5. Values Prediction for Training Layer from Tanagra: (for 20,000 recordset)

Value	Recall	1-Precision
$m < 0.50000000$	0.9994	0.0072
$m \geq 0.50000000$	0.8874	0.0102

Table 6. Confusion Matrix for training layer which uses the common features::

	$m < 0.50000$	$m \geq 0.50000$
$m < 0.50000$	18780 (a)	11 (b)
$m \geq 0.50000$	136 (c)	1072 (d)

Accuracy is calculated using the following formula using Confusion matrix [12]

$$\begin{aligned} \text{Accuracy} &= \frac{a + d}{a + b + c + d} \\ &= 99.26\% \end{aligned}$$

7. Murali, A., Rao, M.: A Survey on Intrusion Detection Approaches. In: Proceedings of First International Conference on Information and Communication Technologies, pp. 233–240. IEEE xplora, Los Alamitos (2005)
8. Ruiz, I.P., del Mar Fernández, M., de Ramón.: An Evaluation of current IDS. Master thesis performed in Information Coding, Linköping (February 7 2008)
9. SANS Institute - Intrusion Detection FAQ,
<http://www.sans.org/resources/idfaq/> (last accessed April 30, 2011)
10. KDD Cup 1999 Intrusion Detection Data,
<http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>
(last accessed April 30, 2011)
11. Han, C.-K., Choi, H.-K.: Effective Discovery of Attacks Using Entropy of Packet Dynamics. The IEEE Network Journal (September/October 2009)
12. Confusion matrix,
http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html (last accessed: April 30, 2011)

A Flow-Level Taxonomy and Prevalence of Brute Force Attacks

Jan Vykopal

Masaryk University
Institute of Computer Science
Botanická 68a, 602 00 Brno
Czech Republic
vykopal@ics.muni.cz

Abstract. Online brute force and dictionary attacks against network services and web applications are ubiquitous. We present their taxonomy from the perspective of network flows. This contributes to clear evaluation of detection methods and provides better understanding of the brute force attacks within the research community. Next, we utilize the formal definitions of attacks in a long-term analysis of SSH traffic from 10 gigabit university network. The results shows that flow-based intrusion detection may profit from traffic observation of the whole network, particularly it can allow more accurate detection of the majority of brute-force attacks in high-speed networks.

1 Introduction

Intrusion detection based on network traffic inspection traditionally attracts the attention of security researchers. Network-based intrusion detection systems originally processed payload of each packet passing by the monitored link. This is a very resource-intensive task, mainly in nowadays multigigabit networks, and not feasible in case of encrypted protocols. To address this issue, flow-based intrusion detection works with traffic aggregates, network flows, that can be acquired only by the inspection of packet headers. A flow is defined as an unidirectional sequence of IP packets that share common properties (the flow key), e. g. source and destination addresses, ports, and protocol type. Although flows do not carry any information about payload, they are sufficient for the detection of many types of attacks such as denial of service, network scans, worms and botnets [10].

In this paper, we focus on flow-based detection of online [1] brute force and dictionary attacks. Many studies show that these attacks are a common type of attacks [9, 11, 12], which exploits the fact that humans select passwords from a small subset of the full password space (e. g., short passwords, dictionary words, proper names, and lowercase strings) [7]. Recent reports by Dragon Research

¹ Attackers verify whether a password is correct or not only by interacting with the login server.

Group [4], HP TippingPoint DV Labs [3] and our operational experience confirms time persistence of these attacks: we have been observing these attack in our /24 network every day for last few years. Attackers try to break in computer systems that allows remote access to i) abuse compromised system as a *stepping stone*, which can hide they malicious activities, ii) gain unauthorized access to user's data, or iii) infect other systems by their worm or botnet that used the same attack vector for its self-propagation [14].

To the best of our knowledge, there is a lack of taxonomy of the brute force and dictionary attacks from the perspective of network flows. However, rigorous definitions of particular attack types are crucial for correct and clear evaluation of detection methods and systems. So, first of all, we propose a basic definition of the following types of attacks and probes: i) simple brute force attack, ii) multiple brute force attack, iii) distributed brute force attack, iv) network port scan, and v) application scan. Second, we are interested in the prevalence of particular attack classes in a real 10 gigabit network. Finally, what is the contribution of flow-based approach to the attack detection.

The paper is organized as follows. Formal definitions of various brute force attack classes are introduced in Section 2. Measurement setup and results of the flow-based analysis, which is focused on identification of the proposed attack classes in the real network traffic, are described in Section 3. Conclusions and plans for further work are presented in Section 4.

2 A Taxonomy of Brute-Force Attacks and Probes

In this section, we propose a basic flow-level taxonomy of brute force attacks and probes. We start with the *simple brute force attack* that is a foundation for the following definitions of *multiple* and *distributed brute force attacks*. We also distinguish and define two types of probes, the *network port scan* and the *application scan*, because they are often mixed up by both researchers and practitioners.

The following definitions rely on accurate flow monitoring (use number of packets and bytes etc.) so we do not consider the use of packet sampling that would seriously distort the acquired flows.

2.1 Simple Brute Force Attack

The elementary type of the brute force attack is a repetitive interactive communication between single source and single destination providing a particular service via the defined network protocol and destination port. Because the attacker uses the same network protocol and only permutates login credentials in consecutive break-in attempts, the amounts of transferred bytes and packets are *similar*.

We use the notion of a bidirectional flow (biflow in short) [12] to formally define this basic type of the attack. Let $b = (start_fwd, start_rvs, srcIP, dstIP, proto, srcPrt, dstPrt, pkt_fwd, pkt_rvs, byt_fwd, byt_rvs)$ is a biflow, where $start_fwd$ is a timestamp when the request starts, $start_rvs$ a timestamp when

the response starts, *srcIP* is the source IP address, *dstIP* the destination IP address, *proto* used network protocol, *srcPrt* source TCP/UDP port, *dstPrt* destination TCP/UDP port, *pkt_fwd* and *byt_fwd* amount of transferred packets and bytes in the forward direction and *pkt_rvs* and *byt_rvs* amount of packets and bytes in the reverse direction. Then a *simple brute force attack* is a ordered set of biflows *SBA*:

$$\begin{aligned}
 SBA &= \{b_i | i \in [1, n]\} \\
 (\forall b_i, b_j \in SBA) : & (b_i(proto) = b_j(proto) \wedge b_i(dstPrt) = b_j(dstPrt) \wedge \\
 b_i(srcIP) &= b_j(srcIP) \wedge b_i(dstIP) = b_j(dstIP)) \\
 (\forall b_i, b_{i+1} \in SBA) : & (b_i(start_fwd) < b_{i+1}(start_fwd)) \\
 (\forall b_i, b_j \in SBA, b_i \neq b_j) : & (d(b_i, b_j) \leq threshold)
 \end{aligned}$$

srcIP is the attacker, *dstIP* victim, *n* denotes the *cardinality* of the attack, $d(x, y)$ is a distance metric function [13], $T = b_n(start_fwd) - b_1(start_fwd)$, where T is a duration of the attack.

2.2 Multiple Brute Force Attack

This type of the attack is comprised of simple attacks from one source against two or more destinations at the same time. Formally, the *multiple brute force attack MBA* is a set of simple attacks *SBA*:

$$\begin{aligned}
 MBA &= \{SBA_i | i \in [1, n]\} \\
 (\forall SBA_i, SBA_j \in MBA, SBA_i \neq SBA_j) : \\
 (SBA_i(srcIP) &= SBA_j(srcIP) \wedge SBA_i(dstIP) \neq SBA_j(dstIP))
 \end{aligned}$$

2.3 Distributed Brute Force Attack

To make the attack more stealthy or more efficient, two or more sources send authentication requests to a single destination at the same time. We define the *distributed brute force attack DBA* as follows:

$$\begin{aligned}
 DBA &= \{SBA_i | i \in [1, n]\} \\
 (\forall SBA_i, SBA_j \in DBA, SBA_i \neq SBA_j) : \\
 (SBA_i(dstIP) &= SBA_j(dstIP) \wedge SBA_i(srcIP) \neq SBA_j(srcIP))
 \end{aligned}$$

2.4 Network Port Scan

Port scanning is a very popular technique for probing running network services. As opposed to the attacks defined above, such probes need not be answered by target hosts. As a result, we can observe only unidirectional traffic. We define the *network port scan NS* using an unidirectional flow [2] $f = (start, srcIP, dstIP, proto, srcPrt, dstPrt, flags, pkt, byt)$:

$$\begin{aligned}
NS &= \{f_i | i \in [1, n]\} \\
(\forall f_i, f_j \in NS) : & (f_i(dstIP) \neq f_j(dstIP) \wedge f_i(dstPrt) = f_j(dstPrt) \wedge \\
& f_i(protocol) \in \{TCP, UDP\} \wedge f_i(byt)/f_i(pkt) \in [min, max] \wedge \\
& f_i(flags) \in SCAN_TYPES) \\
(\forall f_i, f_{i+1} \in NS) : & (f_i(start) < f_{i+1}(start) \wedge f_{i+1}(start) - f_i(start) \leq t_{diff})
\end{aligned}$$

$f(bytes)/f(packets)$ represents the value of *bytes per packet* that is known for particular scan techniques. *SCAN_TYPES* defines combinations of sent TCP flags used for scanning (it is an empty set for UDP scans) and t_{diff} the maximum time difference between two consecutive flows.

For example, TCP SYN scan is characterized by i) the length of the very first TCP packet that depends on the used operating system (typically from 48 to 64B) and ii) settings of the single TCP flag (SYN). Note, we generally cannot distinguish between an unsuccessful TCP connection establishment and TCP SYN scanning in terms of the definition above.

2.5 Application Scan

Another type of probing is the *application scan*. It is an interactive communication at the application layer (as opposed to the network port scan at the transport layer). The main difference between the network port scan and the application one is that the former is not “visible“ in the application log files whereas the latter is. Attackers may use this scan to confirm that an expected service is virtually up and running at the given network port. At flow level, it can be understood as a special type of the single brute force attack, which is comprised of a small number of biflows. Opposed to the SBA, these application scan biflows are typically formed with distinct amounts of transferred bytes and packets:

$$\begin{aligned}
AS &= \{SBA_i | i \in \langle 1, n \rangle\} \\
n &< small_number
\end{aligned}$$

The requirement of the distinct amount of transferred traffic can be adjusted by the suitable distance metric function of SBA_i .

3 Flow-Based Analysis

To enlighten real attack scenarios, we analyzed network traffic in real 10 gigabit network according to the described definitions. We focused on SSH flows collected at two 10 gigabit uplinks of the Masaryk University network, which directly connects approximately 15 000 hosts to the Internet. We chose SSH because it is constantly very popular service among attackers (as outlined in Introduction).

The main analysis covers the period of 39 days in two sets: the first one lasted from May 10th to June 10th, 2010 and the second one from October 8th to 14th, 2010. We chose data from May and October because they are the most busiest in comparison to other months. The purpose of these two sets is to show if (and eventually how) the behavior and attacks evolves in time. Additionally, we also analyzed the entire time period between May 10th to October 10th with respect to port scanning to answer the question if the actual attacks are preceded by scanning and probing.

3.1 Measurement Setup

The campus 10 gigabit uplinks were monitored by two hardware-accelerated FlowMon probes [6] that monitor the uplinks without any packet loss and export non-sampled NetFlow v9 data to the `nfdump` collector [5]. The timeouts used for NetFlow export were set as follows: the active timeout to 300 seconds and the inactive one to 30 seconds. Note that timeout settings essentially influences exported flows (e.g., if these values are too small, the flows may be exported prematurely and thus some attacks may not fit the definitions).

The acquired NetFlow data were processed in 5-minute time windows in automated way to identify the defined attack classes. We searched for *SBA*s against SSH services that satisfies the following criteria (service running on TCP port 22, time difference between consecutive biflows is up to 3 seconds and transferred amount of data is similar and in the defined interval):

1. $proto = TCP$,
2. $dstPort = 22$,
3. $d(b_i, b_j) < 3$, where $d(b_i, b_j) = |b_i(start_fwd) - b_j(start_fwd)| + byt_dif + pkt_dif$

$$byt_dif = \begin{cases} 0 & |b_i(byt) - b_j(byt)| < 4000 \wedge b_i(byt), b_j(byt) \in [1000, 5000] \\ 1 & \text{otherwise} \end{cases}$$

$$pkt_dif = \begin{cases} 0 & |b_i(pkt) - b_j(pkt)| < 20 \wedge b_i(pkt), b_j(pkt) \in [10, 30] \\ 1 & \text{otherwise} \end{cases}$$

In case of *MBA*s and *DBA*s, we used the same parameters except the time criteria $|b_i(start_fwd) - b_j(start_fwd)|$. We consider the time difference between **all** biflows of the *SBA*s from the attack source (in case of *MBA*s) or to the attack destination (in case of *DBA*s).

Concerning *NS*, we searched for TCP SYN probes: $f(bytes)/f(packet) \in [40, 64]$ B, $SCAN_TYPES = \{SYN\}$. In our experience, the vast majority of scanning probes satisfies this *bytes per packet* rate. Application scans are not analyzed in this work.

Then we manually correlated attacks identified by our definitions to the entries in log files of SSH daemons to obtain the ground truth. We searched for messages

describing a break-in attempt: e.g. *Invalid user webmaster from attacker's_IP* or *Failed password for invalid user root from attacker's_IP port number ssh2*. Unfortunately, we could not directly access all SSH logs in our campus network. Therefore it is sufficient for validation if the attacker of MBAs is found in a log file of one victim.

3.2 Results

We observed in total 20 885 simple brute force attacks (SBA): 16 819 in the first data set and 4 443 in the second one. The vast majority of SBAs, 20 793 (99.55%), form 107 MBAs and 73 SBAs (0.35%) a single DBA. Only 21 SBAs (0.1%) cannot be aggregated to any MBA or DBA. The prevalence of the attack classes is similar in both data set even they are 4 months distant (see Figure 1). To support the flow-level analysis, we inspected log files at all SSH servers that we could access and found 99 attacks², i.e., 76.7% of attacks observed at the flow level (note we could not confirm neither deny remaining 23.3% of attacks). We do not consider the DBA in further text because a single appearance is not statistically relevant and it was not found in logs.

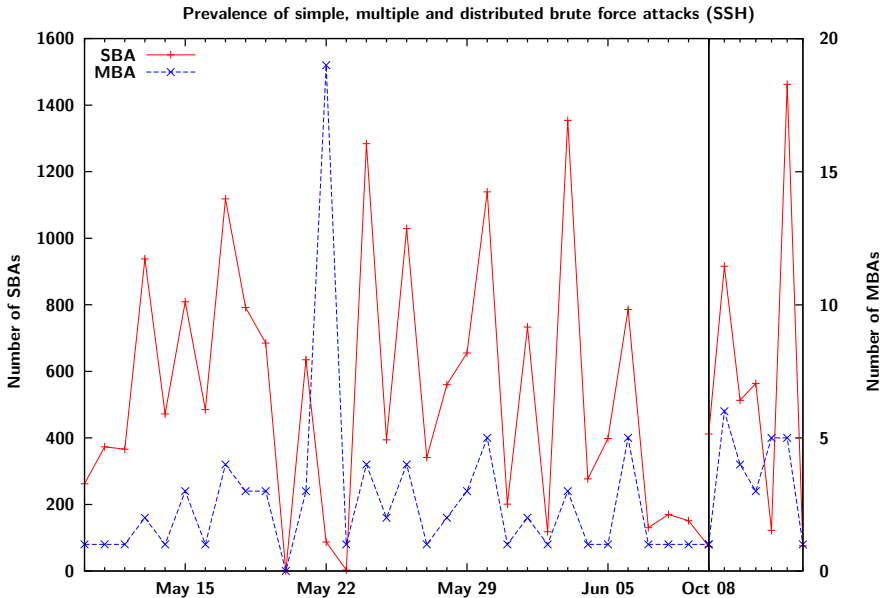


Fig. 1. Daily sums of particular attack classes against SSH in the /16 campus network in two periods in 2010

² MBAs, a DBA and SBAs that do not form any other attack type.

Table 1. Number of unique attackers and attacks

	Total attacks	Unique attackers	Unique attackers [%]
Non-aggregable SBA	21	10	47.6
MBA	107	79	73.8
DBA	1	73	100

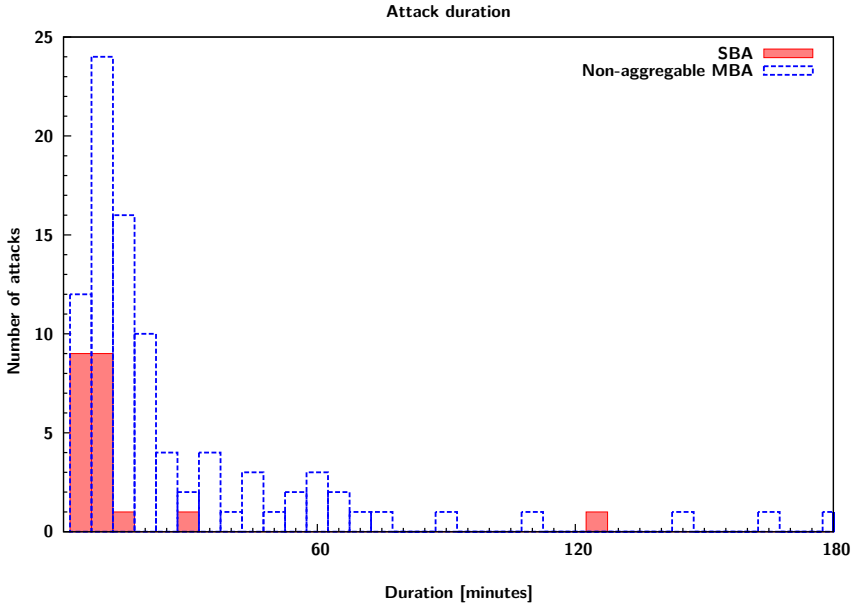


Fig. 2. Distribution of attack durations – detailed view from 5 minutes to 3 hours

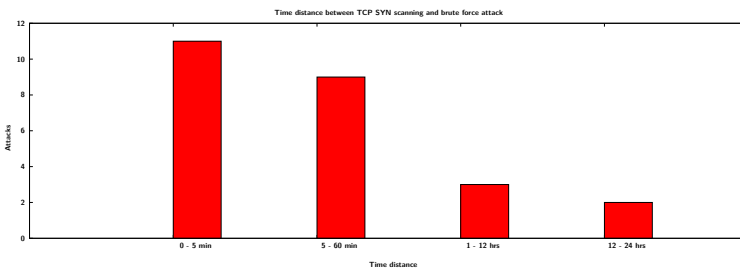


Fig. 3. Number of MBAs that were preceded by NS grouped by the time distance between them

For attack detection purposes, we are interested in attackers behaviour as well as their origin (location of the attacking host). Table 1 shows unique attackers of particular attack classes. The table also says that there is a recidivism of some attackers. Next, attack duration varied, Figure 2 depicts number of attacks

Table 2. Top 10 countries of attackers' origin (location)

Rank	Country	Unique attackers
1	China	29
2	South Korea	10
3	USA	8
4	Russia	5
5	Taiwan	4
6	Poland	4
7	Italy	3
8	Spain	3
9	Romania	2
10	Netherlands	2

Table 3. Top 10 autonomous system numbers of attackers' origin (location)

Rank	ASN	Unique attackers
1	4134	12
2	4837	7
3	9318	4
4	4766	4
5	4808	3
6	5617	2
7	4847	2
8	4538	2
9	42116	2
10	31334	2

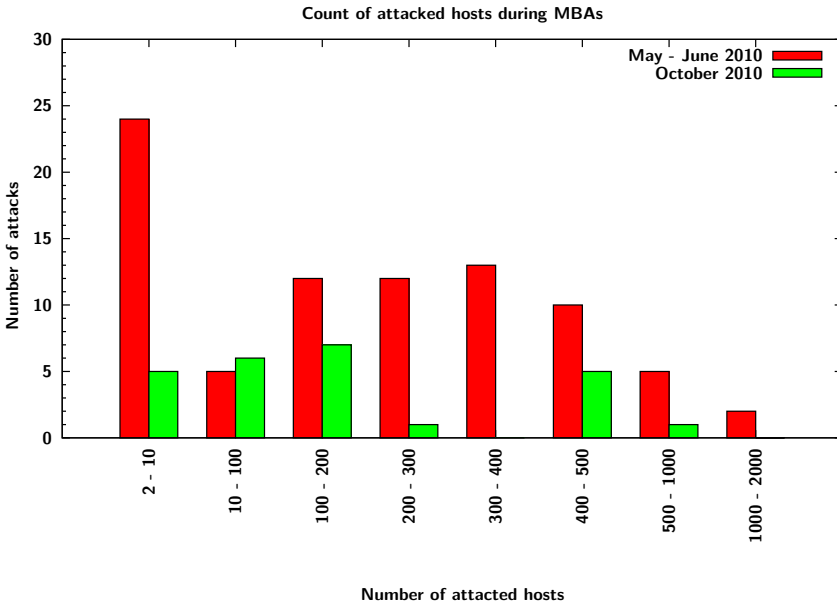


Fig. 4. Count of hosts attacked in one MBA

lasting from 5 minutes to 3 hours; the longest attack lasts 14 hours and 50 minutes (the distribution seems to be long-tailed).

Another important question concerning attackers' behavior is *Are brute force attacks preceded by network port scanning?* We consider attacks from the second data set (7 days in October 2010) and search for the closest preceding TCP SYN port scans (NS, as defined above) from the attacker's IP address in time period from May 10th, 2010 to October 2010. All 25 MBAs attacks were preceded by NS that occurred from 0 to 1360 minutes (22 hrs 40 mins) before the MBAs; 11 NS were spot up to 5 minutes before the attacks (see Figure 3).

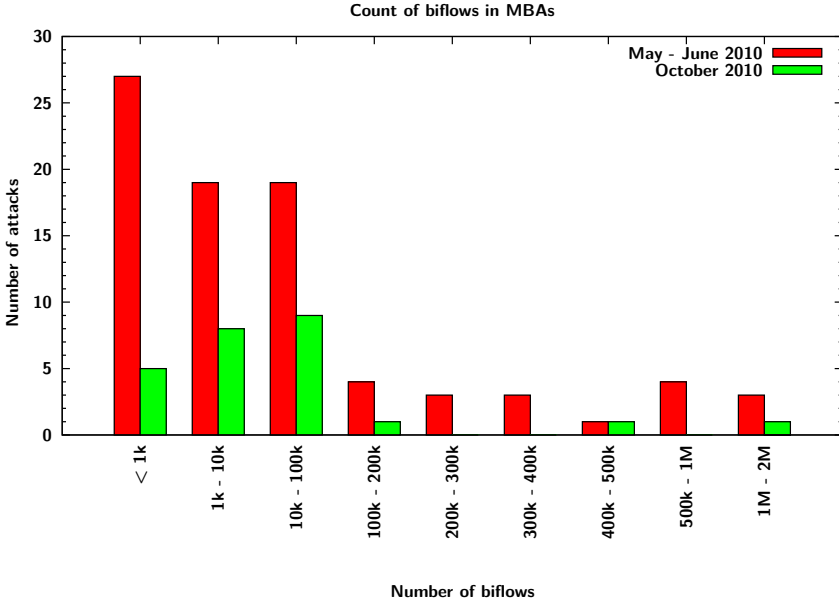


Fig. 5. Count of biflows forming one MBA

Next, we queried WHOIS database to obtain attackers' origin (autonomous system and country code) and resolved reverse DNS for attackers' IP addresses. Attackers came from 24 different countries and 58 autonomous systems. Table 2 depicts top 10 countries of origin and Table 3 top 10 autonomous systems. The vast majority of attackers conducting MBAs (63 attackers, 79.7%) did not have set any DNS reverse record, but in case of attackers of non-aggregable SBAs we found only two empty records (out of ten, 20%).

High occurrence of MBAs in both data sets opens other crucial questions relevant to the network- and flow-based detection: How many hosts are attacked? What is the cardinality of attacks? We observed 1 850 617 biflows that form 107 MBAs against 4 271 hosts. The average number attacked hosts in one MBA was 219 using 107 882 biflows, while the median was 151 hosts using 7 780 biflows. Histograms are shown in Figure 4 and 5.

4 Conclusions and Future Work

We propose the first flow-level taxonomy of brute force attacks (SBA, MBA and DBA) and probes (NS and AS) based on both biflows and unidirectional flows. The proposed attack definitions contain several parameters so they can be utilized in various types of validations of brute force attack detection.

Next, the analysis of brute force attacks against SSH servers based on the long-term measurement of a 10 gigabit university network is presented. We

analyzed non-sampled NetFlow data of SSH traffic in the time period covering 39 days. Results show very interesting phenomena: i) the overwhelming majority of biflows formed multiple brute force attacks and ii) network port scanning always(!) preceded the actual attacks. These findings can influence design of detection techniques, particularly flow-based detection of brute force attacks. In contrast to the host-based detection, flow-based one can capture more precisely multiple attacks.

The proposed taxonomy can be further developed to reflect various aspects of attacks, e. g., similarly to [8]. Another direction for future work is to undertake similar extensive measurement for other network protocols such as RDP, FTP or HTTP.

References

1. Alata, E., Nicomette, V., Kaaniche, M., Dacier, M., Herrb, M.: Lessons learned from the deployment of a high-interaction honeypot. In: EDCC 2006: Proceedings of the Sixth European Dependable Computing Conference, pp. 39–46. IEEE Computer Society Press, Washington, DC, USA (2006)
2. Claise, B.: Cisco Systems NetFlow Services Export Version 9. RFC 3954 (Informational) (October 2004)
3. Hewlett-Packard Development Company. Top Cyber Security Risks Threat Report for (2010), <http://dvlabs.tippingpoint.com/toprisks2010>
4. Dragon Research Group. sshpwauth report (2010), <http://www.dragonresearchgroup.org/insight/sshpwauth.txt>
5. Haag, P.: NFDUMP - NetFlow processing tools (2009), <http://nfdump.sourceforge.net/>
6. INVEA-TECH. Standard FlowMon Probe (2009), <http://www.invea-tech.com/>
7. Menezes, A.J., Van Oorschot, P.C., Vanstone, S.A., Rivest, R.L.: Identification and Entity Authentication. In: Handbook of Applied Cryptography. CRC Press, Boca Raton (1997)
8. Mirkovic, J., Reiher, P.: A Taxonomy of DDoS Attack and DDoS Defense Mechanisms. SIGCOMM Comput. Commun. Rev. 34(2), 39–53 (2004)
9. C. Seifert. Analyzing Malicious SSH Login Attempts (2006), <http://www.securityfocus.com/infocus/1876> (retrieved online January 3, 2010)
10. Sperotto, A., Schaffrath, G., Sadre, R., Morariu, C., Pras, A., Stiller, B.: An Overview of IP Flow-Based Intrusion Detection. Communications Surveys Tutorials 12(3), 343–356 (2010)
11. Thames, J.L., Abler, R., Keeling, D.: A Distributed Active Response Architecture for Preventing SSH Dictionary Attacks. In: IEEE Southeastcon 2008, pp. 84–89 (2008)
12. Trammell, B., Boschi, E.: Bidirectional Flow Export Using IP Flow Information Export (IPFIX). RFC 5103 (Proposed Standard) (January 2008)
13. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search - The Metric Space Approach, vol. 32. Springer, Heidelberg (2006)
14. Čeleda, P., Krejčí, R.: Embedded Malware An Analysis of the Chuck Norris Botnet. In: To appear in European Conference on Computer Network Defense (EC2ND) (October 2010)

Multi Application User Profiling for Masquerade Attack Detection

Hamed Saljooghinejad and Wilson Naik Rathore

Department of Computer and Information Science,
University of Hyderabad, Hyderabad, India
hamed.saljooghinejad@gmail.com,
naikcs@uohyd.ernet.in

Abstract. Masquerade attack or Impersonation attack refers to an act of illegitimate user gaining unauthorized privileges of the system. Detecting these attacks is more complex due to the fact that the insiders carry out most of these attacks. Masquerade attack is detected by profiling users system usage. If his/her normal profile deviates from his/her original behavior, he is detected as a masquerader. Most of the research was done using command line data & GUI Usage analysis. The command line data which contains commands, logs, system calls and the GUI profiling using keyboard and mouse activities, can not capture the complete event behavior of the users, Due to the reason that users are not fixed to a single application in their usage period. Hence it is very difficult to detect masquerader in the existing systems. In this paper we have proposed a new framework to capture the data across multiple applications to build the user profile. We have developed our own tool to capture the event data across multiple applications. Our experimental result shows that our framework is better in detection than the existing methods. We have applied four different classifiers, K-Nearest Neighbor, SVM, BayesNet and NaïveBase on the collected user profiles. Our results show that K-NN is the best classifier for the collected Multi application GUI data.

Keywords: Masquerade Detection, Intrusion Detection System, Anomaly Detection, User Profiling.

1 Introduction

According to the 2010 cyber crime watch survey [15], 35% of the surveyed executives and law enforcement officials experienced unauthorized access and use of their information, systems, and networks. Masquerade attack is second in the top five list of electronic crimes perpetrated after virus, worms and other malicious code attacks. When an insider impersonates another person inside the organization most of his actions may be technically legal for the system and hence it is more difficult to detect such violations. Also, the insider has enough knowledge about the system as well as the behavior of the victims so that he can escape detection for a longer period of time. The only information, which can be used to detect masquerade attacks is

contained in the actions a masquerader performs. This set of actions is known as behavioral profile. In the absence of a real-world data set for the study of masquerade attacks, we have developed our own data collection logger tool. Most of the existing profiling techniques are profiling data for single application, but our proposed approach of profiling multi application shows better results due to the fact that users use multiple applications in a particular session, if the system is not designed to log and profile all the activities, the detection rate would be very low and in some cases system may not detect the masquerader. Masquerade detection techniques are based on the premise that when a masquerader attacks the system, he will sufficiently deviate from the users behavior and thus can be recognized by using machine learning techniques .[9][10]

2 Background and Related Work

Masquerade detection was reported by observing the command line data by profiling user command line data and then finding anomalies in his usage. Researchers also experimented using GUI data with keyboard events, mouse events and shortcuts for masquerade detection. Following 2.1 and 2.2 sections, we have mentioned previous masquerade detection techniques.

2.1 Command Line Data for Detecting Masquerade

Most of the command line approaches follow data collection, data preprocessing and user profile construction, user profile updating, size of testing blocks (number of commands). The initial research in this area done by Schonlau et al. [6], which collected a dataset of Unix command line data of 50 users which called SEA dataset. Later The Naive Bayes classifier was first applied on Schonlau's dataset by Roy A. Maxion [7]. Maxion et al. extended their previous work by applying the Naives Bayes classifier on Greenberg's enriched command line data [8]. Wang and Stolfo's work [11] introduced the application of one-class training for masquerader detection.

Some other efforts but not significant have been done by monitoring system calls [21],[22],[23],[24] ; analyzing the audit log [25][26][27][28][29]; program execution traces [30][31][32][33][34][35]; and call stack information[36]. These above approaches either rely on limited set of information or use the data for purposes other than masquerade detection, such as user authentication.

2.2 GUI Based Data for Detecting Masquerade

However command line data detection mechanisms could not truly detect masquerade in the modern graphical user interface(GUI) systems like windows and variant of Unix like Linux or Mac OSx. Nowadays working with GUI systems is more common and studying of different aspects of them is crucial.

GUI base data mostly related to data, which comes from the interaction of users with mouse and keyboard. Some efforts were done for user authentication techniques in the area of keystrokes dynamic[13][14] as well as mouse usage [12]. Poursa and Broadly [12] consider Analysis of mouse data which was taken from users who worked with browsers only. This approach had some disadvantages because they

focused on the browser data, though maybe users work with some application other than browsers.

Researchers tried to collect, comprehensive GUI behavior data for masquerade detection. For this purpose, Garg and Upadhyaya [1] developed an active system logger by using of Microsoft .NET framework and C# language on Windows XP System. GUI event data is captured from users and useful parameters are extracted to construct the feature vectors. This profiling method was good but the most important disadvantage of this approach is that, they implemented it only for Microsoft GUI systems with much focus only on mouse usage. Their methodology is not scalable to Unix variants like(Linux, Sun and MacOS) GUI systems nor they considered multi applications. Moreover, their detection rate was not impressive, since they used Two-class SVM approach. Bhukya [2] later designed logger in KDE environment. The disadvantages of their work were that they collected data only from a particular application and did not consider the complexity of profiling multi applications. They did their experiment just with one single application and their framework was for only the KDE in Linux version. The other activity in GUI based area was Imsand work [3] which was based on the notion of how the current user interacts with the graphical user interface. This method does not use mouse movements or keystroke dynamics, rather profiles how the user manipulates the windows, icons, menus, and pointers that comprise a graphical user interface. The use of time factor is not stated clearly in their work. They do not appear to consider time as a factor nor the application details, which is crucial for intrusion analysis.

3 Proposed Method

We have proposed a new approach for detecting masquerader across multiple applications. In our approach we have captured the user behavior from all the applications which the users frequently use . Existing methods [1][2] have only captured the data from a single application. The advantages of our approach are that it can profile all the events across applications rather than collecting events from single application. User usually uses different applications like Firefox, Office, Desktop Explorer etc. and his/her behaviors are different with respect to each application. So, in the existing methods the system can not be able to capture all the events across the applications. In our approach we have designed an event logger tool, which collects all the events at the application level. In this approach all the events across the applications are profiled and used for masquerade detection. For example one of the unique features is the number of user switchings across multiple applications, this unique feature can give the expertise of the user.

4 Experimental Setup

We have designed the logger tool to collect all the event data for the users across multiple applications which he/she uses at a particular sessions. In the following section 3.1 and 3.2 we have described the data collection and feature extraction processes.

4.1 Data Collection

For collecting the data we developed our own logger to collect the information in X window systems. Our logger can be run on operating systems which are capable of being run on X Window system. The collected event details include id of the particular window, application name of that window, time of occurrence along with different attributes of that particular window.

We collected three sorts of data regarding all applications which are run by user:

1)Window Data

Data regarding user interactions with a particular window is captured in this part. Users in the GUI environment try to switch between windows, Maximizing, Minimizing, opening, closing and etc. The following is a samples of window data:

Event Occured at: Tue Feb 8 10:03:13 2011

WID=65011715--WName=Nautilus--msg:The active window changed from previous WID=69206091—WName=Firefox

2)Mouse Data

Mouse data is also important and meaningful for detecting masquerader. because it carries considerable communication data between user and applications. In this category mouse-related user activities like mouse click, mouse right click, mouse movement to the menu bar area and etc, were captured for every application. Following is an example of mouse data events:

Event Occured at: Tue Feb 8 09:58:22 2011

WID=65011737--WName=Firefox--msg:Mouse left button clicked

3)Keyboard Data

All the keyboard events are logged and stored separately for each application. Different keyboard events are key pressed, key released, number of time the key is pressed, number of time the shortcut key pressed (Ctrl, Alt, shift modifier) in a particular session. The following is an example of keyboard data event:

Event Occured at: Tue Feb 8 10:00:33 2011

WID=69206091--WName=Firefox--msg:Shortcut Pressed—Ctrl+z

4.2 Feature Extraction and Preprocessing

After all the events are captured, we have preprocessed, parsed and extracted only meaningful and unique features of the users.

The extracted features are divided into three categories:

1)Window features : The features, which were generated when the user interacts with a particular window. Totally 9 features were extracted.

2)Mouse Data : Total 8 features were extracted from the mouse events of different applications.

3)Keyboard Data : 5 features are extracted from the keyboard events of different applications.

Following section gives all the feature details:

4.3 Calculation of Features

22 features were generated as follow:

1) window Data (9 features)

1.1)Window Coordination (2) : The average number of changes in x and y coordinates of window, per user session.

1.2)Window Size (2) : The average number of changes in width and height of window, per user session.

1.3)Window Maximize,Minimize (2) : The average number of times that user minimizes or maximizes the window, per user session.

1.4)Window Restore Maximizing,Minimizing (2) : The average number of times that user restores the minimized window or restores from maximized window, per user session.

1.5)Window Switching (1) : The average number of times that a user switches between windows, per user session.

2) Mouse Data (8 features)

2.1)Mouse Enter and Exit (2) : The average number of mouse entrances and exits into and from each window, with respect to each application per user session.

2.2)Mouse Movement (2) : The average number of entrances to menu bar and working area, with respect to each application per user session.

2.3)Mouse Clicks (2) : The average number of left and right mouse clicks with respect to each application per user session.

2.4)Mouse Scroll Up, Down(2) : The average number of mouse scroll up and scroll down with respect to each application per user session.

3) Keyboard Data (5 features)

3.1)Key Pressed (1) : The average number of keys pressed per user session with respect to each application.

3.2)Shortcut key Pressed(1): The average number of shortcut keys pressed with respect to each application per user session.

3.3)Ctrl, Alt, Shift Modifier (3): The average number of Ctrl, Alt and Shift modifier pressed with respect to each application per user session.

4.4 Learning and Classification Methods

We have experimented the extracted features with multiple classifiers to choose the best classifier with high detection rate. We have chosen SVM[16][17], K-Nearest Neighbor[18], BayesNet and NaïveBase classifiers [19] . The Collected feature vectors were divided into different training and testing sessions for each and every user. Training sessions were used for learning and test sessions of other users were used as masquerade records for that particular user.

5 Results and Discussions

For experimentation we have collected data from 3 users across multiple applications. The collected data contain multiple sessions 9, 19, 32 and each session for 10 minutes. This data was fed to the parsing engine mentioned above to sanitize and extract 22 features for each application in every session. The methodology to train and test the data was as following:

- Datasets were obtained for three distinct users, A(32 sessions), B(19 sessions), C(9 sessions) .
- Data divided into training and test sets as following:

Table 1. Data Sets

<i>User</i>	<i>Training Sessions</i>	<i>Test Sessions</i>	<i>Total Sessions</i>
<i>A</i>	<i>18</i>	<i>10</i>	<i>32</i>
<i>B</i>	<i>10</i>	<i>9</i>	<i>19</i>
<i>C</i>	<i>6</i>	<i>3</i>	<i>9</i>

- The training and test sets are given as inputs to the classifiers {K-NN, SVM, BayesNet, NaïveBayes} for classification . We have used Weka tool [20] to perform the classification.

- Following table gives the classification accuracies of different classifiers:

Table 2. Detection Rate For user A,B and C(A-B,C : A as normal and B,C as masquerade records)

Model	K-NN		SVM		BayesNet		NaiveBase	
	DR	FP	DR	FP	DR	FP	DR	FP
A - B,C	91.70%	6.60%	91.20%	10.30%	93.17%	10.10%	88.29%	12.50%
B – A,C	91.66%	8.30%	88.24%	11.80%	87.75%	12.30%	86.27%	13.70%
C – B,A	91.25%	13.00%	86.89%	18.30%	83.06%	54.20%	50.82%	16.00%

As we can see K-NN gives better detection rate with lower false positives.

5.1 Detection Evaluation by Different Number of Training and Test Sets

In order to show a better view, we performed the previous experiment with different number of training sets and test sets for user A. As it shows, the detection rate of K-NN is better than the other three classifiers. The K-NN detection rate is being better with more number of training sets.

Table 3. Detection Rate For user A with Different Training and Test sets Size(Tr:training set;Te:Test set)

Classifier	5 Tr 27 Te	8 Tr 24 Te	10 Tr 22 Te	12Tr 20Te	15 Tr 17 Te	17 Tr 15 Te	20Tr 12Te	22 Tr 10 Te	24 Tr 8 Te	27 Tr 5 Te
K-NN	85.2	88.7	91.4	92.6	89.3	91.7	96.2	97.4	97.3	98.8
SVM	79.5	81.9	83.1	83.6	90.5	91.2	87.6	88.7	83.8	83.6
BayesNet	79.8	80.9	87.9	87.9	92.9	93.1	89	87.7	87.1	85.5
NaiveBase	79.8	81.9	72.7	78.9	82.1	88.3	84.7	83.1	83.9	85.5

The following graph shows the result which achieved from Table 3:

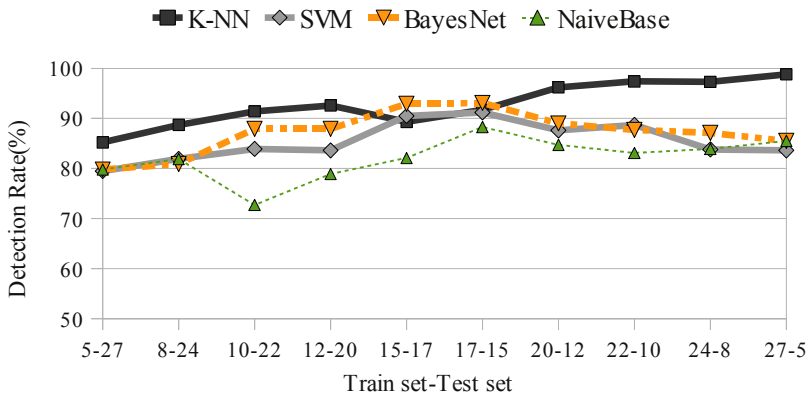


Fig. 1. Detection Rate For user A with Different Training and Test sets Size

6 Conclusion and Future Work

We have proposed a new framework to capture Multi application based GUI profiling which works across multiple Operating Systems. After capturing the events we have processed and extracted relevant features of each application and constructed feature vectors. These feature vectors which classified for masquerade detection using multiple classifiers namely K-NN, SVM, Naïve Bayes and BayesNet. Our experiments show that GUI multi application profiling is well classified with less false positives using K-NN classifier. We also observed that K-NN improves detection rate with increased number of training sets. In future we want to work with high number of users across multiple layers and test the classification accuracies.

References

1. Garg, A., Rahalkar, R., Upadhyaya, S.: Kevin Kwait: Profiling Users in GUI Based Systems for Masquerade Detection. In: Proceedings of 7th Annual IEEE Information Assurance Workshop (IAW 2006), June 21-23, United States Military Academy, West Point (2006)
2. Bhukya, W.N., Kommuru, S.K., Negi, A.: Masquerade Detection Based Upon GUI User Profiling in Linux Systems. In: Cervesato, I. (ed.) ASIAN 2007. LNCS, vol. 4846, pp. 228–239. Springer, Heidelberg (2007)
3. Imsand, E.S., Hamilton Jr., J.A.: GUI Usage Analysis for Masquerade Detection. In: Proceedings of 2007 IEEE, Information Assurance Workshop (IAW 2007), June 21-23, United States Military Academy, West Point (2007)
4. Li, L.: Manikopoulos.: Windows NT One-class Masquerade Detection. In: Proceedings of 2004 IEEE, Information Assurance Workshop (IAW 2004), June 2004, United States Military Academy, West Point (2004)
5. Pusara, M., Brodley, C.: User Re-authentication via mouse movements. In: Proceedings of the 2004 ACM workshop on visualization and data mining for computer security, Washington D.C., USA, October 29 (2004)
6. Schonlau, M., DuMouchel, W., Ju, W.-H., Karr, A.F., Vardi, M.T.: Computer Intrusion: Detecting Masquerades. *Statistical Science* 16, 58–74 (2001)
7. Maxion, R.A., Townsend, T.N.: Masquerade Detection Using Truncated Command Lines. In: Proceedings of International Conference on Dependable Systems and Networks (DSN 2002), pp. 219–228 (2002)
8. Maxion, R.A.: Masquerade Detection Using Enriched Command Lines. In: Proceedings of International Conference on Dependable Systems and Networks (DSN 2003), San Francisco, CA (June 2003)
9. Lane, T., Brodley, C.E.: An Application of Machine Learning to Anomaly Detection. In: Proceedings of Twentieth National Information Systems Security Conference, Gaithersburgh, MD, vol. 1, pp. 366–380 (1997)
10. Lane, T., Brodley, C.: Sequence Matching and Learning in Anomaly Detection for Computer Security. In: Proceedings of AAAI 1997 Workshop on AI Approaches to Fraud Detection and Risk Management, pp. 43–49 (1997)
11. Wang, K., Stolfo, S.J.: One Class Training for Masquerade Detection. In: ICDM Workshop on Data Mining for Computer Security, DMSEC 2003 (2003)
12. Pusara, M., Brodley, C.: User Re-authentication via mouse movements. In: Proceedings of the 2004 ACM workshop on visualization and data mining for computer security, Washington D.C., USA, October 29 (2004)
13. Monroe, F., Rubin, A.: Authentication via Keystroke Dynamics. In: ACM Conference on Computer and Communications Security, pp. 48–56 (1997)
14. Shavlik, J., Shavlik, M., Fahland, M.: Evaluating Software Sensors for Actively Profiling Windows 2000 Computer Users. In: Lee, W., Mé, L., Wespi, A. (eds.) RAID 2001. LNCS, vol. 2212, Springer, Heidelberg (2001)
15. CERT. 2010 e-crimes watch survey (2010)
16. Platt, J.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Schlkopf, B., Burges, C., Smola, A. (eds.) *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge (1998)

17. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO Algorithm for SVM Classifier Design. Technical Report CD-99-14. Control Division, Dept of Mechanical and Production Engineering, National University of Singapore (1999)
18. Aha, D., Kibler, D.: Instance-based learning algorithms. *Machine Learning* 6, 37–66 (1991)
19. John, G.H., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338–345. Morgan Kaufmann, San Mateo (1995)
20. <http://www.cs.waikato.ac.nz/ml/weka/>
21. Hofmeyr, S., Forrest, S., Somayaji, A.: Intrusion Detection Using Sequences of System Calls. *Journal of Computer Security* 6(3), 151–180 (1998)
22. Lee, W., Stolfo, S., Mok, K.: A Data Mining Framework for Building Intrusion Detection Models. In: *IEEE Symposium on Security and Privacy*, pp. 120–132 (1999)
23. Forrest, S., Hofmeyr, S.A., Somayaji, A.: Computer Immunology. *Communications of the ACM* 40(10), 88–96 (1997)
24. Warrender, C., Forrest, S., Pearlmuter, B.: Detecting Intrusions using System Calls: Alternative Data Models. In: *IEEE Symposium on Security and Privacy (Oakland, CA)*, pp. 133–145 (1999)
25. Ilgun, K., Kemmerer, R., Porras, P.: State Transition Analysis: A Rule-Based Intrusion Detection Approach. *Software Engineering* 21(3), 181–199 (1995)
26. Li, Y., Wu, N., Jajodia, S., Wang, S.: Enhancing Profiles for Anomaly Detection Using Time Granularities. *Journal of Computer Security* 10(1,2), 137–157 (2002)
27. Javitz, H.S., Valdes, A.: The SRI IDES Statistical Anomaly Detector. In: *Proceedings of the IEEE Research in Security and Privacy (Oakland, CA)*, pp. 316–376 (May 1991)
28. Wespi, A., Dacier, M., Debar, H.: Intrusion detection using variable-length audit trail patterns. In: Debar, H., Mé, L., Wu, S.F. (eds.) *RAID 2000*. LNCS, vol. 1907, p. 110. Springer, Heidelberg (2000)
29. Ye, N.: A Markov Chain Model of Temporal Behavior for Anomaly Detection. In: *Proceedings of the 2000 IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop*, pp. 171–174 (2000)
30. Ghosh, A., Schwartzbard, Schatz, M.: Learning Program Behavior Profiles for Intrusion Detection. *First USENIX Workshop on Intrusion Detection and Network Monitoring*, 51–62 (1999)
31. Levitt, K., Ko, C., Fink, G.: Automated Detection of Vulnerabilities in Privileged Programs by Execution Monitoring. In: *Computer Security Application Conference* (1994)
32. Marceau, C.: Characterizing the behavior of a program using multiple-length N-grams. In: *Proceedings of the 2000 workshop on New security Paradigms, Ballycotton, County Cork, Ireland*, pp. 101–110 (2000)
33. Michael, C., Ghosh, A.: Using Finite Automata to Mine Execution Data for Intrusion Detection: A Preliminary Report. In: Debar, H., Mé, L., Wu, S.F. (eds.) *RAID 2000*. LNCS, vol. 1907, p. 66. Springer, Heidelberg (2000)
34. Wagner, D., Dean, D.: Intrusion Detection via Static Analysis. In: *IEEE Symposium on Security and Privacy*, pp. 156–169 (2001)
35. Rajagopalan, M., Debray, S., Hiltunen, M., Schlichting, R.: Profile-directed Optimization of Event-based Programs. In: *Proceedings of ACM SIGPLAN* (2002)
36. Feng, H., Kolesnikov, O., Fogla, P., Lee, W., Gong, W.: Anomaly Detection using Call Stack Information. In: *Proceedings of IEEE Symposium on Security and Privacy, Oakland, California* (May 2003)

A Novel Technique for Defeating Virtual Keyboards - Exploiting Insecure Features of Modern Browsers

Tanusha S. Nadkarni, Radhesh Mohandas, and Alwyn R. Pais

Information Security Research Lab,
Department of Computer Science and Engineering,
National Institute of Technology Karnataka,
Surathkal, India
{tanushanadkarni, radhesh, alwyn.pais}@gmail.com

Abstract. Advancement in technology is a necessity of time, but as new techniques are introduced, new security vulnerabilities are discovered and exploited in practice. In this paper we are presenting a new approach to defeat virtual keyboards using a new method for capturing parts of a browser screen. The page rendered in the browser is captured by using the canvas element provided by HTML5. We have specified the technical details of how this functionality is exploited and created a malicious extension for Mozilla Firefox browser. This extension captures screenshots of web pages rendered in the browser and sends them to a remote server. In addition, we have suggested mitigation strategies to prevent misuse of such browser functionalities.

Keywords: Virtual keyboards, screen capture, Mozilla Firefox browser, Extension, malware, HTML5, canvas, keylogger.

1 Introduction

A keylogger is a program designed to secretly monitor and log user keystrokes. It can be used to intercept passwords and other confidential information entered using a keyboard. To avert stealing of passwords using a keylogger, many websites provide a virtual keyboard. Virtual keyboard is an on-screen keyboard for entering sensitive login credentials by clicking on the keys appearing on the screen using a mouse. Though using virtual keyboards raises the bar, methods to get around them have been known right from the time they were introduced. Using screencasting software, the screen can be recorded, and password entered using virtual keyboard can be stolen and even be sent to a remote server.

The methods which are commonly used for performing screen capture are Graphic Device Interface (GDI), DirectX and Windows Media API [7]. These methods use Windows APIs to capture what is being displayed on the screen.

Security software such as Trusteer Rapport [8], SnoopFree Privacy Shield [9] can block screen capture when the above methods are used. Hence we present a new way of capturing virtual keyboard using HTML5 canvas and its Javascript methods, which cannot be prevented by these anti-screen capture software products.

This paper is organized as follows. Section 2 briefs up on related work, where we have mentioned about the malware created in the past for the similar purpose. Section 3 throws light on the functionality and implementation details of the extension we made for Firefox browser. In Section 4 we have listed the results of testing our extension against Indian Banking sites. We have also explained anti-screen capture tools and their failure to prevent browser content capture. Section 5 illustrates possible mitigation strategies. Finally we conclude the paper in Section 6.

2 Related Work

Defeating virtual keyboards by capturing 10 x 10 pixels around mouse clicks has been known since 1997 and is extensively used by several advanced keyloggers and malware [5]. One well known worm using this technique was W32/Dumaru that was an attack against the e-Gold keypad [6].

Some malware capture the screen only on the first click, and simply record the coordinates of the mouse click on all subsequent clicks. But this kind of simple password stealing can be avoided by using a mutating virtual keyboard where the key placement changes randomly every click.

Hispasec labs describes a banking trojan X, which captures user's screen in a video clip [1]. The attacker then receives a video clip of the victim's screen and can then repeat all the steps followed by the user to gain access to his banking account. The video clip covers only a small portion of the screen, using the cursor as the reference, but is large enough for the attacker to watch the legitimate user's movements and typing when using the virtual keyboard. The Trojan monitors the browser windows and their titles passively, until the user visits any of the banking institutions it monitors. This Trojan monitors several Brazilian banking institutions and uses two Microsoft Windows standard libraries (msvfw.dll and avifil32.dll) to perform the video capture. The Trojan uses the standard methods exported from the library gdi32.dll included in the Microsoft Windows graphic engine.

Hispacec labs illustrates details of another Trojan Y, which takes screenshots per click [2]. Every time the user clicks the virtual keyboard, the Trojan performs a series of small screen captures of a small area that surrounds the cursor. It adds a small red arrow that pinpoints the exact place the user clicked, so that the attacker can see clearly the key the user selected. It has been specifically designed for bank institutions in Argentina, Bolivia, Brazil, Cape Verde, Spain, USA, Paraguay, Portugal, Uruguay, and Venezuela.

3 Our Screen Capture Application and Its Implementation

We have created an extension for Mozilla Firefox to illustrate this new technique for defeating virtual keyboard using screen capture. Similar extensions can be built for Internet Explorer, Google Chrome and other popular browsers. The attacker needs to install the extension on the victim's computer using one of the standard backdoor techniques. The extension is customized to defeat the virtual keyboards of popular banks operating in India. It silently monitors the hostname and patterns in the URL for a set of target websites. When the victim activates the virtual keyboard on one of these sites, the visible section of the webpage is captured. A series of screenshots captured on consecutive mouse clicks is used to trace the entered password as shown in Figure 1. The data URL (Base64 encoded string) of the captured image is then posted to a remote server, which can be decoded back into an image. In the following sections we have provided the programmatic details of our extension.



Fig. 1. The sequence of images showing the clicked keys reveal the password

3.1 HTML5 Canvas Element

HTML 5 defines the `<canvas>` element as "a resolution-dependent bitmap canvas which can be used for rendering graphs, game graphics, or other visual images on the fly." [10]. A canvas is a rectangle in your page where graphics can be drawn using JavaScript. `<canvas>` creates a fixed size drawing surface that exposes one or more rendering contexts. `<canvas>` was first introduced by Apple for the Mac OS X Dashboard and later implemented in Safari. Gecko 1.8-based browsers, such as Firefox 1.5, also support this element. The `<canvas>` element is a part of the WhatWG Web applications 1.0 specification also known as HTML 5 [11]. The XPCOM interface *nsIDOMHTMLCanvasElement* is the interface to a HTML `<canvas>` element [16].

The `<canvas>` is initially blank, and to display something a script first needs to access the rendering context and draw on it. The canvas element has a DOM method called *getContext()* used to obtain the rendering context and its drawing functions [12]. The *drawWindow()* method draws a snapshot of the contents of a DOM window into the canvas [13]. This is the main function which is used for saving webpage as an image and is a method of the XPCOM interface *nsIDOMCanvasRenderingContext2D* [17]. The *context.toDataURL()* method returns a data: URL containing a representation of the image in the specified format, which is .PNG format by default [14]. This function is used to get a data: URL that has the Base-64 encoded image.

3.2 Sending Images to Remote Server

The data: URL returned by the method *toDataURL()* can be used for various purposes. It can be used for converting the canvas to an image file, and saved on the local machine. This string can be used to display image on a webpage or it can be converted to a file using PHP function *createimagefromstring()*. We sent this Base64 encoded string over the Internet to a remote server using asynchronous XMLHttpRequest. The strings are stored on the remote system and can be converted back into images to obtain the password. Hence when a user installs this extension and uses virtual keyboard on banking site, his login credentials can be obtained in the form of screenshots.

4 Test and Results

We tested our extension on Mozilla Firefox 3.6.13 for 25 Indian Banks. We found that there were 4 variations of Virtual Keyboard used by these banks.

1. Basic non-mutating virtual keyboards
2. Advanced non-mutating virtual keyboards - In these the key changes to * on mouse click and then changes back to the same key. We reason that the programmers thought that this was a security feature so that a password stealer would only record * if it just recording the pixels on mouse clicks. On closer analysis we say that the key changes to * on mousedown event, and then changes back on mouse up event. So this hurdle could be easily gotten around by listening to the appropriate events.
3. Mutating keyboards, in which the placement of keys changes on every click i.e. a new key is present at the same position after selecting a key by clicking it.
4. Advanced Mutating keyboards which incorporates features of both type 2 and 3.

Our extension could record the correct key sequence for all these types of virtual keyboards, the details of which have been presented in Table 1. We analyzed their Javascript source code and could easily program around their event handlers for shuffling the keys.

Table 1. Number of banks having each type of Virtual Keyboard

Number of Banks	Type 1	Type 2	Type 3	Type 4	Virtual Keyboard Defeated
25	19	3	2	1	25

4.1 Config Options

Once installed this addon is not visible anywhere in the browser GUI. i.e. it is hidden from the Tools>Addons list and no option to access the addon is present on menu list/status bar thus making it stealthy. By manipulating the rdf datasource that the Extension Manager depends upon, we could hide extension from the list [15]. The GUI can be accessed only by using Ctrl+Shift+V shortcut, which will be known only to the extension author.

1. Configurable pixel size - We could configure pixel size for our extension and found that with an area size of 30x30 pixels we could capture the virtual keyboard keys for all the banks.
2. The stolen data is posted to a free hosting site. Utilizing the Base64 encoding feature, only text data is posted, and no images are displayed. This means that nothing will be visible.
3. We then import the data posted on the free hosting website into our local machine, and convert the strings back into images. The password table is as shown in Figure 2. We tested with passwords consisting of uppercase letters, lowercase letters, digits and special symbols.

No.	IP Address	Date & Time	Webpage Title	URL	Password
1	117.211.83.4	2011-03-07 19.59.19	Welcome to Canara Bank Internet Banking	https://netbanking.canarabank.in/netbanking/RetailLogin.html	
2	117.211.83.4	2011-03-07 20.02.00	No Title	https://inet.idbibank.co.in/corp/BANKAWAY?Action.RetUser.Init.001=Y	
3	117.211.83.4	2011-03-07 20.05.07	ING Vysya Internet Banking	https://online.ingvysyabank.com/auth/jsp/authindex.jsp	
4	112.110.89.188	2011-03-07 21.53.47	Internet Banking Retail	https://www.kvnet.co.in/retail/entry?	
5	112.110.89.188	2011-03-07 21.59.10	State Bank of India	https://www.onlinesbi.com/retail/login.htm	

Fig. 2. The password table

4.2 Testing against Trusteer Rapport and SnoopFree Privacy Shield

Rapport is web security software developed by Trusteer, a company that provides safe communication between business websites and customers. Rapport is a lightweight browser security plug-in. It protects a user’s browsing sessions while visiting specific websites such as e-commerce and banking websites. When visiting any protected site, Rapport blocks any attempt to take control of the session by malware, which includes keylogging and screen capture, session hijacking, and DNS redirection hijacks. Rapport prevents taking screen shots while you are connected to protected websites and uses API blocking to prevent this type of behaviour, alerting users if any such activities are attempted.

We tested our extension against Rapport, but it was not able to prevent screen capture for any of the sites protected by it. Rapport prevents screen capture by blocking windows APIs, and our screen capture method performs DOM to image conversion using Javascript.

SnoopFree Privacy Shield is a security guard that watches for programs that try to invade privacy. If any program tries to access potentially sensitive information, SnoopFree Privacy Shield stops the offending program and asks the user how to handle. Whenever a screen capture application tries to capture screen, SnoopFree generates a warning message where the user can either allow or deny access to the application. If the user denies the application then the screen capture application will not be able to capture the screen, and in most cases have a blank screen recording. However SnoopFree did not generate any warnings when our extension was executed.

This proves that the existing prevention mechanisms are not aware of these novel methods, which specifically capture browser content.

4.3 Testing against Indian Banks

We programmed our extension to filter URLs and capture screenshots only for login pages of 25 Indian bank sites. For every bank against which we tested our extension, we collected password, IP address of the system which was accessing the bank site, and date on which the login page was being accessed. These informative details along with the Base64 Encoded string of captured image are posted on a site hosted on a remote server. We saved all these details on our server in order to demonstrate how passwords can be collected by taking advantage of this new browser vulnerability.

5 Mitigation Strategies

As we have demonstrated with our extension, a hacker with a little bit of programming experience can write a password stealer that logs keys and screenshots using the rich set of features provided by the modern browsers and their programming tools. But to protect the end user, all the related software products need to inculcate secure computing strategies in them.

1. Currently we see that most anti-malware and anti-virus products check only executables for malicious activity or signatures. Browser plugins doing the same kind of activity run under the browser space which is a trusted platform for them to launch all kinds of attacks. Hence these security products should be upgraded to spot malicious plugins using both signatures and behavior analysis.
2. The procedure of uploading browser addons in open source communities like Mozilla is pretty lax. We suggest that a thorough code review process and signing process be put in place for the extensions for Mozilla Firefox before releasing them to the general public. Unless the security reviewers go through each and every line of extension source code and understand the functionality, it may not possible to recognize malicious activity. Since extensions are given same privileges as a browser, hackers can effortlessly produce such malicious extensions and distribute them by piggybacking to other extensions promising some functionality.

3. We propose an idea of a sandbox for untrusted browser extensions. In Firefox, extensions have the same privileges as the browser. Extensions can access the file system, edit the registry, connect to the Internet, and have access to password and bookmarks managers. Therefore, the main aim of this solution is to prevent users being fooled by extensions which promise some functionality and silently perform malicious activity. The extension sandbox will allow user to grant privileges per extension. This will prevent extensions from silently carrying out malevolent actions other than what they claim to do.

6 Conclusion

In this paper we have presented a new method for defeating virtual keyboards by capturing browser content as screenshots. This new mechanism does not use Windows screen capture APIs and cannot be prevented by any of the available prevention software products. After conducting tests on 25 Indian banks, we have found that all of them are vulnerable to such a kind of screen capture attack, and none of them have employed proper anti-screen capture solutions. We demonstrate how hackers may get hold of passwords using malicious browser extensions which can be easily created using HTML5, Javascript methods and XPCOM interfaces available. Hence this paper basically intends to add new a dimension to the work of security researchers working on preventing password stealing through key stroke capture.

References

1. Banking Trojan Captures User's Screen in Video Clip, Hispasec/VirusTotal (September 05, 2006), <http://www.hispasec.com/laboratorio/banking-trojan-capture-video-clip.pdf>
2. New technique against virtual keyboards, Hispasec/VirusTotal, Hispasec / VirusTotal (September 26, 2006), <http://www.hispasec.com/laboratorio/New-technique-against-virtual-keyboards.pdf>
3. Debasis Mohanty: Defeating Virtual Keyboard Protection, <http://www.coffeeandsecurity.com/resources/papers/defeat-vk.pdf>
4. Cracking On-Screen Keyboards with Visual Keyloggers, <http://mrooney.blogspot.com/2009/02/cracking-on-screen-keyboards-with.html>
5. Virtual Keyboard and the Fight Against Keyloggers, <http://palisade.plynt.com/issues/2009Feb/fight-against-keyloggers/>
6. W32/Dumaru, <http://www.mcafee.com/threat-intelligence/malware/default.aspx?id=100980>
7. Screenshot, <http://en.wikipedia.org/wiki/Screenshot>
8. Trusteer Rapport, <http://www.trusteer.com/product/trusteer-rapport>

9. SnoopFree Privacy Shield, <http://www.snoopfree.com/>
10. HTML5 Canvas, <http://diveintohtml5.org/canvas.html>
11. Canvas Tutorial, https://developer.mozilla.org/en/canvas_tutorial
12. Basic Usage of Canvas Element, https://developer.mozilla.org/en/Canvas_tutorial/Basic_usage
13. Drawing Graphics with Canvas, https://developer.mozilla.org/en/drawing_graphics_with_canvas
14. HTMLCanvasElement, <https://developer.mozilla.org/en/DOM/HTMLCanvasElement>
15. OnHacks Firefox Malware Tutorial, <http://onhacks.org/lang/en/2009/02/11/firefox-malware-tutorial-1/>
16. nsIDOMHTMLCanvasElement interface, <http://www.oxymoronical.com/experiments/xpcomref/applications/Firefox/3.5/interfaces/nsIDOMHTMLCanvasElement>
17. nsIDOMCanvasRenderingContext2D, https://developer.mozilla.org/en/XPCOM_Interface_Reference/nsIDOMCanvasRenderingContext2D

SQL Injection Disclosure Using BLAH Algorithm

Justy Jameson* and K.K. Sherly**

ToCH Institute of science and technology, Arakkunnam, Kerala, 682313, India
justyjameson@gmail.com, shrly_shilu@yahoo.com

Abstract. Data security has become a topic of primary discussion for security expert. Vulnerabilities are pervasive resulting in exposure of organization and firm to a wide array of risk. In recent years, a large number of software systems are being ported towards the Web, and platforms providing new kinds of services over the Internet are becoming more and more popular: e-health, e-commerce, and e-government. Code injection attack is a major concern for web security. This paper is a new approach for detecting SQL injection. We use static analysis based SQL injection detection technique. In which we identify a hot spot and BLAHE algorithm for SQL injection disclosure. The BLAHE algorithm approach has high accuracy. At the same time, the processing speed is fast enough to enable online detection of SQL injection.

Keywords: SQL injection, BLAH algorithm, Security.

1 Introduction

Today's modern web era, expects the organization to concentrate more on web application security. This is the major challenge faced by all the organization to protect their precious data against malicious access or corruptions. An input validation issue is a security issue if an attacker finds that an application makes unfounded assumptions about the type, length, format, or range of input data. The attacker can then supply a malicious input that compromises an application. The cross site scripting attacks, SQL Injections [9] attacks and Buffer Overflow are the major threat in the web application security through this input validation security issues.

The most worrying aspect of SQL Injection attack is; it is very easy to perform, even if the developers of the application are well known about this type of attacks. The basic idea behind in this attack is that the malicious user counterfeits the data that a web application sends to the database aiming at the modification of SQL Query that will be executed by DBMS software. Input validation issues can allow the attackers to gain complete access to such a database. Researchers have proposed a different technique to provide a solution for SQLIAs (SQL Injection attacks), but many of these solutions have limitations that affect their effectiveness and practicality. Researchers have indicated that solution to these types of attacks may be based on defense coding practices. But it's not efficient because of three reasons. First, it is very

* Justy Jameson is doing m-tech in ToCH institute of science and technology.

** Sherly K K is an associate professor in ToCH institute of science and technology.

hard to bring out a rigorous defensive coding discipline. Second, many solutions based on defensive coding address only a subset of the possible attacks. Third, legacy software poses a particularly difficult problem because of the cost and complexity of retrofitting existing code so that it is compliant with defensive coding practices. In this work, an attempt has been made to increase the efficiency of the above techniques by the empirical method for protecting web application against SQL Injection attacks.

The remainder of the paper is organized as follows: Section 2 contains the basic idea; Section 3 describes related works and section 4 contains BLAHE(BLAH) algorithm.

2 Basic Idea

The simplest SQL injection technique is bypassing form-based logins. The web application code in PHP is like this: `$query=" select *from users where user_name='$stdname' and password='$password'"; $result=mysql_query ($query);`. Here's what happen when a user submits a username and password. The query will go the users table to see if there is a row where the username and password in the row match those supplied by the user. If such a row is found, the `$result` must be true otherwise false. `$query=" select*from users where username='or1=1'and password='1=1'". $result=mysql_query ($query);`. This query returns true when `user_name` and `password` are not on the table. Therefore the attacker can login into the web application.

Many researchers proposed various solutions to the SQL injection attacks. Mainly four types are present. The following section describes about the solution.

2.1 Tainted Data Tracking

The main idea of this run-time protection mechanism is to "taint" and track the data that comes from user input. This can be done via instrumenting the runtime environment of web application or interpreter of the back-end scripting language, e.g., PHP.

2.2 Static Analysis Based Intrusion Detection

Hotspots, i.e., Locations in the back-end program that submits the SQL statements, can be easily identified by examining the source code or byte code of a web application. Using static string analysis technique [1], it is possible to construct a regular expression that conservatively approximates the set of SQL statements generated at a hotspot.

2.3 Black-Box Testing

By collecting a library of attack patterns, Y. W. Huang et al applied black-box testing of web applications. The defect of the approach is that without prior knowledge of source code, it is not as effective as white-box testing to discover non-trivial attacks.

2.4 SQL Randomization

SQL randomization is basically an extension of the instruction randomization technique to defend against code injection attacks. The key idea is to instrument a web application and append a random integer number after each SQL keyword in the constant string fragments that are used to dynamically build SQL statements. In addition, the SQL parser employed by the web application is rewritten to accept the randomized SQL keywords and reject SQL keywords that are not appended with randomized integer numbers. At run-time, if a user tries to inject SQL code in data input, the injected SQL code will be rejected due to a syntax error.

3 Related Works

Xiang Fu et al [1], propose the design of a static analysis framework (SAFELI) for identifying SQL Injection attack vulnerabilities at compile time. SAFELI statically monitoring the Microsoft Symbolic intermediate language byte code of an ASP.NET Web application, using symbolic execution. The main limitation of Xiang et al's work is that this approach can discover the SQL injection attacks only on Microsoft based product.

Buehrer et al [12], propose the mechanism which filters the SQL Injection in a static manner. The SQL statements by comparing the parse tree of a SQL statement before and after input and only allowing to SQL statements to execute if the parse trees match. They conducted a study using one real world web application and applied their SQLGUARD solution for each application.

R. Ezumalai proposes a system against SQLIA is based on signature based approach, which has been used to address security problems related to input validation. This approach uses Hirschberg algorithm [10] to compare the statement from the specification

Angelo Ciampa et al [4] propose a heuristic based approach for detecting SQL injection vulnerabilities in web application. They propose an approach and a tool named VIp3R for web application penetration testing. This approach is based on pattern matching of error messages.

William G.J et al [2] propose Protecting Web Applications Using Positive Tainting and Syntax-Aware Evaluation. This approach has both conceptual and practical advantages. This paper introduces a tool WASP for SQL injection identification.

Marco Cova et al [14], propose a mechanism to the anomaly-based detection of attacks against web applications. Swaddler analyzes the internal state of a web application and learns the relationships between the application's critical execution points and the application's internal state. By doing this, Swaddler is able to identify attacks that attempt to bring an application in an inconsistent, anomalous state, such as violations of the intended workflow of a web application

Panagiotis Manolios et al [15], proposed the mechanism to keep track of the positive taints and negative taints. This work outlined a new automated technique for preventing SQLIAs based on the novel concept of positive tainting and a flexible syntax aware evaluation. It will check the SQL statement with this taints and if it finds any suspicious activity, it will generate the alarm.

4 BLAHE Algorithm

BLAHE algorithm is used for identifying the SQL injection in the web application. This section explains about the blahe algorithm. The blahe algorithm is the combination of the three concepts that are blast, ssah and equal checker.

BLAST [18] comprises three steps. 1) It compiles a list of high-scoring words (neighboring words) from a given query sequence. 2) each neighboring word is compared with the database sequences. If the neighboring word is identical to a word (sequence fragment) in the database, a hit is recorded. 3) every hit sequence is extended in both directions and the extension is stopped as soon as the similarity score becomes less than a threshold value.

SSAHA [18] is a two-stage algorithm. 1) A hash table is constructed from sequences in the database 2) Query words are searched appropriately from the hash table in the second stage. Let Q be the query sequence and D be the available sequences in the database. Each database sequence $S_i \in D$ is divided into k-tuple. The hash table is stored in memory as two data structures—a list of positions L and an array of pointers A into L. The pointer A [w] point to an entry of L which describes the position of the first occurrence of the k-tuple in the database D. The positions of all occurrences of w in D can be obtained by traversing L. In the second stage, the hash table is used to search for occurrences of a query sequence Q within the database. A list of hits is computed and added to a master list M. M is next sorted, first by index and then by shift. The final searching process is done by scanning through M.

BLAH is a hybridized algorithm which combines the advantages of the BLAST and SSAH. It is a two-stage algorithm. 1) A clustered k-tuple table is created. 2) Find database similarity regions using k-tuple table.

The database D is converted into a k-tuple table (KT) in the first stage. The KT consists of three attributes, namely, Tuple-weight, Sequence-index and Sequence-offset.

Definition 1. K-tuple: Let $S = \langle s_1, s_2, s_3, \dots, s_n \rangle$ be a sequence of length n. Then, any consecutive k elements of this sequence form a k-tuple. Two k-tuples are called overlapping if they share some elements between them.

Definition 2. Sequence-base: Number of distinct elements present in the sequences of the database is called the Sequence-base.

Definition 3. Tuple-offset: The position in a sequence where a k-tuple starts is called a tuple-offset.

Definition 4. Sequence-index: If a database has n number of sequences, then the sequence index of the i th sequence in the database is i.

Definition 5. Tuple-weight: Every distinct k-tuple is assigned a unique integer value which is called Tuple-weight

$$W = \sum_{i=1}^k w^{k-i} * ti \tag{1}$$

The database D is converted into a k-tuple table (KT) in the first stage. The KT consists of three attributes, namely, Tuple-weight, Sequence-index, and Sequence-offset. This table has a cluster index on Tuple-weight. There can be k^k Sequence-base distinct Tuple-weight values in the table. We can obtain the positions of all occurrences of a k-tuple in D from KT. The KT is constructed by making only one scan through D. Each sequence S_i of length n is broken into $n - k + 1$ number of overlapping k-tuples. The Tuple-weight W is calculated using for each such k-tuple with offset O . Finally, a row $[W,i,O]$ is inserted into KT. For eg $S_1 = \langle ABCABCAC \rangle$, $S_2 = \langle CCACACC \rangle$, $k=2$ Overlapped sequences: $S_1 \rightarrow AB, BC, CA, AB, BC, CA, AC$ $S_2 \rightarrow CC, CA, AC, CA, AC, CC$.

Table 1. Database K Tuple Table

Tuple-weight	Sequence-index	Sequence-offset
AB1	1	0
AB1	1	3
AC2	1	6
AC2	2	2
AC2	2	4
BC5	1	1
BC5	1	4
CA6	1	2
CA6	1	5
CA6	2	1
CA6	2	3
CC8	2	0
CC8	2	5

A query sequence is aligned with the existing sequences using BLAST. The k-tuple table is used here to choose some database regions on which the alignment is performed. KT, thus, is useful in reducing the search space for the alignment process. The query sequence is broken into overlapping k-tuples and its Tuple-weight is evaluated. A list of Sequence-index and Sequence-offset is obtained from KT for each k-tuple in query sequence. The sequence-indexes are ordered according to the number of distinct k-tuples present in that sequence. Let us consider a query sequence $\langle ABCACB \rangle$ which gives five overlapping 2-tuples $\langle AB \rangle$, $\langle BC \rangle$, $\langle CA \rangle$, $\langle AC \rangle$, and $\langle CB \rangle$. Find the positions of these 2-tuples in the existing sequence. As $\langle BC \rangle$ exists in S_1 at offsets 1 and 4, Sequence-index contains $\{1, 1\}$ and Sequence-offset column contains $\{1, 4\}$ for $\langle BC \rangle$ in new Table 2. The information shown in Table 2 is generated from the k-tuple table.

Table 2. Query K Tuples And Their Occurrence In Database

Tuple-weight	Sequence-index	Sequence-offset
AB1	$\{1,1\}$	$\{0,3\}$
BC5	$\{1,1\}$	$\{1,4\}$
CA6	$\{1,1,2,2\}$	$\{2,5,1,3\}$
AC2	$\{1,2,2\}$	$\{6,2,4\}$

Next, the Sequence-indexes are arranged according to the number of distinct query k-tuples present in the database sequence. The ordered Sequence-indexes along with the positions of k-tuples for the above example are shown in Table 3.

Table 3. Ordered Sequence Indexes

Sequence index	No of distinct tuples	Offset
1	4	< {0,3} {1,4} {2,5} {6} >
2	2	< {} {} {1,3} {2,4} >

Equal checkers specialty is that check whether any equal symbol in the string and also check the right and left similarity.

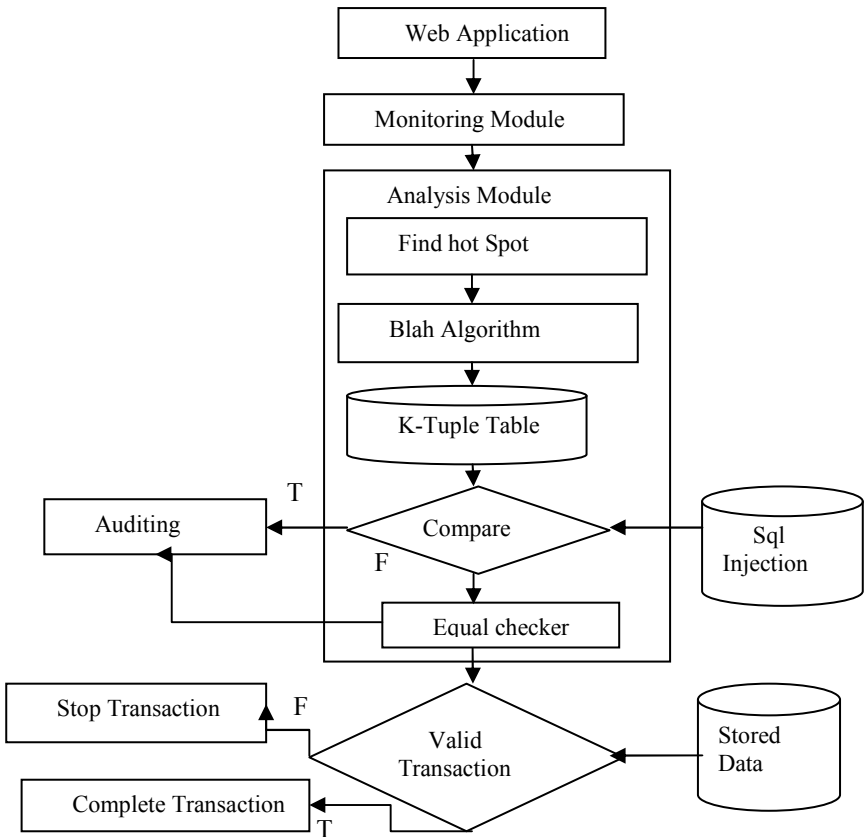


Fig. 1. System Architecture

5 System Architecture

Fig. 1. SQL injection identification’s system architecture is shown in the figure. For the identification it uses 4 modules ie, monitoring module, analysis module, transaction module and auditing module.

System architecture is based on static analysis intrusion detection. In this architecture has four modules to detect security issues. A monitoring module gets the input from web application and sends to the analysis module. An analysis module which finds out the hot spot from web application, it uses blah algorithm. BLAH is a sequence alignment algorithm. It uses for string comparison. Processing speed and accuracy of this algorithm are high. It compares stored SQL injection database. If any SQL injection identified it stop the transaction, otherwise it allows the transaction.

5.1 Monitoring Module

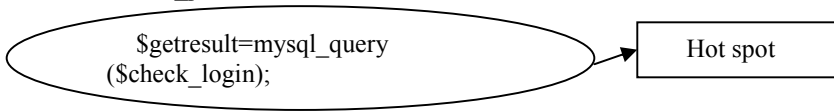
The monitoring module gets an input from the web application and sends it to the analysis module for further checking.

5.2 Analysis Module

5.2.1 Hot Spot

Hotspot is that line where it gets the input from the user and vulnerable in execution. This step performs a simple scanning of the application code to identify hotspots. A hotspot need to be recognized on the web application code page. To find a hotspot, we need to find the mysql_query () function. This function sends the query to the currently active database on the server.

```
<?php require_once('Connection/conn.php');?>
<? Mysql_select_db($database_conn); ?>
<? If(isset($_POST['submit'])) {
    $user_name=$_POST['username']; $user_pwd=$_POST['password'];
    $check_login="SELECT * from student where uname='$user_name' AND
password='$user_pwd'";
```



```
?>
```

5.2.2 Blahe Algorithm

Previous section explain about blahe algorithm in detailed .Here shows the example of the one sql injected query and there subsequence creation .

Input of the blah algorithm:

```
"select * from studentreg where username='$jesna' or '1=1'"; //and
password='$password'";
```

Table 4. Subsequences

sel	st	her	'je
ele	stu	ere	jes
lec	tud	re	en
ect	ude	e u	sa
ct	den	us	na'
t *	ent	use	a'
*	ntr	ser	'o
* f	tre	ern	or
fr	reg	rna	or
or	eg	nam	r '
rom	g w	ame	'l
om	wh	me=	'l=
ms	whe	e='	l=l
		= 'j	= 'l'

Match

After the blah algorithm execution it checks in equal checker for right and left similarity.

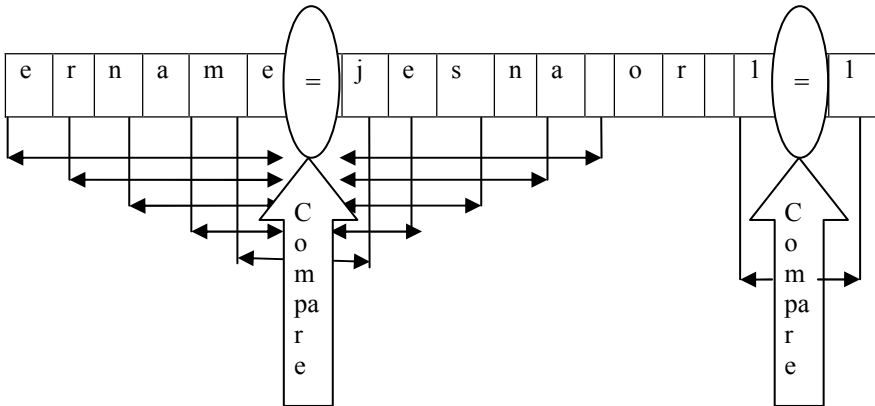


Fig. 2. Equal checker

Fig. 2. Shows the equal checker working on this algorithm

5.3 Auditing Module

Auditing module is used to identify which type of SQL injection attack is done in the query. Information about that injection is recorded. This module generates an alert for administrator and also stops the transaction. The auditing module also identifies which user will generate the SQL injection and warning them or prohibit them from the site access.

5.4 Transaction Module

If we find any SQL injection in the hot spot then we stop the transaction otherwise we continue the transaction. BLAH algorithm is used for finding SQL injection in a web application. In BLAH algorithm we compare with k-tuple table and predefined SQL injection database. If there is no SQL injection in the hot spot then they will check for that transaction is valid or not. If the transaction is valid, they allow the complete transaction otherwise stop the transaction.

6 Experimental Results

We implemented our method SQL Injection Disclosure Using BLAH Algorithm in a web application having a MySQL database in the back end. We tested the proposed system. About the execution of the 100 sql query about 15 query will be identified as a sql injected. Most challenging part of this algorithm is that there is no false negative value because the usage of blahe algorithm.

Table 5. Performance Analysis

Number of sql queries executed in the web application	Total number of subsequences in queries	Average no of subsequence per query	No of valid query	No.of sql injected query
100	6000	60	85	15

Table 6. Accuracy Result

No. Of query	Total false positive	Total false negative
85(valid)+15(sql inection)	3	0

7 Conclusion

This paper will present a novel highly automated approach for protecting Web applications from SQLIAs. The hybrid algorithm named as BLAH which combines the advantages of BLAST and SSAHA algorithms. This approach has high accuracy. BLAHE algorithm is used for the disclosure of the sql injection. At the same time; the processing speed is fast enough to enable online detection of SQL injection. The most challenging part of experiment result is that there is no false negative value.

References

- [1] Fu, X., Lu, X., Peltsverger, B., Chen, S.: A Static Analysis Framework For Detecting SQL Injection Vulnerabilities. IEEE, Los Alamitos (2007)
- [2] Halfond, W.G.J., Orso, A., Manolios, P.: WASP: Protecting Web Applications Using Positive Tainting and Syntax-Aware Evaluation. IEEE, Los Alamitos (2003)
- [3] Thomas, S., Williams, L.: Using Automated Fix Generation to Secure SQL Statements. In: International Workshop on Software Engineering and Secure System. IEEE, Los Alamitos (2006)
- [4] A heuristic-based approach for detecting SQL-injection vulnerabilities in Web applications by Angelo Ciampa, Corrado Aaron Visaggio, Massimiliano Di Penta
- [5] Bandhakavi, S.: CANDID: Preventing SQL Injection Attacks using Dynamic Candidate Evaluations. ACM, New York (2007)
- [6] Kamra, A., Bertino, E., Lebanon, G.: Mechanisms for database intrusion detection and response. In: Data Security & Privacy, pp. 31–36. ACM, New York (2008)
- [7] Ali, S., Rauf, A., Javed, H.: SQLIPA: An Authentication Mechanism Against SQL Injection
- [8] Su, Z., Wassermann, G.: The Essence of Command Injection Attacks in Web Applications. In: 33rd ACM (2006)
- [9] Anely, C.: Advanced SQL injection in sql server application. Next generation security software ltd. White paper (2002)
- [10] A linear Algorithm for Computing Maximal Common Subsequences by D.S. Hirschberg Princeton University
- [11] Livshits, V.B., Lam, M.S.: Finding Security vulnerability in java applications with static analysis. In: Proceedings of the 14th Usenix Security Symposium (August 2005)
- [12] Buehrer, G.T., Weide, B.W., Sivilotti, P.A.G.: Using Parse tree validation to prevent SQL Injection attacks. In: Proc. Of the 5th International Workshop on Software Engineering and Middleware (SEM 2005), pp. 106–113 (September 2005)
- [13] Nguyen-tuong, A., Guarnieri, S., Greene, D., Shirley, J., Evans, D.: Automatically hardening web applications using Precise Tainting. In: Twentieth IFIP Intl, Information security conference (SEC 2005) (May 2005)
- [14] Kiani, M., Clark, A., Mohay, G.: Evaluation of anomaly based character Distribution models in the detection of SQL injection attack
- [15] Halfond, W.G.J., Orso, A.: Combining Static Analysis and Runtime monitoring to counter SQL Injection attacks. In: 3rd International Workshop on Dynamic Analysis, St. Louis, Missouri, p. 1 (2005)
- [16] Halfond, W.G.J., Orso, A., Manolios, P.: WASP: Protecting Web Applications Using Positive Tainting and Syntax-Aware Evaluation. IEEE Transaction of Software Engineering 34(1) (January/February 2008)
- [17] Ezumalai, R., Aghila, G.: Combinatorial Approach for Preventing SQL Injection Attacks. IEEE, Los Alamitos (2009)
- [18] Kundu, A., Panigrahi, S., Sural, S., Majumdar, A.K.: Senior Member, IEEE, BLAST-SSAHA Hybridization for Credit Card Fraud Detection

Author Index

- Abbadi, Imad M. IV-406, IV-557
Abbas, Ash Mohammad II-307
Abraham, Anuj III-503
Abraham, John T. III-168
Abraham, Siby I-328
Achuthan, Krishnashree I-488, II-337
AdiSrikanth, III-570
Aditya, T. I-446
Adusumalli, Sri Krishna IV-572
Agarwal, Vikas II-595
Aghila, G. II-327, IV-98
Agrawal, P.K. IV-244
Agrawal, Rohit II-162
Agrawal, Shaishav III-452
Agushinta R., Dewi II-130, II-138,
II-146
Ahmed, Imran II-317
Ahn, Do-Seob II-595
Aishwarya, Nandakumar II-490, II-498,
III-269
Akhtar, Zahid II-604
Al-Sadi, Azzat A. II-535
Alam, Md. Mahabubul III-349
Alam Kotwal, Mohammed Rokibul
II-154
Ananthi, S. I-480
Andres, Frederic IV-79
Anisha, K.K. III-315
Anita, E.A. Mary I-111
Anju, S.S. II-490, II-498, III-269
Annappa, B. IV-396
Anto, P. Babu III-406
Anusiya, M. IV-155
Aradhya, V.N. Manjunath III-289,
III-297
Arifuzzaman, Md. III-349
Asif Naeem, M. II-30
Asokan, Shimmi IV-63
Athira, B. II-80
Awais, Muhammad II-374
Awasthi, Lalit Kr III-609
Azeem, Mukhtar II-525
Azeez, A.A. Arifa IV-145
Babu, Korra Sathya II-1
Babu, K. Suresh II-636
Babu, L.D. Dhinesh I-223
Babu, M. Rajasekhara I-182
Baburaj, E. I-172
Badache, N. IV-593
Bagan, K. Bhoopathy IV-524
Bajwa, Imran Sarwar II-30
Bakshi, Sambit III-178
Balasubramanian, Aswath I-411
Banati, Hema II-273
Banerjee, Indrajit III-68
Banerjee, Joydeep III-82
Banerjee, Pradipta K. II-480
Banerjee, Usha II-648
Bansal, Roli III-259
Banu, R.S.D. Wahida II-545
Baruah, P.K. I-446
Basak, Dibyajnan I-519
Basil Morris, Peter Joseph II-577
Baskaran, R. II-234, IV-269
Bastos, Carlos Alberto Malcher
IV-195
Batra, Neera I-572
Bedi, Punam II-273, III-259
Bedi, R.K. II-397
Behl, Abhishek II-273
Bhadoria, P.B.S. IV-211
Bhardwaj, Ved Prakash II-568
Bharti, Brijendra K. IV-358
Bhat, Veena H. III-522
Bhattacharyya, Abhijan I-242
Bhosale, Arvind IV-512
Bhuvanagiri, Kiran Kumar IV-293
Bhuvanewary, A. II-327
Biji, C.L. IV-300
Binu, A. I-399
Biswas, G.P. II-628
Biswas, Subir III-54
Biswas, Suparna II-417
Biswas, Sushanta II-612, II-620
Biswash, Sanjay Kumar I-11
Boddu, Bhaskara Rao II-296
Borah, Samarjeet III-35

- Borkar, Meenal A. IV-25
 Boutekkouk, Fateh II-40

 Chaganty, Aparna IV-19
 Chaitanya, N. Sandeep IV-70
 Chakraborty, Suchetana II-585
 Chakravorty, Debaditya III-35
 Challa, Rama Krishna IV-608
 Chanak, Prasenjit III-68
 Chand, Narottam III-122, III-609
 Chandra, Deka Ganesh II-210
 Chandra, Jayanta K. II-480
 Chandran, K.R. I-631
 Chandrika, I-704
 Chaniyani, S.S. Mozaffari III-289
 Chaoub, Abdelaali I-529
 Chaudhary, Ankit III-488
 Chauhan, Durg Singh I-21
 Chawhan, Chandan III-35
 Chawla, Suneeta II-430
 Chia, Tsorng-Lin III-334
 Chintapalli, Venkatarami Reddy IV-455
 Chitrakala, S. III-415
 Chittineni, Suresh III-543
 Choudhary, Surendra Singh I-54
 Chouhan, Madhu I-119
 Chowdhury, Chandreyee I-129
 Chowdhury, Roy Saikat II-577

 Dadhich, Reena I-54
 Dahiya, Ratna III-157
 Dandapat, S. IV-165
 Das, Madhabananda IV-113
 Das, Satya Ranjan II-172
 Das, Subhalaxmi IV-549
 Datta, Asit K. II-480
 Dawoud, Wesam I-431
 Deb, Debasish II-577
 Dedavath, Saritha I-34
 Deepa, S.N. III-503
 Dehalwar, Vasudev I-153
 Dehuri, Satchidananda IV-113
 Desai, Sharmishta II-397
 Devakumari, D. II-358
 Devani, Mahesh I-213
 Dhanya, P.M. IV-126
 Dhar, Pawan K. I-284
 Dharanyadevi, P. II-234
 Dhavachelvan, P. II-234
 Dhivya, M. II-99

 Dilna, K.T. III-185
 Dimililer, Kamil III-357
 Diwakar, Shyam II-337
 Doke, Pankaj I-607, II-430
 Dongardive, Jyotshna I-328
 Donoso, Yezid II-386
 Doraipandian, Manivannan III-111
 Dorizzi, Bernadette III-20
 Durga Bhavani, S. III-1
 Dutta, Paramartha I-83
 Dutta, Ratna IV-223

 El Abbadi, Jamal I-529
 El-Alfy, El Sayed M. II-535
 Elhaj, Elhassane Ibn I-529
 Elizabeth, Indu I-302
 Elumalai, Ezhilarasi I-1

 Ferreira, Ana Elisa IV-195
 Ferri, Fernando IV-79

 Gadia, Shashi II-191
 Gaiti, Dominique II-471
 Ganeshan, Kathiravelu IV-501
 Garcia, Andrés III-664
 Garcia, Anilton Salles IV-195
 Gaur, Manoj Singh I-44, I-162, I-562,
 II-183, II-452, III-478, III-644
 Gaur, Vibha II-284
 Gautam, Gopal Chand I-421
 Geetha, V. II-48
 Geevar, C.Z. III-460
 Ghosh, Pradipta III-82
 Ghosh, Saswati II-620
 Giluka, Mukesh Kumar I-153
 Gindi, Sanjyot IV-349
 Gireesh Kumar, T. II-506
 Giuliani, Alessandro I-284
 Godavarthi, Dinesh III-543
 Gómez-Skarmeta, Antonio Fernando
 III-664
 Gondane, Sneha G. II-99
 Gopakumar, G. I-320
 Gopalan, Kaliappan IV-463
 Gore, Kushal I-607
 Gosain, Anjana I-691
 Gosalia, Jenish IV-378
 Govardhan, A. I-581
 Govindan, Geetha I-294
 Govindarajan, Karthik I-192

- Grifoni, Patrizia IV-79
 Grover, Jyoti III-644
 Gualotuña, Tatiana IV-481
 Guerroumi, M. IV-593
 Gunaraj, G. I-192
 Gunjan, Reena III-478
 Gupta, Ankur I-501
 Gupta, B.B. IV-244
 Gupta, Deepika II-183
 Gupta, J.P. I-260
 Gupta, Juhi IV-205
 Gupta, Priya IV-512
- Habib, Sami J. II-349
 Hafizul Islam, SK II-628
 Harivinod, N. III-396
 Harmya, P. II-490, II-498, III-269
 Harshith, C. II-506
 Hassan, Foyzul II-154, III-349
 Hati, Sumanta III-580
 Hazarika, Shyamanta M. II-109, II-119
 Hazra, Sayantan III-601
 Hemamalini, M. IV-175
 Hivarkar, Umesh N. IV-358
 Hsieh, Chaur-Heh III-334
 Huang, Chin-Pan III-334
 Huang, Ping S. III-334
- Ibrahim, S.P. Syed I-631
 Indira, K. I-639
 Isaac, Elizabeth IV-145
- Jaganathan, P. I-683
 Jagdale, B.N. II-397
 Jain, Jitendra III-326
 Jain, Kavindra R. III-239
 Jain, Kavita I-328
 Jalal, Anand Singh II-516, IV-329
 Jameson, Justy II-693
 Janani, S. IV-175
 Jaya, IV-233
 Jayakumar, S.K.V. II-234
 Jayaprakash, R. II-656
 Jena, Sanjay Kumar II-1
 Jia, Lulu IV-421
 Jiménez, Gustavo II-386
 Jisha, G. IV-1, IV-137
 Joseph, Shijo M. III-406
 Juluru, Tarun Kumar I-34, III-590
- Kacholiya, Anil IV-205
 Kahlon, K.S. II-58
 Kakoty, Nayan M. II-119
 Kakulapati, Vijayalaxmi IV-284
 Kalaivaani, P.T. III-143
 Kale, Sandeep II-604
 Kanade, Sanjay Ganesh III-20
 Kanavalli, Anita I-141
 Kancharla, Tarun IV-349, IV-368
 Kanitkar, Aditya R. IV-358
 Kanivadhana, P. IV-155
 Kankacharla, Anitha Sheela I-34, III-590
 Kanmani, S. I-639, II-69
 Kannan, A. II-19
 Kannan, Rajkumar IV-79
 Kapoor, Lohit I-501
 Karamoy, Jennifer Sabrina Karla II-138
 Karande, Vishal M. IV-386
 Karmakar, Sushanta II-585
 Karthi, R. III-552
 Karthik, S. I-480
 Karunanithi, D. IV-284
 Karunanithi, Priya III-624
 Karuppanan, Komathy III-425, III-615, III-624, III-634
 Katiyar, Vivek III-122
 Kaur, Rajbir I-44, I-162
 Kaushal, Sakshi IV-445
 Kavalcioglu, Cemal III-357
 Kayarvizhy, N. II-69
 Keromytis, Angelos D. III-44
 Khajaria, Krishna II-9
 Khalid, M. I-182
 Khan, Majid Iqbal II-471, II-525
 Khan, Srabani II-620
 Khan Jehad, Abdur Rahman III-349
 Khanna, Rajesh III-205
 Khattak, Zubair Ahmad IV-250
 Khilar, P.M. I-119
 Kim, Pansoo II-595
 Kimbahune, Sanjay I-607, II-430
 Kiran, N. Chandra I-141
 Kishore, J.K. II-460
 Ko, Ryan K.L. IV-432
 Kolikipogu, Ramakrishna IV-284
 Kopparapu, Sunil Kumar II-317, IV-293
 Koschnicke, Sven I-371
 Kothari, Nikhil I-213

- Krishna, Gutha Jaya I-382
 Krishna, P. Venkata I-182
 Krishna, S. III-522
 Krishnan, Saranya D. IV-63
 Krishnan, Suraj III-374
 Kopparapu, Sunil Kumar III-230
 Kulkarni, Nandakishore J. III-570
 Kumar, Chiranjeev I-11
 Kumar, C. Sasi II-162
 Kumar, G.H. III-289
 Kumar, G. Ravi IV-70
 Kumar, G. Santhosh I-399
 Kumar, Ishan IV-205
 Kumar, K.R. Ananda I-704
 Kumar, K. Vinod IV-19
 Kumar, Manish I-44
 Kumar, Manoj II-9
 Kumar, Naveen I-461
 Kumar, Padam I-461
 Kumar, Ravindra II-307
 Kumar, Santosh I-619
 Kumar, Santhosh G. III-93
 Kumar, Saumesh I-461
 Kumar, Sumit I-619
 Kumaraswamy, Rajeev IV-339
 Kumari, M. Sharmila III-396
 Kumari, V. Valli IV-572
 Kumar Pandey, Vinod III-230
 Kumar Sarma, Kandarpa III-512
 Kurakula, Sudheer IV-165
 Kussmaul, Clifton III-533

 Lachiri, Zied IV-318
 Lal, Chhagan II-452
 Latif, Md. Abdul II-154
 Laxmi, V. II-183, II-452
 Laxmi, Vijay I-44, I-162, I-562, III-478,
 III-644
 Lee, Bu Sung IV-432
 Li, Tiantian IV-421
 Limachia, Mitesh I-213
 Lincoln Z.S., Ricky II-130
 Linganagouda, K. III-444
 Lingeswarara, C. II-19
 Liu, Chenglian IV-534
 Lobiyal, D.K. III-132, III-654
 Londhe, Priyadarshini IV-512
 López, Elsa Macías IV-481

 Madheswari, A. Neela II-545
 Madhusudhan, Mishra III-365

 Mahalakshmi, T. I-310
 Mahalingam, P.R. III-562, IV-137
 Maheshwari, Saurabh III-478
 Maiti, Santa II-172
 Maity, G.K. III-249
 Maity, Santi P. I-519, III-249, III-580
 Maity, Seba I-519
 Majhi, Banshidhar III-178
 Majhi, Bansidhar IV-549
 Maji, Sumit Kumar I-649
 Malay, Nath III-365
 Malaya, Dutta Borah II-210
 Malik, Jyoti III-157
 Mallya, Anita I-302
 Manan, Jamalul-lail Ab IV-250
 Mandava, Ajay K. I-351
 Mannava, Vishnuvardhan I-250
 Manomathi, M. III-415
 Maralappanavar, Meena S. III-444
 Marcillo, Diego IV-481
 Marimuthu, Paulvanna N. II-349
 Mary, S. Roselin IV-9
 Masera, Guido II-374
 Mastan, J. Mohamedmoideen Kader
 IV-524
 Mehrotra, Hunny III-178
 Meinel, Christoph I-431
 Mendiratta, Varun II-273
 Menta, Sudhanshu III-205
 Mishra, A. IV-244
 Mishra, Ashok II-223
 Mishra, Dheerendra IV-223
 Mishra, Shivendu II-407
 Misra, Rajiv I-101
 Missaoui, Ibrahim IV-318
 Mitra, Abhijit III-512, III-601
 Mitra, Swarup Kumar III-82
 Mittal, Puneet II-58
 Modi, Chintan K. III-239
 Mohammadi, M. III-289
 Mohandas, Neethu IV-187
 Mohandas, Radhesh II-685, III-10
 Mohanty, Sujata IV-549
 Mol, P.M. Ameera III-193
 Moodgal, Darshan II-162
 Moragón, Antonio III-664
 More, Seema I-361
 Moussaoui, S. IV-593
 Mubarak, T. Mohamed III-102
 Mukhopadhyay, Sourav IV-223

- Mukkamala, R. I-446
 Muniraj, N.J.R. I-270, III-168
 Murthy, G. Rama IV-19

 Nadarajan, R. II-366
 Nadkarni, Tanusha S. II-685
 Nag, Amitava II-612, II-620
 Nagalakshmi, R. I-683
 Nagaradjane, Prabagarane III-374
 Nair, Achuthsankar S. I-284, I-294,
 I-302, I-320
 Nair, Bipin II-337
 Nair, Madhu S. III-193, III-276
 Nair, Smita IV-368
 Nair, Vrinda V. I-302
 Namboodiri, Saritha I-284
 Namritha, R. III-634
 Nandi, Sukumar I-619
 Narayanan, Hari I-488
 Nasiruddin, Mohammad II-154
 Naskar, Mrinal Kanti III-82
 Nataraj, R.V. I-631
 Naveen, K. Venkat III-570, III-615
 Naveena, C. III-297
 Nazir, Arfan II-525
 Neelamegam, P. III-111
 Neogy, Sarmistha I-129, II-417
 Nigam, Apurv II-430
 Nimi, P.U. IV-46
 Niranjana, S.K. III-297
 Nirmala, M. I-223
 Nirmala, S.R. III-365
 Nitin, I-21, II-568, IV-25
 Noopa, Jagadeesh II-490, II-498, III-269
 Nurul Huda, Mohammad II-154, III-349

 Oh, Deock-Gil II-595
 Okab, Mustapha II-40
 Oliya, Mohammad I-232
 Olsen, Rasmus L. IV-37

 Padmanabhan, Jayashree I-1, IV-541
 Padmavathi, B. IV-70
 Pai, P.S. Sreejith IV-339
 Pai, Radhika M. II-460
 Paily, Roy IV-165
 Pais, Alwyn R. II-685, IV-386
 Pais, Alwyn Roshan III-10
 Pal, Arindarjit I-83
 Palaniappan, Ramaswamy IV-378

 Palaty, Abel IV-56
 Pandey, Kumar Sambhav IV-56
 Panicker, Asha IV-300
 Panneerselvam, S. I-223
 Pappas, Vasilis III-44
 Parasuram, Harilal II-337
 Parmar, Rohit R. III-239
 Parthasarathy, Magesh Kannan I-192
 Parvathy, B. I-204
 PatilKulkarni, Sudarshan III-342
 Patnaik, L.M. I-141, II-636, III-522
 Patra, Prashanta Kumar I-649
 Pattanshetti, M.K. IV-244
 Paul, Anu II-201
 Paul, Richu III-213
 Paul, Varghese II-201
 Paulsen, Niklas I-371
 Pavithran, Vipin I-488
 Pearson, Siani IV-432
 Perumal, V. I-471
 Petrovska-Delacrétaz, Dijana III-20
 Phani, G. Lakshmi IV-19
 Ponpandiyan, Vigneswaran IV-541
 Poornalatha, G. II-243
 Povar, Digambar I-544
 Prabha, S. Lakshmi I-192
 Prabhu, Lekhesh V. IV-339
 Pradeep, A.N.S. III-543
 Pradeepa, J. I-471
 Prajapati, Nitesh Kumar III-644
 Prakasam, Kumaresh IV-541
 Pramod, K. III-444
 Prasad, Ramjee IV-37
 Prasath, Rajendra II-555
 Prasanna, S.R. Mahadeva III-326
 Prasanth Kumar, M. Lakshmi I-11
 Pratheepraj, E. III-503
 Priya, K.H. I-471
 Priyadharshini, M. IV-269
 Priyadharshini, V. IV-175
 Pung, Hung Keng I-232

 Qadeer, Mohammed Abdul II-442

 Radhamani, A.S. I-172
 Rafsanjani, Marjan Kuchaki IV-534
 Raghavendra, Prakash S. II-243
 Raghuvanshi, Rahul I-153
 Rahaman, Hafizur III-68

- Raheja, J.L. III-488
 Raheja, Shekhar III-488
 Rahiman, M. Abdul III-304
 Rahman, Md. Mostafizur II-154
 Rai, Anjani Kumar II-407
 Rai, Anuj Kumar III-111
 Rai, Mahendra K. III-469
 Raja, K.B. II-636
 Rajapackiyam, Ezhilarasie III-111
 Rajasekhar, Ch. I-78
 Rajasree, M.S. III-304
 Rajendran, C. III-552
 Rajesh, R. III-497
 Rajeswari, A. III-143
 Rajimol, A. II-253
 Rajkumar, K.K. III-435
 Rajkumar, N. I-683
 Raju, C.K. II-223, IV-211
 Raju, G. I-671, II-253, III-435
 Ramachandram, S. IV-70
 Ramamohanreddy, A. I-581
 Ramaraju, Chithra I-661
 Ramasubbareddy, B. I-581
 Ramaswamy, Aravindh I-411
 Ramesh, Sunanda I-1
 Ramesh, T. I-250
 Rameshkumar, K. III-552
 Rana, Sanjeev I-91
 Rani, Prathuri Jhansi III-1
 Rao, Appa III-102
 Rao, Avani I-213
 rao, D. Srinivasa I-78
 Rao, Prasanth G. III-522
 Rastogi, Ravi I-21
 Rathi, Manisha I-260
 Rathore, Wilson Naik II-676
 Razi, Muhammad II-146
 Reddy, B. Vivekavardhana IV-309
 Reddy, G. Ram Mohana IV-473
 Reddy, P.V.G.D. Prasad III-543
 Reddy, Sateesh II-460
 Regentova, Emma E. I-351
 Reji, J. III-276
 Revathy, P. IV-284
 Revett, Kenneth IV-378
 Roberta, Kezia Velda II-146
 Rodrigues, Paul IV-9, IV-269
 Rokibul Alam Kotwal, Mohammed
 III-349
 Roopalakshmi, R. IV-473
 Roy, J.N. III-249
 Roy, Rahul IV-113
 Sabu, M.K. I-671
 Saha, Aritra III-35
 Sahaya, Nuniek Nur II-138
 Sahoo, Manmath Narayan I-119
 Sahoo, Soyuj Kumar III-326
 Saikia, Adity II-109, II-119
 Sainarayanan, G. III-157
 Sajeev, J. I-310
 Sajitha, M. III-102
 Saljooghinejad, Hamed II-676
 Samad, Sumi A. III-93
 Samanta, Debasis II-172
 Sambyal, Rakesh IV-608
 Samerendra, Dandapat III-365
 Samraj, Andrews IV-378
 Samuel, Philip II-80, IV-1
 Sandhya, S. II-88
 Santa, José III-664
 Santhi, K. III-221
 SanthoshKumar, G. II-263
 Santhoshkumar, S. I-223
 Saralaya, Vikram II-460
 Sarangdevot, S.S. I-592
 Saraswathi, S. IV-155, IV-175
 Sardana, Anjali IV-233
 Saritha, S. II-263
 Sarkar, D. II-612, II-620
 Sarkar, Partha Pratim II-612, II-620
 Sarma, Monalisa II-172
 Saruladha, K. II-327
 Sasho, Ai I-340
 Sasidharan, Satheesh Kumar I-552
 Satapathy, Chandra Suresh III-543
 Sathisha, N. II-636
 Sathishkumar, G.A. IV-524
 Sathiya, S. IV-155
 Sathu, Hira IV-491, IV-501
 Satria, Denny II-138
 Sattar, Syed Abdul III-102
 Savarimuthu, Nickolas I-661
 Sayeesh, K. Venkat IV-19
 Schatz, Florian I-371
 Schimmler, Manfred I-371
 Sebastian, Bhavya I-302
 Sehgal, Priti III-259
 Selvan, A. Muthamizh III-497

- Selvathi, D. IV-300
 Sen, Jaydip IV-580
 Sendil, M. Sadish I-480
 Senthilkumar, Radha II-19
 Senthilkumar, T.D. III-185
 Shah, Mohib A. IV-491, IV-501
 Shahram, Latifi I-351
 Shajan, P.X. III-168
 Sharma, Amita I-592
 Sharma, Dharendra Kumar I-11
 Sharma, Divya I-511
 Sharma, H. Meena I-162
 Sharma, Neeraj Kumar II-284
 Sharma, Ritu I-511
 Sharma, Sattvik II-506
 Sharma, Sugam II-191
 Sharma, Surbhi III-205
 Sharma, T.P. I-421
 Shekar, B.H. III-396
 Shenoy, P. Deepa I-141, III-522
 Shenoy, S.K. III-93
 Sherly, K.K. II-693
 Shringar Raw, Ram III-654
 Shukla, Shailendra I-101
 Shyam, D. II-99
 Sikdar, Biplab Kumar III-68
 Singal, Kunal III-488
 Singh, Anurag III-609
 Singh, Ashwani II-374
 Singh, Jai Prakash IV-89
 Singh, Jyoti Prakash I-83, II-612, II-620
 Singh, Manpreet I-91, I-572
 Singh, Puneet III-570
 Singh, Rahul I-340
 Singh, Sanjay II-460
 Singh, Satwinder II-58
 Singh, Vijander I-54
 Singh, Vrijendra II-516, IV-329
 Singh, Preety II-183
 Sinha, Adwitiya III-132
 Sivakumar, N. II-88
 Skandha, S. Shiva IV-70
 Smith, Patrick II-191
 Sojan Lal, P. III-460
 Song, Jie IV-421
 Soni, Surender III-122
 Sood, Manu I-511
 Soumya, H.D. I-361
 Sreenath, N. II-48
 Sreenu, G. IV-126
 Sreevathsan, R. II-506
 Srikanth, M.V.V.N.S. II-506
 Srinivasan, Avinash IV-260
 Srinivasan, Madhan Kumar IV-269
 Srivastava, Praveen Ranjan III-570
 Srivastava, Shweta I-260
 Starke, Christoph I-371
 Suaib, Mohammad IV-56
 Suárez-Sarmiento, Alvaro IV-481
 Subramaniam, Tamil Selvan Raman IV-541
 Suchithra, K. IV-339
 Sudarsan, Dhanya IV-137
 Sudhansh, A.S.D.P. IV-165
 Sujana, N. I-361
 Sukumar, Abhinaya I-1
 Sulaiman, Suziah IV-250
 Sundararajan, Sudharsan I-488
 Swaminathan, A. II-648
 Swaminathan, Shriram III-374
 Swamy, Y.S. Kumara IV-309
 Tahir, Muhammad II-471
 Takouna, Ibrahim I-431
 Thakur, Garima I-691
 Thampi, Sabu M. I-64, IV-126, IV-145, IV-187
 Thangavel, K. II-358
 Thilagu, M. II-366
 Thiyagarajan, P. IV-98
 Thomas, Diya I-64
 Thomas, K.L. I-544, I-552
 Thomas, Likewin IV-396
 Thomas, Lincy III-425
 Thomas, Lisha III-221
 Thukral, Anjali II-273
 Tim, U.S. II-191
 Tiwary, U.S. III-452, III-469
 Tobgay, Sonam IV-37
 Tolba, Zakaria II-40
 Tripathi, Pramod Narayan II-407
 Tripathi, Rajeev I-11
 Tripathy, Animesh I-649
 Tripti, C. IV-46
 Tyagi, Neeraj I-11
 Tyagi, Vipin II-568
 Uma, V. II-656
 Umber, Ashfa II-30
 Unnikrishnan, C. III-562

- Usha, N. IV-309
 Utomo, Bima Shakti Ramadhan II-138

 Vanaja, M. I-78
 Varalakshmi, P. I-411, I-471
 Varghese, Elizabeth B. III-383
 Varshney, Abhishek II-442
 Vasanthi, S. III-213
 Vatsavayi, Valli Kumari II-296
 Venkatachalapathy, V.S.K. II-234
 Venkatesan, V. Prasanna IV-98
 Venugopal, K.R. I-141, II-636, III-522
 Verma, Amandeep IV-445
 Verma, Chandra I-284
 Verma, Gyanendra K. III-452, III-469
 Verma, Rohit I-21
 Vidya, M. I-361
 Vidyadharan, Divya S. I-544
 Vijay, K. I-78
 Vijaykumar, Palaniappan I-411
 VijayLakshmi, H.C. III-342
 Vinod, P. I-562
 Vipeesh, P. I-270

 Vishnani, Kalpa III-10
 Vivekanandan, K. II-88
 Vorungati, Kaladhar I-488
 Vykopal, Jan II-666

 Wadhai, V.M. II-397
 Wankar, Rajeev I-382
 Wattal, Manisha I-501
 William, II-130
 Wilscy, M. III-315, III-383
 Wirjono, Adityo Ashari II-130
 Wisudawati, Lulu Mawaddah II-146
 Wu, Jie IV-260

 Xavier, Agnes I-328

 Yadav, Gaurav Kumar IV-368
 Yu, Fan III-54
 Yuvaraj, V. III-503

 Zaeri, Naser II-349
 Zheng, Liyun IV-534
 Zhu, Shenhaochen I-340
 Zhu, Zhiliang IV-421