ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING 177

Natarajan Meghanathan
Dhinaharan Nagamalai
Nabendu Chaki (Eds.)

Advances in Computing and Information Technology

Springer

Advances in Intelligent Systems and Computing

177

Editor-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

For further volumes:
http://www.springer.com/series/11156

Natarajan Meghanathan, Dhinaharan Nagamalai,
and Nabendu Chaki (Eds.)

Advances in Computing and Information Technology

Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY) July 13–15, 2012, Chennai, India – Volume 2

Springer

Editors
Dr. Natarajan Meghanathan
Department of Computer Science
Jackson State University
Jackson
USA

Dr. Nabendu Chaki
Department of Computer Science &
Engineering
University of Calcutta
Calcutta
India

Dr. Dhinaharan Nagamalai
Wireilla Net Solutions PTY Ltd
Melbourne
VIC
Australia

ISSN 2194-5357
e-ISSN 2194-5365
ISBN 978-3-642-31551-0
e-ISBN 978-3-642-31552-7
DOI 10.1007/978-3-642-31552-7
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012940793

© Springer-Verlag Berlin Heidelberg 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Preface

The Second International Conference on Advances in Computing and Information Technology (ACITY-2012) was held in Chennai, India, during July 13–15, 2012. ACITY attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West. The goal of this conference series is to bring together researchers and practitioners from academia and industry and share cutting-edge development in the field. The conference will provide an excellent international forum for sharing knowledge and results in theory, methodology and applications of Computer Science and Information Technology. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of Computer Science and Information Technology.

The ACITY-2012 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the conference. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer-review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. The overall acceptance rate of ACITY-2012 is less than 20%. Extended versions of selected papers from the conference will be invited for publication in several international journals. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in various research areas of Computer Science and Information Technology. In closing, ACITY-2012 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. We would like to thank the General and Program Chairs, organization staff, the members of the Technical

Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research.

It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Natarajan Meghanathan
Dhinaharan Nagamalai
Nabendu Chaki

Organization

General Chairs

David C. Wyld	Southeastern Louisiana University, USA
E.V. Krishnamurthy	Australian National University, Australia
Jae Kwang Lee	Hannam University, South Korea
Jan Zizka	SoNet/DI, FBE, Mendel University in Brno, Czech Republic
V.L. Narasimhan	Pentagram R&D Intl. Inc., New Bern, USA
Michal Wozniak	Wroclaw University of Technology, Poland

Steering Committee

Abdul Kadhir Ozcan	Karatay University, Turkey
Brajesh Kumar Kaushik	Indian Institute of Technology-Roorkee, India
Dhinaharan Nagamalai	Wireilla Net Solutions PTY LTD, Australia
Eric Renault	Institut Telecom - Telecom SudParis, Evry, France
Jacques Demerjian	Communication & Systems, France
James Henrydoss	AT&T and University of Colorado, USA
Krzysztof Walkowiak	Wroclaw University of Technology,Poland
Murugan D.	Manonmaniam Sundaranar University, India
Nabendu Chaki	University of Calcutta, India
Natarajan Meghanathan	Jackson State University, USA
Raja Kumar M.	Taylor's University, Malaysia
Salah Al-Majeed	University of Essex, UK
Selma Boumerdassi	Conservatoire National des Arts Et Metiers (CNAM), France
Sundarapandian Vaidyanathan	VelTech Dr. RR & Dr. SR Technical University, India

Program Committee Members

A.H.T. Mohammad	University of Bradford, UK
A.P. Sathish Kumar	PSG Institute of Advanced Studies, India
AAA. Atayero	Covenant University, Nigeria
Abdul Aziz	University of Central Punjab, Pakistan
Abdul Kadhir Ozcan	Karatay University, Turkey
Abdul Kadir Ozcan	The American University, Cyprus
Abdulbaset Mohammad	University of Bradford, United Kingdom
Ahmad Saad Al-Mogren	King Saud University, Saudi Arabia
Ahmed M. Khedr	Sharjah University, Sharjah, UAE
Ahmed Nada	Al-Quds University, Palestinian
Ajay K. Sharma	Dr. B R Ambedkar National Institute of Technology, India
Alaa Ismail Elnashar	Taif University, KSA
Alejandro Garces	Jaume I University, Spain
Alejandro Regalado Mendez	Universidad del Mar - México, USA
Alfio Lombardo	University of Catania, Italy
Ali El-Rashedy	University of Bridgeport, CT, USA
Ali M.	University of Bradford, United Kingdom
Ali Maqousi	Petra University, Jordan
Alireza Mahini	Islamic Azad University-Gorgan, Iran
Alvin Lim	Auburn University, USA
Amandeep Singh Thethi	Guru Nanak Dev University Amritsar, India
Amit Choudhary	Maharaja Surajmal Institute,India
Anand Sharma	MITS-Rajasthan, India
Anjan K.	RVCE-Bangalore, India
Ankit Thakkar	Nirma University, India
Ankit	BITS, PILANI India
Anthony Atayero	Covenant University, Nigeria
Aravind P.A.	Amrita School of Engineering India
Arun Pujari	Sambalpur University, India
Arunita Jaekel	University of Windsor, Canada
Ashok Kumar Das	IIT Hyderabad, India
Ashok kumar Sharma	YMCA Institute of Engineering, India
Ashutosh Dubey	NRI Institute of Science & Technology, Bhopal
Ashutosh Gupta	MJP Rohilkhand University, Bareilly
Athanasios Vasilakos	University of Western Macedonia, Greece
Azween Bin Abdullah	Universiti Teknologi Petronas, Malaysia
B. Srinivasan	Monash University, Australia
Babak Khosravifar	Concordia University, Canada
Balakannan S.P.	Chonbuk Nat. Univ., Jeonju
Balasubramanian K.	Lefke European University, Cyprus
Balasubramanian Karuppiah	Dr. MGR University, India
Bari A.	University of Western Ontario, Canada

Beatrice Cynthia Dhinakaran	TCIS, South Korea
Bela Genge	European Commission Joint Research Centre, Belgium
Bharat Bhushan Agarwal	I.F.T.M University, India
Bhupendra Suman	IIT Roorkee , India
Biju Pattnaik	University of Technology, India
Bikash singh	Islamic University-Kushtia, Bangladesh
Binod Kumar Pattanayak	Siksha O Anusandhan University, India
Bobby Barua	Ahsanullah University of Science and Technology, Bangladesh
Bong-Han	Kim, Chongju University, South Korea
Boo-Hyung Lee	KongJu National University, South Korea
Brajesh Kumar Kaushik	Indian Institute of Technology, India
Buket Barkana	University of Bridgeport, USA
Carlos E. Otero	University of South Florida Polytechnic, USA
Charalampos Z. Patrikakis	National Technical University of Athens, Greece
Chin-Chih Chang	Chung Hua University ,Taiwan
Cho Han Jin	Far East University, South Korea
Choudhari	Bhagwati Chaturvedi College of Engineering, India
Christos Politis	Kingston University, UK
Cristina Ribeiro	University of Waterloo, Canada
Cristina Serban	Ovidius University of Constantza, Romania
Danda B. Rawat	Old Dominion University, USA
David C. Wyld	Southeastern Louisiana University, USA
Debasis Giri	Haldia Institute of Technology, India
Debdatta Kandar	Sikkim Manipal University, India
Dhinaharan Nagamalai	Wirella Net Solutions PTY Ltd, Australia
Diego Reforgiato	University of Catania, Italy
Dimitris Kotzinos	Technical Educational Institution of Serres, Greece
Doreswamyh hosahalli	Mangalore University, India
Durga Toshniwal	Indian Institute of Technology, India
E. Martin	University of California, Berkeley, USA
E.V. Krishnamurthy	ANU College of Engg & Computer Science, Austraila
Emmanuel Bouix	iKlax Media, France
Eric Renault	Institut Telecom - Telecom SudParis, Evry, France
Ermatita Zuhairi	Sriwijaya University, Indonesia
Farag M. Sallabi	United Arab Emirates University, UAE
Farshad Safaei	Shahid Beheshti University, Iran
Ford Lumban Gaol	University of Indonesia
Genge Bela	Joint Research Centre, European Commission, Italy
Ghalem Belalem	University of Oran, Algeria
Giovanni Cordeiro Barroso	Universidade Federal do Ceara, Brasil
Giovanni Schembra	University of Catania, Italy
Girija Chetty	University of Canberra, Australia

Gomathi Kandasamy	Avinashilingam Deemed University for Women, India
Gopalakrishnan Kaliaperumal	Anna University, Chennai
Govardhan A.	JNTUH College of Engineering, India
Guo Bin	Institute TELECOM SudParis, France
H.V. Ramakrishnan	Dr. MGR University, India
Haider M. Alsabbagh	Basra University, Iraq
Haller Piroska	Petru Maior University-Tirgu Mures, Romania
Hao Shi	Victoria University, Australia
Hao-En Chueh	yuanpei University, Taiwan
Hari Chavan	National Institute of Technology, Jamshedpur, India
Henrique J.A. Holanda	UERN - Universidade do Estado do Rio Grande do Norte, Brasil
Henrique Joao Lopes Domingos	University of Lisbon, Portugal
Hiroyuki Hisamatsu	Osaka Electro-Communication University, Japan
Ho Dac Tu	Waseda University, Japan
Homam Reda El-Taj	Universiti Sains Malaysia, Malaysia
Hong yu	Capitol College, USA
Huosheng Hu	University of Essex, UK
Hussein Al-Bahadili	Petra University, Jordan
Hussein Ismail Khalaf Al-Bahadili	Petra University, Jordan
Hwangjun Song	Pohang University of Science and Technology,South Korea
Ignacio Gonzalez Alonso	University of Oviedo, Europe
Indrajit Bhattacharya	Kalyani Govt. Engg. College, India
Intisar Al-Mejibli	University of Essex, UK
Ioannis Karamitsos	Itokk Communications, Canada
J.K. Mandal	University of Kalyani, India
Jacques Demerjian	Communications & Systems, France
Jae Kwang Lee	Hannam University, South Korea
Jalel Akaichi	University of Tunis, Tunisia
Jan Zizka	SoNet/DI, FBE, Mendel University in Brno, Czech Republic
Jeong-Hyun Park	Electronics Telecommunication Research Institute, South Korea
Jeyanthy N.	VIT University, India
Jifeng Wang	University of Illinois at Urbana Champaign, USA
Johann Groschdl	University of Bristol, UK
Jose Enrique Armendariz-Inigo	Universidad Publica de Navarra, Spain
Juan Li	North Dakota State University, USA
Jyoti Singhai	Electronics and Communication Deptt-MANIT, India
Jyotirmay Gadewadikar	Alcorn State University, USA
Kai Xu	University of Bradford, United Kingdom
Kamalrulnizam Abu Bakar	Universiti Teknologi Malaysia, Malaysia

Karim Konate	University Cheikh Anta DIOP, Dakar
Kaushik Chakraborty	Jadavpur University, India
Kayhan Erciyes	Izmir University, Turkey
Khaled Shuaib	United Arab Emirates University, UAE
Khamish Malhotra	University of Glamorgan, UK
Khoa N. Le	University of Western Sydney, Australia
Krishnamurthy E.V.	ANU College of Engg & Computer Science, Australia
Krzysztof Walkowiak	Wroclaw University of Technology, Poland
Kuribayashi	Seikei University, Japan
L. Nirmala Devi	Osmania University - Hyderabad, India
Laiali Almazaydeh	University of Bridgeport, USA
Lu Yan	University of Hertfordshire, UK
Lus Veiga	Technical University of Lisbon, Portugal
Lylia Abrouk	University of Burgundy, France
M. Aqeel Iqbal	FUIEMS, Pakistan
M. Rajarajan	City University, UK
M. Ali	University of Bradford, UK
Maode Ma	Nanyang Technological University, Singapore
Marco Folli	University of Pavia, Italy
Marco Roccetti	Universty of Bologna, Italy
Massimo Esposito	ICAR-CNR, Italy
Md. Sipon Miah	Islamic University-Kushtia, Bangladesh
Michal Wozniak	Wroclaw University of Technology, Poland
Michel Owayjan	American University of Science & Technology, Lebanon
Miguel A. Wister	Juarez Autonomous University of Tabasco, Mexico
Mohamed Hassan	American University of Sharjah, UAE
Mohammad Ali Jabreil Jamali	Islamic Azad University, Iran
Mohammad Hadi Zahedi	Ferdowsi University of Mashhad, Iran
Mohammad Hajjar	Lebanese University, Lebanon
Mohammad Kaghazgaran	Islamic Azad University, Iran
Mohammad Mehdi Farhangia	Universiti Teknologi Malaysia, Malaysian
Mohammad Momani	University of technology Sydney, Australia
Mohammad Talib	University of Botswana, Botswana
Mohammad Zaidul Karim	Daffodil International University, Bangladesh
Mohammed Feham	University of Tlemcen, Algeria
Mohammed M. Alkhawlani	University of Science and Technology, Yemen
Mohsen Sharifi	Iran University of Science and Technology, Iran
Muhammad Sajjadur Rahim	University of Rajshahi, Bangladesh
Murty	Ch A S, JNTU, Hyderabad
Murugan D.	Manonmaniam Sundaranar University, India
Mydhili Nair	M S Ramaiah Institute of Technology, India
N. Krishnan	Manonmaniam Sundaranar University, India
Nabendu Chaki	University of Calcutta, India

Nadine Akkari	King abdulaziz University, Saudi Arabia
Naohiro Ishii	Aichi Institute of Technology, Japan
Nasrollah M. Charkari	Tarbiat Modares University, Iran
Natarajan Meghanathan	Jackson State University, USA
Nicolas Sklavos	Technological Educational Institute of Patras, Greece
Nidaa Abdual Muhsin Abbas	University of Babylon, Iraq
Nour Eldin Elmadany	Arab Acadmy for Science and Technology, Egypt
Ognjen Kuljaca	Alcorn State University, USA
Olakanmi Oladayo	University of Ibadan, Nigeria
Omar Almomani	Universiti Utara Malaysia, Malaysia
Orhan Dagdeviren	Izmir University, Turkey
Osman B. Ghazali	Universiti Utara Malaysia, Malaysia
Othon Marcelo Nunes Batista	Universidade Salvador, Brazil
Padmalochan Bera	Indian Institute of Technology, Kharagpur, India
Partha Pratim Bhattacharya	Mody Institute of Technology & Science, India
Patricia Marcu	Leibniz Supercomputing Centre, Germany
Patrick Seeling	University of Wisconsin - Stevens Point, USA
R. Thandeeswaran	VIT University, India
Phan Cong Vinh	London South Bank University, UK
Pinaki Sarkar	Jadavpur University, India
Polgar Zsolt Alfred	Technical University of Cluj Napoca, Romania
Ponpit Wongthongtham	Curtin University of Technology, Australia
Quan (Alex) Yuan	University of Wisconsin-Stevens Point, USA
Rafael Timoteo	University of Brasilia - UnB, Brazil
Raied Salman	Virginia Commonwealth University, USA
Rajendra Akerkar	Technomathematics Research Foundation, India
Rajeswari Balasubramaniam	Dr. MGR University, India
Rajkumar Kannan	Bishop Heber College, India
Rakhesh Singh Kshetrimayum	Indian Institute of Technology, Guwahati, India
Raman Maini	Punjabi University, India
Ramayah Thurasamy	Universiti Sains Malaysia, Malaysia
Ramayah	Universiti Sains Malaysia, Malaysia
Ramin karimi	University Technology Malaysia
Razvan Deaconescu	University Politehnica of Bucharest, Romania
Reena Dadhich	Govt. Engineering College Ajmer
Reshmi Maulik	University of Calcutta, India
Reza Ebrahimi Atani	University of Guilan, Iran
Rituparna Chaki	West Bengal University of Technology, India
Robert C. Hsu	Chung Hua University, Taiwan
Roberts Masillamani	Hindustan University, India
Rohitha Goonatilake	Texas A&M International University, USA
Rushed Kanawati	LIPN - Universite Paris 13, France
S. Geetha	Anna University - Tiruchirappalli, India
S. Hariharan	B.S. Abdur Rahman University, India

S. Venkatesan	University of Texas at Dallas - Richardson, USA
S.A.V. Satyamurty	Indira Gandhi Centre for Atomic Research, India
S. Arivazhagan	Mepco Schlenk Engineering College, India
S. Li	Swansea University, UK
S. Senthil Kumar	Universiti Sains Malaysia, Malaysia
Sajid Hussain	Acadia University, Canada
Salah M. Saleh Al-Majeed	University of Essex, United Kingdom
Saleena Ameen	B.S.Abdur Rahman University, India
Salem Nasri	ENIM, Monastir University, Tunisia
Salim Lahmiri	University of Qubec at Montreal, Canada
Salini P.	Pondichery Engineering College, India
Salman Abdul Moiz	Centre for Development of Advanced Computing, India
Samarendra Nath Sur	Sikkim Manipal University, India
Sami Ouali	ENSI, Compus of Manouba, Manouba, Tunisia
Samiran Chattopadhyay	Jadavpur University, India
Samodar reddy	India school of mines , India
Samuel Falaki	Federal University of Technology-Akure, Nigeria
Sanjay Singh	Manipal Institute of Technology, India
Sara Najafzadeh	University Technology Malaysia
Sarada Prasad Dakua	IIT-Bombay, India
Sarmistha Neogy	Jadavpur University, India
Satish Mittal	Punjabi University, India
S.C. SHARMA	IIT - Roorkee, India
Seetha Maddala	CBIT, Hyderabad
Selma Boumerdassi	Cnam/Cedric, France
Sergio Ilarri	University of Zaragoza, Spain
Serguei A. Mokhov	Concordia University, Canada
Shaoen Wu	The University of Southern Mississippi, USA
Sharvani G.S.	RV College of Engineering, Inida
Sherif S. Rashad	Morehead State University, USA
Shin-ichi Kuribayashi	Seikei University, Japan
Shivan Haran	Arizona state University, USA
Shobha Shankar	Vidya vardhaka College of Engineering, India
Shrikant K. Bodhe	Bosh Technologies, India
Shriram Vasudevan	VIT University, India
Shrirang Ambaji Kulkarni	National Institute of Engineering, India
Shubhamoy Dey	Indian Institute of Management Indore, India
Solange Rito Lima	University of Minho, Portugal
Souad Zid	National Engineering School of Tunis, Tunisia
Soumyabrata Saha	Guru Tegh Bahadur Institute of Technology, India
Sridharan	CEG Campus - Anna University, India
Sriman Narayana Iyengar	VIT University, India
Srinivasulu Pamidi	V R Siddhartha Engineering College Vijayawada, India

Sriram Maturi	Osmania University, India
Subhabrata Mukherjee	Jadavpur University, India
Subir Sarkar	Jadavpur University, India
Sundarapandian Vaidyanathan	VelTech Dr. RR & Dr. SR Technical University, India
Sunil Singh	Bharati vidyapeeth's College of Engineering, India
Sunilkumar S. Manvi	REVA Institute of Technology and Management Kattigenhalli, India
SunYoung Han	Konkuk University, South Korea
Susana Sargento	University of Aveiro, Portugal
Swarup Mitra	Jadavpur University, Kolkata, India
T. Ambaji Venkat Narayana Rao	Hyderabad Institution of Technology and Management , India
T.G. Basavaraju	National Institute of Technology Karnataka (NITK), India
Thomas Yang	Embry Riddle Aeronautical University, USA
Tri Kurniawan Wijaya	Technische Universitat Dresden, Germany
Tsung Teng Chen	National Taipei Univ., Taiwan
Utpal Biswas	University of Kalyani, India
V.M. Pandharipande	Dr. Babasaheb Ambedkar Marathwada University, India
Valli Kumari Vatsavayi	AU College of Engineering, India
Vijayalakshmi S.	VIT University, India
Virgil Dobrota	Technical University of Cluj-Napoca, Romania
Vishal Sharma	Metanoia Inc., USA
Wei Jie	University of Manchester, UK
Wichian Sittiprapaporn	Mahasarakham University, Thailand
Wided Oueslati	l'institut Superieur de Gestion de Tunis, Tunisia
William R. Simpson	Institute for Defense Analyses, USA
Wojciech Mazurczyk	Warsaw University of Technology, Poland
Xiaohong Yuan	North Carolina A & T State University, USA
Xin Bai	The City University of New York, USA
Yahya Slimani	Faculty of Sciences of Tunis, Tunisia
Yannick Le Moullec	Aalborg University, Denmark
Yaser M. Khamayseh	Jordan University of Science and Technology, Jordan
Yedehalli Kumara Swamy	Dayanand Sagar College of Engineering, India
Yeong Deok Kim	Woosong University, South Korea
Yogeshwar Kosta	Marwadi Education Foundations Group of Institutions, India
Yuh-Shyan Chen	National Taipei University, Taiwan
Yung-Fa Huang	Chaoyang University of Technology, Taiwan
Zaier Aida	National Engeneering School of GABES, Tunisia
Zakaria Moudam	Université sidi mohammed ben Abdellah, Morocco
Zuqing Zhu	Cisco Systems, USA

External Reviewers

A. Kannan	K.L.N. College of Engineering, India
Martin	Sri Manakula Vinayagar Engineering College, India
Abhishek Samanta	Jadavpur University, Kolkata, India
Ayman Khalil	Institute of Electronics and Telecommunications of Rennes, France
Cauvery Giri	RVCE, India
Ch. V. Rama Rao	Gudlavalleru Engineering College, India
Chandra Mohan	Bapatla Engineering College, India
E.P. Ephzibah	VIT University-Vellore, India
Hameem Shanavas	Vivekananda Institute of Technolgy, India
Kota Sunitha	G. Narayanamma Institute of Technology and Science, Hyderabad
Kunjal B. Mankad	ISTAR, Gujarat, India
Lakshmi Rajamani	Osmania University, India
Lavanya	Blekinge Institute of Technology, Sweden
M.P. Singh	National Institute of Technology, Patna
M. Tariq Banday	University of Kashmir, India
M.M.A. Hashem	Khulna University of Engineering and Technology, Bangladesh
Mahalinga V. Mandi	Dr. Ambedkar Institute of Technology, Bangalore, Karnataka, India
Mahesh Goyani	G H Patel College of Engineering and Technology, India
Maragathavalli P.	Pondicherry Engineering College, India
M.P. Singh	National Institute of Technology, Patna
M. Tariq Banday	University of Kashmir, India
M.M.A. Hashem	Khulna University of Engineering and Technology, Bangladesh
Mahalinga V. Mandi	Dr. Ambedkar Institute of Technology, India
Monika Verma	Punjab Technical University, India
Moses Ekpenyong	University of Uyo, Nigeria
Mini Patel	Malwa Institute of Technology, India
N. Kaliammal	NPR College of Engg &Tech, India
N. Adhikari	Biju Pattnaik University of Technology, India
N.K. Choudhari	Bhagwati Chaturvedi College of Engineering, India
Naga Prasad Bandaru	PVP Siddartha Institute of Technology, India
Nagamanjula Prasad	Padmasri Institute of Technology, India
Nagaraj Aitha	I.T, Kamala Institute of Tech & Science, India
Nana Patil	NIT Surat, Gujrat
Nitiket N. Mhala	B.D. College of Engineering - Sewagram, India
P. Ashok Babu	Narsimhareddy Engineering College, India
P. Sheik Abdul Khader	B.S. Abdur Rahman University, India

Pesn Krishna Prasad	Aditya Engineering College, India
Partha Pratim Bhattacharya	Mody Institute of Technology & Science, India
Pappa Rajan	Anna University, India
Pooja Jain	JUIT Waknaghat, India
Prabu Dorairaj	NetApp Inc, India
Pradeepini Gera	Jawaharlal Nehru Technological University, India
Pravin P. Karde	HVPM's College of Engineering & Technology - Amravati, India
Premanand K. Kadbe	Vidya Pratishthan's College of Engineering, India
Priti Sajja	Sardar Patel University, India
R. Baskaran	Anna University - Chennai, India
R. Murali	Dr. Ambedkar Institute of Technology, Bangalore
R.M. Suresh	Mysore University
Rabindranath Bera	Sikkim Manipal Inst. of Technol., India
Rajashree Biradar	Ballari Institute of Technology and Management, India
Rajesh Kumar Krishnan	Bannari Amman Inst. of Technol., India
Rajesh Kumar P.	The Best International, Australia
Rajeshwari Hegde	BMS College of Engineering, India
Rajagopal Palsonkennedy	Dr. MGR University, India
Selvakumar Ramachandran	Blekinge Institute of Technology, Sweden

Contents

Computing and Information Technology Algorithms and Applications

Analysis, Control and Synchronization of Hyperchaotic Zhou System via Adaptive Control 1
Sundarapandian Vaidyanathan

Secured Ontology Matching Using Graph Matching 11
K. Manjula Shenoy, K.C. Shet, U. Dinesh Acharya

Image Restoration Using Knowledge from the Image 19
S. Padmavathi, K.P. Soman, R. Aarthi

Harmony-Based Feature Weighting to Improve the Nearest Neighbor Classification 27
Ali Adeli, Mehrnoosh Sinaee, M. Javad Zomorodian, Ali Hamzeh

Effectiveness of Different Partition Based Clustering Algorithms for Estimation of Missing Values in Microarray Gene Expression Data 37
Shilpi Bose, Chandra Das, Abirlal Chakraborty, Samiran Chattopadhyay

MACREE – A Modern Approach for Classification and Recognition of Earthquakes and Explosions 49
K. Vijay Krishnan, S. Viginesh, G. Vijayraghavan

Building Concept System from the Perspective of Development of the Concept 57
Cheng Xian-yi, Shen Xue-hua, Shi Quan

E-Government Adoption: A Conceptual Demarcation 67
Rahmath Safeena, Abdullah Kammani

Expert Webest Tool: A Web Based Application, Estimate the Cost and Risk of Software Project Using Function Points 77
Ajay Jaiswal, Meena Sharma

A Web-Based Adaptive and Intelligent Tutor by Expert Systems 87
Hossein Movafegh Ghadirli, Maryam Rastgarpour

Traffic Light Synchronization 97
K. Sankara Nayaki, N.U. Nithin Krishnan, Vivek Joby, R. Sreelakshmi

Partition Sort versus Quick Sort: A Comparative Average Case Analysis with Special Emphasis on Parameterized Complexity 107
Niraj Kumar Singh, Soubhik Chakraborty

Removal of Inconsistent Training Data in Electronic Nose Using Rough Set 115
Anil Kumar Bag, Bipan Tudu, Nabarun Bhattacharyya, Rajib Bandyopadhyay

A Decentralised Routing Algorithm for Time Critical Applications Using Wireless Sensor Networks 125
R.A. Roseline, P. Sumathi

Application Security in Mobile Devices Using Unified Communications 135
Ramakrishna Josyula

Iterative Image Fusion Using Fuzzy Logic with Applications 145
Srinivasa Rao Dammavalam, Seetha Maddala, M.H.M. Krishna Prasad

A Review on Clustering of Web Search Result 153
Mansaf Alam, Kishwar Sadaf

Application of Information Theory for Understanding of HLA Gene Regulation in Leukemia 161
Durjoy Majumder

Improving Security in Digital Images through Watermarking Using Enhanced Histogram Modification 175
Vidya Hari, A. Neela Madheswari

Survey on Co-operative P2P Information Exchange in Large P2P Networks 181
S. Nithya, K. Palanivel

SNetRS: Social Networking in Recommendation System 195
Jyoti Pareek, Maitri Jhaveri, Abbas Kapasi, Malhar Trivedi

Linguistic Conversion of Syntactic to Semantic Web Page 207
G. Nagarajan, K.K. Thyagarajan

Two-Stage Rejection Algorithm to Reduce Search Space for Character Recognition in OCR 215
Srivardhini Mandipati, Gottumukkala Asisha, S. Preethi Raj, S. Chitrakala

Transforming Election Polling from Electronic Voting to Cloud as a Software Service in India 225
P. Vidhya

Area Efficient Architecture for Frequency Domain Multi Channel Digital Down Conversion for Randomly Spaced Signals 233
Latha Sahukar, M. Madhavi Latha

New Distances for Improving Progressive Alignment Algorithm 243
Ahmed Mokaddem, Mourad Elloumi

Service-Oriented Architecture (SOA) and Semantic Web Services for Web Portal Integration 253
T.G.K. Vasista, Mohammed A.T. AlSudairi

Evolutionary Multi-Objective Optimization for Data-Flow Testing of Object-Oriented Programs 263
P. Maragathavalli, S. Kanmani

The Simulation of the Assembling Production Process 273
Róbert Pauliček, Tomáš Haluška, Pavel Važan

Parallel Performance of Numerical Algorithms on Multi-core System Using OpenMP 279
Sanjay Kumar Sharma, Kusum Gupta

Efficiency Improvement of a-Si:H/μc-Si:H Tandem Solar Cells by Adding a-SiOx:H Layer between Ag Nano Particles and the Active Layer 289
Ozcan Abdulkadir, Dhinaharan Nagamalai

Unsupervised Hidden Topic Framework for Extracting Keywords (Synonym, Homonym, Hyponymy and Polysemy) and Topics in Meeting Transcripts 299
J.I. Sheeba, K. Vivekanandan, G. Sabitha, P. Padmavathi

Comprehensive Study of Estimation of Path Duration in Vehicular Ad Hoc Network 309
R.S. Raw, Vikas Toor, N. Singh

Securing DICOM Format Image Archives Using Improved Chaotic Cat Map Method 319
M. Arun Fera, Suresh Jaganathan

Brain Tumor Segmentation Using Genetic Algorithm and Artificial Neural Network Fuzzy Inference System (ANFIS) 329
Minakshi Sharma, Sourabh Mukharjee

Model Oriented Security Requirements Engineering (MOSRE) Framework for Web Applications 341
P. Salini, S. Kanmani

Axial T2 Weighted MR Brain Image Retrieval Using Moment Features 355
Abraham Varghese, Reji Rajan Varghese, Kannan Balakrishnan, J.S. Paul

Securing Data from Black Hole Attack Using AODV Routing for Mobile Ad Hoc Networks 365
V. Kamatchi, Rajeswari Mukesh, Rajakumar

ISim: A Novel Power Aware Discrete Event Simulation Framework for Dynamic Workload Consolidation and Scheduling in Infrastructure Clouds 375
R. Jeyarani, N. Nagaveni, S. Srinivasan, C. Ishwarya

Processing RDF Using Hadoop 385
Mehreen Ali, K. Sriram Bharat, C. Ranichandra

An Ant Colony Optimization Based Load Sharing Technique for Meta Task Scheduling in Grid Computing 395
T. Kokilavani, D.I. George Amalarethinam

Optimum Sub-station Positioning Using Hierarchial Clustering 405
Shabbiruddin, Sandeep Chakravorty, Amitava Ray

Comparison of PCA, LDA and Gabor Features for Face Recognition Using Fuzzy Neural Network 413
Dhanya S. Pankaj, M. Wilscy

Investigations on Object Constraints in Unified Software Development Process 423
Meena Sharma, Rajeev G. Vishwakarma

Vehicle Safety Device (Airbag) Specific Classification of Road Traffic Accident Patterns through Data Mining Techniques 433
S. Shanthi, R. Geetha Ramani

Design of Video Conferencing Solution in Grid 445
Soumilla Sen, Ajanta De Sarkar

Survey of Computer-Aided Diagnosis of Thyroid Nodules in Medical Ultrasound Images 459
Deepika Koundal, Savita Gupta, Sukhwinder Singh

A Brief Review of Data Mining Application Involving Protein Sequence Classification 469
Suprativ Saha, Rituparna Chaki

Hybrid Algorithm for Job Scheduling: Combining the Benefits of ACO and Cuckoo Search 479
R.G. Babukarthik, R. Raju, P. Dhavachelvan

Hybrid Ant Colony Optimization and Cuckoo Search Algorithm for Job Scheduling 491
R. Raju, R.G. Babukarthik, P. Dhavachelvan

Particle Based Fluid Animation Using CUDA 503
Uday A. Nuli, P.J. Kulkarni

A Novel Model for Encryption of Telugu Text Using Visual Cryptography Scheme 511
G. Lakshmeeswari, D. Rajya Lakshmi, Y. Srinivas, G. Hima Bindu

Frequent Queries Selection for View Materialization 521
T.V. Vijay Kumar, Gaurav Dubey, Archana Singh

HURI – A Novel Algorithm for Mining High Utility Rare Itemsets 531
Jyothi Pillai, O.P. Vyas, Maybin Muyeba

SMOTE Based Protein Fold Prediction Classification 541
K. Suvarna Vani, S. Durga Bhavani

Conceptual Application of List Theory to Data Structures 551
B.K. Tripathy, S.S. Gantayat

A Comprehensive Study on Multifactor Authentication Schemes 561
Kumar Abhishek, Sahana Roshan, Prabhat Kumar, Rajeev Ranjan

Quality Assessment Based Fingerprint Segmentation 569
Kumud Arora, Poonam Garg

Improved Algorithms for Anonymization of Set-Valued Data 581
B.K. Tripathy, A. Jayaram Reddy, G.V. Manusha, G.S. Mohisin

An Efficient Flash Crowd Attack Detection to Internet Threat Monitors (ITM) Using Honeypots 595
K. Munivara Prasad, M. Ganesh Karthik, E.S. Phalguna Krishna

Hierarchical Directed Acyclic Graph (HDAG) Based Preprocessing Technique for Session Construction 611
S. Chitra, B. Kalpana

Extending Application of Non-verbal Communication to Effective Requirement Elicitation 623
Md. Rizwan Beg, Md. Muqeem, Md. Faizan Farooqui

A Critical Review of Migrating Parallel Web Crawler 631
Md. Faizan Farooqui, Md. Rizwan Beg, Md. Qasim Rafiq

Parallel Character Reconstruction Expending Compute Unified Device Architecture 639
Anita Pal, Kamal Kumar Srivastava, Atul Kumar

A Novel Algorithm for Obstacle Aware RMST Construction during Routing in 3D ICs 649
Prasun Ghosal, Satrajit Das, Arindam Das

Digital Image Processing and Pattern Recognition

Study of the EEG Signals of Human Brain for the Analysis of Emotions 659
Ashish R. Panat, Anita S. Patil

A Dynamic Approach for Mining Generalised Sequential Patterns in Time Series Clinical Data Sets 667
M. Rasheeda Shameem, M. Razia Naseem, N.K. Subanivedhi, R. Sethukkarasi

Motion Correction in Physical Action of Human Body Applied to Yogasana 675
Geetanjali Kale, Varsha Patil

Improvement the Bag of Words Image Representation Using Spatial Information 681
Mohammad Mehdi Farhangi, Mohsen Soryani, Mahmood Fathy

Left Object Detection with Reduced False Positives in Real-Time 691
Aditya Piratla, Monotosh Das, Jayalakshmi Surendran

Adaptive Fusion Based Hybrid Denoising Method for Texture Images 699
Preety D. Swami, Alok Jain

An Efficient De-noising Technique for Fingerprint Image Using Wavelet Transformation 709
Ashish Kumar Dass, Rabindra Kumar Shial

Detection of Optic Disc by Line Filter Operator Approach in Retinal Images 719
R. Murugan, Reeba Korah

An Extension of FFT Based Image Registration 729
P. Thangavel, R. Kokila

Scene Text Extraction from Videos Using Hybrid Approach 739
A. Thilagavathy, K. Aarthi, A. Chilambuchelvan

MSB Based New Hybrid Image Compression Technique for Wireless Transmission 749
S.N. Muralikrishna, Meghana Ajith, K.B. Ajitha Shenoy

Multi-temporal Satellite Image Analysis Using Unsupervised Techniques . . . 757
C.S. Arvind, Ashoka Vanjare, S.N. Omkar, J. Senthilnath, V. Mani, P.G. Diwakar

Separable Discrete Hartley Transform Based Invisible Watermarking for Color Image Authentication (SDHTIWCIA) 767
J.K. Mandal, S.K. Ghosal

Normalised Euclidean Distance Based Image Retrieval Using Coefficient Analysis 777
Nilofar Khan, Wasim Khan

SoC Modeling for Video Coding with Superscalar Projection 787
S.K. Fairooz, B.K. Madhavi

Invisible Image Watermarking Using Z Transforms (IIWZT) 797
J.K. Mandal, Rakesh Kanji

Region Identification of Infected Rice Images Using the Concept of Fermi Energy 805
Santanu Phadikar, Jaya Sil, Asit KumarDas

Unsymmetrical Trimmed Midpoint as Detector for Salt and Pepper Noise Removal 813
K. Vasanth, V. Jawahar Senthil Kumar, V. Elanangai

Identification of Handwritten Text in Machine Printed Document Images 823
Sandipan Banerjee

Optimization of Integration Weights for a Multibiometric System with Score Level Fusion 833
S.M. Anzar, P.S. Sathidevi

A Two Stage Combinational Approach for Image Encryption 843
H.S. Sharath Kumar, H.T. Panduranga, S.K. Naveen Kumar

Wavelet SIFT Feature Descriptors for Robust Face Recognition 851
Nerella Arun Mani Kumar, P.S. Sathidevi

Image Denoising Based on Neutrosophic Wiener Filtering 861
J. Mohan, A.P. Thilaga Shri Chandra, V. Krishnaveni, Yanhui Guo

Modified Difference Expansion for Reversible Watermarking Using Fuzzy Logic Based Distortion Control 871
Hirak Kumar Maity, Santi P. Maity, Debashis Maity

A Novel Genetic Algorithm Based Data Embedding Technique in Frequency Domain Using Z Transform (ANGAFDZT) 885
J.K. Mandal, A. Khamrui, S. Chakraborty, P. Sur, S.K. Datta, I. RoyChoudhury

Speckle Noise Reduction Using Fourth Order Complex Diffusion Based Homomorphic Filter 895
Jyothisha J. Nair, V.K. Govindan

A New Hybrid Approach for Denoising Medical Images 905
Deepa Bharathi, Sumithra Manimegalai Govindan

Detection of Cerebral Aneurysm by Performing Thresholding-Spatial Filtering-Thresholding Operations on Digital Subtraction Angiogram 915
Jubin Mitra, Abhijit Chandra

Author Index 923

Analysis, Control and Synchronization of Hyperchaotic Zhou System via Adaptive Control

Sundarapandian Vaidyanathan

R & D Centre, Vel Tech Dr. RR & Dr. SR Technical University
Avadi-Alamathi Road, Avadi, Chennai-600 062, India
sundarvtu@gmail.com
http://www.vel-tech.org/

Abstract. This paper investigates the analysis, control and synchronization of the hyperchaotic Zhou system (2009) via adaptive control. First, an adaptive control scheme is derived to stabilize the hyperchaotic Zhou system with unknown parameters to its unstable equilibrium at the origin. Then an adaptive synchronization scheme is derived to achieve global chaos synchronization of the identical hyperchaotic Zhou systems with unknown parameters. The results derived for adaptive stabilization and synchronization for the hyperchaotic system are established using the Lyapunov stability theory. Numerical simulations are shown to demonstrate the effectiveness of the adaptive control and synchronization schemes derived in this paper.

Keywords: Adaptive control, hyperchaos, synchronization, hyperchaotic Zhou system.

1 Introduction

Hyperchaotic system is defined as a chaotic system with more than one positive Lyapunov exponent. Hyperchaotic system has the characteristics of high capacity, high security and high efficiency. Typical examples of hyperchaotic systems are hyperchaotic Rössler system [1], hyperchaotic Lorenz-Haken system [2] and hyperchaotic Chua's circuit [3].

The problem of controlling a chaotic system was introduced by Ott *et al.* ([4], 1990). The control of chaotic systems is basically to design state feedback control laws that stabilizes the chaotic systems around the unstable equilibrium points. Active control method is used when the system parameters are known and adaptive control method is used when some or all of the system parameters are unknown ([4]-[6]).

Chaos synchronization is a phenomenon that may occur when two or more chaotic oscillators are coupled or when a chaotic oscillator drives another chaotic oscillator. In most of the chaos synchronization approaches, the *master-slave* or *drive-response* formalism is used. If a particular chaotic system is called the *master* or *drive* system and another chaotic system is called the *slave* or *response* system, then the idea of chaos synchronization is to use the output of the master system to control the slave system so that the output of the slave system tracks the output of the master system asymptotically.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 1–10.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Since the pioneering work by Pecora and Carroll ([7], 1990), several approaches have been proposed for chaos synchronization such as the active control method ([8]-[9]), adaptive control method ([10]-[12]), sampled-data control method [13], backstepping method [14], sliding mode control method ([15]-[16]), etc.

This paper investigates the analysis, control and synchronization for the hyperchaotic Zhou system (Zhou *et al.* [17], 2009). First, we derive adaptive feedback control for the hyperchaotic Zhou system about its unstable equilibrium at the origin. Then we derive adaptive synchronization scheme for the identical hyperchaotic Zhou systems. The adaptive control and synchronization results derived in this paper are established using Lyapunov stability theory [18].

This paper has been organized as follows. In Section 2, we give a system description and qualitative analysis of the hyperchaotic Zhou system. In Section 3, we derive results for the adaptive control of the hyperchaotic Zhou system with unknown parameters. In Section 4, we derive results for the adaptive synchronization of the identical hyperchaotic Zhou systems with unknown parameters. In Section 5, we summarize the main results obtained in this paper.

2 Analysis of the Hyperchaotic Zhou System

The hyperchaotic Zhou system ([17], 2009) is described by the 4D dynamics

$$\begin{aligned}\dot{x}_1 &= a(x_2 - x_1) + x_4\\ \dot{x}_2 &= cx_2 - x_1x_3\\ \dot{x}_3 &= -bx_3 + x_1x_2\\ \dot{x}_4 &= dx_1 + 0.5x_2x_3\end{aligned} \tag{1}$$

where x_1, x_2, x_3, x_4 are the state variables of the system and a, b, c, d are constant, positive parameters of the system.

The system (1) is symmetrical about the x_3-axis because it is invariant under the coordinate transformation

$$(x_1, x_2, x_3, x_4) \to (-x_1, -x_2, x_3, -x_4)$$

and the phase portrait of the system (1) in $x_1x_2x_4$-three dimensional space is symmetrical about the origin.

The system (1) is *hyperchaotic* when

$$a = 35, \;\; b = 3, \;\; c = 12 \;\text{ and }\; 0 < d \leq 34.8 \tag{2}$$

Figure 1 describes the phase portrait of the hyperchaotic system (1) where the parameters are takes as in (2) with $d = 1$.

Obviously, the hyperchaotic system (1) has only an equilibrium point at the origin.

The linearization matrix of the system (1) at the origin is given by

$$A = \begin{bmatrix} -a & a & 0 & 1\\ 0 & c & 0 & 0\\ 0 & 0 & -b & 0\\ d & 0 & 0 & 0\end{bmatrix}$$

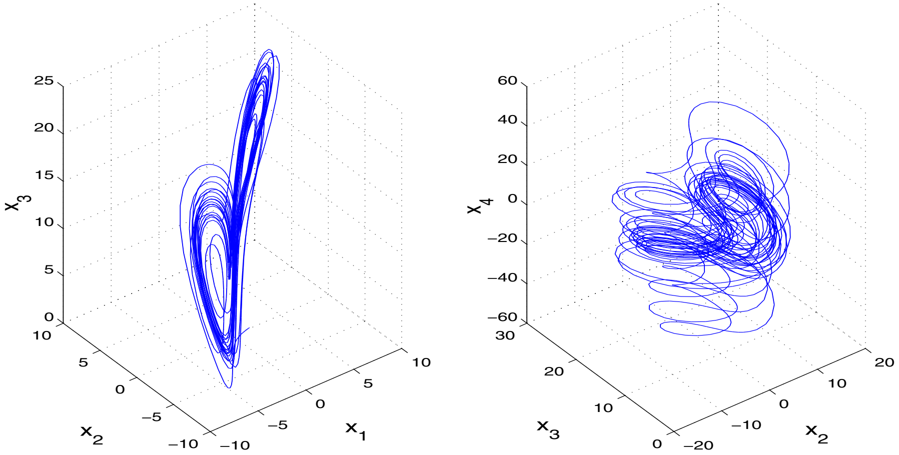


Fig. 1. Phase Portrait of the Hyperchaotic Zhou System

The eigenvalues of A are

$$\lambda_1 = c, \lambda_2 = -b, \lambda_3 = \frac{-a + \sqrt{a^2 + 4d}}{2}, \lambda_4 = \frac{-a - \sqrt{a^2 + 4d}}{2}$$

Since $\lambda_1 = c > 0$, it follows by the Lyapunov stability theory [18] that the origin is an unstable equilibrium of the system (1).

3 Adaptive Control of the Hyperchaotic Zhou System

3.1 Main Results

In this section, we discuss the adaptive controller design for globally stabilizing the hyperchaotic Zhou system (2009), when the parameter values are unknown.

Thus, we consider the controlled hyperchaotic Zhou system described by the dynamics

$$\begin{aligned}\dot{x}_1 &= a(x_2 - x_1) + x_4 + u_1 \\ \dot{x}_2 &= cx_2 - x_1 x_3 + u_2 \\ \dot{x}_3 &= -bx_3 + x_1 x_2 + u_3 \\ \dot{x}_4 &= dx_1 + 0.5 x_2 x_3 + u_4\end{aligned} \tag{3}$$

where u_1, u_2, u_3, u_4 are feedback controllers to be designed using the states x_1, x_2, x_3, x_4 and estimates $\hat{a}, \hat{b}, \hat{c}, \hat{d}$ of the unknown system parameters a, b, c, d of the system.

Next, we consider the following adaptive control functions

$$\begin{aligned}u_1 &= -\hat{a}(x_2 - x_1) - x_4 - k_1 x_1 \\ u_2 &= -\hat{c} x_2 + x_1 x_3 - k_2 x_2 \\ u_3 &= \hat{b} x_3 - x_1 x_2 - k_3 x_3 \\ u_4 &= -\hat{d} x_1 - 0.5 x_2 x_3 - k_4 x_4\end{aligned} \tag{4}$$

where $\hat{a}$, $\hat{b}$, $\hat{c}$ and $\hat{d}$ are the estimates of the system parameters a, b, c and d, respectively, and $k_i, (i = 1, 2, 3, 4)$ are positive constants.

Substituting the control law (4) into the plant equation (3), we obtain

$$\begin{aligned}\dot{x}_1 &= (a - \hat{a})(x_2 - x_1) - k_1 x_1\\ \dot{x}_2 &= (c - \hat{c})x_2 - k_2 x_2\\ \dot{x}_3 &= -(b - \hat{b})x_3 - k_3 x_3\\ \dot{x}_4 &= (d - \hat{d})x_1 - k_4 x_4\end{aligned} \tag{5}$$

We define the parameter estimation error as

$$e_a = a - \hat{a}, \;\; e_b = b - \hat{b}, \;\; e_c = c - \hat{c}, \;\; e_d = d - \hat{d} \tag{6}$$

Using (6), the state dynamics (3) can be simplified as

$$\begin{aligned}\dot{x}_1 &= e_a(x_2 - x_1) - k_1 x_1\\ \dot{x}_2 &= e_c x_2 - k_2 x_2\\ \dot{x}_3 &= -e_b x_3 - k_3 x_3\\ \dot{x}_4 &= e_d x_1 - k_4 x_4\end{aligned} \tag{7}$$

We use Lyapunov approach for the derivation of the update law for adjusting the parameter estimates $\hat{a}$, $\hat{b}$, $\hat{c}$ and $\hat{d}$.

Consider the quadratic Lyapunov function defined by

$$V(x_1, x_2, x_3, x_4, e_a, e_b, e_c, e_d) = \frac{1}{2}\left(x_1^2 + x_2^2 + x_3^2 + x_4^2 + e_a^2 + e_b^2 + e_c^2 + e_d^2\right) \tag{8}$$

which is a positive definite function on $\mathbb{R}^8$.

Note that

$$\dot{e}_a = -\dot{\hat{a}}, \;\; \dot{e}_b = -\dot{\hat{b}}, \;\; \dot{e}_c = -\dot{\hat{c}}, \;\; \dot{e}_d = -\dot{\hat{d}} \tag{9}$$

Differentiating V along the trajectories of (5) and using (9), we obtain

$$\begin{aligned}\dot{V} = &-k_1 x_1^2 - k_2 x_2^2 - k_3 x_3^2 - k_4 x_4^2 + e_a\left[x_1(x_2 - x_1) - \dot{\hat{a}}\right]\\ &+ e_b\left[-x_3^2 - \dot{\hat{b}}\right] + e_c\left[x_2^2 - \dot{\hat{c}}\right] + e_d\left[x_1 x_4 - \dot{\hat{d}}\right]\end{aligned} \tag{10}$$

In view of Eq. (10), the estimated parameters are updated by the following law:

$$\begin{aligned}\dot{\hat{a}} &= x_1(x_2 - x_1) + k_5 e_a\\ \dot{\hat{b}} &= -x_3^2 + k_6 e_b\\ \dot{\hat{c}} &= x_2^2 + k_7 e_c\\ \dot{\hat{d}} &= x_1 x_4 + k_8 e_d\end{aligned} \tag{11}$$

where k_5, k_6, k_7 and k_8 are positive constants.

Next, we prove the following result.

Theorem 1. *The hyperchaotic Zhou system (3) with unknown parameters is globally and exponentially stabilized by the adaptive control law (4), where the update law for the parameters is given by (11) and* $k_i, (i = 1, 2, \ldots, 8)$ *are positive constants.*

Proof. Substituting (11) into (10), we obtain

$$\dot{V} = -k_1 x_1^2 - k_2 x_2^2 - k_3 x_3^2 - k_4 x_4^2 - k_5 e_a^2 - k_6 e_b^2 - k_7 e_c^2 - k_8 e_d^2 \qquad (12)$$

which is a negative definite function on $\mathbb{R}^8$.

Thus, by Lyapunov stability theory [18], it follows that the plant dynamics (7) is globally exponentially stable and also that the parameter estimate errors e_a, e_b, e_c, e_d converge to zero exponentially with time. □

3.2 Numerical Results

For the simulations, the fourth order Runge-Kutta method with step-size $h = 10^{-8}$ is used to solve the hyperchaotic Zhou system (3) with the adaptive control law (4) and the parameter update law (11). The parameters of the system (3) are selected as $a = 35,\ b = 3,\ c = 12$ and $d = 1$. We also take $k_i = 4$ for $i = 1, 2, \ldots, 8$.

Suppose that the initial values of the estimated parameters are

$$\hat{a}(0) = 8,\ \ \hat{b}(0) = 24,\ \ \hat{c}(0) = 30,\ \ \ \hat{d}(0) = 17$$

Suppose that we take the initial values of the states of the system (3) as

$$x_1(0) = -6,\ \ x_2(0) = 7,\ \ x_3(0) = 20,\ \ \ x_4(0) = -18$$

Figure 2 shows that the states of the closed-loop system (7) converge to the equilibrium $E_0 = (0, 0, 0, 0)$ exponentially with time. Figure 3 shows that the estimates $\hat{a}, \hat{b}, \hat{c}, \hat{d}$ converge to the system parameters a, b, c, d exponentially with time.

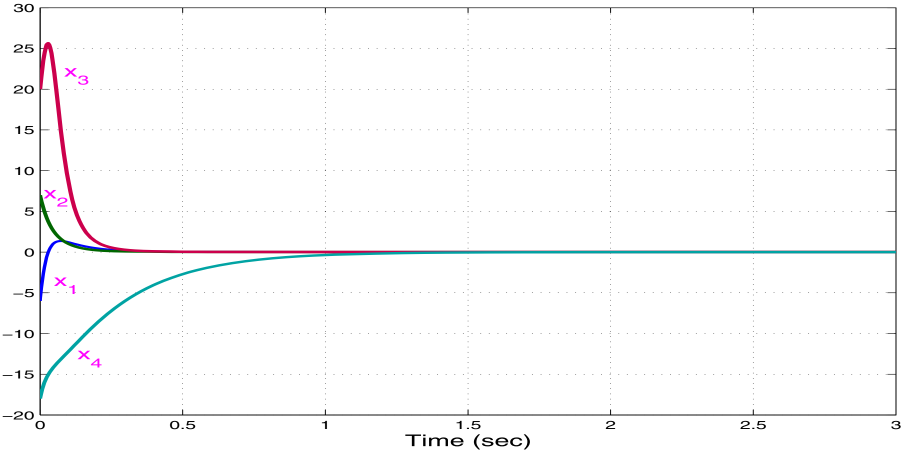


Fig. 2. Time History of the States of the Controlled Hyperchaotic Zhou System

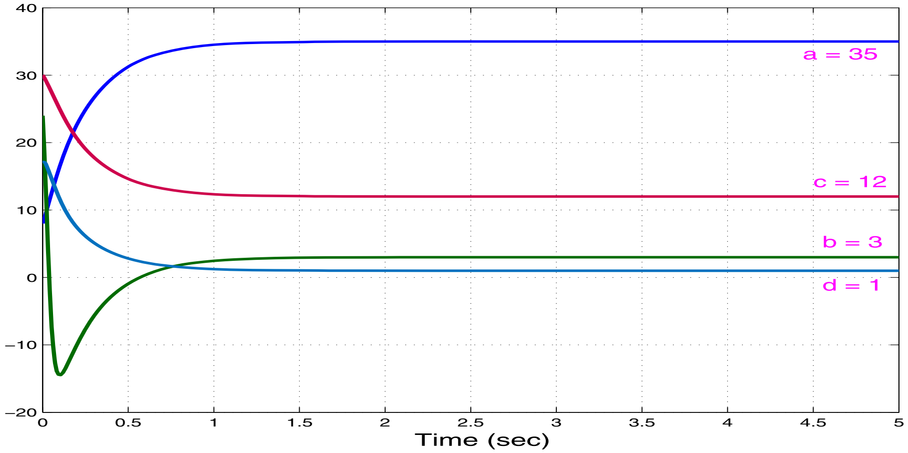


Fig. 3. Time History of the Parameter Estimates $\hat{a}(t), \hat{b}(t), \hat{c}(t), \hat{d}(t)$

4 Adaptive Synchronization of the Hyperchaotic Zhou System

4.1 Main Results

In this section, we discuss the adaptive synchronizer design for the identical hyperchaotic Zhou systems (2009) with unknown parameters.

As the master system, we consider the hyperchaotic Zhou dynamics described by

$$\begin{aligned}\dot{x}_1 &= a(x_2 - x_1) + x_4\\ \dot{x}_2 &= cx_2 - x_1x_3\\ \dot{x}_3 &= -bx_3 + x_1x_2\\ \dot{x}_4 &= dx_1 + 0.5x_2x_3\end{aligned} \tag{13}$$

where x_1, x_2, x_3, x_4 are the state variables and a, b, c, d are unknown system parameters.

As the slave system, we consider the controlled hyperchaotic Zhou dynamics described by

$$\begin{aligned}\dot{y}_1 &= a(y_2 - y_1) + y_4 + u_1\\ \dot{y}_2 &= cy_2 - y_1y_3 + u_2\\ \dot{y}_3 &= -by_3 + y_1y_2 + u_3\\ \dot{y}_4 &= dy_1 + 0.5y_2y_3 + u_4\end{aligned} \tag{14}$$

where y_1, y_2, y_3 are the state variables and u_1, u_2, u_3 are the nonlinear controllers to be designed.

The synchronization error e is defined by

$$e_i = y_i - x_i, \quad (i = 1, 2, 3, 4) \tag{15}$$

Then the error dynamics is obtained as

$$\begin{aligned} \dot{e}_1 &= a(e_2 - e_1) + e_4 + u_1 \\ \dot{e}_2 &= ce_2 - y_1y_3 + x_1x_3 + u_2 \\ \dot{e}_3 &= -be_3 + y_1y_2 - x_1x_2 + u_3 \\ \dot{e}_4 &= de_1 + 0.5(y_2y_3 - x_2x_3) + u_4 \end{aligned} \tag{16}$$

We define the adaptive synchronizing law

$$\begin{aligned} u_1 &= -\hat{a}(e_2 - e_1) - e_4 - k_1e_1 \\ u_2 &= -\hat{c}e_2 + y_1y_3 - x_1x_3 - k_2e_2 \\ u_3 &= \hat{b}e_3 - y_1y_2 + x_1x_2 - k_3e_3 \\ u_4 &= -\hat{d}e_1 - 0.5(y_2y_3 - x_2x_3) - k_4e_4 \end{aligned} \tag{17}$$

where $\hat{a}, \hat{b}, \hat{c}$ and $\hat{d}$ are estimates of the system parameters a, b, c and d, respectively, and $k_i, (i = 1, 2, 3, 4)$ are positive constants.

Substituting (17) into (16), we obtain the closed-loop error dynamics as

$$\begin{aligned} \dot{e}_1 &= (a - \hat{a})(e_2 - e_1) - k_1e_1 \\ \dot{e}_2 &= (c - \hat{c})e_2 - k_2e_2 \\ \dot{e}_3 &= -(b - \hat{b})e_3 - k_3e_3 \\ \dot{e}_4 &= (d - \hat{d})e_1 - k_4e_4 \end{aligned} \tag{18}$$

We define the parameter estimation error as

$$e_a = a - \hat{a}, \;\; e_b = b - \hat{b}, \;\; e_c = c - \hat{c}\text{and} \;\; e_d = d - \hat{d} \tag{19}$$

Substituting (19) into (18), the error dynamics (18) can be simplified as

$$\begin{aligned} \dot{e}_1 &= e_a(e_2 - e_1) - k_1e_1 \\ \dot{e}_2 &= e_ce_2 - k_2e_2 \\ \dot{e}_3 &= -e_be_3 - k_3e_3 \\ \dot{e}_4 &= e_de_1 - k_4e_4 \end{aligned} \tag{20}$$

We use Lyapunov approach for the derivation of the update law for adjusting the parameter estimates $\hat{a}$, $\hat{b}$, $\hat{c}$ and $\hat{d}$.

Consider the quadratic Lyapunov function defined by

$$V(e_1, e_2, e_3, e_4, e_a, e_b, e_c, e_d) = \frac{1}{2}\left(e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_a^2 + e_b^2 + e_c^2 + e_d^2\right) \tag{21}$$

which is a positive definite function on $\mathbb{R}^8$.

Note that

$$\dot{e}_a = -\dot{\hat{a}}, \;\; \dot{e}_b = -\dot{\hat{b}}, \;\; \dot{e}_c = -\dot{\hat{c}}, \;\; \dot{e}_d = -\dot{\hat{d}}. \tag{22}$$

Differentiating V along the trajectories of (20) and using (22), we obtain

$$\begin{aligned} \dot{V} = -k_1e_1^2 - k_2e_2^2 - k_3e_3^2 - k_4e_4^2 + e_a\left[e_1(e_2 - e_1) - \dot{\hat{a}}\right] \\ +e_b\left[-e_3^2 - \dot{\hat{b}}\right] + e_c\left[e_2^2 - \dot{\hat{c}}\right] + e_d\left[e_1e_4 - \dot{\hat{d}}\right] \end{aligned} \tag{23}$$

In view of Eq. (23), the estimated parameters are updated by the following law:

$$\begin{aligned}\dot{\hat{a}} &= e_1(e_2 - e_1) + k_5 e_a \\ \dot{\hat{b}} &= -e_3^2 + k_6 e_b \\ \dot{\hat{c}} &= e_2^2 + k_7 e_c \\ \dot{\hat{d}} &= e_1 e_4 + k_8 e_d\end{aligned} \tag{24}$$

where k_5, k_6, k_7 and k_8 are positive constants.

Theorem 2. *The identical hyperchaotic Zhou systems (13) and (14) with unknown parameters are globally and exponentially synchronized by the adaptive control law (17), where the update law for the parameters is given by (24) and* $k_i, (i = 1, 2, \ldots, 6)$ *are positive constants.*

Proof. Substituting (24) into (23), we obtain

$$\dot{V} = -k_1 e_1^2 - k_2 e_2^2 - k_3 e_3^2 - k_4 e_4^2 - k_5 e_a^2 - k_6 e_b^2 - k_7 e_c^2 - k_8 e_d^2 \tag{25}$$

which is a negative definite function on $\mathbb{R}^8$.

Thus, by Lyapunov stability theory [18], it follows that the error dynamics (20) is globally exponentially stable and also that the parameter estimate errors e_a, e_b, e_c, e_d converge to zero exponentially with time. □

4.2 Numerical Results

For the numerical simulations, the fourth order Runge-Kutta method with step-size $h = 10^{-8}$ is used to solve the hyperchaotic Zhou systems (13) and (14) with the adaptive control law (17) and the parameter update law (24).

The parameters of the hyperchaotic Zhou systems are selected as $a = 35, b = 3, c = 12$ and $d = 1$.

We also take $k_i = 4$ for $i = 1, 2, \ldots, 8$.

Suppose that the initial values of the estimated parameters are

$$\hat{a}(0) = 7, \ \ \hat{b}(0) = 16, \ \ \hat{c}(0) = 20, \ \ \hat{d}(0) = 9$$

Suppose that the initial values of the master system (13) are taken as

$$x_1(0) = 7, \ \ x_2(0) = -14, \ \ x_3(0) = 12, \ \ x_4(0) = -17$$

Suppose that the initial values of the slave system (14) are taken as

$$y_1(0) = -8, \ \ y_2(0) = 25, \ \ y_3(0) = 9, \ \ y_4(0) = 12$$

Figure 4 shows that the identical hyperchaotic Zhou systems (13) and (14) are exponentially synchronized with time. Figure 5 shows that the parameter estimates $\hat{a}, \hat{b}, \hat{c}, \hat{d}$ converge to the system parameters a, b, c, d exponentially with time.

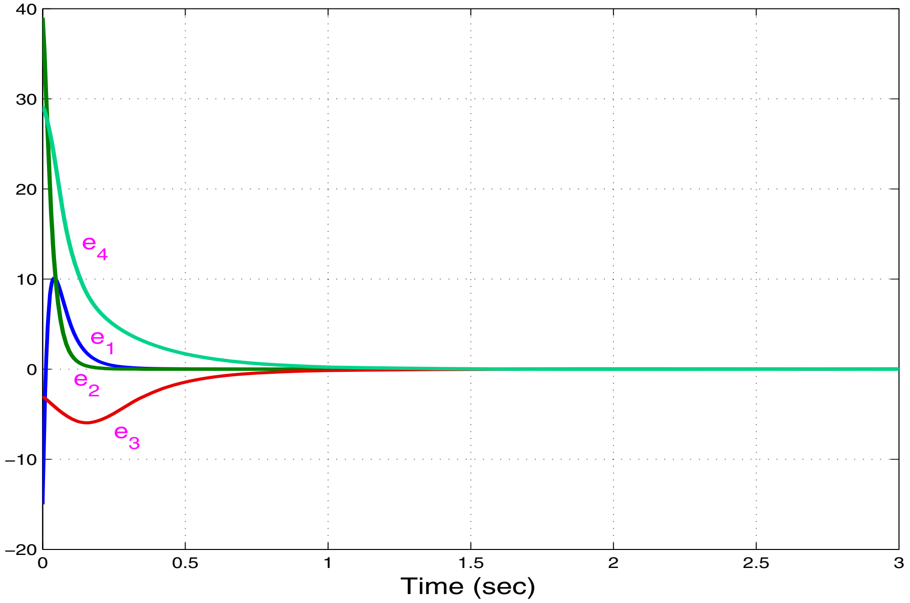


Fig. 4. Time History of the Synchronization Error for Identical Hyperchaotic Zhou Systems

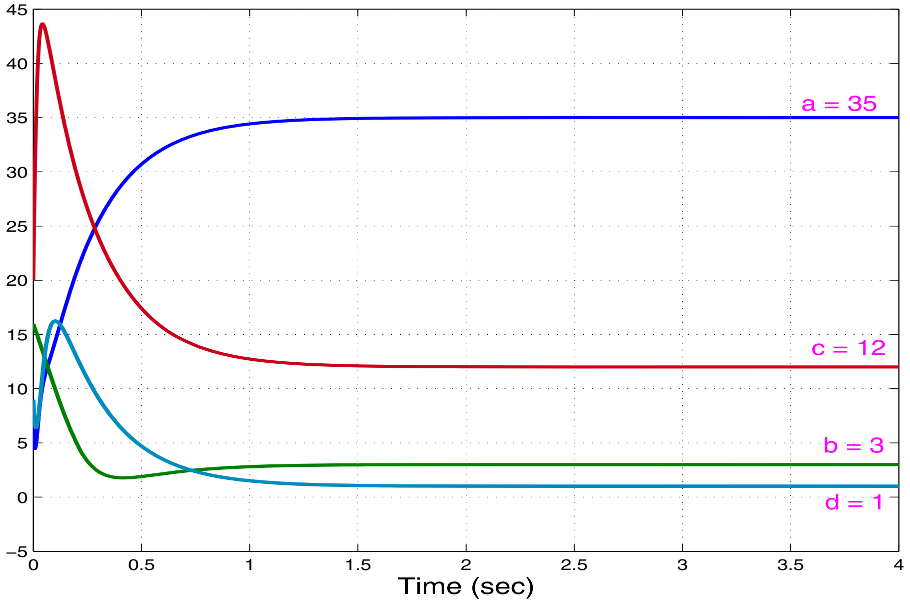


Fig. 5. Time History of the Parameter Estimates $\hat{a}(t), \hat{b}(t), \hat{c}(t), \hat{d}(t)$

5 Conclusions

In this paper, we derived results for the adaptive controller and synchronizer for the hyperchaotic Zhou system (Zhou *et al.* 2009) with unknown parameters. First, we designed an adaptive control scheme to stabilize the hyperchaotic Zhou system to its unstable equilibrium point at the origin based on the Lyapunov stability theory. Then we designed an adaptive synchronization scheme for the global chaos synchronization of the identical hyperchaotic Zhou systems with unknown parameters. Our synchronization results were established using the Lyapunov stability theory. Numerical simulations are presented to demonstrate the effectiveness of the adaptive controller and synchronizer schemes derived for the hyperchaotic Zhou system (2009).

References

1. Rössler, O.E.: An equation for hyperchaos. Phys. Lett. A 71, 155–157 (1979)
2. Ning, C.Z., Haken, H.: Detuned lasers and the complex Lorenz equations: Subcritical and supercritical Hopf bifurcations. Phys. Rev. A 41, 3826–3837 (1990)
3. Kapitaniak, T., Chua, L.O.: Hyperchaotic attractor of unidirectionally coupled Chua's circuit. Int. J. Bifurcat. Chaos 4, 477–482 (1994)
4. Ott, E., Grebogi, C., Yorke, J.A.: Controlling chaos. Phys. Rev. Lett. 64, 1196–1199 (1990)
5. Ge, S.S., Wang, C., Lee, T.H.: Adaptive backstepping control of a class of chaotic systems. Internat. J. Bifur. Chaos 10, 1149–1156 (2000)
6. Sun, M., Tian, L., Jiang, S., Xun, J.: Feedback control and adaptive control of the energy resource chaotic system. Chaos, Solitons & Fractals 32, 168–180 (2007)
7. Pecora, L.M., Carroll, T.L.: Synchronization in chaotic systems. Phys. Rev. Lett. 64, 821–824 (1990)
8. Lu, L., Zhang, C., Guo, Z.A.: Synchronization between two different chaotic systems with nonlinear feedback control. Chinese Physics 16(6), 1603–1607 (2007)
9. Sundarapandian, V.: Hybrid chaos synchronization of hyperchaotic Liu and hyperchaotic Chen systems by active nonlinear control. Internat. J. Computer Sci. Eng. Inform. Tech. 1(2), 1–14 (2011)
10. Liao, T.L., Tsai, S.H.: Adaptive synchronization of chaotic systems and its applications to secure communications. Chaos, Solitons & Fractals 11, 1387–1396 (2000)
11. Sundarapandian, V.: Adaptive synchronization of hyperchaotic Lorenz and hyperchaotic Lü systems. Internat. J. Instrument. Control Sys. 1(1), 1–18 (2011)
12. Sundarapandian, V.: Adaptive stabilization and synchronization of Lü-like chaotic attractor. Internat. J. Comp. Sci. Eng. Informat. Tech. 1(4), 15–26 (2011)
13. Yang, T., Chua, L.O.: Control of chaos using sampled-data feedback control. Internat. J. Bifur. Chaos 9, 215–219 (1999)
14. Yu, Y.G., Zhang, S.C.: Adaptive backstepping synchronization of uncertain chaotic systems. Chaos, Solitons & Fractals 27, 1369–1375 (2006)
15. Konishi, K., Hirai, M., Kokame, H.: Sliding mode control for a class of chaotic systems. Phys. Lett. A 245, 511–517 (1998)
16. Sundarapandian, V.: Global chaos synchronization of Pehlivan systems by sliding mode control. Internat. J. Comp. Sci. Eng. 3(5), 2163–2169 (2011)
17. Zhou, P., Cao, Y.X., Cheng, X.F.: A new hyperchaos system and its circuit simulation by EWB. Chinese Physics B 18(4), 1394–1398 (2009)
18. Hahn, W.: The Stability of Motion. Springer, New York (1967)

Secured Ontology Matching Using Graph Matching

K. Manjula Shenoy[1], K.C. Shet[2], and U. Dinesh Acharya[1]

[1] Department of Computer Science and Engineering, MIT, Manipal University, Manipal
{manju.shenoy,dinesh.acharya}@manipal.edu
[2] Department of Computer Engineering, NITK, Suratkal
kcshet@rediffmail.com

Abstract. Today's market evolution and high volatility of business requirements put an increasing emphasis on the ability for systems to accommodate the changes required by new organizational needs while maintaining security objectives satisfiability. This is all the more true in case of collaboration and interoperability between different organizations and thus between their information systems. Ontology mapping has been used for interoperability and several mapping systems have evolved to support the same. Usual solutions do not take care of security. That is almost all systems do a mapping of ontologies which are unsecured. We have developed a system for mapping secured ontologies using graph similarity concept. Here we give no importance to the strings that describe ontology concepts, properties etc. Because these strings may be encrypted in the secured ontology. Instead we use the pure graphical structure to determine mapping between various concepts of given two secured ontologies. The paper also gives the measure of accuracy of experiment in a tabular form in terms of precision, recall and F-measure.

1 Introduction

Researchers have developed several tools that enable organizations to share information, largely, most of these have not taken into the account the necessity of maintaining privacy and confidentiality of data and metadata of the organizations who want to share information. Consider the scenario of two different country military wanting to share information about a mission at hand while preserving the privacy of their systems. To the best of our knowledge current systems do not allow this type of information sharing.

Need for secured information sharing also exists for intra organizational information sharing too. Within the organizations different departments may use different systems which are autonomously constructed. The secure interoperability may be required here too.

Privacy should be maintained for both data and metadata. Metadata describes how data is organized (data schema), how access are controlled in the organization(the internal access control policy and role hierarchies) and the semantics of the data used in the organization(ontology).

Organizations looking to interoperate are largely using metadata like ontologies to capture the semantics of the terms used in the information sources maintained by the

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 11–18.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

organizations. Normally it has been assumed that these ontologies will be published by the organizations. Published ontologies from different organizations are mapped and matching rules are generated. Queries to information sources are rewritten using these matching rules so that vocabulary used in the query matches with the vocabulary of information source.

Unlike the traditional way some organizations may not like to publish their metadata or share it with other external users. Yet they want interoperation. In this case the privacy of the metadata must be preserved. The external user should not have access to ontologies in clear text. So ontologies may be encrypted and then published. The mapping system should now be able to recognize mapping in this encrypted ontology. Here we present one such system.

2 Related Work

The present ontology mapping systems can be classified into the following categories.

1. Word Similarity based: Here matching is performed based on similarity of words describing concepts, properties or names of concepts and properties occurring in the ontology.[4]

2. Structure based: Here structure of ontologies has been used for matching concepts.[5][6][7].

3. Instance based: These take the instances under concepts to find matching.[8]. These methods are further subdivided into Opaque and pattern based. In Opaque instance matching we use statistical properties like distribution, entropy and mutual information etc. In Pattern based method instance pattern are matched.

4. Inference Based: The semantics of concepts under ontologies are expressed as rules in a logical language and then the matching is performed using an inference engine.

There are also hybrid algorithms for matching ontologies.

[1] discusses need for secured data sharing in or among organization and [2] explains need for secured data mining. [3] proposes two methods for privacy preserving ontology matching. One of which is semi-automatic. And the other requires the dictionaries or thesauri or corpuses to be encrypted. Our method falls purely under structure based ontology matching which can be applied to encrypted ontologies. [4] defines a graph matching technique we used, in the literature.

3 Graph Matching Technique Used

3.1 Generalizing Hubs and Authorities[17]

Efficient web search engines such as Google are often based on the idea of characterizing the most important vertices in a graph representing the connections or links between pages on the web. One such method, proposed by Kleinberg [16], identifies in a set of pages relevant to a query search the subset of pages that are good *hubs* or the subset of pages that are good *authorities*. For example, for the query “university,” the home-pages of Oxford, Harvard, and other universities are good authorities, whereas web-pages that point to these home-pages are good hubs. Good hubs are pages that point to good authorities, and good authorities are pages that are pointed to by good hubs. From these implicit relations, Kleinberg derives an iterative method that assigns an "authority score" and a "hub score" to every vertex of a given graph. These scores can be obtained as the limit of a converging iterative process, which is described in section below.

Let $G = (V,E)$ be a graph with vertex set V and with edge set E and let hj and aj be the hub and authority scores of vertex j. We let these scores be initialized by some positive values and then update them simultaneously for all vertices according to the following *mutually reinforcing relation*: the hub score of vertex j is set equal to the sum of the authority scores of all vertices pointed to by j, and, similarly, the authority score of vertex j is set equal to the sum of the hub scores of all vertices pointing to j:

$$\begin{cases} h_j & \leftarrow \quad \sum_{i:(j,i)\in E} a_i, \\ a_j & \leftarrow \quad \sum_{i:(i,j)\in E} h_i. \end{cases}$$

Let B be the matrix whose entry (i, j) is equal to the number of edges between the vertices i and j in G (the *adjacency matrix* of G), and let h and a be the vectors of hub and authority scores. The above updating equations then take the simple form

$$\begin{bmatrix} h \\ a \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} h \\ a \end{bmatrix}_k, \qquad k = 0, 1, \ldots,$$

which we denote in compact form by

$$x_{k+1} = M\, x_k, \qquad k = 0, 1, \ldots,$$

Where

$$x_k = \begin{bmatrix} h \\ a \end{bmatrix}_k, \quad M = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}.$$

Notice that the matrix M is symmetric and nonnegative. We are interested only in the relative scores and we will therefore consider the *normalized* vector sequence

$$z_0 = x_0 > 0, \quad z_{k+1} = \frac{M z_k}{\|M z_k\|_2}, \quad k = 0, 1, \ldots,$$

Where$\|..\|_2$ is the Euclidean vector norm. Notice that the above matrix M has the property that

$$M^2 = \begin{bmatrix} BB^T & 0 \\ 0 & B^T B \end{bmatrix},$$

and from this equality it follows that, if the dominant invariant subspaces associated with BB^T and B^TB have dimension 1, then the normalized hub and authority scores are simply given by the normalized dominant eigenvectors of BB^T and B^TB. This is the definition used in [16] for the authority and hub scores of the vertices of G. The arbitrary choice of *z*0 = **1** made in [16] is shown here to have an extrenal norm justification. Notice that when the invariant subspace has dimension 1, then there is nothing particular about the starting vector **1**, since any other positive vector *z*0 would give the same result. We now generalize this construction. The authority score of vertex *j* of *G* can be thought of as a similarity score between vertex *j* of *G* and vertex *authority* of the graph

hub→authority

and, similarly, the hub score of vertex *j* of *G* can be seen as a similarity score between vertex *j* and vertex *hub*. The mutually reinforcing updating iteration used above can be generalized to graphs that are different from the hub–authority structure graph.

The idea of this generalization is easier to grasp with an example; we illustrate it first on the path graph with three vertices and then provide a definition for arbitrary graphs. Let *G* be a graph with edge set *E* and adjacency matrix *B* and consider the *structure graph*

1→2→3

With each vertex *j* of *G* we now associate three scores *xi*1, *xi*2, and *xi*3, one for each vertex of the structure graph. We initialize these scores with some positive value and then update them according to the following mutually reinforcing relation:

$$\begin{cases} x_{i1} \leftarrow & & \sum_{j:(i,j)\in E} x_{i2}, & \\ x_{i2} \leftarrow & \sum_{j:(j,i)\in E} x_{i1} & & +\sum_{j:(i,j)\in E} x_{i3}, \\ x_{i3} \leftarrow & & \sum_{j:(j,i)\in E} x_{i2}, & \end{cases}$$

or, in matrix form (we denote by **x***j* the column vector with entries *xij*),

$$\begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \\ \mathrm{x}_3 \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B & 0 \\ B^T & 0 & B \\ 0 & B^T & 0 \end{bmatrix} \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \\ \mathrm{x}_3 \end{bmatrix}_{k}, \qquad k = 0, 1, \ldots,$$

which we again denote *xk*+1 = *Mxk*. The situation is now identical to that of the previous example and all convergence arguments given there apply here as well. We now come to a description of the general case. Assume that we have two directed graphs *GA* and *GB* with *nA* and *nB* vertices and edge sets *EA* and *EB*. We think of *GA* as a structure graph that plays the role of the graphs *hub* −→ *authority* and 1 −→ 2 −→ 3 in the above examples. We consider real scores *xij* for *i* = 1, . . . , *nB* and *j* = 1, . . . , *nA* and simultaneously update all scores according to the following updating equations:

$$x_{ij} \leftarrow \sum_{r:(r,i)\in E_B,\ s:(s,j)\in E_A} x_{rs} + \sum_{r:(i,r)\in E_B,\ s:(j,s)\in E_A} x_{rs}.$$

This equation can be given an interpretation in terms of the product graph of *GA* and *GB*. The *product graph* of *GA* and *GB* is a graph that has *nA.nB* vertices and that has an edge between vertices (*i*1, *j*1) and (*i*2, *j*2) if there is an edge between *i*1 and *i*2 in *GA* and there is an edge between *j*1 and *j*2 in *GB*. The above updating equation is then equivalent to replacing the scores of all vertices of the product graph by the sum of the scores of the vertices linked by an outgoing or incoming edge. Equation can also be written in more compact matrix form. Let *Xk* be the *nB* × *nA* matrix of entries *xij* at iteration *k*. Then the updating equations take the simple form

$$X_{k+1} = BX_kA^T + B^TX_kA, \qquad k = 0, 1, \ldots,$$

where *A* and *B* are the adjacency matrices of *GA* and *GB*. This equation is further revised by Laure Ninove [18] as follows Where X_K is replaced by S_k

$$\frac{BS_kA^t + B^tS_kA}{\| BS_kA^t + B^tS_kA \|}$$

4 Secured Ontology Mapping Using Graph Matching

First we explain the graph matching technique we used. Consider the two graphs Ga and Gb shown in Figure 1. Suppose we want to match vertex 1 of Ga with vertex 4 of Gb , we need to find how much similar the vertices 2 of Ga and 2 of Gb , and 2 of Ga and 1 of Gb.

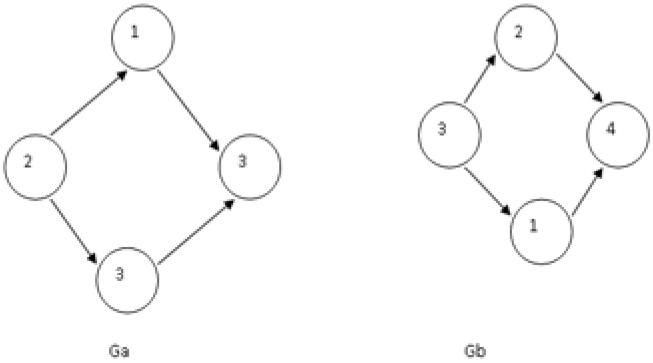


Fig. 1. Graphs to be matched

If A is the adjacency matrix of Ga and B is the adjacency matrix of Gb and S is the similarity matrix defined as follows between vertices we can get the total similarity matrix between individual vertices can be calculated using the formula

$$\frac{BSA^t + B^tSA}{\| BSA^t + B^tSA \|}$$

Here A^t stands for transpose of A. S is the initial similarity matrix. The size of S is nXm.

Where m is number of concepts in first ontology and n is number of ontology concepts in second.

The secured mapping method generates adjacency matrices based on hierarchical relationship of concepts of the encrypted ontologies as per the following algorithm. S is the unity matrix initially.

Algorithm 1. Generating Adjacency matrix for the encrypted ontology given

Let O be the ontology given and A for adjacency matrix. If n is the number of concepts in ontology O then A has order nXn.

```
1. Initialize A [i][j]=0 for all i and j between 0 and n.
2. For i= 1to n
       Begin
           Str=get i^th concept of O
           Collection = get all super classes of Str.
           For each Object x  in the  Collection
               Begin
                   For  j = 1 to n
                   If jth concept of O matches with x then
                       A[i][j]=1;
               End
       End
```

5 Results

The evaluation of the proposed system above is carried out for OAEI systematic benchmark suite. Since we compare for equality of names, and give importance to structure we need not encrypt the ontology for study of evaluation measures. The evaluation measures we considered are Precision, Recall and F-measure. Precision gives the ratio of correctly found correspondences over the total number of returned correspondences. If R is the reference alignment and A is the found alignment then the ratio for precision is

$$P(A,R) = \frac{|R \cap A|}{|A|}.$$

Recall is the ratio of correctly found correspondences to the total number of expected correspondences. The formula is

$$R(A,R) = \frac{|R \cap A|}{|R|}.$$

The following formula is used for finding F-measure.

$$M_\alpha(A,R) = \frac{P(A,R) \times R(A,R)}{(1-\alpha) \times P(A,R) + \alpha \times R(A,R)}.$$

Here α is between 0 and 1. If α is 1 F-measure is same as precision otherwise if it is 0 then F-measure is same as recall. Usually it is taken as 0.5.

Table 1 gives the dataset and the results of experiments in terms of evaluation measures stated above.

Table 1. Result analysis

Benchmark test no	Precision	Recall	F-measure
101	1	0.8	0.88
103	1	0.8	0.88
104	1	0.8	0.88

6 Conclusion

Maintaining privacy in interoperation systems is becoming increasingly important. Ontology matching is the primary means of resolving semantic heterogeneity. Ontology matching helps establish semantic correspondence rules that are used for query rewriting and translation in interoperation systems. For information systems that want maximum privacy, the privacy of their ontologies must be maintained. Our system gives a method to map ontologies which are secured. Limitations of our system is accuracy in terms of precision and recall. The future work would be to improve this.

References

[1] Clifton, C., Doan, A., Kantarcioglu, M., Schadow, G., Vaidya, J., Elma Garmid, A., Suciu, D.: Privacy preserving data integration and sharing. In: Proc. DMKD 2004 (2004)

[2] Agrawal, R., Srikant, R.: Privacy preserving data mining. In: Proc. SIGMOD 2000 (2000)

[3] Mitra, P., Liu, P., Pan, C.C.: Privacy preserving ontology matching. In: Proc. AAAI Workshop (2005)

[4] Li, J.: LOM:Lexicon based ontology mapping tool. In: Proc. PerMIS 2004 (2004)

[5] Noy, N.F., Musen, M.A.: Anchor-Prompt: Using non local context for semantic matching. In: Proc. IJCAI 2001 (2001)

[6] Noy, N.F., Musen, M.A.: The Prompt Suite: Interactive tools for ontology mapping and merging. International Journal of Human Computer Studies 59(6) (2003)

[7] Melnik, S., Molina, H.G., Rahm, E.: Similarity flooding a versatile graph matching algorithm and its application to schema matching. In: Proc. ICDE 2002 (2002)

[8] Doan, A., Madhavan, J., Domingos, P., Halevy, A.: Learning to map be tween ontologies on the semantic web. In: Proc. WWW 2002 (2002)

[9] Melnik, S., Molina, H.G., Rahm, E.: Similarity flooding a versatile graph matching algorithm. In: Proc. ICDE 2007 (2007)

[10] Choi, N., Song, I.Y., Han, H.: A survey on ontology mapping. In: SIGMOD RECORD 2006 (2006)

[11] Shvaiko, P., Euzenat, J.: Ten Challenges for ontology matching. In: Proc. ICODAS 2008 (2008)

[12] Kalfoglou, Y., Schorelmmer, M.: Ontology mapping:The State of the Art. The Knowledge Engineering Review 18(1) (2003)

[13] Ehrig, M., Staab, S.: QOM: Quick Ontology mapping. GI Jahrestagung (1) (2004)

[14] Rahm, E., Bernstein, P.: A survey of approaches to automatic schema matching. The VLDB Journal 10(4) (2001)
[15] Shvaiko, P., Euzenat, J.: A Survey of Schema-Based Matching Approaches. In: Spaccapietra, S. (ed.) Journal on Data Semantics IV. LNCS, vol. 3730, pp. 146–171. Springer, Heidelberg (2005)
[16] Kleinberg, J.M.: Authoritive sources in a hyper linked environment. Journal of ACM (1999)
[17] Van Dooren, B., et al.: A measure of similarity between graph vertices: Application to synonym extraction and we searching. SIAM Review (2004)
[18] Ninove, L., et al.: Graph similarity algorithms. Seminar Presented at Department of Mathematical Engineering. University of Catholique de Louvain (2007)

Image Restoration Using Knowledge from the Image

S. Padmavathi, K.P. Soman, and R. Aarthi

Amrita School of Engineering, Coimbatore, Tamil Nadu, India
{s_padmavathi,r_aarthi}@cb.amrita.edu, kp_soman@amrita.edu

Abstract. There are various real world situations where, a portion of the image is lost or damaged which needs an image restoration. A Prior knowledge of the image may not be available for restoring the image, which demands for a knowledge derivation from the image itself. Restoring the lost portions of the image based on the knowledge obtained from the image area surrounding the lost area is called as Digital Image Inpainting. The information content in the lost area could contain structural information like edges or textural information like repeating patterns. This knowledge is derived from the boundary area surrounding the lost area. Based on this, the lost area is restored by looking at similar information in the same image. Experimentation have been done on various images and observed that the algorithm restores the image in a visually plausible way.

1 Introduction

The restoration process can be viewed as an algorithm that fills the gap with the information obtained from the rest of the image. The results would look natural enough that observer without prior knowledge of the original image will not notice the gaps. A Preliminary version of this is "Cloning Brush tool" in Adobe Photoshop where the user has to provide the information of what to fill in. There are two major methods of restoring the missing data. The traditional method concentrates on the structural information called the isophotes [3], which are lines of similar color. A series of differential equations are used iteratively to extend these lines in to the missing area using the information obtained from the boundary pixels. Another section of algorithms are based on Texture synthesis [1]. This method restores the missing data from an initial seed. Before a pixel is synthesized, its neighbors are sampled. Then the whole image is queried to find out a source pixel with similar neighbors. At this point, the source pixel is copied to the pixel to be synthesized. This is called as pixel based texture synthesis. Alternatively Patch based texture synthesis[6,11] could be used for increased computational efficiency, where a small area (patch) is synthesized rather than a single pixel.

Both structural and texture synthesis methods have their strengths and weaknesses. The former extends the linear structures well but introduces artifacts such as blur. The later avoids the blur but does not extend the linear structures. In general, if the thickness of the area to be restored is very less than the structural methods are advantageous. On the other hand if the thickness is large the texture methods give better results. This paper uses an algorithm similar to exemplar based method as discussed in [6]. It propagates both the texture and the structure and hence restores the missing data with a good visual quality.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 19–25.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

The algorithm first fills the information in lost area that lies near the structural area and the boundary and then proceeds to other areas. The image texture is propagated by direct sampling of the source region. For such restoration the whole image is searched for an area that closely matches the boundary. The restoration is done from the boundary towards the center. This kind of filling from the boundary towards the center is called as “fill front”. Moreover the filling is done in a patch-wise manner. So every time when a patch is filled, the boundary changes and the boundary is detected again and updated. To ensure a better quality of the image the algorithm checks the boundary area for sharp changes like edges and assigns more weights to the unknown pixels closer to the edges. The algorithm also gives more weights to the pixels near the boundary. The pixel with highest weight is considered for filling first. The algorithm stops when all pixels in the damaged area are restored or synthesized. The results have been obtained for various parameters.

2 Restoration Algorithm

The inpainting algorithm discussed in this paper accepts a damaged image as input with the damaged areas marked in special color. The algorithm first groups the pixels with the special color as unknown. For ease of comparison, we adopt notation similar to that used in the inpainting literature. The lost area which is the region to be filled is called as the *target* region indicated by Ω as shown in the fig. 1. The boundary between the known and the unknown area or the contour is denoted by $\delta\Omega$. The unaffected image area is called the source region, Φ.

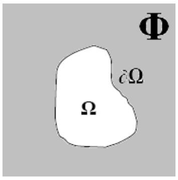


Fig. 1. Inpainting Problem

The algorithm in this paper accomplishes restoration on a patch wise manner where a patch is a small window or a square matrix of pixels. The terminologies used in this paper are derived from Crimsini et.al. For a pixel ‘*p*’ the patch ψ_p is formed with ‘*p*’ as the centre as shown in fig. 2a. The patch contains some unknown pixels from Ω and some known pixels from Φ with ‘*p*’ belonging to the boundary. A patch that is similar to the known pixels is searched in the entire image (excluding the unknown pixels), (i.e.) patch that closely matches with the known pixels of ψ_p is searched in Φ. The patch that yields minimum SSD (Sum of Squared differences) value is taken as the best match. This is called as the exemplar patch. The best-match sample from the source region comes from the patch ψ_q as shown in fig. 2b. The values corresponding to the unknown pixels are copied from the best matched patch. The unknown pixels of ψ_p are copied from corresponding locations of $\psi_{q.}$ A higher number of known pixels in the patch ψ_p increase the confidence of accuracy. Since the pixels are copied as such blurring of edges is avoided. If any other patch near ‘p’ is considered first for restoration, edges(structure) will not be extended correctly. Hence the pixels on the boundary and the pixels near the edges have to be given a higher priority for restoration process. These priorities are termed as Confidence and Data terms respectively.

The patch priority is computed as the product of confidence term and data term. The patch with maximum priority value is restored first. Once when the highest priority patch is filled, there will be a change in the confidence values and the boundary as shown in fig. 2c. The confidence term of that particular patch will be updated as the sum of all the confidence values of the newly filled pixels divided by the total number of pixels in the patch. The new boundary is detected and the process is repeated until all the patches are filled and till the number of boundary pixels becomes zero.

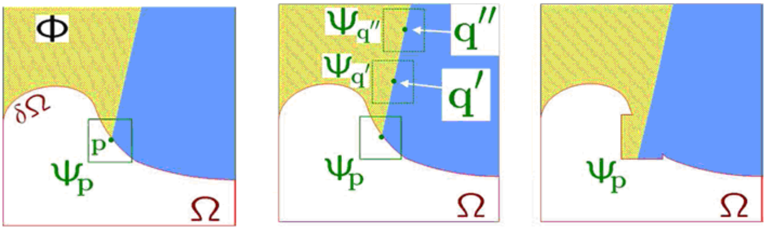


Fig. 2a. Higher Priority patch ψ_p **Fig. 2b.** Matching patches ψ_q **Fig. 2c.** ψ_p filled

Fig. 2. Inpainting of a patch ψ_p

Confidence Term C(p)

The confidence term of a patch denotes the amount of maximum reliable information about the source region in a particular patch. Initially the confidence value is assigned as 0 for the target region(Ω) and 1 for the source region(Φ). The confidence value for a patch ψ_p is calculated as in equation 1

$$C(p) = \frac{\sum_{q \in \Phi \cap \Psi_p} C(q)}{|\Psi_p|} \tag{1}$$

Where, C(q) is the confidence value of those pixels belonging to the source region and the patch ψ_p.The denominator is the area of the patch, which is the total number of pixels in the patch.

Data term D(p)

The data term gives the structure information of the image. The structure is measured in terms of magnitude of image gradient. The data term for the target region is set as 0 initially. The data term for a patch ψ_p is calculated as in equation 2

$$D(p) = \frac{\nabla\Phi_p}{\alpha} \tag{2}$$

Where $\nabla\Phi_p$ is the maximum value of the image gradient in the patch, α is a normalization factor. Sobel’s Gradient operator [9] is used to calculate the image gradient.

The restoration process is given by the following algorithm:

1. Extract the manually selected target region and its initial front (boundary).
2. Repeat until there is no boundary pixel:

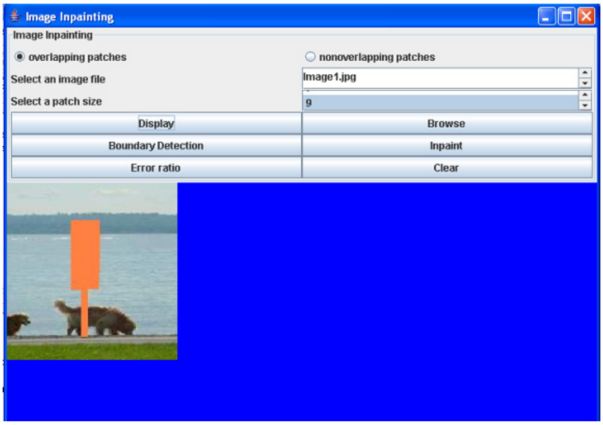
a. Identify the $\delta\Omega$. If the target region exit.
b. Compute priorities P(p)
P(p) = C(p) * D(p), where C(p) is the confidence term and D(p) is the data term.
c. Find the patch Ψp with the maximum priority,
i.e., p = arg max p for all P(p).
d. Find the best matching patch Ψq Є Φ in source region that minimizes d(Ψ p*;* Ψ q), where d is the Sum of Squared Distance(SSD).
e. Copy image data from Ψ q to Ψ p for all pixels belonging to the target region.
f. Update C(p) and the D(p) for the newly filled pixels.

3 Experimentation

The algorithm is implemented in Java. Using an image manipulation software the user specifies the area to be inpainted with a special color. This image is then given to the inpainting algorithm which extracts this unique colored area as mask. The boundary of the mask and the confidence term for all pixels in the mask area are calculated. While inpainting a colored image, spurious colors are generated by the algorithm if different colored channels are inpainted separately. In order to avoid this, the data term is calculated individually for each channels and its average is considered for the total priority. The patch with highest priority is chosen for filling. For any patch in the source region the distance is calculated for each color channel and a sum of the three is stored for that corresponding patch. The patch with a minimum total distance is considered as best match. The algorithm proceeds as specified in the previous section.

For any image the patch size should be slightly greater than the smallest texture element (Texel). Hence, size of the patch is a constraint of acute interest. Change in patch size can bring about visible changes in the output of the system and its performance. The experimentation is done for different patch sizes like 3,5,7,9. The patch size once selected would be maintained through the whole process. While searching for the best samples in the source region, patches are considered in two different format: overlapping (O) and non- overlapping (NO). Experimentation is done for both cases Fig. 3 is a snapshot of the implementation showing the image to be inpainted. The user is allowed to choose the image, the patch size and the type of patches. The output of the modules of the algorithm could also be viewed. Fig 4 shows a snapshot where the image has been restored partially.

The performance of the algorithm is measured by the speed and the accuracy of restoration. The accuracy of the restoration is a usually a subjective process where an observer looks at the inpainted image and able to locate where it is modified. The objective measure can be used when a reference image is available so that the area to be inpainted is already known. When the area is not known the image is restored using Adobe Photoshop where the user specifies what has to be filled in the inpainting area and this is taken as the Reference image. In either case the accuracy and the error in reconstruction is calculated with respect to the reference image.

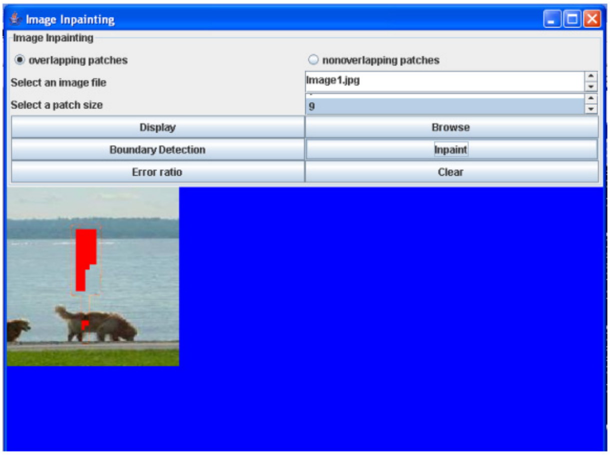


● overlapping patches ○ nonoverlapping patches

Select an image file: Image1.jpg

Select a patch size: 9

Fig. 3. Screen shot of the implementation showing the image to be inpainted



● overlapping patches ○ nonoverlapping patches

Select an image file: Image1.jpg

Select a patch size: 9

Fig. 4. Screen shot of the implementation showing a partially restored image

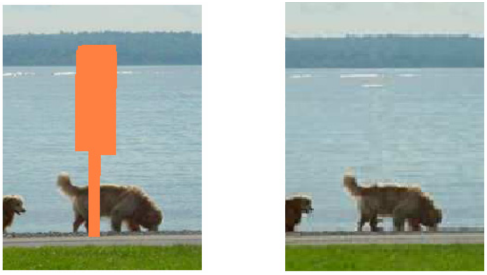


Fig. 5a. Before inpainting **Fig. 5b.**After inpainting

Fig. 5. Inpainting larger area

Experimentation is done for various sizes of target area. Fig. 5a and 6a shows a set of input images with larger (L) and smaller (S) target size respectively. Fig. 5b and 6b shows the corresponding restored images. The time and accuracy for these images are tabulated in Table 1 for various patch sizes and patch types.

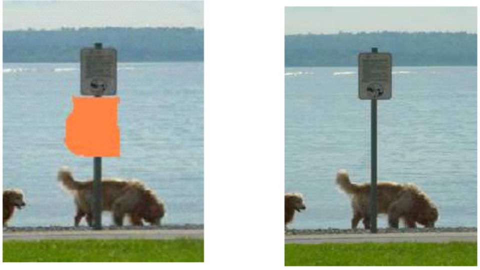


Fig. 6a. Before inpainting **Fig. 6b.** After inpainting

Fig. 6. Inpainting smaller area

Table 1. Performance of Restoration algorithm for images in Fig 5 and 6

Image	Nature & Size	Patch size	Color Image		Error Measure (in %)
			Time	Accuracy (in %)	
Image 1	NO/L	9	1m 45s	96.21	3.78
Image1	NO/L	7	2m 35s	96.08	3.91
Image1	NO/L	5	5m 8s	96.01	3.98
Image1	NO/L	3	7m 8s	95.83	4.17
Image1	O/L	9	1m 50s	97.08	2.92
Image1	O/L	7	2m 41s	96.68	3.32
Image1	O/L	5	5m 50s	96.05	3.95
Image1	O/L	3	7m 17s	95.40	4.60
Image1	NO/S	9	1m 30s	95.04	4.96
Image1	NO/S	7	2m 9s	96.60	3.40
Image1	NO/S	5	4m 59s	96.76	3.24
Image1	NO/S	3	6m 53s	97.42	2.58
Image1	O/S	9	1m 43 s	95.67	4.33
Image1	O/S	7	2m 32s	96.16	3.84
Image1	O/S	5	4m 23s	96.87	3.13
Image1	O/S	3	6 m 30s	97.02	2.98

In Table 1 NO and O represents non overlapping patches and Overlapping patches respectively. L and S represents larger and smaller target area respectively. It could be observed that the overlapping patches gives better accuracy and lesser error than Non-Overlapping patches which implies that reconstruction is better in the former case. The time consumed for reconstruction is more for Larger and Overlapping patches when compared to Smaller and Non-Overlapping Patches.

4 Conclusion

The experiments have been conducted on various images involving natural scenes, regular structured images, different shapes and sizes of the area to be restored. The algorithm is robust towards changes in shape and topology of the region to be restored. It preserves edge sharpness and avoids spurious colors. It is observed from the results that, overlapping small patches gives better results even for a composite textured image such as a natural scene image. But when the image has uniform texture or no structure, non overlapping patches of any size gives good results. The system gives moderate result when the information for filling is not available anywhere in the rest of the image.

References

[1] Ashikhmin, M.: Synthesizing natural textures. In: Proc. ACM Symposium on Interactive 3D Graphics, pp. 217–226. Research Triangle Park, NC (2001)

[2] Chan, T.F., Shen, J.: Non-texture inpainting by curvature-driven diffusions (CDD). Journal of Visual Communication and Image Representation 4(12), 436–449 (2001)

[3] Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proc. ACM Conference Computer Graphics, SIGGRAPH, New Orleans, LU, pp. 417–424 (July 2000), `http://mountains.ece.umn.edu/`

[4] de Bonet, J.S.: Multi resolution sampling procedure for analysis and synthesis of texture images. In: Proc. ACM Conference Computer Graphics, SIGGRAPH, vol. 31, pp. 361–368 (1997)

[5] Harrison, P.: A non-hierarchical procedure for re-synthesis of complex texture. In: Proc. Int. Conf. Central Europe Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic (February 2001)

[6] Criminisi, A., Pérez, P., Toyama, K.: Region Filling and Object Removal by Exemplar-Based Image Inpainting, Microsoft Research, Cambridge (UK) and Redmond (US)

[7] Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proc. ACM Conference Computer Graphics, SIGGRAPH, New Orleans, LU, pp. 417–424 (July 2000)

[8] Bertalmio, M., Vese, L., Sapiro, G., Osher, S.: Simultaneous structure and texture image inpainting. In: Proc. Conference Computer Vision Pattern Recognition, Madison, WI (2003)

[9] Gonzalez, R.C., Woods, R.E.: Digital Image Processing

[10] Rutherford, H.: A Practial Introduction to Image Processing using Java. Pearson University

[11] Hertzmann, A., Jacobs, C., Oliver, N., Curless, B., Salesin, D.: Image analogies. In: Proc. ACM Conf. Comp. Graphics, SIGGRAPH, Eugene Fiume (August 2001)

Harmony-Based Feature Weighting to Improve the Nearest Neighbor Classification

Ali Adeli[1,2], Mehrnoosh Sinaee[1], M. Javad Zomorodian[1,3], and Ali Hamzeh[1]

[1] Department of Computer Science & Engineering, Shiraz University, Shiraz, Iran
[2] Institute of Computer Science, Bojnurd Darolfonoun Technical College, Bojnurd, Iran
[3] Institute of Computer Science, Shiraz Bahonar Technical College, Shiraz, Iran
{aliadeli,mehrnooshsinaee,jzomorodian}@shirazu.ac.ir,
ali@cse.shirazu.ac.ir

Abstract. This paper introduces the use of Harmony Search with novel fitness function in order to assign higher weights to informative features while noisy irrelevant features are given low weights. The fitness function is based on the Area Under the receiver operating characteristics Curve (AUC). The aim of this feature weighting is to improve the performance of the *k*-NN algorithm. Experimental results show that the proposed method can improve the classification performance of the *k*-NN algorithm in comparison with the other important method in realm of feature weighting such as Mutual Information, Genetic Algorithm, Tabu Search and chi-squared ($\chi 2$). Furthermore, on synthetic data sets, this method is able to allocate very low weight to the noisy irrelevant features which may be considered as the eliminated features from the data set.

Keywords: AUC, Harmony Search, Feature weighting, Noisy feature elimination, *k*-NN.

1 Introduction

In non-parametric density estimation algorithms, the distribution of data is calculated without any particular assumption on its parameters. Two popular approaches in these algorithms are: Kernel Density Estimation (KDE) and k-Nearest Neighbor (*k*-NN). *K*-NN is a simple classifier that has been used in various real world applications. In some cases *k*-NN is vulnerable with some problems, such as few instances, noisy data and too many features, which decrease the performance of *k*-NN. To improve the performance of this classifier, many solutions are introduced. One of the approaches to solve those mentioned problems is searching in feature space to find optimal subset of features that can improve the classification accuracy of *k*-NN. This goal can be achieved by assigning the weight to all features in order to eliminate irrelevant ones from noisy data sets. In this paper, we attack this problem and introduce a novel algorithm to deal with. At first, proposed approach assigns different weight to feature using Harmony Search and AUC measure as the fitness function. In other words, Harmony Search with statistical measure as a fitness function has been used to allocate optimal weights to all features to achieve better classification accuracy.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 27–36.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

This paper is organized as follows: in section 2, related work in this domain is reviewed. AUC and Receiver Operating Characteristics (ROC) curve are described in section 3. The proposed method is presented in section 4. Section 5 includes data set, experimental results and discussion, and conclusion is mentioned in the last section of paper.

2 Related Works

There is much research considering the problem of feature selection. It has been used when I) the number of features is larger than the number of training data, II) the number of features is too large for feasible computation III) many features include noisy value. These issues can result in a significant drop in classification performance.

Many methods have been proposed for feature ranking, i.e. Weight Adjusted k Nearest Neighbor (WAKNN) which is introduced by Han [7] to overcome the problem of curse of dimensionality. He implemented his idea on the text classification using *k*-NN. In his work, each attribute takes a weight using the Mutual Information (MI) between each word and the class variable. In the domain of feature weighting, another work refers to Weighted Artificial Immune Recognition System (WAIRS) [10]. In this paper, MI is the main algorithm for feature weighting. Note that the weighted attributes were added to the AIRS. Classification is the final step of AIRS algorithm that is performed by *k*-NN.

Jankowski and Copernicus recommended weighted *k* Nearest Neighbor (WkNN) idea [8]. In each fold of their algorithm, the initial weights for all features are set to 1. During each fold, the values of the weights are summed (subtracted) with Δ value. If the upsdated value can improve the accuracy of the *k*-NN, the new value is replaced with old one for corresponding feature. After each fold, weighting procedure returns a vector of weights. After all folds, i.e. 10 folds, the algorithm computes a normalized vector which is a summation of 10 vectors.

GAW is a common solution for weighting attributes that is suggested by Tang and Tseng [11]. GAW is based on the Genetic Algorithm (GA) with real representation. In this paper, weighing approach is used to improve the accuracy of Weighted Fuzzy *k*-NN (WFKNN) classifier. Guvenir and Akkus studied on Weighted k Nearest Neighbor Feature Projection (WkNNFP) [6]. In WkNNFP, Single Feature Accuracy (SFA) procedure is utilized for feature weighting. In SFA, weight of each feature is determined according to accuracy which is obtained by considering only this feature.

Tabu Search (TS) is proposed as a weighting method in [13]. In this paper, a Hybrid Tabu Search/*K*-NN algorithm is proposed to perform both feature selection and feature weighting simultaneously. In other words, *k*-NN is used each weight set generated by TS. It searches heuristically in a local neighborhood area and moves from a solution to its best admissible neighbor.

The proposed chi-squared ($\chi2$) Feature Weighting ($\chi2$FW) method can be classified as a mutual information approach for assigning features weights [12]. In this sense, the mutual information (the Chi-Squared statistical score) between the values of a feature and the class of the training instances are used to assign feature weights [12].

The algorithm uses Sequential Weighting as the weighting criteria in order to give weights to the features. The weighting criteria ranks features according to their $\chi 2$ scores. In other words, the features having the lowest $\chi 2$ score have their weights set to 1, those with the second lowest-scored features have their weight set to 2 and so on. The process goes on until weights are assigned to the highest $\chi 2$ scored features [12]. In the wide range of weighting approach, algorithm processes the usefulness of each feature independently. So, the non-linear interaction between features has been ignored. While in the proposed method, each Harmony vector takes into account this interaction and also their importance (not its importance) on the classification problem.

3 Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) curve is a two dimensional illustration of the classifier performance. It is suitable solution to analyze the classification accuracy in the binary class problem. For this purpose, the ROC curve plots the True Positive rate (Sensitivity) versus False Positive rate (1- Specificity). The ROC is a strong and statistical tool to compare binary classifiers. Sensitivity and Specificity are described in (1) and (2). To plot the ROC curve, the sensitivity and specificity need to be calculated as follows [3]:

- True Positive (TP) = number of predicted positive cases that are actually positive.
- True Negative (TN) = number of predicted negative cases that are actually negative.
- False Positive (FP) = number of predicted positive cases that are actually negative.
- False Negative (FN) = number of predicted negative cases that are actually positive.

$$Sensitivity = \frac{TP}{TP+FP} \quad (1)$$

$$Specifity = \frac{TN}{TN+FP} \quad (2)$$

The AUC is a part of the area of the unit square. The AUC is a scalar value, in interval [0--1], to show the discriminative power of binary classifiers. If the value of AUC is less than "0.5", it shows undesirable result, but if the AUC value of classifier is close to "1", it shows a remarkable performance for binary classification. Equation (3) shows the AUC formula.

$$AUC = \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k - Y_{k-1}) \quad (3)$$

3 Proposed Method

The aim of this study is to improve the classification accuracy of the *k*-NN algorithm. One of the best solutions to improve the *k*-NN classification is that the informative features are given large weights while noisy irrelevant features are given low weights.

In this respect, a great approach is needed to select the best features easily. For this purpose, first contribution of the paper has focused on the Harmony-based feature weighting. The Harmony Search is one of the best and popular search tools. Harmony Search as a weighting procedure is a novel approach which can classify input instances with informative and relevant feature. The second contribution of the paper is to use the AUC as a fitness function for Harmony vectors.

First of all, the data set has been split to the unseen data and training sets. Next, the 10-fold cross validation function has been used to validate the *k*-NN. Then in each fold, the Harmony Search procedure is called. In Harmony algorithm, a population of n Harmony vector has been produced randomly and the fitness function of population (Harmony Vectors) is computed. Note that Harmony vectors includes real value in range [0--1]. After that, fitness value of each Harmony vector is calculated using AUC function. Then evaluated Harmony vectors are used in evolutionary progress. The cycle of Harmony Search will be described later. The evolutionary process has continued until the conditions are satisfied, i.e. variance of fitness value for the best Harmony vector is lower than a predefined threshold. After each fold, the training error should be computed with the validation set. For this propose, Harmony algorithm returns a vector (in size of features) with the best real values in range [0--1] (each value refers to corresponding feature). After that, the weighted features are stored to the *k*-NN algorithm for classification. Note that all weights will be employed in edited version of Minkowski metric (8) to compute the distance between training and testing instances. After 10 folds, best features are given higher weights while the irrelevant ones are given low weights and then they have been used in the *k*-NN. Finally, the testing error has been computed.

3.1 Harmony Search (HS)

Harmony Search (HS) is a metaheuristic algorithm that is proposed by Geem et al. [2]. HS is inspired by improvisation process of music player and mimicking its phenomenon [2]. In improvisation process, musician plays a note to find best Harmony. Similar to this process, in engineering problem, a decision variable generates a value to find global optimum. HS is a derivative-free, does not require initialization setting for decisions values, free from divergence and can deal with both types of variables (discrete and continuous) [2,9].

A new Harmony vector can be made by choosing a pitch from following rules: 1) playing one pitch from memory (musician's memory); 2) playing a pitch near to pitches in the memory; and 3) playing a random pitch from possible range of pitches. Likewise, in HS algorithm, the value of a decision variable is selected according to one of the following rules: 1) selecting a value from Harmony Memory (HM); 2) selecting a value near to values of HM; and 3) selecting a random value in the possible range of values. HS includes some parameters which are described as follow:

HM (Harmony Memory) encloses all the generated Harmony vectors. Equation (4) shows the memory. **HMS** (HM Size) determines the number of Harmony vectors in HM. Each vector is a solution for optimization problem. Note that in experimental results, value of HMS is set to 100.

$$HM = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_n^1 \\ x_1^2 & x_2^2 & \dots & x_n^2 \\ & & & \\ x_1^{HMS} & x_2^{HMS} & \dots & x_n^{HMS} \end{bmatrix} \tag{4}$$

HMCR (HM Considering Rate) is a probability number to select a value from HM for decision variable (rule 1 and 2) that is illustrated in (5). To satisfy the third above rule, we use (1-HMCR) to choose a random value out of the HM but in possible range of values. HMCR is a high value because in the music domain, each musician has a specific methodology and follows its method in the most melodies. Note that in experimental results, value of HMCR is set to 0.9.

$$x_i^{'} = \begin{cases} x_i^{'} \in \{x_i^1, x_i^2, \dots, x_i^{HMS}\} & \text{with probability HMCR} \\ x_i^{'} \in X_i & \text{with probability (1-HMCR)} \end{cases} \tag{5}$$

PAR (Pitch-Adjusting Rate) is a probability number to determine the rate of small changes in values of the variable. We will see in Algorithm 1 that the probability of PAR is checked inside the condition of HMCR. In other words, if the random generated value is lower than the HMCR, the condition of PAR will be checked, otherwise the PAR condition is not checked. This probability is used to satisfy second above rule. The PAR value is calculated according to (6). The value of the PAR is small because musician follows its method and selects rarely a random Harmony. Note that in experimental results, value of PAR is set to 0.1.

$$x_i^{'} = \begin{cases} x_i^{'} \pm rand(0,1)bw & \text{with probability PAR} \\ x_i^{'} & \text{with probability (1-PAR)} \end{cases} \tag{6}$$

bw is an arbitrary distance bandwidth that shows the range of small changes in values of variable. **MaxImp** is the maximum number of iterations. Note that in experimental results, value of *bw* is set to 0.001.

3.2 Fitness Function

The fitness function used in the HS is based on the AUC [13]. In the AUC algorithm, each instance takes a probability score based on its label of neighbors. To compute the probability, first, the distance of each sample to the others has been calculated using (8) which uses the weight of all features. Then, the labels of k nearest neighbors to this sample are considered. The score of each sample has been computed with (7).

$$\text{Score(i)}=\frac{\text{no. of positive NN to Sample i}}{\text{k}} \tag{7}$$

The Minkowski metric for the weighted feature is changed to (8) which considers the weights of attributes. In (8), w_i is the weight of i^{th} feature.

$$L_k(a,b) = \left(\sum_{i=1}^{d} w_i \left| a_i(x) - b_i(y) \right|^k \right)^{\frac{1}{k}} \tag{8}$$

It is also referred to as the L_k norm. So, the Euclidean distance is the L_2 norm, and the L_1 norm refers to the Manhattan one. After the computation of scores, TP and FP rates are measured, and the ROC curve is plotted. At last, to compute the value of the AUC, the area of all trapezoids, which are located under the ROC curve, is calculated. The summation of all these areas can be considered as the AUC and fitness of Harmony vector.

4 Experimental Results

In the testing phase, 10-fold cross validation used to validate empirical results. After each fold, the validation set has been used to compute the training error of the *k*-NN classification. After 10 fold, the testing error of the *k*-NN classification has been calculated using the testing set (unseen data). The results of the testing have been reported in Tables 2 and 3. The data sets used for the analysis of the model have been indexed in Table 1. Experimental results were achieved in two types. In the first type, our presented method deals with 9 binary class (multi-class) distribution of data which are mentioned in Table 3. All data sets were chosen from UCI repository [1]. Next type of testing is applied in order to analyze the behavior of the proposed method on generated irrelevant features (Table 2). For the second type, our method tested on the seven synthetic data sets which are randomly generated with some relevant and irrelevant features [12, 13]. All the synthetic data sets contain 500 samples in the binary class distribution. The values of all features (relevant/irrelevant) are randomly picked from distribution in interval [0--1]. In all synthetic data sets, a data point belongs to positive class if the average value of relevant features for this instance is smaller than the threshold; otherwise it belongs to negative class. The threshold is set as the average values of all features in whole data. So, for each data set, the threshold is deterministic. Table 2 compares the performance of the proposed method with the simple *k*-NN on the synthetic data sets.

4.1 Discussion

In this section, results of the proposed method are compared with some important feature weighting methods, Mutual Information (MI), Genetic Algorithm (GA), Tabu Search (TS) and chi-squared Feature Weighting (χ2FW). Note that the basic classifier used in the all mentioned weighting methods is the *k*-NN. Experimental results show that the proposed method can improve the classification performance of the *k*-NN. Furthermore, in a number of cases, *k*-NN classifier with the Harmony-based weighting method can perform better than the simple *k*-NN (without feature ranking). In Table 2, effecting of weighting method on *k*-NN classification is presented. For this purpose, we generated some data sets with different number of relevant and irrelevant features. Experimental results show that in all cases of generated data sets, the proposed method outperforms the simple *k*-NN without weighting mechanism.

Algorithm 1 HS procedure

Input(s): Parameter setting

Output: Best Harmony solution (vector)

```
Initial new vector ($x'$) with zero values for all features
for solution :=1 to HMS
    Generate random Harmony vector in range [0 1]
    call the fitness function to compute the fitness of solution
end for
for iteration :=1 to MaxImp
    for feature :=1 to no. of features
        if rand(0,1) < HMCR then
            $x_{iteration}' \leftarrow$ select randomly a value of column feature from HM
            if rand(0,1) < PAR then
                $x_{iteration}' \leftarrow x_{iteration}' + \text{rand}(0,1).bw(\text{i})$;
            end if
        else
            $x_{iteration}' \leftarrow$ randomly select any pitch within bounds
        end if
    end for
    calculate the fitness of new vector $x'$
    if fitness($x'$) > fitness (worst vector)
        replace the new vector $x'$ with the worst one
    end if
end for
```

Table 1. Data sets is used in this experiment. Number of features and number of samples in each data set is mentioned.

Dataset	# features	# samples	# class
Glass	10	214	6
Ionosphere	34	351	2
Iris	4	150	3
Hepatitis	19	155	2
Pima	8	760	2
Sonar	60	208	2
Soybean	35	307	19
Vote	16	435	2
WBC	10	699	2

Table 2. Empirical results of classic *k*-NN and the proposed approach when irrelevant features incorporated into the synthetic data sets

# Relevant features	# Irrelevant Features	Feature weighting	Original *k*-NN
4	6	**16.21±1.24**	25.01±0.67
5	5	**22.12±1.36**	24.54±1.13
5	8	**16.48±0.86**	22.87±1.32
6	4	**24.72±1.01**	26.65±0.63
8	5	**18.54±0.58**	20.72±1.78
10	10	**22.04±0.53**	26.98±1.54

Table 3. Comparison of classification errors between the proposed method and the other weighting methods. The best results on each data set is highlighted in bold face. Note that the last column refers to results of the proposed method.

	k-NN without weighting	Weighting based				
		MI	GA	TS	χ2 FW	Proposed method
WBC	93.99	95.56	94.52	95.02	96.63	**97.82**
Glass	88.43	92.78	89.56	90.40	90.56	**94.03**
Hepatitis	84.51	81.29	**87.72**	84.25	80.74	85.58
Ionosphere	89.74	89.46	85.48	93.80	90.35	**95.72**
Iris	93.33	92.67	96.84	96.70	95.02	**97.73**
Pima	69.02	75.01	67.36	74.59	75.97	**77.43**
Sonar	85.03	95.95	89.85	94.20	92.51	**97.10**
Soybean	89.01	91.55	**92.08**	90.78	89.65	85.19
Vote	92.93	95.64	92.65	94.03	94.52	**96.13**

In Table 3, the proposed method outperforms the rest on the data sets such as WBC, Glass, Ionosphere, Iris, Pima, Sonar and Vote. In comparison with the mentioned weighting methods for *k*-NN classifier, experimental results prove that the feature weighting scheme based on Harmony Search is an impressive solution to improve the accuracy of *k*-NN classifier.

WAKNN computes the weight of each feature according to value of MI between this feature and class label [7]. The reason of WAKNN's weak performance is that they process the usefulness of each feature independently. So, the nonlinear interaction between features has been ignored. In other words, WAKNN considers the correlation between each feature and the class label independently from other features. In some cases, there are two features which the correlation between each feature and the class label is low, but the correlation between the combination of them as a features subset and the class label is high. However in the proposed method, each Harmony vector considers the contribution of all features on the classification problem. The Table compares the proposed method with GAW which gives weight to features according to GA [11]. In our approach, one of the important advantages is AUC fitness function. In GAW, classification accuracy rate of the test set (known instances which were tested) is employed for fitness function. Comparison between fitness functions of GA and HS illustrates that the AUC is a dominant function and assign fitness to Harmony vector with high confidence because of its statistical property. So, k-NN classifier with the HS-based weighting algorithm and the AUC fitness function outperforms the k-NN with Genetic-based feature weighting.

The problem of TS is that this kind of search causes the objective function to deteriorate [13]. In other words, it may be fall into local optimum without any reaching to the best set of weights (in task of feature weighting). Chi-squared Feature Weighting (χ2FW) is a weighting method that is calculated according to (9). In (9), i and j are discrete variables which can assume l and c possible values, respectively. n_{ij} and e_{ij} are the observed frequency and the expected frequency, respectively. Similar to WAKNN, the chi-squared method processes the usefulness of each feature independently. So, the nonlinear interaction between features has been ignored. In some cases, we need to analyze the effect of a group of features on the classification problem instead of considering just one feature.

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(n_{ij}-e_{ij})^2}{e_{ij}} \tag{9}$$

Our algorithm shows poor performance on the Soybean and Hepatitis data sets. The reason of poor performance is that the weights are inefficient for features.

5 Conclusion

In this paper, a novel method has been introduced for feature weighting. The proposed method of feature weighting is based on HS. The best Harmony vector returns real values in range [0--1] for all features. The weighting of features is based on the fitness of the Harmony vector. The fitness of Harmony vector has been calculated using the AUC, a statistical measure to compare classifiers. The experimental results show that the proposed method improves the k-NN classification. In some cases, the proposed method helps the k-NN to result in more accurate classification than some other method in the realm of feature weighting such as MI, TS and GA. Furthermore, in the synthetic datasets, this method is able to allocate very low weight to the noisy irrelevant features which may be considered as the eliminated features from the dataset.

References

1. Blake, L., Merz, C.J.: UCI repository of machine learning databases, `http://www.ics.uci.edu/~mlearn/MLRepository.html`
2. Das, S., Mukhopadhyay, A., Roy, A., Abraham, A., Panigrahi, B.K.: Exploratory power of the harmony search algorithm: analysis and improvements for global numerical optimization. IEEE Transactions on Systems Man and Cybernetics Part B 41(1), 89–106 (2011)
3. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis, vol. 7. Wiley (1973)
4. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing, vol. 12. Springer (2003), `http://www.mitpressjournals.org/doi/abs/10.1162/evco.2004.12.2.269`
5. Fawcett, T.: Roc graphs, Notes and practical considerations for data mining researchers ROC graphs. Intelligent Enterprise 31 (HPL-2003-4) 28 (2003)
6. Guvenir, H.A., Akkus, A.: Weighted k nearest neighbor classification feature projections. In: Proc. of the Twelfth International Symposium on Computer and Information Sciences, ISCIS XII, pp. 44–51 (1997)
7. Han, E.-H(S.), Karypis, G., Kumar, V.: Text Categorization Using Weight Adjusted *k*-Nearest Neighbor Classification. In: Cheung, D., Williams, G.J., Li, Q. (eds.) PAKDD 2001. LNCS (LNAI), vol. 2035, pp. 53–65. Springer, Heidelberg (2001), `http://dl.acm.org/citation.cfm?id=646419.693652`
8. Jankowski, N.: Discrete feature weighting selection algorithm. In: Proceedings of the International Joint Conference on Neural Networks, vol. 1, pp. 636–641 (2003)
9. Mahdavi, M., Fesanghary, M., Damangir, E.: An improved harmony search algorithm for solving optimization problems. Applied Mathematics and Computation 188(2), 1567–1579 (2007)
10. Seeker, A., Freitas, A.: Wairs improving classification accuracy by weighting attributes in the airs classifier. In: IEEE Congress on Evolutionary Computation, CEC 2007, pp. 3759–3765 (2007)
11. Tang, P.H., Tseng, M.H.: Medical data mining using BGA and RGA for weighting of features in fuzzy k-NN classification. In: International Conference on Machine Learning and Cybernetics, vol. 5, pp. 3070–3075 (2009)
12. Vivencio, D., Hruschka, E., Nicoletti, M., dos Santos, E., Galvao, S.: Feature-weighted k-nearest neighbor classifier. In: FOCI 2007, pp. 481–486 (2007)
13. Zomorodian, M.J., Adeli, A., Sinaee, M., Hashemi, S.: Improving Nearest Neighbor Classification by Elimination of Noisy Irrelevant Features. In: Horng, M.-F. (ed.) ACIIDS 2012, Part II. LNCS, vol. 7197, pp. 11–21. Springer, Heidelberg (2012)

Effectiveness of Different Partition Based Clustering Algorithms for Estimation of Missing Values in Microarray Gene Expression Data

Shilpi Bose[1], Chandra Das[1], Abirlal Chakraborty[2], and Samiran Chattopadhyay[2]

[1] Department of Computer Science and Engineering, Netaji Subhash Engineering College,
Kolkata- 700 152, India
`bose.shilpi08@gmail.com, chandradas@hotmail.com`
[2] Department of Information Technology, Jadavpur University, Kolkata- 700 092, India
`abir126@gmail.com, samiranc@it.jusl.ac.in`

Abstract. Microarray experiments normally produce data sets with multiple missing expression values, due to various experimental problems. Unfortunately, many algorithms for gene expression analysis require a complete matrix of gene expression values as input. Therefore, effective missing value estimation methods are needed to minimize the effect of incomplete data during analysis of gene expression data using these algorithms. In this paper, missing values in different microarray data sets are estimated using different partition-based clustering algorithms to emphasize the fact that clustering based methods are also useful tool for prediction of missing values. However, clustering approaches have not been yet highlighted to predict missing values in gene expression data. The estimation accuracy of different clustering methods are compared with the widely used KNNimpute and SKNNimpute methods on various microarray data sets with different rate of missing entries. The experimental results show the effectiveness of clustering based methods compared to other existing methods in terms of Root Mean Square error.

Keywords: Microarray analysis, missing value estimation, c-means, fuzzy c-means, possibilistic c-means, fuzzy possibilistic c-means.

1 Introduction

Recent advancement of microarray technologies has made the experimental study of gene expression data faster and more efficient. Microarray techniques, such as DNA chip and high density oligonucleotide chip are powerful biotechnologies as they are able to record the expression levels of thousands of genes simultaneously [1].

The data generated in a set of microarray experiments are usually gathered in a matrix with genes in rows and experimental conditions in columns. Frequently, these matrices contain missing values (MVs). This is due to the occurrence of imperfections during the microarray experiment (e.g. insufficient resolution, spotting problems, deposition of dust or scratches on the slide, hybridization failures etc.) that create

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 37–47.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

suspected values, which are usually thrown away and set as missing [2]. In large-scale studies involving thousands to tens of genes and dozens to hundreds of experiments, the problem of missing values may be severe. Virtually every experiment contains some missing entries and more than 90% of genes are effected. The presence of missing gene expression values constitutes a problem for downstream data analysis, since many of the methods employed, such as principal component analysis [3] or singular value decomposition [4] (e.g. classification and model-based clustering techniques) require complete matrices . Due to economic reasons or biological sample availability, repeating the microarray experiments in order to obtain a complete gene expression matrix is usually not feasible and also analysis results can be influenced by the estimation of replacing the missing values. Thus, in order to minimize the effect of missing values on analysis and avoid improper analysis, missing value estimation is an important preprocess.

Generally, the procedures for dealing with the randomly present missing data can be grouped into three categories [5], [2]: (1) Ignorance-based procedures: This is the most trivial approach to deal with data sets when the proportion of complete data is small, but the elimination brings a loss of information; (2) Model-based procedures: This is a missing data recovery method, which defines a model for the partially missing data. However, the complexity of the method prevents the applications of large data sets; (3) Imputation-based procedures: This is the type of missing data substitution methods, which fill the missing values by certain means of approximation. Statistical imputation belongs to this category, where the missing values are substituted by a statistically inspired value that has a high likelihood for the true occurrence, for example the mean values computed from the set of non-missing data records.

There are several simple ways to deal with missing values such as deleting genes with missing values from further analysis, filling the missing entires with zeroes, or imputing missing values of the average expression level for the gene ('row average') [2] etc. Two advanced estimation methods for missing value estimation in microarray data have been proposed by Troyankaya et al. [5]; a weighted K-nearest neighbor method (KNNimpute) and a singular value decomposition method (SVDimpute). KNNimpute method is proposed as a robust and sensitive method for missing value estimation. It uses the KNN procedure to select genes, and uses weighted linear combinations to predict missing values. Recently, there is an estimation method called sequential K-nearest neighbor method (SKNNimpute) [6] for microarray data. This imputes missing values sequentially from the gene having least missing values and uses the imputed value for the latter imputation. Efficiencies of KNNimpute and SKNNimpute are better than the above mentioned simple methods in terms of missing value prediction error on non time series or noisy data. SVDimpute that takes all gene profile correlation information into consideration yields best results on time series data with low noise levels. However, estimation abilities of KNNimpute and SKNNimpute depend on the important model parameter K-value, the number of gene neighbor used to estimate the missing value. The parameter is usually specified by the user, which requires the user have some domain knowledge. There is no theoretical way, however, to determine these parameters appropriately. Several other methods have also been developed to estimate missing values. Bayesian principal component analysis (BPCA) [7] is shown to perform exceptionally well [8], [9]. However, BPCA is a sophisticated method that is highly dependent on the number of principal axes [8]. The fixed-rank approximation algorithm (FRAA) proposed by Friedland et al. [10] carries out the estimation of all missing entries in the gene expression data matrix simultaneously based on the singular value decomposition (SVD) method. Local least-squares imputation (LLSimpute) by Kim et al. [11] exploits the local similarity structures in the data and uses the least-squares optimization method to find the missing values that are represented as a linear combination of similar genes. However, the prediction error generated using these methods still impacts on the performance of statistical and machine learning algorithms including class prediction, class discovery, and differential gene identification algorithms [12]. There is, thus, considerable potential to develop new techniques that will provide minimal prediction errors for different types of microarray data including both time and non-time series sequences.

Current research demonstrates that if the correlation/similarity between genes is exploited then missing value prediction error can be reduced significantly [13] in gene expression data. Cluster analysis [14], which partitions the given data set into distinct subgroups, is also applied to predict missing values in microarray data. Intuitively, objects in a cluster are more similar to each other than those belonging to different clusters. In this sense, objects in a cluster are more correlated with each other, whereas objects in different clusters are less correlated. As it can partitions different objects into groups, based on some similarity/dissimilarity criterion, it can also be used to discover structures based on similarity/dissimilarity in gene expression data without providing any interpretation. After clustering, missing values present in a gene can be predicted more accurately from other similar genes belonging to the same cluster.

In this paper, prediction accuracies are given for estimation of missing values in microarray gene expression data with respect to RMS error, using different partition based clustering algorithms. The effectiveness of the partition-based clustering methods, along with a comparison with SKNN and KNN imputation methods, is demonstrated on three microarray data sets.

2 Different Partition-Based Clustering Algorithms for Estimaion of Missing Values

In this section different partition-based clustering algorithms are described and then a new imputation method has been demonstrated to predict missing values in microarray gene expression data.

2.1 Notation

Throughout this paper, microarray data are represented by matrices with rows corresponding to genes and columns to experimental conditions. In particular, G represents original data matrix (with real MVs), while S is a complete gene expression matrix without any missing values with N genes and E experiments (with N >> E) after preprocessing G. In this S matrix, data are randomly deleted to create test data matrix T. X represents a set containing N number of genes. Every gene contains E number of attributes.

2.2 c-Means Clustering Algorithm

The algorithm proceeds by partitioning N number of objects into c nonempty subsets. During each partition, the centroids or means of the clusters are computed. This process iterates until the criterion function converges. Typically, the square-error criterion is used, defined as

$$E = \sum_{i=1}^{K} \sum_{x_k \in U_i} |x_k - m_i|^2 \tag{1}$$

The main steps of the c-means algorithm [15] are as follows:

1) Arbitrarily choose c number of object from X and they are assigned in m_i, i = 1 to c as initial cluster means.
2) Assign each data object x_k to the cluster U_i for the closest mean.
3) Compute new mean for each cluster using

$$m_i = \frac{\sum_{x_k \in U_i} x_k}{|U_i|} \tag{2}$$

where $|U_i|$ is the number of objects in cluster U_i.
4) Iterate until criterion function converges, i.e., there are no more new assignments.

2.3 Fuzzy c-Means (FCM) Clustering Algorithm

This is a fuzzification of the c-means clustering algorithm. It partitions a set of N objects $\{x_k\}$ into c clusters by minimizing the objective function

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^p \|x_k - m_i\|^2 \tag{3}$$

where $1 \leq p < 1$ is the fuzzifier, m_i is the i^{th} cluster center, $\mu_{ik} \in [0, 1]$ is the membership of the k^{th} pattern and $\|.\|$ is the distance norm, such that

$$m_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^p x_k}{\sum_{k=1}^{N} (\mu_{ik})^p} \tag{4}$$

and

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{p-1}}} \tag{5}$$

$\forall i$, $d_{ik} = \|x_k - m_i\|^2$, subject to $\sum_{i=1}^{c} \mu_{ik} = 1, \forall k$, and $0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i.$

The algorithm [16] proceeds as follows:

1) Pick the initial means mi, i = 1, · · · , c. choose value for fuzzifier p and threshold ε. Set the iteration counter t = 1.
2) Repeat Steps 3-4, by incrementing t, until $|\mu_{ik}(t) - \mu_{ik}(t-1)| > \varepsilon$.
3) Compute μ_{ik} by eqn. (5) for c clusters and N data objects.
4) Update means m_i by eqn. (4).

2.4 Possibilistic c-Means (PCM) Clustering Algorithm

It partitions a set of N objects $\{x_k\}$ into c clusters by minimizing the objective function

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} (t_{ik})^q ||x_k - m_i||^2 + \sum_{i=1}^{c} \eta_i \sum_{k=1}^{N} (1 - t_{ik})^q \tag{6}$$

where $1 \leq q < 1$ is the fuzzifier, mi is the ith cluster center, t_{ik} ϵ[0, 1] is the typical membership of the kth pattern, η_i are suitable positive integers and ||.|| is the distance norm, such that

$$m_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^q x_k}{\sum_{k=1}^{N} (t_{ik})^q} \tag{7}$$

and

$$t_{ik} = \frac{1}{1 + \left(\frac{d_{ik}^2}{\eta_i} \right)^{\frac{1}{q-1}}} \tag{8}$$

and

$$\eta_i = K \frac{\sum_{k=1}^{N} t_{ik}^q d_{ik}^2}{\sum_{k=1}^{N} t_{ik}^q} \tag{9}$$

here typically K is chosen to be 1. The main steps of the PCM algorithm [17] are as follows:

1) Pick the initial means mi, i = 1, · · · , c. choose value for fuzzifier p and threshold ε. Set the iteration counter it = 1.
2) Repeat Steps 3-4, by incrementing it, until $|t_{ik}(it) - t_{ik}(it-1)| > \varepsilon$.
3) Compute t_{ik} by eqn. (8) for c clusters and N data objects.
4) Update means m_i by eqn. (7).

2.5 Fuzzy-Possibilistic c-Means (FPCM) Clustering Algorithm

It partitions a set of N objects $\{x_k\}$ into c clusters by minimizing the objective function

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik}^{p} + t_{ik}^{q}) \, ||x_k - m_i||^2 \tag{10}$$

subject to the constraints $p > 1$, $q > 1$, $0 \leq \mu_{ik}, t_{ik} \leq 1$, and

$$\sum_{i=1}^{c} \mu_{ik} = 1, \forall k \tag{12}$$

and

$$\sum_{k=1}^{N} t_{ik} = 1, \forall i \tag{12}$$

and

$$m_i = \frac{\sum_{k=1}^{N} (\mu_{ik}^{p} + t_{ik}^{q})\, x_k}{\sum_{k=1}^{N} (\mu_{ik}^{p} + t_{ik}^{q})} \tag{13}$$

here μ_{ik} is the fuzzy membership value given in eqn. (5) and t_{ik} is the typical or possibilistic membership value given in eqn.(8), p and q are fuzzifiers.

The main steps of the FPCM algorithm [18] are as follows:

1) Pick the initial means m_i, $i = 1, \cdots, c$. choose value for fuzzifier p, q and threshold ε. Set the iteration counter it = 1.
2) Repeat Steps 3-4, by incrementing it, until $|\mu_{ik}(it) + t_{ik}(it) - \mu_{ik}(it - 1) + t_{ik}(it - 1)| > \varepsilon$.
3) Compute μ_{ik} by eqn. (5) and t_{ik} by eqn. (8) for c clusters and N data objects.
4) Update means m_i by eqn. (13).

2.6 Imputation of Missing Values

Initially, all missing values in T are replaced by the estimation given by row (gene) averages to obtain a complete matrix. Specially, this step of gene average substitution, performed in all clustering methods, provides the possibility of contributing the maximum number of genes for estimating the missing values. Then any one of the above mentioned clustering algorithms are executed on this complete matrix. The missing values are imputed by making use of the weighted mean of the values of the corresponding attribute over all clusters. The weighting factors are the membership degrees u_{ik} of a gene x_k to the i^{th} cluster. The missing gene expression value x_{kj} is imputed by:

$$x_{kj} = \frac{\left(\sum_{i=1}^{c} u_{ik}^{l} v_{ij}\right)}{\sum_{i=1}^{c} u_{ik}^{l}} \tag{14}$$

where u_{ik} is the membership value of kth gene in the ith cluster. v_{ij} represents value of jth attribute of mean of ith cluster and l is the fuzzifier. For hard c-mean clustering membership values are either 0 or 1.

The main steps of the imputation algorithm is as follows:

1) Initially all missing values in T are replaced by the estimation given by row (gene) averages for obtaining a complete matrix.
2) Apply any one of the above mentioned clustering algorithm to cluster genes.
3) Estimate missing values by using eqn.(14) with the means obtained from clustering result.
4) Repeat steps 1 and 2 for different number of clusters.

3 Experimental Results

The above mentioned different partition-based clustering algorithms are compared with the previously developed KNNimpute and SKNNimpute methods by imputation of microarray data. Data sets used in this work are selected from publically available microarray data. Three microarray data sets are used: cluster analysis and display of genome-wide expression patterns (data 1) [19], Genomic expression programs in the response of yeast cells to environmental changes (data 2) [20] and the transcriptional program in the response of human fibroblast to serum (data 3)[21]. The metric used to assess the accuracy of estimation is Root Mean Squared (RMS) error which is calculated as follows:

$$RMS_{error} = \frac{\sum_{h=1}^{n} (R_h - I_h)^2}{n} \tag{15}$$

where R_h is the real value, I_h is the imputed value, and n is the number of missing values.

Before any further process, each data set is preprocessed for the evaluation, by removing rows and columns containing missing expression values greater than 50% and rest are replaced by row average values, yielding complete matrices. For every data set between 1 and 20% of the data are deleted at random to create test data set. Each method is then used to recover the introduced missing values for each data set, and the estimated values are compared to those in the original data set.

Every clustering method is executed for c = 5 to 50, where c is the number of clusters. The experiments show that for c > 50 the clustering results detoriates. The value of fuzzifier is varied from 1.2 to 2. For every clustering method best result (i.e. minimum RMS error) is taken for different values of fuzzifier as well as for different values of number of clusters (c).The result is shown for different rate of missing entries present in every data set.

The efficiency of different partition-based clustering methods mentioned here are compared with the KNNimpute and the SKNNimpute methods by applying them to three microarray data sets with different missing rates. Both KNNimpute and SKNNimpute methods require the value of k which is the number of nearest neighbors used in imputation. When k is between 5 and 20, they have given good performances. Accordingly, minimal RMS errors of these two methods are shown by varying k between 5 to 20 in every data set with different rates of missing values. In Table 1, prediction accuracies of different clustering methods are shown for different number of clusters. In Table 2, only best results are shown for data 2 and Data 3 using different clustering algorithms.

In figure 1, it is found that c-mean has given best results compared to all other partition-based clustering algorithms mentioned here and also with respect to KNNimpute and SKNNimpute methods for all different rates of missing entries in data 1 and data 2. FCM, PCM, and FPCM clustering methods also have given better results with respect to KNNimpute and SKNNimpute methods for all cases in data 1 and data 2. For data 3, FCM gives best results for all rates of missing. The other clustering methods have also given better results compared to KNNimpute and SKNNimpute methods for data 3.

Table 1. Comparative Performance Analysis of Different Clustering Methods on Data 1

Rate of data missing (%)	No. of clusters	Prediction Accuracy			
		c-Mean	FCM	PCM	FPCM
1	5	0.484	0.523	0.519	0.523
	10	0.474	0.522	0.524	0.483
	20	0.459	0.502	0.518	0.448
	30	0.45	0.499	0.522	0.454
	50	0.437	0.563	0.523	0.443
5	5	0.534	0.531	0.533	0.543
	10	0.47	0.529	0.529	0.494
	20	0.46	0.527	0.527	0.463
	30	0.452	0.502	0.525	0.464
	50	0.441	0.503	0.525	0.503
10	5	0.544	0.555	0.543	0.531
	10	0.501	0.524	0.513	0.494
	20	0.474	0.523	0.513	0.473
	30	0.463	0.515	0.524	0.502
	50	0.441	0.564	0.535	0.532
20	5	0.583	0.576	0.587	0.594
	10	0.556	0.563	0.542	0.542
	20	0.494	0.547	0.542	0.502
	30	0.463	0.542	0.564	0.524
	50	0.468	0.548	0.564	0.524

Table 2. Best Performance of Different Clustering Methods on Data 2 and Data 3

Data Set	Rate of data missing(%)	Prediction Accuracy							
		No. of clusters	c-Mean	No. of clusters	FCM	No. of clusters	PCM	No. of clusters	FPCM
Data 2	1	50	0.51	30	0.59	10	0.7	20	0.61
	5	50	0.53	30	0.62	10	0.71	10	0.61
	10	50	0.54	25	0.64	20	0.73	10	0.62
	20	50	0.55	30	0.65	20	0.74	20	0.65
Data 3	1	30	0.74	15	0.68	50	0.8	10	0.73
	5	20	0.75	20	0.7	15	0.8	25	0.76
	10	30	0.76	10	0.71	10	0.85	30	0.75
	20	30	0.8	10	0.75	10	0.86	30	0.8

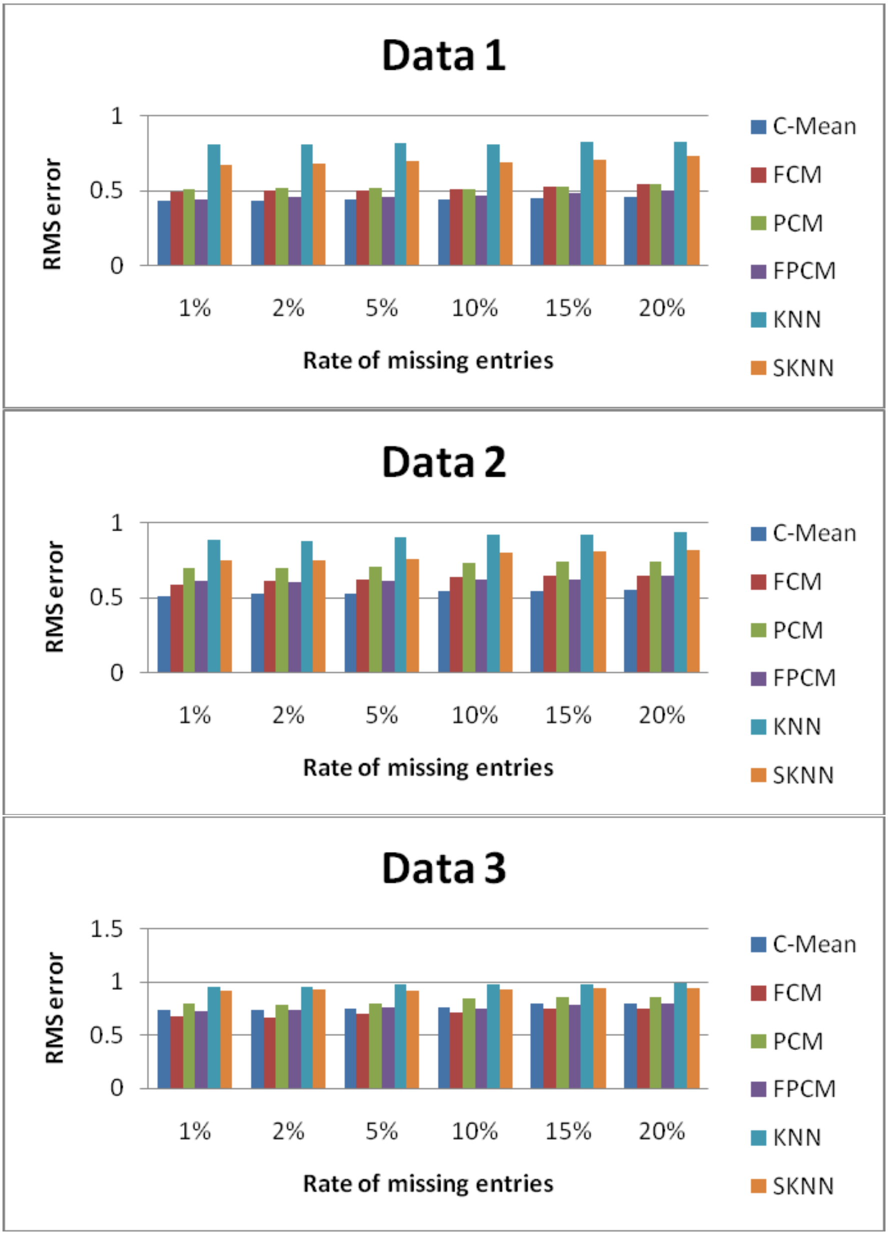


Fig. 1. Comparison of accuracy of Different clustering methods with KNNimpute and SKNNimpute methods for three types of data sets over 1 to 20% data missing. The accuracies are evaluated by RMS error.

4 Conclusion

In this paper, the performance accuracy of different partition-based clustering algorithms for missing value estimation in microarray data sets are compared with KNNimpute and SKNNimpute methods. The experimental results show that in all cases clustering methods have given better results than KNNimpute and SKNNimpute methods in terms of RMS error. So, it can be concluded that clustering methods are also very effective for missing value estimation in microarray gene expression data.

References

1. Schulze, A., Downward, J.: Navigating gene expression using microarrays - a technology review. Nat. Cell Biol. 3, E190–E195 (2001)
2. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J.J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Staudt, L.M.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503–511 (2000)
3. Raychaudhuri, S., Stuart, J.M., Altman, R.B.: Principal component analysis to summarize microarray experiments: application to sporulation time series. In: Pac. Symp. Biocomputing, pp. 455–466 (2000)
4. Alter, O., Brown, P.O., Bostein, D.: Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl Acad. Sci. USA 97, 10101–10106 (2000)
5. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Bostein, D., Altman, R.B.: Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525 (2001)
6. Kim, K.Y., Kim, B.J., Yi, G.S.: Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics 5(160) (2004)
7. Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K.I., Ishii, S.: A bayseian missing value estimation method for gene exression profile data. Bioinformatics 19, 2088–2096 (2003)
8. Wang, X., Li, A., Jiang, Z., Feng, H.: Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme. BMC Bioinformatics 7, 1–10 (2006)
9. Wong, D.S.V., Wong, F.K., Wood, G.R.: A multi-stage approach to clustering and imputation of gene expression profiles. Bioinformatics 23, 998–1005 (2007)
10. Friedland, S., Niknejad, A., Chihara, L.: A simultaneous reconstruction of missing data in DNA microarrays. Linear Algebra Appl. 416, 8–28 (2006)
11. Kim, H., Golub, G.H., Park, H.: Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 21, 187–198 (2005)
12. Sehgal, M.S.B., et al.: Statistical neural networks and support vector machine for the classification of genetic mutations in ovarian cancer. In: IEEE CIBCB 2004, USA (2004)
13. Sehgal, M.S., et al.: K-ranked covarience based missing values estimation for microarray data classification. In: HIS (2004)

14. Au, W.-H., Chan, K.C.C., Wong, A.K.C., Wang, Y.: Attribute clustering for grouping, selection, and classification of gene expression data. IEEE Trans. on Computational Biology and Bioinformatics 2(2) (2005)
15. Tou, J.T., Gonzalez, R.C.: Pattern recognition principles. Addison-Wesley, London (1974)
16. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York (1981)
17. Krishnapuram, R., Keller, J.: A possibilistic approach to clustering. IEEE Trans. Fuzzy Syst. 4(3), 393–396 (1993)
18. Pal, N.R., Pal, K., Bezdek, J.C.: A mixed c-means clustering model. In: IEEE Int. Conf. Fuzzy Systems, Spain, pp. 11–21 (1997)
19. Eisen, M., Spellman, P., Brown, P., Bostein, D.: Cluster analysis and display of genome wide expression patterns. Proc. Natl Acad. Sci., USA 95, 14863–14868 (1998)
20. Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Bostein, D., Brown, P.: Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell. 11, 4241–4257 (2000)
21. Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J.J., Bogosk, M.S., et al.: The transcriptional program in the response of human fibroblast to serum. Science 283, 83–87 (1999)

MACREE – A Modern Approach for Classification and Recognition of Earthquakes and Explosions

K. Vijay Krishnan, S. Viginesh, and G. Vijayraghavan

Anna University, Chennai – 600025, Tamil Nadu, India
{vkrish.krishna,vigineshsankararaman}@gmail.com

Abstract. Though many systems are available for discrimination between earthquakes and explosions, our introduces new advances and some rudimental results of our ongoing research project. To discriminate between earthquakes and explosions, temporal and spectral features extracted from seismic waves, additional some seismological parameters (such as epicenter depth, location, magnitude) are crux for rapid and correct recognizing event sources (earthquakes or explosions). Seismological parameters are used as the first step to screen out obvious earthquake events. Fourier transforms (FFT), chirp-Z transforms, wavelet transforms have been conducted and some prominent features are acquired by present experimental dataset. In some experiments, wavelet features plus support vector classification (SVC) have reached very high correct recognition rate (>90%). This proposed paper can be used in evolving scenarios.

Keywords: Classification, Recognition, Earthquake, Explosion, Tempo-Spectral features, Support Vector Machines (SVC).

1 Introduction

The research on seismic signal processing, analysis, and further discrimination of earthquakes and explosions plays a fundamental role in the development of seismology, and is also indispensable for public welfare and world peace. Modern digital seismographs may record seismic waves of earthquakes and significant explosions occurring sequentially or simultaneously. These sequential or simultaneous occurring characteristics would be harmful to properly explain the recorded seismic waves and might beget some false conclusions. So it is very meaningful to separate earthquake events and explosion events which may occur sequentially or simultaneously from recorded seismic waves. The separation of simultaneously occurring earthquake and explosion events is one aim of our next researches, and should be investigated in further researches by some special signal processing techniques such as independent component analysis [1].

We are presently focused on the separation of sequentially occurring earthquake and explosion events or unrelated earthquake and explosion events. Sequentially occurring continuous or intermittent events have been processed, and saved in seismic wave files so that each file contains only one event, earthquake or explosion.

N. Meghanathan et al. (Eds.): Advances in Computing \& Inform. Technology, AISC 177, pp. 49–56.
springerlink.com
© Springer-Verlag Berlin Heidelberg 2013

2 Discriminative Features

It has been shown [2][3] that the hypocenters of earthquakes are almost deeper than those of explosions. This may impact the travel ways of seismic waves. In addition, the origin mechanisms of the 2 event types differ essentially. Many different temporal and spectral features (the ratio of different magnitude scales [4], seismic phase, P-wave initial arrival time, the direction of P-wave initial arrival, the ratio of P/S-wave magnitude values, relationship of waves, complexity of wave, ratio of spectrums, cestrum, instantaneous spectrum, etc) have been proposed and investigated, but no ideal feature(s) is widely accepted due to the problem's complexity and proposed and some encouraging rudimentary results have been acquired [5]. Initially by Fourier transform (FFT), overall spectrum layout is acquired. Fourier transform transfers a time domain signal into frequency domain. In the discrete form, its definition is as following:

$$X(e^{j\omega}) = \sum_{n=0}^{N-1} x(n) e^{-j\omega nk} - 1, 0 \le k \le N$$

Where $x(n)$ is the discrete time signal or sampled continuous time signal with length N, $\omega = 2\pi / N$ is the angular frequency, $X(e^{j\omega})$ is the Fourier Spectrum. Chirp-Z transform is a classical algorithm with thinning spectrum [6]. This algorithm is formulated from discrete Fourier transform. For N - points length time signal $x(n)$, the Chirp-Z transform is defined as [7]:

$$\mathrm{X(z_r) = CZT\,[x(n)] = \sum_{n=0}^{N-1} x(n) z^{-n} = \sum_{n=0}^{N-1} x(n) A^{-n} W^{nr}}$$

Where $A = A_0 e^{j}, W = W_0 e^{-j\wp}$; A, W both is Real numbers. If $A0$ >1, the integral or accumulative path of CZT is outside the unit circle; otherwise, the path is inside the unit circle. If W_0 >1, the path rotates inwards; otherwise, rotates outwards. If A_0 = W_0 = 1, the path is an arc along the unit circle. Θ_0 is the onset angular frequency, $\wp$ is the sampling interval also known as angular increment.

Let $\theta_r = \theta_0 + r\varphi_0$, $r = 0, 1, L, M - 1$ are interested points of frequency range. If $\theta_0 = 0$, $M = N$, Chirp-z transform and discrete Fourier transform are the same. If $\varphi_0 = 2\pi / M < 2\pi / N$, *CZT* sampling spectrum X (z r) is finer than DFT, it may acquire more precise spectrum characters.

The result [5] of an experiment for discriminating 40 earthquakes and 40 explosions with 0.01 Hz-5 Hz frequency range are listed in Figure 1. In this figure, abscissa axis represents the logarithm value of average energy (normalized), the ordinate axis means the logarithm value of dominant frequencies; and **Red circles represent explosion events, Blue asterisks represent earthquake events.**

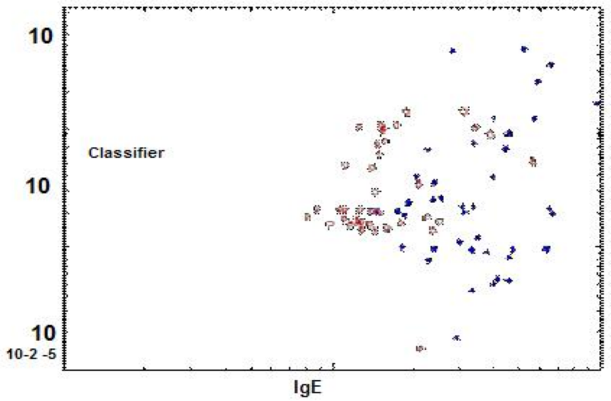


Fig. 1. Rudimentary result [5] of chirp-z transforms

This **means** that appropriate$_{n=0}$ window length of seismic signal may also be an important role in the classification of earthquake and explosion. But these seemingly encouraging results are very limited when applied to more other sites events.

Features extracted from chirp-z transform and wavelet transform are actually temporal-spectral quantities. How to combine theses temporal-spectral features with classical event features, wave temporal and spectral features for acquiring more robust classification and recognition result is imperative but challenging due to heterogeneity of earth structure and complexity of event behaviors. Thus, location comparability and magnitude scale comparability are fundamental requirements for high accurate explosion recognition.

3 Recognition Features

In pattern recognition field, recognition and classification are often interchangeably used though their meanings are subtle different from each other. Strictly speaking, recognition is identified a new sample as one of some several presumable classes, and classification is the process of designing some rule(s), then designating each of a group of samples to one of correspondent class. For a pattern recognition system, recognition is often considered in testing and practical application phases, and classification is often considered in training and learning phases. A typical pattern recognition system is sketched in Figure 2.

The transform of 4-wavelet packet has also already been investigated to extract 3 types of wave features - energy ratio, Shannon entropy and logarithmic energy entropy:

$$E_{wt}(i\ )=100*\sum_{n=1}^{N}x^2(i,j)/\sum_{m=1}^{J}x^2(n,m)$$

$$E_{shannon}(i)=-\sum_{j=1}^{J}x^2(i,j)\log(x^2(i,j))$$

$$E_{\log}(i) = -\sum_{j=1}^{J} \log(x^2(i,j))$$

These features are supplied to a classifier of v-SVC (support vector classifier) for verifying the capabilities of these features. In three approaches of pattern recognition: syntactic, neural and statistical, the last one is main force in many practical recognition applications. In the statistical approach of pattern recognition, the main focus of researches is acquiring some technique for best generalization of decision rules which derived from training samples in experimental data sets. This approach requires a powerful computational capability, demanding some flexible use of numerical programs for studying the data set as well as tools for evaluating the data analysis procedures themselves. As many new techniques are still being proposed in the literature, an easy, robust one.

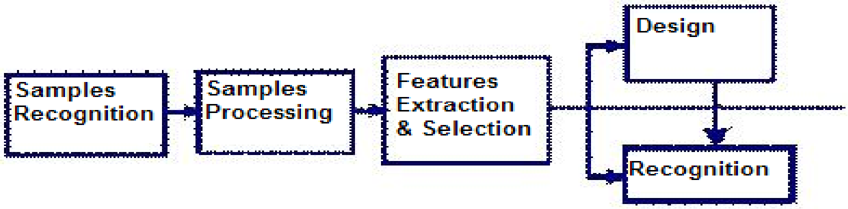


Fig. 2. Sketch of Pattern Recognition system

The results [8] showed that the feature of Shannon entropy is the best candidature to discriminate earthquake and explosion among the above three features. Classifications by v -SVC are carried out for more elaborate recognition tests. The results show that window length is also an important factor for recognition rate. The recognition rates of several different window lengths are ranged from 81% to 98%. The best window length is 2000 sampling points which achieves 98% programming tool or platform is needed that enables a fast and flexible algorithm implementation. Hereby the use of a widely available numerical toolset like Matlab may be profitable for both, the use of existing techniques, as well as for the study of new algorithms. Moreover, because of its general nature in comparison with more specialized statistical environments, it offers an easy integration with the preprocessing of data of any nature. This may certainly be facilitated by the large set of toolboxes available in Matlab. So the recognition algorithms are current implemented in Matlab.

Because of abounding pattern recognition algorithm existing in literature, each different algorithm may be best for its suitable situation and sample data structure. But none of any single algorithm is good for any structural samples. So, basing upon the careful analysis of present problem's sample structure, several typical algorithms have been selected to accommodate the classification of earthquake and explosion. They are Fisher's classifier, a classic linear classifier which suitable to linear classifiable problem; ICHAM (Improved Continuous Hamming's Method) classifier, which suitable to interleaving conglobation sample structure problem; general linear classifier, different from Fisher's by no need to calculate covariance and ever applicable to some non-linear classifiable problem if incurring errors is acceptable; Parzen's classifier, which not needs any presumption and completely learned from data; and a v-SVC(support vector classifier), which entirely is in a new frame and theoretically can be used to formulate almost any shape of delimitating boundary line for any structure's samples, linear classifiable and non-linear classifiable.

The parameters [8] for ν-SVC are set as follows:

ν =0.5; Kernel is Sigmoid

$$k(x_i, x_j) = \tanh((\Upsilon x_i - x_j + c))$$

Where x_i, x_j are $i-thj-th$ sample vector(s) respectively. Parameters in Sigmoid kernel: $\gamma = 1/100$, $c = 0$.

4 Decision Support System

Construction of suitable decision support system is the best solution for explosion recognition and earthquake vs. explosion events classification. The schematic layout for the decision support system of recognizing natural earthquake events and explosion events is displayed in Figure 3 and a snapshot of user operation interface is displayed in Figure 4. The developing platform of this system is MS Visual Studio 2008, programming language is C#. Most seismic signal processing algorithms are coding and developing in Matlab (v7.1) and packed as dll (dynamic-link library) files.

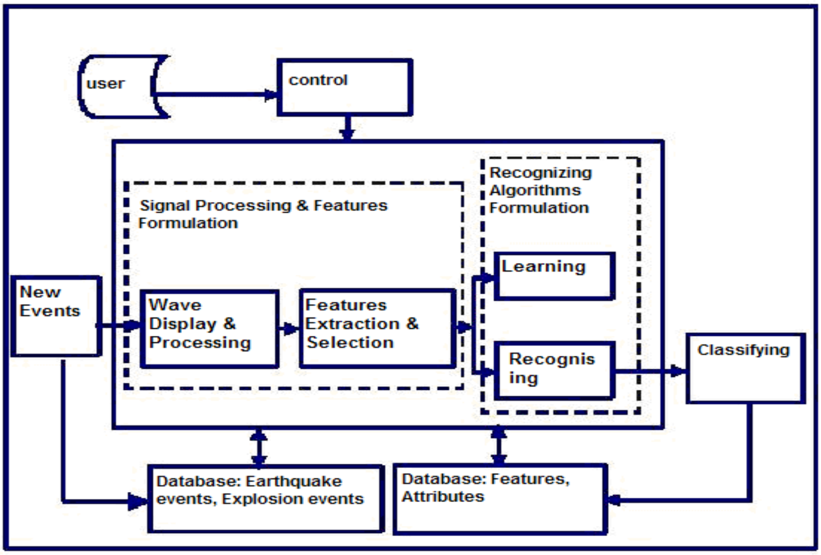

user
control
Signal Processing & Features Formulation
Recognizing Algorithms Formulation
Learning
New Events
Wave Display & Processing
Features Extraction & Selection
Recognis ing
Classifying
Database: Earthquake events, Explosion events
Database: Features, Attributes

Fig. 3. Layout of the decision support system

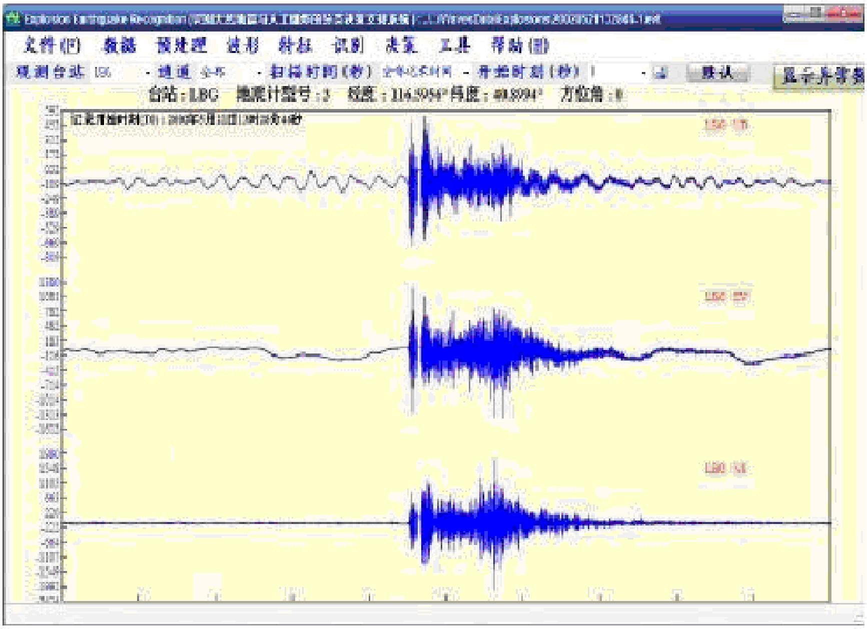


Fig. 4. Interface of the decision support system

E Event wave data that stored in popular seismic data formats such as evt, sed, sac, and txt can be read in and processing in current developing phase, more data formats, especially seed (Standard for the Exchange of Earthquake Data), are planning to add into the system. Prevailing seismic wave data are stored as one file for one event, which may contain more than 100 observatory stations, and generally each station has 3 channels (UD, EW, NS). Thus, the size of one wave data file may larger than 500MB for some events.

Raw seismic wave records may contain a great deal of silent void data, outburst disturbing data, environmental noising data, device trending excursion, etc. So some appropriate data preprocessing must be applied before any meaningful signal processing and wave feature extraction.

Event features such as event magnitude(s) (Ms, Mb, etc), hypocenter depth, epicenter longitude/latitude, must be inputted into the system by user(s), and stored in an event catalog database. However, the raw event wave file cannot be directly stored in database due to its huge size (may > 500MB/single file). But the name of raw event wave file must be associated with event catalog, and can be simultaneously stored in the database.

Temporal features extracted from event waves shall be acquired by user's interacting operation with the waves in the software interface. P-wave initial arrive time, S -wave initial arrive time, P-wave initial magnitude, P-wave maximal magnitude, S-wave maximal magnitude etc, these temporal-related wave signal values can be observed and measured in seismic wave graph. Temporal features can be gotten or calculated from these measurements.

Spectral features are computation-consuming quantities. In order to utilize Matlab's powerful scientific computing capabilities, all spectral features extraction algorithms(chirp- Z, wavelet, Hilbert-Huang etc) are coded and completed in Matlab and these algorithms are packed in several DLL files for convenient called by the system's user.

Explosion recognition and classification of earthquake and explosion events are the core parts of the system. Several pattern classification algorithms (SVC, Fisher's, ICHAM, Linear, and Parzen's Classifier) have been implemented in the system. Of these algorithms, SVC (support vector classifier) is the most robust one for our present experiments.

5 The Working of Next Step

The main and core parts of system's software have basically been completed, but there are still many problems waiting to be solved. Algorithms' bugs may unceasingly be found, and algorithms' limitations also perhaps will be faced when more event data arriving and more experiments being conducted. So algorithms improvements and ever adaptations are always needed. Statistical decision module is also needed due to different pattern recognition algorithm may derive inconsistent conclusions for just one event source (explosion or earthquake).

For clarification, the workings of next step are briefly listed as follows:

- Refining user interface, wave operation, wave signal filtering;
- Debugging temporal, spectral, tempo-spectral feature extraction algorithms;
- Implementing features browse and update;
- Refining explosion recognition algorithm;
- Completing statistical decision and inference module.

References

[1] Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley and Sons (2001)

[2] Tjøstheim, D.: Improved seismic discrimination using pattern recognition. Physics of the Earth and Planetary Interiors, 85–108 (1978)

[3] Bian, Y.J.: Application of genetic BP network to discriminating earthquakes and explosions. Acta Seismologica Sinica 15(5), 540–549 (2002)

[4] Bian, Y.J.: Application of Fisher method to discriminating earthquakes and explosions using criterion mb/Ms. Acta Seismologica Sinica 18(4), 441–450 (2005)

[5] Huang, H.M., Bian, Y.J., Li, R., Lu, S.J.: Discrimination of earthquakes and explosions using chirp-z transform spectrum features. In: Proceedings of the WRI World Congress on Computer Science and Information Engineering, vol. 7, pp. 210–214 (2009)

[6] Rabiner, L.R., Schafer, R.W., Rader, C.M.: The chirp z-transform algorithm. IEEE Transactions on Audio and Electro-acoustics (June 1969)

[7] Hu, G.S.: Digital Signal Processing Theory, Algorithms and Implementations. Qsinghua Unversity Press, Beijing (2003)

[8] Huang, H.M., Bian, Y.J., Lu, S.J., Jiang, Z.F., Li, R.: A research on seismic signal wavelet features of earthquake and explosion. Acta Seismologica Sinica (2010) (in press)
[9] Huang, H.M., Bian, Y.J., Lu, S.J., Li, R., Jiang, Z.F.: v SVC algorithm applied in earthquake and explosion recognition and the choice of window length. Seismological and Geomagnetic Observation and Research (2010) (in press)

Building Concept System from the Perspective of Development of the Concept

Cheng Xian-yi, Shen Xue-hua, and Shi Quan

School of Computer Science and Technology, Nantong University,
Nantong Jiangsu, 226019, China

Abstract. The concept system with rich content is the key to improve the performance of knowledge-based artificial intelligence knowledge system. And a sufficient number of concepts, rich in semantic association, to meet the multi-tasking and developed concept system are one of the major challenges of knowledge engineering. It is the fundamental goal of conceptualization of knowledge, too. In this paper, for the study of natural language processing, from the perspective of development of the concept, a framework is proposed to building concept system.

Keywords: Natural Language Processing, concept system, ontology, semantics, knowledge.

1 Activities of Senior Intelligence

The Original intention of artificial intelligence is to simulate person's intelligence. However, because we know little about the generating process of living beings' (especially human) intelligent behavior, we can only explore the problem that is easier to solve by existing tools. The knowledge expressions that are used by existing systems of artificial intelligence are all aimed at particular application. That is problem oriented. Both Symbolisms and Connectivism are all approximation of knowledge representation. With the development of artificial intelligence in deep and application, many problems put forward a challenge for the existing technology of artificial intelligence. The applications that had been paid more attention are as the following several aspects [1].

(1)Although we have find much new applicable technology about machine translation, automatic abstract, question-answering system and so on in the field of natural language processing, it is the calculation of concept which is not the keyword matching even in a small category, such as what is the meaning of “Li go tomorrow”.

(2)In the field of problem solving, being different from weak method and the method of expert system which put the state space search as its core ,many new applications put forward higher request to the method which is based on knowledge ,and it outstands more and more knowledge intensive features. For example, the low temperature disaster happened in the middle and low areas of the Yangzi River before the Festival Spring in 2008 years. Its rescue involved the weather, traffic, electric power, medical

N. Meghanathan et al. (Eds.): Advances in Computing \& Inform. Technology, AISC 177, pp. 57–65.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

treatment, civil affairs, and many other fields of knowledge. It no longer meets the reasoning realization of the classic expert system but the concept association.

(3)In the field of Web information processing, for the extraction of the interested information and the filter of the harmful information online, we have developed the text mining, public opinion monitor, vertical search and other new demand, and the user hope to put the Web as a problem solver but not a resource pool.

(4)In the filed of scene analysis, for the computer vision and pattern recognition, we apply ourselves to the understanding of the image consistently and we hope to find the public emergencies, to recognize and alarm the internal theft of self-service bank, to find the old man's abnormal step appearance by accompanying robot and to forecast the production of food, and so on, through the image analysis which is captured by camera or remote sensing satellite.

(5)In the field of distributed computing, such as MAS (Multi-Agent Systems) and the robot football, it is not simple the task distribution or data distribution but needs fusion, collaboration and coordination.

(6)The challenge of NP problem. For example, the problem of protein folding is the process to study how a string of amino acids could fold almost in a moment and form a kind of very complex three-dimensional structure protein. It would take 10^{127} years for a super computer which used the verisimilitude for this project calculation to look for the final folding form of a short sequence which consists of only 100 amino acids [2]. There are a lot of problems like this.

The common between previous six problems and the problems of traditional artificial intelligence is the needing of knowledge, but they have the absent distinction. To solve them, senior intelligence system has to deal with a lot of variables. It is not a stack of simple concept (knowledge), but needs to understand the dependent of concept and highlights the knowledge hidden behind them.

In order to realize the senior intelligence, we need to construct a knowledge system which is rich in content, used neatly, explicit and can be stated. It is concept system that supports the system.

2 The Development of Concept System

In daily life, a seemingly simple common sense issue may involve a lot of complex association of concept, and these concepts and the connections between the concept is gained in the process of people's cognitive development. In the process of cognitive development, a child gets new concepts constantly and finds the dependent relationship between them and the existing concepts. The child's concept structure become complicated and complete gradually through such a long process of accumulation.

Here is a list of several influential concept systems:

2.1 FramNet

FramNet treats the frame as its core and it is based on the real corpus. It puts many lemmas that have the same semantic roles in the same frame and uses frame elements which have individual character to describe the protean semanteme of nature language. Then it reveals the various semanteme of each word in each meaning and the possibility of syntactic integrating by the marked sentence[3]. For example, the word "hit" can be expressed as "zonk" and can also be treated as "create an accident or bad effect". How can we distinguish its meanings? Generally speaking, a word's different meanings are associated with the different meaning framework that the word participates in. When the meaning of a word is based on a particular frame, we would say that the word has activated a framework. Accordingly, the word "hit" activates a "hit the target" framework in a certain context environment and it also may activates a specific "cause harm" framework in another context environment[2].

2.2 Semantic Web

In brief, semantic Web is a kind of intelligent network that can understand human language. It can not only understand the human language, but also can make the communication between people and computer as easy as the communication between people. The core of it: it can add the semanteme (Meta data) that the computer can understand by the documents on the World Wide Web (such as the HTML), so that the entire Internet could become a general medium of information exchange. We can improve the resources' usability and usefulness of the World Wide Web and its interconnection by the following methods.

Although semantic web is a better network, its realization is a complex and huge engineering. The realization of semantic web is finished by XML Language and framework of resource description (RDF). XML is a tool used to define markup language. It includes XML declaration, DTD (document type definition) which is used to define the grammar of language, the detail instruction which is used to describe marks and the document itself. The document itself contains marks and content too. RDF is used to express the content of webpage.

The system structure of semantic web is shown in figure 1.

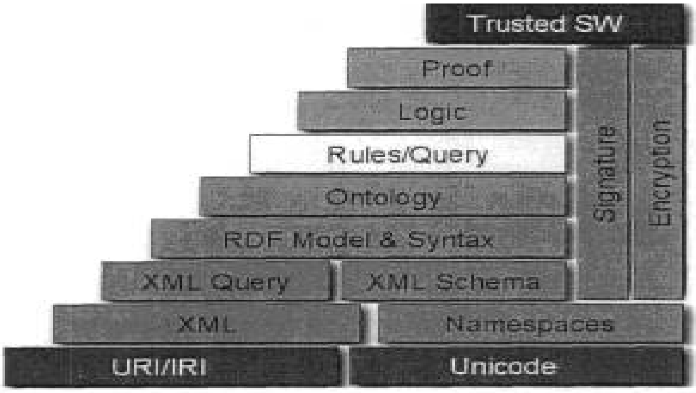


Fig. 1. The system structure of semantic web

The first layer: Unicode +URI (international code+ uniform resource logo). It is the foundation of the whole system structure. Unicode is a character set and is the code that is responsible for handling resources. URI is responsible for the identification of resources .it makes the precise retrieval of information and intelligence possible.

The second layer: xML + NS (Namespace) + XMLSchema (extensible markup language + name domain + program of extensible markup language). It is responsible for representing the content and structure of data from grammar. XML uses a set of elements which is programmed prior to mark on the data and provides convenience for the computer processing. Name domain can distinguish the data elements' belong and can transform the synonyms between different name domains.

The third layer: RDF + RDFSchema (framework of resource description+ framework program of resource description). It is responsible for providing semantic model to describe the content and structure of information on the Web. RDF is a language of describing the information resources on the Web and its goal is to establish a framework which is used for the coexistence of a variety of metadata standard. The framework can make full use of all sorts of metadata and start exchange and use of data which is based on Web. RDFSchema uses the expression system that machine can understand to define the vocabulary of description resources.

The forth layer: Ontologyvocabula (the collection of ontology vocabulary). It is responsible for the definition of sharing knowledge and the description of the relationship between the various resources.

The fifth floor: Logic. It is responsible for providing the inference rule of justice and logic and the basis for the intelligence service.

The sixth layer: Proof. It provides support for the signature of mutual validation and data exchange among intelligence agents.

The seventh layer: Trust. It provides trust guarantee.

The reasoning from the fifth layer to the seventh layer is on the based of the nether 4 layers.

The system structure of semantic web is under construction and the research of this system structure has not formed a logical description and theory system which is satisfying and strict at current international scope.

2.3 Concept Map

Concept map is a more modern method of knowledge representation. It was first proposed by John f. Sowa in 1984 and was a model of expressing language in network [6]. Concept map is a directed and connected graph that is represented by graph. It includes two nodes: concept node and concept relation node. The direction of the arc represents the relationship between concept knot and concept relation knot. Concept node represents a concrete or abstract entity in the field of question .concept relations note points out a kind of relationship that involves in one or more concept nodes.

In the concept map, concept node is represented with a rectangular and concept relation node is represented with an ellipse. Directed arc marks up the adjacent concept nodes of the concept relation node. Each concept map can represent a proposition and a typical knowledge base contains a lot of figures like this. For example, the concept map in picture 3 represents the proposition of "zhangsan gives lisi a red book ".

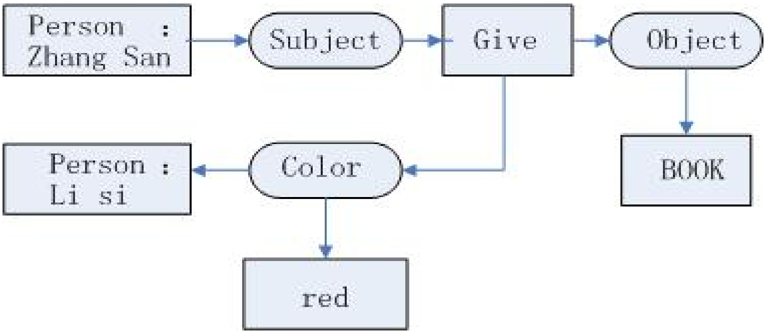


Fig. 2. Concept map examples

We can operate on concept map and create a new concept map. The formal rules of operation: copy, restrict, connect and simplify.

Concept map is a logic system which is based on semantic network. It is not only straightforward but also easy to operate for concept map to express the knowledge. It can produce new concept relevance and inference rule with the operating of concept map. In addition, the concept map can also create a mapping relationship with natural language directly. These advantages that concept map involves make it more suitable for expressing concept structure. Concept map has been favored by many researchers since it was putted forward and they have applied it in many different fields (such as knowledge engineering, information retrieval). Comparing with the classic method of knowledge representation, Concept map is more coincident with human's thinking and language habits. But it can only express some simple concept relations and is not applicable to express commonsense knowledge that contains complex concept structure. Conducting knowledge representing with concept map needs to analyze the structure of knowledge said, so its acquisition needs the participation of experts and it can not be gained automatically by intelligent system. In addition, for the solving of a complex question, reasoning based on concept map is easy to produce redundancy or cause the disaccord of reasoning result. Therefore, an intelligent system based on concept map can only do some simple problem solving and it is not competent for commonsense problem solving that contains a large amount of complex concept relevance.

2.4 Concept Lattice

Concept lattice is a complete concept hierarchical structure that reflects the contact between objects and attributes and the relations between generalization and specialization. Each node of Concept lattice is a formal concept and it consists of two parts: the denotation and the connotation. Denotation is the examples that concept covers and connotation is the description of concept which covers the common features of examples. In addition, concept lattice manifests the relationship between generalization and specialization of these concepts concisely and vividly by Hasse diagram. Concept lattice is regarded as a powerful tool for data analysis. The process of creating concept lattice from data set (which is called formal background in concept lattice) is a kind of process of concept clustering in essence. However, concept lattice can be used for many tasks of machine learning. Concept lattice has been applied extensively and successfully since it is putted forward. For example, we can extract multiple types of knowledge from concept lattice such as some properties and rules [8]. We can organize and manage a lot of information effectively by the concept lattice in information retrieval.

We can represent knowledge by concept lattice. We can also add the new obtainment of concept to the existing concept lattice by the structure algorithm of concept lattice, so as to we can update knowledge base constantly. Concept lattice can describe the category and hierarchical relationships of concept lattice clearly. It also can extract many common characteristics or rules of practical examples and organize them effectively. However, concept lattice is not suitable for expressing concept which has dynamic characteristics. We can obtain common properties of relevant concepts by concept lattice that is the connotation of concept. It is not good at expressing knowledge of common sense because the concept relations of commonsense knowledge contains are too complex and the concept relations that concept lattice can describe are too simple—they are only some relations of generalization and specialization. In addition, concept lattice has some problems itself. Such as the constructing efficiency of concept lattice, if the constructing algorithm is inappropriate, the number of concept lattice nodes may grow according to index level and this will cause a large number of redundant data.

2.5 HowNet

HowNet is a common sense knowledge which regards the concept that a Chinese word or an English word represents as its object and treats the relations between concepts or attributes of concept as its basic content [4]. With the investigation and analysis about 6000 Chinese characters, HowNet abstracts more than one thousand mete-senses. Mete-sense is the minimum unit which is the most basic and whose meaning can't be segmented in HowNet. It is the basic factor to explain the dictionary of knowledge and other vocabulary entries are all defined by it. Computerization is the important characteristic of know nets. HowNet is oriented to computer and it is established by computer. It may be an intelligent component of computer in the future. As a knowledge system, HowNet is a net but not a tree in fact [5].

HowNet also focus on reflecting the relationships between concepts or various attributes of concept. HowNet teaches the knowledge network system of figure 3 to the computer clearly and makes knowledge operable for the computer.

In general, the relationships that HowNet described between mete-senses mainly include hyponymy and, synonymy, amtonymy, the part-whole relation, the attributes-host relation, the material–product relation, the agency/experience/ subject of relation-event relation, the recipient/content/subjacency etc event relation, the tool-event relation, the place–event relation, the time-event relation, the value-attribute relation, the entity-value relation, the role-event relation and the correlative relationship and so on.

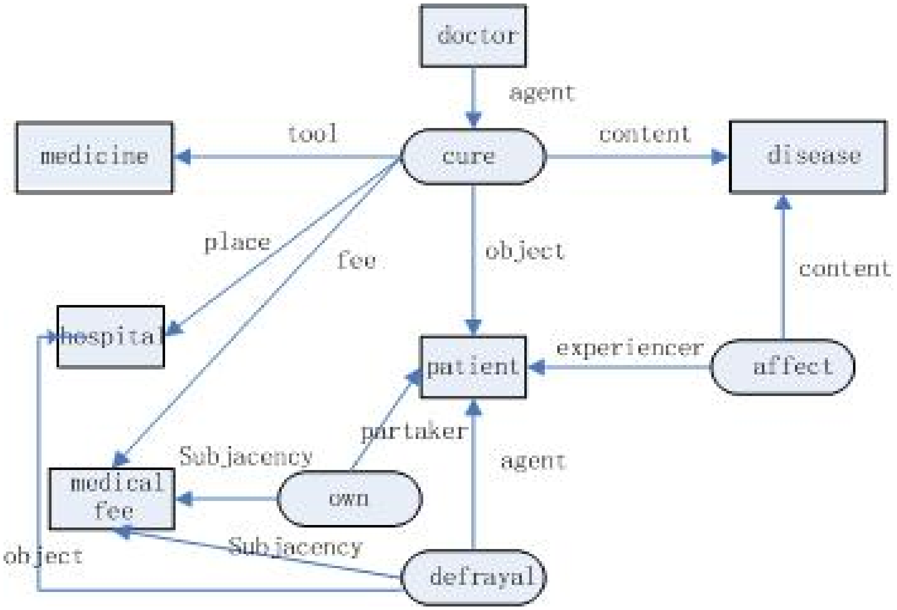


Fig. 3. Knowledge representations which is based on HowNet

2.6 HNC

HNC theory is a theory system about the understanding of natural language. It is a theory system that is based on semantic expression and mixes semanteme, grammar and pragmatic together. [7] HNC theory's goal is taking the association venation of concept as the main line, establishing a kind of natural language expression pattern and computer pattern of understanding and processing that can simulate the perception process of brain language and making the computer obtain the ability of digesting fuzzy.

HNC is based on the following two hypotheses:

Hypothesis 1: symbols of language space are different in thousands ways. However, there is only a kind of symbol of language concept space.

Hypothesis 2: role is in the internal and mutual space of all things and it must produce some role. It must be accompanying with a process or transfer before achieving final effect and it must be appearing a new relation or state after achieving final effect. New effect can induce new role and it recycles and recycles. This is the basic rule that all things exists and develops and this is also the basic rule of language expressing and concept reasoning (it becomes role and effect chain rule).

It introduces the language concept space. The meaning of the words can be mapped to the system of concept sign and it can be expressed by the combination of the concepts. HNC generalizes the symbolic expression of the natural language concept as the followings:

{Category string} {Hierarchical string} {Symbol of combination structure}

It means: the lambda expression of concept is consisted of category symbols, hierarchical symbols and combination symbols. Category string and hierarchical string constitute a lambda expression of concept primitive. Two or more concept primitives constitute a new concept by the combination of the combination structure symbol. The simple concept is consisted of a single concept primitive and the Compound concept is consisted of several concept primitives.

Example: Sun Zhongshan led the Xinhai revolution and overthrew the autocratic monarchy system which rules people thousands of years in china. It is great significant to the social progress, but it also fails to change the tragic fate of Chinese and society property of semi-colonial and semi-feudal in china.

Semantic tagging of the example's language concept space:

SG=R011X*20J#R0B1=<R411J>#R0B2=<!24R411X*21J[HEl]>+ReCS0jD1*20J#ReC= {PS*10J} +(lby)XY0*22J [7]

2.7 Ontology

Ontology is the clear specifications of conceptualization. It mainly includes four aspects:

(l)Conceptualization: abstract model of the objective world phenomenon.

(2)Clear: concepts and the relations among them are defined precisely.

(3)Formalization:accurate and mathematical description.It can be read by computer.

(4)Sharing: the knowledge that ontology reflects is recognized by its users.

Although there are many different ways about definition, different researchers have the unified understanding to ontology. They all take ontology as a semantic foundation of different subjects' (person, machine, software system, etc) communicating (dialogue, interoperability, sharing and so on) in the interior of field. That is ontology provides a consensus of clear definition. The goal of this consensus is mainly to service for the machine.

Many tools of ontology construction have been made in the past 10 years, such as from Ontolingua, OntoSaurus, WebOnto to Protege-2000, WebODE, OilEd, OntoEdit, KAON, Text-To-Onto, etc. These tools provide a graphical interface which is friendly and a mechanism of consistency check. With these tools, the users can concentrate on the organization of the ontology content without knowing the details of ontology description language and they also avoid a lot of mistakes and make it convenient for ontology construction. But, these tools only provide the editing function of ontology and what supports it is still the way of constructing ontology by hand. Even with these ontology editing tools, users still need to enter and edit the name, constraint, attributes, etc of each concept. The structure of the ontology is a five tuple ---O: = {C, R, Ao, F, S}. Here the C and R are two disjoint sets. The elements of C are called concepts and the elements of R are called relationship. Ao represents ontology axiom, F represents function and S represents example. We can see that the relation of ontology, example, function and axiom are all based on the basis of the concept and the concept system plays an important role on the construction ontology from the structure of the ontology.

3 Conclusions

Concept system reflects the dependence of the concepts, which is the core of knowledge system, and has played a role lurking in the background. People's cognitive ability all comes from that. Thus the research of concept structure is helpful for improving the problem solving ability of knowledge system, especially the open field. The research of concept system is asked to solve three problems: how to represent and store concept system, how to construct concept system gradually and how to realize all kinds of intelligent task using concept system.

Acknowledgments. This dissertation is supported by National Natural Science Foundation of China (Grant No.: 61171132).

References

1. Wei, H.: Formal concept system of artificial intelligence. Science Press, Beijing (2011)
2. Li, Q.: FrameNet–a thesaurus engineering based on frame semantics. Science and Technology BBS 16, 39 (2005)
3. Jia, J.-Z., Dong, G.: The compare among Frame, WordNet, VerbNet. Library and Information Technology 11, 1682–1686 (2007)
4. Liu, Q., Li, S.-J.: The calculation of lexical semantic similarity based on HowNet. Computer Linguistics and Chinese Information Processing 7, 59–76 (2002)
5. Dong, Z.-D., Dong, Q.: HowNet (EB/OL) (September 23, 1999/March 06, 2004), `http://www.keenage.com`
6. He, W., Wei, H.: Study summarization of Concept structure. Computer Application and Software 27(1), 157–159 (2010)
7. Jin, Y.-H.: HNC Language Understanding Technology and Its Applications. Science Press, Beijing (2006)
8. Wang, D.X., Hu, X.G., et al.: Association rules mining on concept lattice using domain knowledge. In: The Fourth International Conference on Machine Learning and Cybernetics, pp. 2151–2154 (2005)
9. Wooldridge: (Shi, C.-Y., et al. (trans.)) The guidance based on many Agent systems. Publishing house of electronics industry, Beijing (2003)

E-Government Adoption: A Conceptual Demarcation

Rahmath Safeena and Abdullah Kammani

College of Computers and Information Technology
Taif University, Saudi Arabia
safi.abdu@gmail.com, akamani@acm.org

Abstract. The Information and Communication Technologies (ICT) are being increasingly used by various governments to deliver their services at the locations convenient to its citizens. E-government is a kind of governmental administration which is based on ICT Services. The essence of e-government is using information technology to break the boundary of administrative organizations, and build up a virtual electronic government. E-government initiatives are common in most countries as they promise a transparent, citizen-centric government and reduce operational cost. Emerging with E-government, theories and practices of public administration have stepped into a new knowledge era. E-government presents a tremendous impetus to move forward with higher quality, cost-effective, government services and a better relationship between citizens and government. This paper discusses the different issues, challenges, adoption factors for e-government implementation and, presents a conceptual demarcation on these factors.

Keywords: e-government, definitions, services, adoption factors, and challenges.

1 Introduction

The current society had a phenomenal transformation due to the advance of Internet. It has opened a new medium of communication for individuals, business, and government organization, providing more opportunities to communicate and get information in an entirely new way. It has made governmental information and services accessible in ways that could not have been conceived two decades ago [1]. In the past, government organizations paid little attention to service quality or responsiveness to clients, but this changed with the approach of E-Government. E-government refers to the use by state authorities of ICT, in particular, the Internet and web-based technology, to deliver information and services and to encourage civic participation [2]. E-government is simply a facility using Information Technology (IT) to deliver public services directly to the customer, where the customers are citizens, business or other government entity [3, 4]. This phenomenon of e-government is increasingly attracting the attention of community citizens including politicians, economists, decision and policy makers amongst others. It has improved managerial effectiveness, and promoted democratic values of public services. It has the promise of increasing

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 67–76.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

accessibility to information, enhancing efficiency and facilitation of greater access to government officials [5, 6]. It is the medium of delivering improved services to citizens, businesses, and other constituents of society through drastically changing the way governments manage information. However, the e-Government challenge is not a technological one. Rather, the challenge is to use technologies to improve the capacities of government institutions, while improving the quality of life of citizens by redefining the relationship between citizens and their government [1]. The development of e-government also means increased electronic co-operation within and among public organizations which even puts demands on development that is not technology oriented. The development towards e-government involves social changes of work roles, attitudes and new competence needs [7].

2 E-Government

E-Government initially began as an intra-governmental communication tool. Initially the government organizations developed websites with information, then developed to online transactions - which made the citizens to engage in online participation that connect citizens and decision-makers [8–11]. E-government represents a fundamental change in the whole public sector structure, values, culture and the ways of conducting business by utilizing the potential of ICT as a tool in the government agency [12]. The Internet is indeed the most powerful and popular means of delivering the services to the customers or citizens. Hence, Web sites have been employed as a platform for delivering a wide range of government services electronically. E-government websites help citizens to gain information on government processes and services and hence participating in democratic processes from anywhere at any time. E-Government improves the efficiency and effectiveness all government operations, with citizens, as well as with other organizations. E-government applications include online payment of tax, bills, filling and submission of applications for several purposes; e-voting etc. e-Government gives citizens more control on interaction with the government; citizens can avail of the governmental services from anywhere and anytime [1]. E-Government is considered as tool for easy administration of governmental activities. Its success depends on its vast usage and management of its infrastructure. Utilization of e-government will provide benefits to the management philosophy of governments. Thus the citizens can collaboratively participate in decision making [13, 14]. Initially E-Government incurs a great cost in building infrastructure but gradually it implementation results in vast savings towards government's activities. It also increases transparency, and reduce corrupt activities in public service delivery. Table 1 depicts E-Government defined by various related study in the near past.

Table 1. E-government Definitions

Citations	Definition
[16, 17]	E-government is defined as the use of ICT to make government more accessible, effective, and accountable.
[13, 18, 19]	E-Government refers to the delivery of [government] information and services online through the Internet or other digital means.
[20]	E-government refers to strategies, organizational forms and processes, as well as information technology employed so as to enhance access to and delivery of government information and services to citizens, businesses, government employees and other agencies.
[21, 22]	E-Government is the use of ICTs in public administrations combined with organizational change and new skills in order to improve public services and democratic processes and strengthen support to public policies.
[23]	E-government is the process of offering better government service to the public.
[13, 14]	E-government is defined as the combination e-administration and e-democracy to achieve the objective of balanced e-government.
[13, 18]	E-Government is the delivery of fast services to citizens, businesses, and other members of the society.
[24–26]	E-Government refers to the strategic application of ICT to "provide citizens and organizations with more convenient access to government information and services; and to provide delivery of public services to citizens, business partners and suppliers, and those working in the public sector"
[15, 27]	E-government is the continuous optimization of service delivery channel, citizen's participation and governance.
[15]	E-government can be defined as a way for governments to use the most innovative information and communication technologies, particularly web-based Internet applications, to provide citizens and businesses with more convenient access to government information and services, to improve the quality of the services and to provide greater opportunities to participate in democratic institutions and processes.

3 E-Government Services

Like any other electronic services, e-government also constitutes various types of services. According to Fang [15] different types of e-government services are categorized in to eight types. 1) Government-to-Citizen (G2C); 2) Citizen-to-Government (C2G); 3) Government-to-Business (G2B); 4) Business -to-Government (B2G); 5) Government-to-Employee (G2E); 6) Government-to-Government (G2G); 7) Government-to-Nonprofit (G2N); 8) Nonprofit-to-Government (N2G). Table 2 gives definition for these of e-government services.

Table 2. E-Government services

Types	Definition
Government-to-Citizen (G2C)	It is an e-government service, from government to citizen in the form of offering valuable information and know-how's.
Citizen-to-Government (C2G)	It is an e-government service, offered for payment of bills and other valuable feedback from the citizen to government.
Government-to-Business (G2B)	It is an e-government service providing transactions and procurement facilities for government purchases and call for tenders.
Business -to-Government (B2G)	It is an e-government service providing communication, collaboration, transactions and procurement of goods and services for business initiatives.
Government-to-Employee (G2E)	It is an e-government initiative that will facilitate the management of the civil service and internal communication with governmental employees to encourage paperless office.
Government-to-Government (G2G)	It is an e-government initiative to provide the Government's departments or agencies cooperation and communication online. It includes internal exchange of information and commodities.
Government-to-Nonprofit (G2N)	It is an e-government initiative that provides information and communication from government to nonprofit organizations, political parties and social organizations, Legislature, etc.
Nonprofit-to-Government (N2G)	It is an e-government initiative that enable exchange of information and communication from non-profit organization to government organizations, political parties and social organizations, Legislature, etc.

4 Discussion: E-Government Adoption Factors

The adoption factors for e-government services should be thoroughly known before any adoption model is constructed. Many researchers have understood the initiatives that encourage the adoption of e-government services in different environments. These studies have shown that despite different environments having different characteristics, there are general initiatives that promote e-government adoption by ordinary citizens. However, it is worth mentioning that certain situations have unique factors which may either impend or aid the adoption of e-government services. In order to have a basic understanding of these varying factors, this study review's the adoption models that have been studied in different locations. The factors that influence the adoption of e-government websites are information quality, system quality and service quality. Information quality is concerned with the measure of the information that the system produces and delivers i.e., characteristics of information produced by e-government Web sites. Quality of information is believed to be the most salient factor for predicting customer decision-making behavior and user intention to use a particular system [30]. The fundamental dimensions of information quality are composed of five dimensions: accuracy, timeliness, relevance, understandability, and completeness [31–33].

System quality refers to the features and performance characteristics of e-government Web sites regarding the quality in use or the citizen's view of quality. It is an important determinant of user acceptance, user satisfaction and system use. In order for the citizens to continually use the e-government website or for the successfulness of e-government website system quality should be high. Service quality refers to the quality of personal support services provided to citizens through e-government Web sites, such as answering questions, taking requests, and providing sophisticated solutions to citizen's problems. It is an important determinant of customer satisfaction and is needed as citizens differ in knowledge, education and experience [30]. Quality of service is composed of five dimensions: tangibles, empathy, reliability, responsiveness and assurance [34]. Choudrie and Dwivedi [35] found that citizens' awareness as a factor for the adoption of e-government. Citizens with fulltime internet access are more likely to be aware of and adopt e-government services. These authors also add that the demographic characteristics of citizens such as the age, gender, education, and social class have an imperative role in explaining the citizen's awareness and adoption of e-government services. While, Warkentin [36] proposed a e-government adoption conceptual model with citizen trust as the underlying catalyst. The author proposes perceived risk, perceived behavioral control, usefulness, and perceived ease of use. Perceived risk is normally defined as a fear of losing personal information or money, and fear of being spied on the Internet. Warkentin states that Perceived risk is negatively related to adoption. The author posits that the perception that an individual has of control over how personal information will be used, and control over how and when information can be acquired, could encourage adoption. Perceived usefulness on the other hand is simply defined as the utility of the system to the user, and perceived ease of use is termed as a system that is easy to use [36].

Alomari, Woods and Sandhu [37] attempted to identify the main factors that influence citizens' intention to adopt e-government websites in Jordan, using a theoretical framework consisting of Diffusion of Innovation Theory (DOI) and the Technology Acceptance Model (TAM) and, they found that Trust in government, website design, beliefs, complexity and perceived usefulness were significant factors in Jordanian citizens' intention to use e-government websites. Deltor and Hupfer [38] identified internal factors within government that affect the adoption and use of government websites and suggest that Partner cooperation, Ability to change internal work processes, IT workforce, funding, citizen participation in design, portal strategies and policies, leadership, marketing and governance as factors for e-government adoption. Chen et.al [39] uses UTAUT model to study on the factors affecting e-government adoption and found that performance expectancy, effort expectancy, social influence, and facilitating conditions impact citizen satisfaction. Rokhman [40] Conducted citizen's willingness to accept and adopt e-government services and found that relative advantage and compatibility proven as useful factors to predict intention to use e-government services. Gilbert and Balestrini [41] bring attitude-based and service-quality-based approaches together. They propose perceived (confidentiality, ease of use, safety, reliability, visual appeal and enjoyment) and perceived relative benefits as the adoption factors for e-government. Phang et al., [25] made a study on senior citizen's adoption of e-government and they found that compatibility, personal image, perceived ease of use and internet safety perceptions are the main factors for e-government adoption. Table 3 shows brief classification of the adoption factors. These adoption factors of e-government were not classified properly in previous literatures. This study demarcates the adoption factors as technological, financial, website quality, user/human, managerial and political perspectives as described in table 3.

Table 3. E-government adoption factor

Factors	Items	Description	Studies
Technological	Standardization procedures, technical infrastructure, security measures[data and software protection, data transfer over networks, safety of electronic payments	The degree of Technological ability achieved for Effective E-Business adoption.	[10, 13, 43–45]
Financial	Appropriate budget allocation, commitment of funding	The degree of Technological ability achieved for Effective E-Business adoption.	[20, 45]
Web site quality	Information quality, system quality, service quality, perceived usefulness, perceived ease of use, user friendliness of the system, scope of the system, Multi-lingual and multi-cultural Issues, protection of information assets, maintain integrity of electronic records, compatibility, internet safety perceptions.	The degree of Web site quality ability achieved for Effective E-Business adoption.	[46, 47]
Human or user	ICT skills, technology expertise, perceived behavioral control, age, Perceived risk ,Uncertainty avoidance, trust, Security, privacy, Fears for job loss by the adoption of new technologies and procedures, Past experience specific to the project, Communication skills, ease of access to the system, cost of use of the system, local language, personal image.	The degree of user ability achieved for Effective E-Business adoption.	[36, 41]
Managerial	Technology culture of management Personnel, Project familiarization of management Personnel, project management, appropriate hierarchy in management, qualifications of the officials, active support from management, Lack of IT knowledge staffs, lack of IT full time employees, staff, Awareness and training	The degree of Managerial ability achieved for Effective E-Business adoption.	[45]
Political	Long-term, unified support, Technology culture of political leadership, Project familiarization of political leadership, Jurisdiction conflict resolution between government agencies, statutory/legislative requirements, Regulatory barriers; regulatory support.	The degree of Political ability achieved for Effective E-Business adoption.	[45]

5 Discussions: E-Government Challenges

Implementation of e-government projects can cause number of challenges as it is the redefining of complete government processes. There are a number of impediments that potentially block e-Government adoption. Barriers can be any factors that cause hindrance to government in developing new or further improving the existing e-government applications. Initial use of e-government Web sites is an important indicator of e-government success. Some information system research indicated that its eventual success depends on its continued use rather than first-time use [17, 28, 29]. However, the desired outcome is not achieved unless a significant number of citizens move beyond the initial adoption and use e-government web sites on a persistent basis [17]. According to case studies from different countries, there are many challenges and issues that need to be addressed for successful implementation of e-government. There are distinct factors that command the adoption of e-government, and these factors depend on the local context of any country. But there is no clear classification of these adoption factors. Warkentin et al [36] describes e-government adoption as the citizen intention to participate in government activity electronically to receive information and request services from the government. According to Carter and Belanger [42] it is intent to use, while Gilbert and Balestrini [41] measure it as willingness to use e-Government services. Altogether it can be stated as a simple decision to use, or not to use, e-government services. The next level of challenge of e-government is to make it frequently used by the citizen. Using e-government service once in a year would not be considered as a meaningful usage of its application. Citizens technical awareness – on how to adapt to frequent technical changes in the services – is another very important challenge of adoption. The successful adoption of e-government can be achieved by developing a set of e-government competencies/adoption factors and investigating the significant relationships of those factors on its performance.

6 Conclusion

Recent advances in ICTs are giving organizations a new competitive edge. Growing impact of ICT, surge in the usage and adoption of E-government services. The globalization of organizations facilitated by the advent of telecommunications and internet technologies has promoted adoption of E-government. The best practices of E-government are revolutionizing not just technology itself but the whole process through which services are provided. This study helps better understand E-government and identifies various competence factors like technological, managerial, political, user and website quality for E-government adoption. This study would provide researchers to do an empirical examination on the identified factors of E-government with its performance .This study would also help managers decide to what extent their organizations should invest in E-government by matching the E-government attributes to their own organization's characteristics. Our investigation of E- government adoption factors may help both researcher and potential adopters.

References

1. Kumar, V., Mukerji, B., Butt, I., Persaud, A.: Factors for Successful e-Government Adoption: a Conceptual Framework. The Electronic Journal of e-Government 5, 63–76 (2007)
2. Luk, S.C.Y.: The Impact of E-government in Greater China: Case Studies of Hong Kong, Taiwan, and Singapore. Presented at the 17th Biennial Conference of the Asian Studies Association, Australia (July 2008)
3. Ghapanchi, A., Albadvi, A., Zarei, B.: A framework for e-government planning and implementation. An International Journal Electronic Government 5, 71–90 (2008)
4. Metaxiotis, K., Psarras, J.: A conceptual analysis of knowledge management in e-government. An International Journal Electronic Government 2, 77–86 (2005)
5. Koh, C., Ryan, S., Prybutok, V.: Creating Value Through Managing Knowledge in an E-Government to Constituency(g2c) Environment. Journal of Computer lnformation System 45, 32–41 (2005)
6. Groznik, A., Kovacic, A., Trkman, P.: The Role of Business Renovation and Information in E-Government. Journal of Computer lnformation System, 81–89 (2008)
7. Grundén, K.: A Social Perspective on Implementation of e-Government - a Longitudinal Study at the County Administration of Sweden. Electronic Journal of e-Government 7, 65–76 (2009)
8. Calista, D., Melitski, J.: e-Government and E-governance: Converging Constructs of Public Sector Information and Communications Technologies. Public Administration Quarterly 31 (2007)
9. Holzer, M., Melitski, J., Rho, S.-Y., Schwester, R.: Restoring Trust in Government: The Potential of Digital Citizen Participation, Washington, DC. IBM Endowment for the Business of Government (2004)
10. Schwester, R.: Examining the Barriers to e-Government Adoption. Electronic Journal of e-Government 7, 113–122 (2009)
11. Moon, M.: The Evolution of E-Government Among Municipalities: Rhetoric or Reality? Public Administration Review 62, 424–433 (2002)
12. Alshehri, M., Drew, S., Alfarraj, O.: A Comprehensive Analysis of E-Government Services Adoption in Saudi Arbaia: Obstacles and Challenges. International Journal of Advanced Computer Science and Applications 3 (2012)
13. Bwalya, K.: Factors Affecting Adoption of E-Government in Zambia. The Electronic Journal on Information Systems in Developing Countries 38, 1–13 (2009)
14. Coleman, S.: African e-Governance – Opportunities and Challenges. Oxford University Press, University of Oxford (2006)
15. Fang, Z.: E-Government in Digital Era: Concept, Practice, and Development. International Journal of the Computer, The Internet and Management 10, 1–22 (2002)
16. InfoDev and CDT: The E-government Handbook for Developing Countries. InfoDev and the Center for Democracy & Technology (2002)
17. Wangpipatwong, S.: Quality Enhancing the Continued Use of E-Government Web Sites: Evidence from E-Citizens of Thailand. International Journal of Electronic Government Research 5, 19–35 (2009)
18. Kumar, V., Mukerji, B., Butt, I., Persaud, A.: Factors for Successful e-Government Adoption: a Conceptual Framework. The Electronic Journal of e-Government 5, 63–76 (2007)
19. Muir, A., Oppenheim, C.: National Information Policy developments worldwide in electronic government. Journal of Information Science 28, 173–186 (2002)

20. Kefallinos, D., Lambrou, M., Sykas, E.: An Extended Risk Assessment Model for Secure E-Government Projects. International Journal of Electronic Government Research 5, 72–92 (2009)
21. Akesson, M., Skalen, P., Edvardsson, B.: E-government and service orientation: gaps between theory and practice. International Journal of Public Sector Management 21, 74–92 (2008)
22. Commission of the European Communities: The Role of eGovernment for Europe's Future. Communication No. 567, The Commission, Brussels (2003)
23. Sridhar, S.: E-Government - a Proactive Participant for E-Learning in Higher Education. Journal of American Academy of Business 7, 258–268 (2005)
24. Gronlund, A.: Introduction. Electronic Government: Design, Applications & Management. Idea Group Publishing, PA (2001)
25. Phang, C., Sutano, J., Li, Y., Kankanhalli, A.: Senior Citizens. In: Phang, C., Sutano, J., Li, Y., Kankanhalli, A. (eds.) Presented at the Adoption of E-Government: In Quest of the Antecedents of Perceived Usefulness, Hawaii (2005)
26. Turban, E., King, J., Lee, M., Warkentin, M., Chung, H.: Electronic Commerce 2002: A Managerial Perspective. Prentice Hall, Upper Saddle River (2002)
27. Baum, C., Di Maio, A., Caldwell, F.: What Is E-Government? Gartner's Definitions (2000)
28. Bhattacherjee, A.: Understanding Information Systems Continuance: An Expectation-Confirmation Model. MIS Quarterly 25, 351–370 (2001)
29. Limayem, M., Hirt, S., Cheung, C.M.: Habit in the Context of Is Continuance: Theory Extension and Scale Development. Presented at the 11th European Conference on Information Systems (2003)
30. DeLone, W., McLean, E.: Information Systems Success: The Quest for the Dependent Variable. Information Systems Research 3, 60–95 (1992)
31. Bailey, J., Pearson, S.: Developing a Tool for Measuring and Analyzing Computer User Satisfaction. Management Science 29, 530–545 (1983)
32. Doll, W., Torkzadeh, G.: The Measurement of End-User Computing Satisfaction. MIS Quarterly 12, 259–274 (1988)
33. Wang, R., Strong, D.: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12, 5–34 (1996)
34. Parasuraman, A., Zeithaml, V.A., Berry, L.: SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. Journal of Retailing 64, 12–40 (1988)
35. Choudrie, J., Dwivedi, Y.: A Survey of Citizens Adoption and Awareness of E-Government Initiatives, the Government Gateway: A United Kingdom Perspective. Brunel University, West London (2005)
36. Warkentin, M., Gefen, D., Pavlou, P., Rose, G.: Encouraging Citizen Adoption of e-Government by Building Trust. Electronic Markets 12, 157–162 (2002)
37. Alomari, M.K., Woods, P., Sandhu, K.: Predictors for E-government Adoption in Jordan: Deployment of an Empirical Evaluation Based on a Citizen-centric Approach. Information Technology & People 25 (2012)
38. Deltor, B., Hupfer, M.E.: Internal Factors Affecting the Adoption and Use of Government Websites. An International Journal Electronic Government 7 (2010)
39. Chan, F., Thong, J., Venkatesh, V., Brown, S., Hu, P.J., Tam, K.: Modeling Citizen Satisfaction with Mandatory Adoption of an E-Government Technology. Journal of the Association for Information Systems 11, 519–549 (2010)

40. Rokhman, A.: E-Government Adoption in Developing Countries; the Case of Indonesia. Journal of Emerging Trends in Computing and Information Sciences 2, 228–236 (2011)
41. Gilbert, D., Balestrini, P.: Barriers and Benefits in the Adoption of E-Government. International Journal of Public Sector Management 17, 286–301 (2004)
42. Carter, L., Bélanger, F.: The Utilisation of E-Government Services: Citizen Trust, Innovation and Acceptance Factors. Information Systems Journal 15, 5–25 (2005)
43. Andersen, K.V.: e-Government: Five Key Challenges for Management. The Electronic Journal of e-Government 4, 1–8 (2006)
44. Azab, N., Kamel, S., Dafoulas, G.: A Suggested Framework for Assessing Electronic Government Readiness in Egypt. Electronic Journal of e-Government 7, 11–28 (2009)
45. Smith, S., Jamieson, R.: Determining Key Factors in E-Government Information System Security. Information Systems Management 23, 23–32 (2006)
46. Pitt, L., Watson, R., Kavan, C.: Service Quality: A Measure of Information Systems Effectiveness. MIS Quarterly 19, 173–185 (1995)
47. Reichheld, F., Markey, R., Hopton, C.: E-Customer Loyalty – Applying the Traditional Rules of Business for Online Success. European Business Journal 12, 173–179 (2000)

Expert Webest Tool: A Web Based Application, Estimate the Cost and Risk of Software Project Using Function Points

Ajay Jaiswal[1] and Meena Sharma[2]

[1] Department of Computer Science & Engineering
CDSE, Indore (MP)-452008, India
jaiswal.ajay05@gmail.com
[2] Reader (Computer Engineering)
Department of Computer Engineering
IET, DAVV, Indore (MP)-452001, India
meena@myself.com

Abstract. There are several area of the software engineering in which we can use the function point analysis (FPA) like project planning, project construction, software implementation etc. In software development, accuracy and efficiency of cost estimation methodology for a web based application is very important. The proposed web based application (i.e. Expert webest tool), is to produce accurate cost estimation and risk estimation throughout the software development cycle to determine feasibility of software project. Cost of the software projects depends on the project size, project type, cost adjustment factor, cost driven factors, nature and characteristics of the project. Software estimation needs to estimates or predict the software costs and software risk early in the software life-cycle.

In this paper we proposed the Expert webest tool in Java, this tool is used to two different purpose, first to estimate the cost of the software & secondly, to estimate the risk in the software. Most of the software's fails due to over budget, delay in the delivery of the software & so on. Function point is a well known established method to estimate the size of software projects. Its measure of software size that uses logical functional terms, business owners & user, more readily understand.

The management of risks is a central issue in the planning and management of any venture. In the field of software, Risk Management is a critical discipline. The process of risk management embodies the identification, analysis, planning, tracking, controlling, and communication of risk. It gives us a structured mechanism to provide visibility into threats to projects success. Risk management is a discipline for living with the possibility that future events may cause adverse effects. Risk management partly means reducing uncertainty. The propose tool indicates the risk & estimates risk using risk exposure. Management team to estimates the cost & risk within a planned budget and provide a fundamental motivation towards the development of web based application project. Find heuristic risk assessment using cost factors, indicating product & project risk using some risk factors & check some risk management strategies in under estimation development time.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 77–86.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Keywords-component: Cost estimation, Expert judgment, Risk management strategies, Type of risks, Expert webest tool, Risk factors.

1 Introduction

Software estimation is a step by step approach to estimating the cost & risk for every project. Estimation area of the software development is size, effort invested, development time, technology used & quality. Web cost project or application is needed accurate cost estimation, producing accurate result. The recommenced steps are used to identify the risks, determine the risk exposure, develop strategies to mitigate the risks, handle the risks.

Improving the functions of project management is a main concern in software development organizations. Suitable budgeting & cost is essential ingredients of a successful project. In that essence the total cost of product must be known at the early design stage, with the maximum of accuracy in order to simplify the trial & error process [1].The cost of software projects depends on the nature & characteristics of the projects and therefore, the accuracy of the estimation model rely on the data by some evaluation affected by high degree of imprecision & uncertainty.

Albrecht's model [5] of functional specification requires the identification of 5 types of components, namely External input, External output, External inquiry element processes, logical internal elementary processes & logical internal & external interface files. The actual calculation process itself is accomplished in 3 stages:

1. Find out the unadjusted function point (UFP)
2. Find out the value adjusted factor (VAF)
3. Find out the adjusted function points (AFP), other details advised please refer to [2][3].

Estimated cost mapped to function points. FPA is a measurement of functional requirement in terms of business transaction & business data. Transaction can be classified into three types of external inputs, external outputs, & external inquiries [3, 15]. The process of risk management embodies the identification, quantification, response & control of risk. We need to balance the cost & indicating risk & minimize or reducing that risk. There are two dimensions of software risk, project risk and product risk.

We have explained the Expert webest tool is organized as follows. In section 2, we present the background and related work. In section 2, we have explained the Expert webest tool. In section 3, we have explained experimental work carried out. Finally we conclude the paper in section 4.

1 Background and Related Work

The objective of this paper is to produce nearly accurate cost estimation & risk minimization reliable and accurate estimation of software development cost and risk estimation are needed throughout the software development cycle to determine feasibility of software project.

Cost of the software projects depends on the project size, project type, cost adjustment factor, cost driven factors, nature and characteristics of the project [1]. Software cost estimation needs to estimates or predict the project costs early in the software life-cycle. Many different models of the software cost & effort estimation have been developed and used during past decades, they are expressed such predefined functions like size of the product, level of reuse parameters etc., therefore an accurate estimation of the project cost of a software project will most likely lead to more successful results & on time completion.

Function point is a measure of software size that uses logical functional terms inputs and outputs. A risk is in exposure to loss or injury or a factor, thing, element or course that involves uncertain danger. Risk assessment involves risk identification, risk analysis, risk planning & risk controlling. Software risk can be internal or external; the internal risks come from risk factor within the organization. The external risks come from out of the organization & are difficult to control. Software risks can be grouped into project risks, process risks & product risks. [13]**.** The proposed architecture gives the incremental risk as the software progress from phase to phase. In [4] the author have developed a software estimation tool based on software engineering metrics model, but in this tool there is no description regarding the costing of the software using ISBSG (International Software Benchmarking standard Group Release Report).

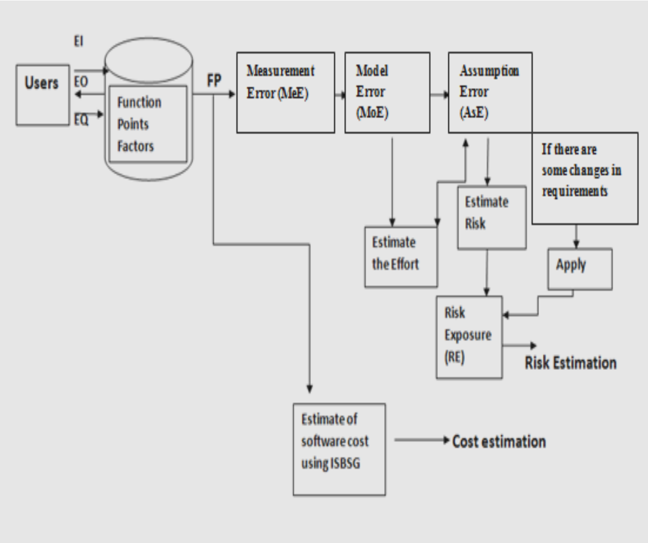


Fig. 1. Architecture of the proposed Expert webest Tool

In this paper we have developed the architecture of the Expert webest tool in order to estimate the cost & risk of software. Architecture of the proposed Expert webest tool is given in fig.(1).

2 Expert Webest Tool

The proposed tool i.e. Expert webest tool, indicate the risk of project & estimate the cost of project. A tool namely Expert webest tool has been developed in the research by using java programming language and java eclipse based on techniques selected which is combination of applying function point & risk management process. Find out the cost of software & estimate the risk of the software.

"Estimation is a prediction that is equally likely above or below the actual results". Estimation uncertainty occurs because an estimate is a probabilistic assessment of future condition. Risk assessment provides a snapshot of the risk situation & is part of a viable risk management program. There are four key factors of risk assessment & these factors are risk identification, risk analysis, risk planning & risk controlling. The first step of this tool is to calculate the function point of an input to the measurement error, model error & assumption error. The architecture of the proposed tool is given in fig.(1) adopted from [5].

2.1 Estimate the Cost of Project

In the proposed tool we have used the International Software Benchmarking Standards Group (ISBSG). The ***ISBSG*** delivers a database of software project history data that is used for estimation, benchmarking and project management. It is an international group of representatives from international metrics organizations who collect project data from countries like, India, Hong Kong, Germany, Japan, and USA. ISBSG Release 6 Report provides the cost value for the software projects. Cost data is derived from 56 projects representing a broad cross section of the software industry. After going through these software projects, the ISBSG conclude that median cost to develop a function point is $US 716, and the average cost is $ US 849 per function point. For more information about the ISBSG please visit: www.ISBSG.org.au[3]. In the calculation of the function point, calculating the value adjustment factor (VAF) is an indicate of the general functionality provided to the user. The VAF is derived from the sum of the degree of influence (DI) of the 14 general system characteristics (GSCs). The 14 GSCs show in table 1. The DI of each one of these characteristics ranges from 0 to 5 as follows:

(i) 0 – no influence;
(ii) 1 – incidental influence;
(iii) 2 – moderate influence;
(iv) 3 – average influence;
(v) 4 – significant influence; and
(vi) 5 – strong influence.

The third and the last stage is the final calculation of the function points. With the help of the following equation we can get the total points of an application.

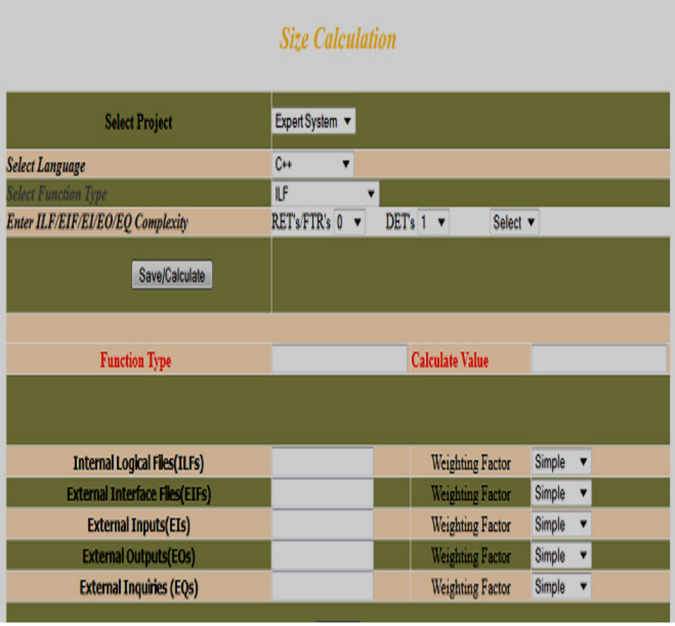
AFP = UFP * VAF [3, 9, 10, and 16]

Table 1. The general characteristics of a system

General system characteristics	
Data communications	Online update
Distributed data processing	Complex processing
Performance	Reusability
Heavily used configuration	Installation ease
Transaction rate	Operational ease
Online data entry	Multiple sites
End user efficiency	Facilitate change

Where **AFP** = adjusted function points; **UFP** = unadjusted function points; and **VAF** = value adjustment factor. [1, 8, 9] Function points are computed by completing the table 1. Five information domain characteristics are determined and counts are provided in appropriate table location [9]. In table 2, the type of components, i.e. External Input (EI), External Output (EO), External Query (EQ), Internal Logical File (ILF), External Interface File (EIF) and UFP is the unadjusted function point. Once these data have been collected, a complexity value is associated with each count. Organization that use FP methods develop criteria for determining whether a particular entity is simple, average, or complex. To compute the AFP the following relationship is used [11, 12, and 13]. Fig.(2), show the size calculation using the function points & fig (3) show the chart of function point verses effort calculation.

$$\mathbf{AFP = UFP * [0.65+0.01 * \Sigma\ i=1\ldots14\ (DI)]}$$


Size Calculation

Select Project: Expert System

Select Language: C++

Select Function Type: ILF

Enter ILF/EIF/EI/EO/EQ Complexity: RET's/FTR's 0, DET's 1, Select

Internal Logical Files(ILFs) — Weighting Factor: Simple

External Interface Files(EIFs) — Weighting Factor: Simple

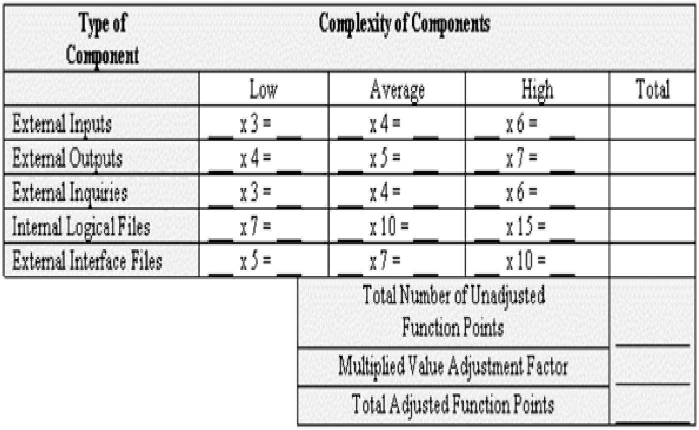
External Inputs(EIs) — Weighting Factor: Simple

External Outputs(EOs) — Weighting Factor: Simple

External Inquiries (EQs) — Weighting Factor: Simple

Fig. 2. Snapshot of how to "Expert webest tool" calculate the function points

Table 2. Computed Unadjusted function point values [15, 16]



Type of Component	Complexity of Components			
	Low	Average	High	Total
External Inputs	__ x 3 = __	__ x 4 = __	__ x 6 = __	
External Outputs	__ x 4 = __	__ x 5 = __	__ x 7 = __	
External Inquiries	__ x 3 = __	__ x 4 = __	__ x 6 = __	
Internal Logical Files	__ x 7 = __	__ x 10 = __	__ x 15 = __	
External Interface Files	__ x 5 = __	__ x 7 = __	__ x 10 = __	
		Total Number of Unadjusted Function Points		____
		Multiplied Value Adjustment Factor		____
		Total Adjusted Function Points		____

2.2 Estimation of the Risk

Each software model has some weakness & also has some advantages. Every software project is exposed to adverse external influences, the so called project risks, which affect the cost and the duration of the project and, possibly, the quality of the products. With a risk analysis it can be determined for a specific project what the risks are. These risks then should be included in a systematic and formal manner in the project estimate in order to obtain a realistic and reliable project estimate and a realistic project plan. There are three dimensions of software risk i.e. technical risk, organization & environmental risk. Risk assessment provides a snapshot of the risk

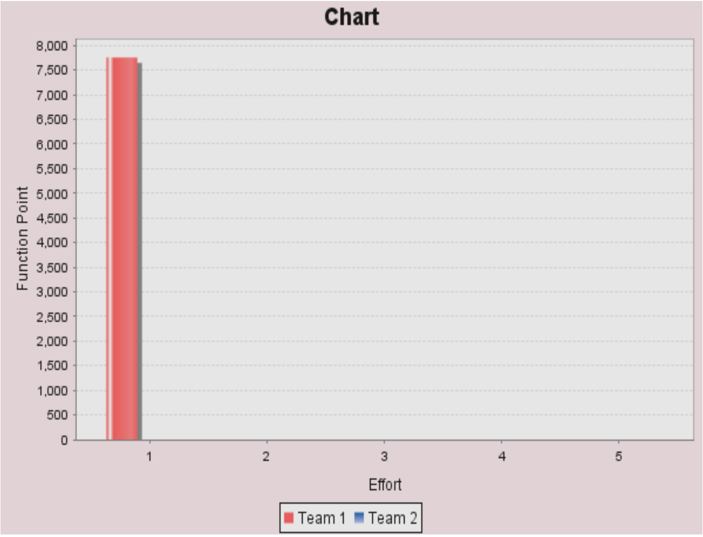


Fig. 3. Graphical representation function point verses effort estimation

situation & is part of a viable risk management program. There are 3 key factors of risk assessment and these factors are risk identification, risk analysis & risk prioritization [14].

- □ **Risk identification** produce list of projects specific risk items, likely to compromise a project success. A typical risk identification technique includes examination of drivers, assumption, analysis & checklist.
- □ **Risk analysis** assesses the loss probability & loss magnitude for each identified risk item and it access compound risk in the risk interaction [11]. Typical technique include performance models, cost models network analysis etc.
- □ **Risk prioritization** produces a ranked ordering of the risk items identified and analyzed. Typical techniques include risk exposure analysis, risk reduction leverage etc.
- □ **Risk planning** helps to prepare you to address each risk items including the coordination of individuals risk item plans with each other & with the overall project plans.
- □ **Risk monitoring** involves tracking the project's program towards resolving its risk items & tracking corrective action where appropriate.

2.3 Estimating Uncertainty Risks

The most critical aspect to any long-term risk management or estimation plan is to recognize and communicate the fact that it is something that will become more exact, or accurate, over time. Initially the model estimates the source of uncertainty using measurement error, model error & assumption error. We have considered the concept of function points to explain the measurement error, model error & assumption error. Function point is an important software metrics which is used to calculate the approximate LOC, cost & effort of software

A. Measurement Error (MeE)

For measurement error (MeE) this is the recognition that some of the input variables in an estimation model might have inherent accuracy limitations. (This is definitely the case with function point-type metrics, which are generally going to be about 12% inaccurate). As a result of Chris F. kemerer [6], function points are assumed to be at least 12% inaccurate. Thus if we estimate a product size of 2000 points, measurement error could mean that the read size is anywhere between 1760 and 1140.

B. Model Error (MoE)

For model errors (MoE), this is the recognition that the model used for estimation cannot generally include all the factors that affect the effort required to produce a software product. Factors that affect effort but are not included explicitly in the model contribute to the model error and this needs to be recognized. This error occurs because all empirical models are abstraction of reality. For eg. Such as 0.5 person-days per function point is usually obtained from results observed for recalled from previous projects. Model is expected to all right on average. If you have a base model on past project data, you should calculate the associated inaccuracy by using the mean magnitude relative error. Thus if you have an estimation model with an inherent 20% inaccuracy & your product is 1000 function points in size, your estimate is likely to be between 400 & 600 person-days.

C. Assumption Error (AsE)

The assumption error (AsE) happens when someone makes incorrect assumptions about the model's input parameters. So what we try to show management is that if you can identify your assumptions, you can investigate the effect of their being invalid by assessing both the probability that an assumption is incorrect and the resulting impact on the estimate. (This is just basic risk analysis but sort of gussied up to make it seem more technical than it is.) For eg. Your assessment that a product size is 1300 function points rests on the assumption that you have correctly identified all the customer requirements. If you can identify your assumption, you can investigate the effect of their being invalid by assessing both the probability that an assumption is incorrect & the resulting impact on the estimate. This is a form of risk analysis.

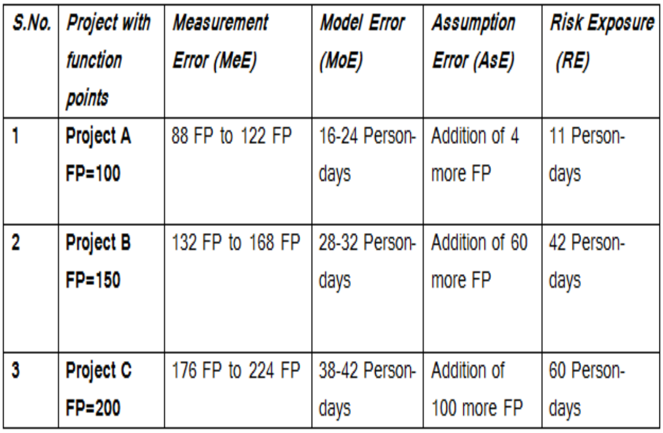
3 Experimental Work

In this section we have presented how the proposed Expert webest tool would be useful to estimate the risk in software and also to estimate the cost of software. For more detail about the Software requirements elicitation and prioritization, please refer to [6, 7, 8, 9 and 10]. According to the measurement error the actual size of the function point varies from 44-56. In order to find out the model error we have assumed that the 0.2 person–days per function point. No estimation model can include all factors that affect the effort required to produce a software product. Suppose we have an estimation model with an inherent 20% inaccuracy and our product is of 50 function points in size, our estimation is likely to be between 8-12 person days. The assumption error occurs when we have some incorrect assumption about the models input parameter. Our assumption is that the product size is of 50 function points rest on the assumption that we have correctly identified all the requirements. If we can identify our assumption, we can investigate the effect of their invalid by assessing both the probability that an assumption is incorrect and the resulting impact on the estimate. This is the called the risk analysis. If we assume that there is 0.4 probabilities that the requirement complexity has been underestimated, so we estimate another 2 function point. We can estimate the risk exposure from the following formula

$$\mathbf{RE = (E2 - E1) * P2}, [11, 13]$$

Here RE means "Risk Exposure", E1 is the effort value if your original assumption is true (1000), E2 is the effort value if your new assumption is true (1100) and P2 is the probability that the new assumption is valid (0.1). So let E1 will then be the most likely estimate originally made (200 person-days) and E2 is the most likely alternative to that (220 = 1,100 * 0.2 person-days). So you get: RE = (220 - 200) * 0.1. This gives you 2 person-days. That is your risk exposure. So this risk exposure of two person-days corresponds to the contingency you need to protect yourself against regarding the assumption error.

Where MeE is Measurement error, **MoE** is Assumption error, **AsE** is Model error & **RE** is Risk Exposure. The results for the given software that we have got from the proposed Expert webest tool are summarized in table 2.



S.No.	Project with function points	Measurement Error (MeE)	Model Error (MoE)	Assumption Error (AsE)	Risk Exposure (RE)
1	**Project A FP=100**	88 FP to 122 FP	16-24 Person-days	Addition of 4 more FP	11 Person-days
2	**Project B FP=150**	132 FP to 168 FP	28-32 Person-days	Addition of 60 more FP	42 Person-days
3	**Project C FP=200**	176 FP to 224 FP	38-42 Person-days	Addition of 100 more FP	60 Person-days

Fig. 4. Table 2, approximately results from Expert webest tool

So this risk exposure of two person-days corresponds to the contingency you need to protect yourself against regarding the assumption error. All in all this is not a bad contingency and is easily placed in the buffer of most project plans. However, the probabilistic nature of risk means that the allowed contingency cannot protect a project if the original assumption is wrong to begin with.

4 Conclusion

In this proposed Expert webest tool focuses on developing estimation tool for web based application. This tool namely (Expert webest tool) is developing by using java as development language & java eclipse as the development tool. Proposed tool easily estimate the risk in software & also to estimate cost of the software.

The cost estimation depends on the calculation of function points, cost adjustment factors & reuse. Function point approach as an input parameter into the "Expert webest tool". This information is needed in the calculation of effort, schedule & total cost for the project.

The risk estimation based on the risk assessment of software projects. Risk identification, risk analysis & risk prioritization are the main subparts of risk assessment. From proposed model it is easy to calculate the risk at different phase as the software projects progresses from phase to phase.

From further research, it is highly recommended that other cost estimating method is considered such as Price-to-win as an added method to cost estimation for web based application & software requirements after adding the thread in to it & then we will prioritize it using analytic hierarchy process & quality function deployment, & after this we will generate the results of that software using the proposed Expert webest tool.

References

[1] Nadan, K.: Practical software project total cost estimation method. In: IEEE MCIT 2010, pp. 7–10. UAE University (2010)

[2] Zuse, H.: Software Metrics-Methods to Investigate and Evaluate Software Complexity Measures. In: Proc. Second Annual Oregon Workshop on Software Metrics, Portland (1991)

[3] International Function Point User Group (IFPUG), Function Point Counting Practices Manual, Release 4.0, IFPUG, Westerville, Ohio (April 1990)

[4] Gupta, D., Sadiq, M.: Software Risk Assessment and Estimation Model. In: International Conference on Computer Science and Information Technology, pp. 963–967. IEEE Computer Society, Singapore (2008)

[5] Symons, C.R.: Function Point Analysis Difficulties and Improvements. IEEE Transaction on Software Engineering 14 (January 1, 1988)

[6] Firesmith, D.: Prioritizing requirements. Journal of Object Technology 3(8) (September 2004)

[7] Karlsson, J.: Software Requirements Prioritizing. In: Proceedings of the International Conference on Requirement Engineering (1996)

[8] Li, Z.-Y., Wang, Z.-X., Yang, Y., Wu, Y., Liu, Y.: Towards multiple ontology Framework for Requirements Elicitation and Reuse. In: 31st IEEE Annual International Computer Software and Application Conference (2007)

[9] Low, G.C., Jeffery, D.R.: Function Point in the Estimation and Evaluation of the Software Process. IEEE Trans. Software Engineering 16(1) (1990)

[10] Sadiq, M., Ghafir, S., Shahid, M.: A Framework to Prioritize the software Requirements using Quality Function Deployment. In: National Conference on Recent Development in Computing and its Application (2009), organized by Jamia Hamdard, Delhi, India

[11] `http://www.aw.com/cseng`

[12] Anda, B., Angelvik, E., Simula, K.: Improving Estimation Practices by Applying Use Case Models. In: Oivo, M., Komi-Sirviö, S. (eds.) PROFES 2002. LNCS, vol. 2559, pp. 383–397. Springer, Heidelberg (2002)

[13] Improving Cost Estimation with Quantitative Risk Analysis Be More Precise by Employing Uncertainty by Gregory Nolder, Vose Consulting, `http://www.voseconsulting.com`

[14] Hoodat, H., Rashidi, H.: Classification and Analysis of Risks in Software Engineering. World Academy of Science, Engineering and Technology 56 (2009)

[15] Meli, R., Santillo, L.: Function point estimation methods: A comparative overview. Data Processing Organization

[16] Meli, R., Satillo, L.: Function Point Measurement Tool for UML Design Specification. Data Processing Organization

A Web-Based Adaptive and Intelligent Tutor by Expert Systems

Hossein Movafegh Ghadirli[1] and Maryam Rastgarpour[2]

[1] Graduate Student in Computer Engineering, Young Researchers Club,
Islamshahr Branch, Islamic Azad University, Islamshahr, Iran
hossein.movafegh@iau-saveh.ac.ir
[2] Faculty of Computer Engineering, Department of Computer,
Science and Research Branch, Islamic Azad University, Saveh, Iran
m.rastgarpour@gmail.com

Abstract. Todays, Intelligent and web-based E-learning is one of regarded topics. So researchers are trying to optimize and expand its application in the field of education. The aim of this paper is developing of E-learning software which is customizable, dynamic, intelligent and adaptive with Pedagogy view for learners in intelligent schools. This system is an integration of adaptive web-based E-learning with expert systems as well. Learning process in this system is as follows. First intelligent tutor determines learning style and characteristics of learner by a questionnaire and then makes his model. After that the expert system simulator plans a pre-test and then calculates his score. If the learner gets the required score, the concept will be trained. Finally the learner will be evaluated by a post-test. The proposed system can improves the education efficiency highly as well as decreases the costs and problems of an expert tutor. As a result, every time and everywhere (ETEW) learning would be provided via web in this system. Moreover the learners can enjoy a cheap remote learning even at home in a virtual simulated physical class. So they can learn thousands courses very simple and fast.

Keywords: Expert Tutor, Intelligent Learning, Adaptive Learning, E-learning, Web-based learning.

1 Introduction

The application of computers in learning began on 1980. Many efforts have been done in order to update and optimize electronic learning (E-learning) yet which great and dramatic advances have been observed in recent years. Generally, E-learning means to improve educational efficiency using information and communication technology [1].

At the first, some Medias like CDs or web applied for E-learning. But these kinds of education are static, non-intelligent and inflexible. Because the course subject had been organized by prior procedure and then trained to different learner in the same style. Diversity of learners leaded to decrease the efficiency of this style. In fact, repeating some lessons was needed for some learners in this method and also some

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 87–95.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

lessons must be removed for some other learners. Later the researchers in pedagogy sciences (training/learning methods) concluded that the learning must be dynamic and intelligent. The fact is that an expert tutor can adapts the sequence of lessons and speed of training with aptitude and characteristics of learner. He can also adjust the expression style with learner's mood as well as cancels the class due to incorporate mental conditions of the learner.

Nowadays, "web-based learning" and "intelligent learning" is one of the most regarded topics in education [2,3]. Moreover, expert tutor is infrequent and expensive. A web-based tutor has some benefits like tirelessly, predominate on concepts, low cost and invariant of time and place. However millions learners of the world can learn by thousands of expert tutor via web in an intelligent and virtual schools.

This paper introduces an intelligent system to apply the abilities of expert systems. So E-learning would be efficient, adaptive and performed by computer and web. Adaptation of web-based contexts is very important, because the contexts would be used by millions variant learners. So the concept, which is developed for one user, isn't applicable for others [3].

The proposed system determines the learning style by a test. Then the learning process starts. Gradually, some characteristics of learner may be change by learner's progress. These improvements would be saved by system in learning process. So learner model gets more accurate step by step. System can receives scientific and mental feedback of learner expertise and intellectually and then change the learning style during the process. This web-based system is developed to facilitate learning every time and everywhere (ETEW). Web-based content is installed and supported in one place while millions learners of the world can use it just via a computer connected to the internet [4]. The aim of proposed system is that to offer the content which the user is not aware about it.

The rest of this paper is organized as follows. Section 2 defines an intelligent E-learning system and presents some available samples. Then it deliberates adaptive E-learning and some learning styles in section 3. Section 4 describes the proposed E-learning system which is intelligent, adaptive, customized and web-based. Finally this paper concludes in section 5.

2 Intelligent E-Learning System

The adaptive intelligent systems are not novel at all. All of these systems are a kind of "Intelligent Tutoring Systems (ITS)" or "Adaptive Hypermedia Systems" [5]. E-learning systems are categorized into two classes: intelligent and non-intelligent. Non-intelligent learning is static, inflexible as mentioned. In these systems, tutor develops course topics previously. Then software engineer presents them in variant methods and the same style to learners. The same style for variant learners is the biggest disadvantage of these systems. Since there are different kinds of learners in E-learning systems in aspect of awareness and mentality, it intensively needs to organize the course contents intelligent and present them to the learner well.

The aim of intelligent E-learning is to realize the customized and adaptive E-learning using of course content, learner type and education method [6]. This system

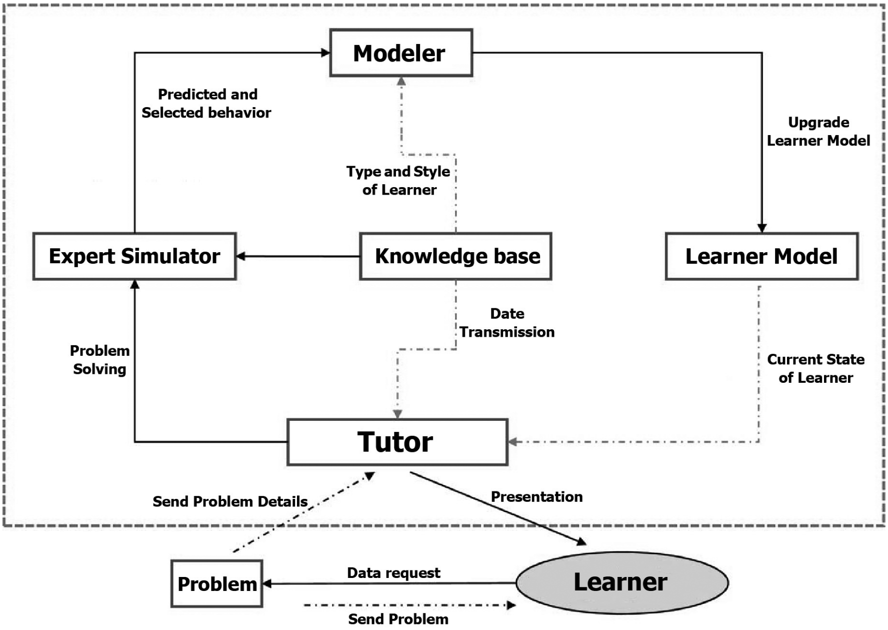


Fig. 1. Process of Intelligent Tutoring System

can recognize the student type. Then it can chooses appropriate course content from knowledge base and present the contents in proper style to the learners. Figure 1 shows the process of ITS.

Some available intelligent E-learning systems are introduced in the following to handle the pragmatics of three elements: content, learner model and education methods for adaptive and customized learning.

- **VCA System** [7]**:** To train the learners, it considers individual differences and talents of learners to develop a virtual classmate agent.
- **SQL-Tutor System** [8]**:** It has been developed by guided exploration. This system selects some questions in basis of learner's model. Then it evaluates learner's answer. It updates the model based on answer validity. Choosing questions would be repeated based on model.
- **Lisp-Tutor system** [9]**:** This system guides the learner intelligently in each step of problem solving without considering his answers. This system tries to teach LISP.
- **DeSIGN System** [10,11]**:** This is a software to teach American Language to deaf learners. This system teaches English words by elements of "train-test" and "teacher" graphically. This system is used in Pittsburg deaf School now.
- **EIAS system** [12]**:** It is as adviser for collaborative learning.
- **CAES system** [13]**:** It has been developed by integration of shipping simulation and intelligent decision system. Its task is to teach shipping to captains in virtual turbulent sea.

- **UC-Links** [14]**:** It is an intelligent system to present the courses in the universities.
- **GENITOR system** [15]**:** It is the generator of training programs.
- **ICATS** [16]**:** This system coordinates the expert system with multimedia system in an intelligent learning system.

3 Adaptive E-Learning

The adaptive E-learning is very important in order to improve efficiency and effectiveness of educational environments. These systems can also be responsible and compatible with the heterogeneous population of learners. An E-learning which is efficient, adaptive and dynamic can recognize learning style of learner by pedagogy principles. It can adapt the learners with current status of system. Then it changes its behavior dynamically and presents the learning concepts according to learner's model. This way leads to improve learning rate finally.

Some psychologist and pedagogy researchers applied many models in adaptive E-learning systems to model behavior and learning style of learners. This model has many advantages in comparison with others. They are proper analysis, recognition of ideal learning style and application of educational science in modeling [17]. Next section introduces styles of learning based on Jakson model and questionnaire.

3.1 Learning Styles

An adaptive E-learning system is based on accurate recognition of behaviors and individual characteristics. In addition of aptitude, personality and behavior, learning style is very important as well [17].

This paper is based on five learning style which is summarizes in Table 1:

- **Sensation seeking (SS):** These people are impulsive and aren't patient. New situations are exciting so that they can't wait and would like to experience and explore it immediately. They believe to action and perform multiple tasks simultaneously. These people would rather to explore their environment by themselves and also learn by test and error.
- **Goal Oriented achievers (GOA):** They adjust certain and difficult goals. They try to increase their abilities by attaining skills and collecting required cognitive resources to realize their goals. They think that troubles are as instructive challenges. Furthermore they believe that can realize to whatever they want.
- **Emotionally Intelligent Achievers (EIA):** Emotional independence and rational thinking are prominent characteristic of them. They are patient learners who have the best efficiency after perceiving of logic behind a problem. They can generalize well from one problem to others. They often divide a problem to smaller and intelligible ones in this process.
- **Conscientious Achievers (CA):** They are responsible and wise people. They can learn well by collecting, analysis and review some information before

action. They prefer to analyze all problem aspects. Thus they can relate discrete data to each other and avoid making a mistake. These people usually have extensive knowledge in areas of interest.

- **Deep learning Achievers (DLA):** they have deep perception of concepts. They want to know how can use previous taught practically. They may learn well when would be aware of note value. So they can test that theory or idea. In fact, learning is difficult for them, when they don't know the target [17].

Table 1. Summarizes the Mentioned Learning Styles

Learning style	Comment
Sensation Seeking	They believe that experiences create learning.
Goal-oriented Achievers	They set difficult and certain target. They have self-confidence to achieve them.
Emotionally Intelligent Achievers	They are rational and goal-oriented instead of dependent and sentimental.
Conscientious Achievers	They are responsible and insight creator.
Deep learning Achievers	They are interested in learning highly.

4 Proposed System

A web-based, adaptive and intelligent tutor is an E-learning system based on web which can be used remotely and ETEW. It can determine learner type (especially in aspect of learning style), learning content and presentation technique adaptively. So that it be updated automatically with learner's characteristics and behavior. This system uses traditional intelligent E-learning. The first E-learning system which is web-based and intelligent, has been reported on 1995[5,18]. Learning all courses is customized well at home via web in this system. So learners can solve some examples and proper exercises ETEW. Finally he can attend in course examination which would be virtual or physical. Figure 2 shows the elements of intelligent tutor.

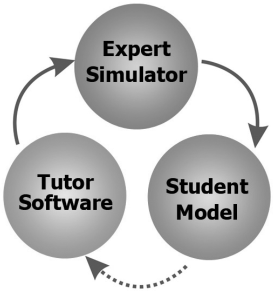


Fig. 2. Elements of Intelligent Tutor

4.1 Learning Environment

Learner can visits website of intelligent virtual school by authentication and logging in the system. An intelligent Graphical User Interface (GUI) is an interface between learners and intelligent tutor. This section of system can affect learning efficiency. An intelligent virtual class has some properties like graphical properties, audio and video to make learning attractive. Moreover, some tools are available to simplify learning process. Learners can communicate well with this inanimate and non-physical system by these tools. Some facilities are:

- Computer games- preferably intellectual games and commensurate with the level of learner
- Frequently asked questions (FAQ)- which consists of commonly questions and proper answers
- Video chat and email- for visual communication between tutor and learners

4.2 Training Method

Knowledge of expert tutor includes of two parts, course knowledge (learning content) and learning technique. Course knowledge is theoretic information, technical content and probably experiments which expert tutor notes. Learning technique is some experiences which he have got during teaching years [19]. An expert tutor determines learner level in according to IQ, understanding, behaviors, talent and individual characteristic like physical class. Learner level consists of "*weak*", "*slow learner*", "*smart*", "*genius*" and so on. Tutor teaches educational content corresponding to learner level in proper method such as film, dynamic view, and game and even bringing up puzzle while he get feedback from learner during training. So learner level may be changed. Tutor helps learner to learn by "*the best way"* in proposed system.

The expert tutor offers an education method based on learner's type. Furthermore each course section has individual significance which is different with the others. Tutor often determine different scores for variant sections according to education method. Moreover he marks highest score to the most important section of course in all education methods.

There are two approaches in E-learning development. In the first one, a problem comes and some examples would be solved then. Finally the learners try an exam. In second approach, content is divided to some parts such as chapter, section, important subsection and so on. The learners take an exam at the end of learning. It is clear the second approach is more effective and has higher level than the first one. In this paper, the proposed system uses the second approach. The smallest part of any topic which can't divide more is called "concept". It is usually equivalent a lesson in physical class. Educational concepts transfer to knowledge base in this system. Then the system can distinguish all concepts and relocate all parts. Sometimes a lesson is needed to repeat, relocate or even remove for a learner. Most of available systems guide a learner to a special aim intelligently in learning process. While only a few intelligent systems provides selecting subsections of a concept for a learner.

This system uses a three layered structure to offer and implement a concept:

1. Pre-test
2. Learning concept
3. Post-test

The pre-test includes of some questions planned by an expert tutor to determine learner's primary knowledge level. The learning concept depends on learner level. So the best method to train a learner is determined. Then learning process starts up. After learning is done, a post-test evaluate the learner by some questions. Figure 3 shows block diagram of proposed system.

4.3 Learning Styles

Learner evaluation is significant. It has two levels, conceptual and objective. Evaluating in concept level refers to learner understanding of lesson concept and evaluating in objective level denotes to learner understanding of lesson topic. Knowledge level of learner is determined with concept level and objective level. The tutor can extract proper questions from question base through an expert system, pre-test and post-test. He notes that a specific score is given to each question.

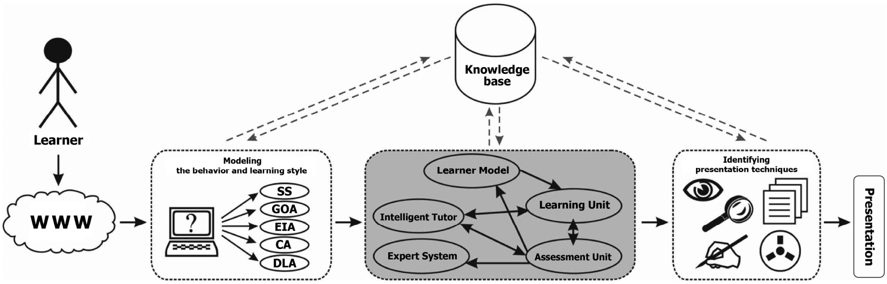


Fig. 3. Block Diagram of Proposed System

Selecting question should satisfy some rules. First, the questions should not be repetitious even if a learner would be trained one concept several times. Second, the question must be planned for all sections of a concept entirely. Third, expert tutor plans questions in all level. Sequence, number and level of questions are determined according to learner level and learning type intelligently. Sum of scores is calculated and learner level is determined after answering the questions.

Table 2 presents five categories of learner's knowledge level about a concept [20, 21]:

Table 2. Categories of Knowledge Level

Knowledge Level	Score
Excellent	86-100
Very good	71-85
Good	51-70
Average	31-50

This system updates the learner's model during progress of question answering. This system can also save last academic status of learner and all his learning records.

5 Conclusion

A web-based, adaptive and intelligent tutor by expert system was presented in this research. Previous E-learning systems offer predefined and static learning concept sequentially to learners. While proposed system can adapts with learning styles (i.e. Sensation Seeking, Goal Oriented achievers, Emotionally Intelligent Achievers and Conscientious Achievers), aptitude, characteristics and behaviors of a learner. It acts as an intelligent tutor which can perform three processes - *pre-test*, *learning concept* and *post-test* - according to characteristic learner. This system uses expert simulator and its knowledge base as well. It is also web-based which leads to be simple learning, low-cost, available everywhere and every time. Consequently thousands of students can learn simultaneous and integrated efficiently.

Nowadays the most educational systems try to be electronic, online, intelligent, adaptive and dynamic. The proposed system tries to get these properties. Moreover it doesn't have any drawback of previous system and human expert tutor. It can improve efficiency of pedagogy and education too. In other words, it helps learners to study in "*the best way*".

References

[1] Eustace, K.: Educational value of E-learning in conventional and complementary computing education. In: Proceedings of the 16th National Advisory Committee on Computing Qualifications, NACCQ, Palmerston North, New Zealand, pp. 53–62 (2003)

[2] Specht, M., Oppermann, R.: ACE - Adaptive Courseware Environment. The New Review of Hypermedia and Multimedia 4, 141–161 (1998)

[3] Brusilovsky, P.: Methods and techniques of adaptive hypermedia. User Modeling and User-Adapted Interaction 6(2-3), 87–129 (1996)

[4] Shaw, E., Johnson, W.L., Ganeshan, R.: Edagogical agents on the Web. In: Proceedings of Third International Conference on Autonomous Agents, pp. 289–290. ACM Press (1999), `http://www.isi.edu/isd/ADE/papers/agents99/agents99.htm`

[5] Brusilovsky, P.: Intelligent tutoring systems for World-Wide Web. In: Holzapfel, R. (ed.) Proceedings of Third International WWW Conference (Posters), pp. 42–45. Fraunhofer Institute for Computer Graphics, Darmstadt (1995)

[6] Brusilovsky, P.: Intelligent tutoring systems for World-Wide Web. In: Holzapfel, R. (ed.) Proceedings of Third International WWW Conference (Posters), pp. 42–45. Fraunhofer Institute for Computer Graphics, Darmstadt (1995)

[7] Fatahi, S., Ghasem-Aghaee, N., Kazemifard, M.: Design an Expert System for Virtual Classmate Agent (VCA). In: Proceedings of the World Congress on Engineering, London, U.K., vol. 1 (2008)

[8] Mitrovic, A.: Learning SQL with a Computerized Tutor. SIGCSE 30(1), 307–311 (1998)

[9] Anderson, J.R., Reiser, B.: The LISP tutor. Byte 10(4), 159–175 (1985)

[10] American Sign Language Project, DePaul University, `http://asl.cs.depaul.edu/`

[11] Bowe, F.: Approaching equality: Education of the deaf. T.J. Publishers, Silver Spring (1991)

[12] Giroux, S., et al.: Epiphyte Advisor Systems for Collaborative Learning. In: Diaz de Ilarraza Sanchez, A., Fernandez de Castro, I. (eds.) CALISCE 1996. LNCS, vol. 1108, pp. 42–50. Springer, Heidelberg (1996)

[13] Yang, C.: An Expert System for Collision Avoidance and Its Application, Ph.D. Thesis (1995)

[14] Chia, Y.H., et al.: Design and implementation of an intelligent educational building system. In: Int. Symposium on Knowledge Acquisition. Representation and Processing, Auburn Univ. (1995)

[15] Kameas, A., Pintelas, P.: The Functional Architecture and Interaction Model of a Generator of Intelligent Tutoring Applications. Journal of Systems and Software (1997)

[16] Ragusa, J.M.: The Synergistic Integration of Expert Systems and Multimedia within an Intelligent Computer Aided environmental tutoring system. In: Proceedings of the 3rd World Congress on Expert Systems (1960)

[17] Movafegh Ghadirli, H., Rastgarpour, M.: A Model for an Intelligent and Adaptive Tutor based on Web by Jackson's Learning Styles Profiler and Expert Systems. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2012, Hong Kong, vol. 1 (2012)

[18] Okazaki, Y., Watanabe, K., Kondo, H.: An Implementation of an intelligent tutoring system (ITS) on the World-Wide Web (WWW). Educational Technology Research 19(1), 35–44 (1996)

[19] Rinne, C.H.: Excellent Classroom Management. Wadsworth Publishing Company (1997)

[20] El-Khouly, M.M., Far, B.H., Koono, Z.: Expert tutoring system for teaching computer programming languages. Expert Systems with Applications 18, 27–32 (2000)

[21] Brusilovsky, P., Schwarz, E., Weber, G.: ELM-ART: An Intelligent Tutoring System on World Wide Web. In: Lesgold, A.M., Frasson, C., Gauthier, G. (eds.) ITS 1996. LNCS, vol. 1086, pp. 261–269. Springer, Heidelberg (1996)

Traffic Light Synchronization

K. Sankara Nayaki[1], N.U. Nithin Krishnan[2], Vivek Joby[2], and R. Sreelakshmi[2]

[1] Lecturer in Department of Information Technology,
Adi Shanakara Institute of Engineering and Technology, Kalady, Kerala
`sankaranayaki@gmail.com`
[2] Students, Department of Information Technology,
Adi Shanakara Institute of Engineering and Technology, Kalady, Kerala

Abstract. Traffic Synchronization mechanisms aim at minimizing traffic bottle necks, controlling traffic flow, by means of dynamic adjustments in Traffic Signal timings. This paper presents a two level synchronization approach, Local Synchronization and Global Synchronization for synchronizing the traffic flow. Focusing upon determining the traffic light timings based not only upon the current lane densities perform the local synchronization which itself is standalone, and also the determination of the green time of the traffic signals based on the densities of the peer junctions at various levels, along with a set of parameters associated with these to provide Global Synchronization, easing out the traffic flow from the heavier density areas to the lower density zones. Hence, implementing two level intelligence platforms, the system serves as a reliable standalone traffic control system control even if either one of the intelligence platform goes down.

1 Introduction

Did you know? Worldwide, each person spends almost 16 minutes daily waiting for the traffic lights to change, contributing 3 years of his entire life, that is 4.3% of a human life being spent simply waiting at the traffic cross points. When it comes to the traffic in India, especially at Kochi city, Kerala, this waiting time would be manifolds.

Kochi city, where the traffic literally overflows roads, in order to cover 11.8 Km from The Ernakulum South to The Aluva Junction, it should take not more than 17.7 min at a standard allowed speed of 40 Km/hr, "in an ideal case". Practically, the same distance takes more than 2 hours, with the traffic getting stuck at the junctions of The Kaloor, Ernakulam North Bridge, Palarivattom and Edapally. The commuters being left stranded for hours, snailing at a pace of less than 5Kmph.

While travelling home (Kochi) from the college (Kalady) and vice versa, we felt a great deal of time, simply being wasted at the traffic blocks, which turns into a chaos specially at the late evening and the morning hours, with vehicles in a fix and the traffic police trying their head out to resolve the mess. We wished to help out the traffic police officials by providing them with a real time solution for the increasing jams in Kochi. Hence, put forth our efforts towards reducing this time gap due to the traffic blocks which in turn became the driving force behind our project.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 97–106.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

In India, Traffic lights' signaling is pre-set time based switching system and employs control booths equipped with switches to control the traffic lights when the timer based system can no longer control the traffic flow effectively. Especially, when it comes to peak hours, it's a common sight to have police personals on duty waving out to the traffic in lanes, struggling to minimize the chances of blocks. Longer waiting times means long queues of vehicles, pollution, ultimately a chaos. With the limitation posed on widening of roads by the infrastructure in the congested city, the existing technology will descend down in performance with the vehicles increasing day by day. The proposed technology performs local control, based on current lane densities making decisions via the information collected from the IR sensors using algorithm [1] or by making use of a video sensor based traffic density collection [2]. The main focus area of this paper is on traffic synchronization.

2 Existing System

The city of Kochi, since ages has been known as the Queen of the Arabian Sea. It is blessed with water, land and air links amidst greenery making it a beautiful place to live in. The city had seen a rapid growth within a span of a few decades right since Indian invasion by the Britishers. Developed in fragments, sea routes from Kochi were exploited for the exports especially for spices, hardly any attention was given to planning of the city. With the boom of the export businesses, the habitation knew no bounds. More and more houses, buildings, factories came up day by day, with no thought being casted over a proper planning. It was too late when realized. Since then, the Traffic Police have been working hard, devising newer methods to tackle the increasing traffic but none of these is successful in the long run. The earth at Kochi is quite unstable, leading to frequent repairs of roads, rains making conditions worse with lots of money spent by the government in road maintenance. Through a brief interview with the Traffic Assistant Police Commissioner, Edapally, we could devise the Problem Statement as:

1) *Poor public awareness on alternate routes.*
2) *Nil traffic Synchronization.*
3) *Infrastructure Bottle Neck- widening of roads.*
4) *No Real Time Update on traffic Status.*

Traffic lights are set up on a tender basis. Keltron Company has been helping out the traffic controllers in the city by designing traffic lights. No surveys have been carried out for analyzing the traffic flow of the entire area under observation. As per International standards, the traffic should not be kept waiting more than 50 seconds. The best timer based traffic control system is the one which limits the red time to 10 seconds. But as far as the timings at congested junctions in Kochi are concerned, this time extends more than one, one and a half minutes. Due to a lack of techniques that allow a dynamic change in the timings, the signals would be rendered operation-less especially at the peak hours with the traffic police personals manually controlling the traffic which should have been the other way instead. The VHF sets serve as the backbone for the traffic police to communicate among themselves and stay updated with the traffic status of the neighboring junctions. There are no mechanisms, such as a central display at the control room, indicating the current traffic status of the area. The alternate routes available too are not properly utilized to reroute the incoming traffic, in case there arises congestion.

3 Proposed Technology

Off the problems, the area that could be focused well was the traffic synchronization, where density based time calculations for the traffic lights would be a better option.

3.1 Architecture

The overall System would be as specified as in Figure 1, where a number of Junctions having a high level embedded device serve as clients for a central Server, which directs each client with the control Information to synchronize the traffic in the area.

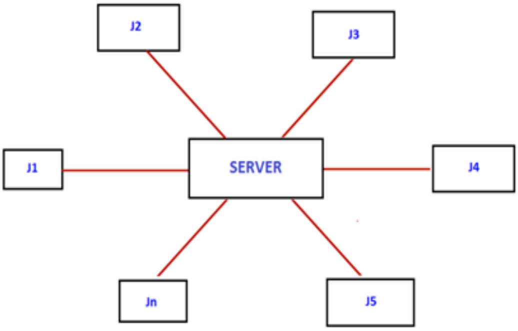


Fig. 1. Overall System Architecture
The junction PCs are connected to a centralized traffic server monitoring the traffic density in each junction.

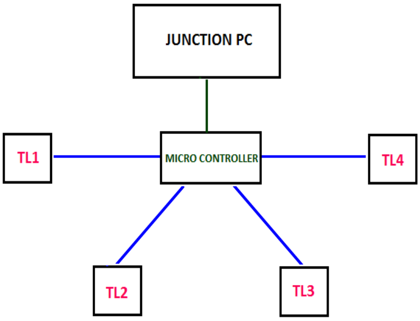


Fig. 2. Architecture of Junction
The junction PCs control the traffic lights using microcontrollers with a difference that the green time for each traffic signal is set by the junction PC.

Each junction consists of a Personal Computer or a High level embedded device connected to a central server. The Junction PC is the one that determines the duration for which a lane should be given green. It hence controls the traffic lights using Microcontrollers.

The proposed technology for can be explained in two levels:

1) Physical Level

The Junction PC collects traffic density information through the sensors placed in each lane, and quantizes it. The density information are passed to the main server every second. The system controls the traffic in a junction by switching the traffic lights based on algorithm in [1] for *local synchronization.*

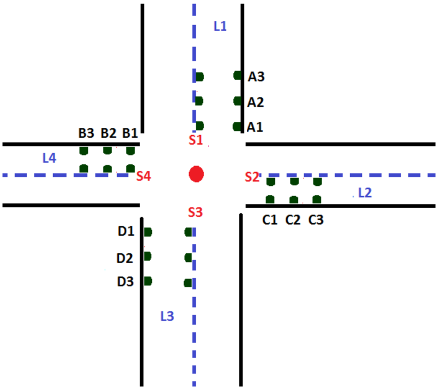


Fig. 3. Local Synchronization

Achieving local synchronization by determining green time using the density values of each lane L1, L2 etc. collected from the sensors A1, A2 etc.

The Algorithm [1] performs the switching of the traffic signals S1, S2, S3 and S4 based on the traffic density of lanes L1, L2, L3 and L4. These function in a cyclic manner as the traditional Traffic Lights with a difference that the switching pattern is based on the densities instead of the pre-determined time based switching, deciding the green time for each lane, dynamically.

2) Logical Level

From a logical view point, the system can be assumed to be composed of two levels of intelligence:

a) Level 1 Intelligence

The level 1 intelligence is provided by the algorithm [1]. The system will hold well even if the server goes down routing traffic in a standalone mode independent of other junctions. It will still function as a man without intelligence, still capable of performing his day to day chores.

b) Level 2 Intelligence

Level-2 intelligence takes into account the densities of the neighboring junctions too. The green time now will have a factor that will help ease out the load at the nearby junctions too. That is the system gets intelligent when connected to the server.

4 Functional Units

1) The Junction PC:

Main task of this unit is to collect the vehicle count from the sensors, quantize it, form it into an *Update packet* (Fig 2a) consisting of Start of packet (SOP), Junction ID (JnID), densities of the four lanes (D1, D2, D3 and D4) numbered clockwise starting from north of the junction, a priority field (P) indicating an orange state in the junction, also it indicates a request seeking the lane densities of the nearby junctions. The density update packet is sent to server periodically.

SOP	**Jn ID**	**D1**	**D2**
D3	**D4**	**P**	**EOP**

Fig. 4. Update Packet Format

SOP-Start of packet JnID – ID of the Junction D1,D2,D3,D4 – Lane densities P – Priority Field EOP – End of Packet

2) Server:

The server works as a high speed router. A sevlet deals with the incoming request and responses. It has mainly two tasks:

a) **Accept Update packet:** Update the *density table* of the *traffic Synchronization* database with the information from the update packet. For a high speed routing, the sevlet maintains a buffer area:

int[n][6] buffer;

n *= Total number of junctions under coverage.*

It stores density values corresponding to each junction with the Junction ID (JnID) serving as the index, next four fields to store the density values and a field to store the time stamp.

b) **Provide Response to request:** To provide the density information of the neighboring junctions when the incoming packet has the priority field set. For this purpose it maintains another buffer area:

String [n][6] buffer;

n *= Total number of junctions under coverage.*

The buffer contains the Junction Id as the index, followed by associated 4 junctions and of which level the association is. The understanding of the levels can be analyzed with the following figures:

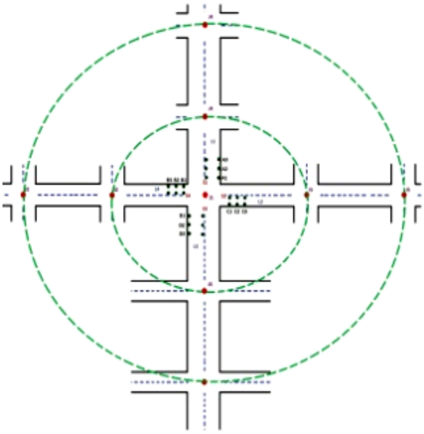
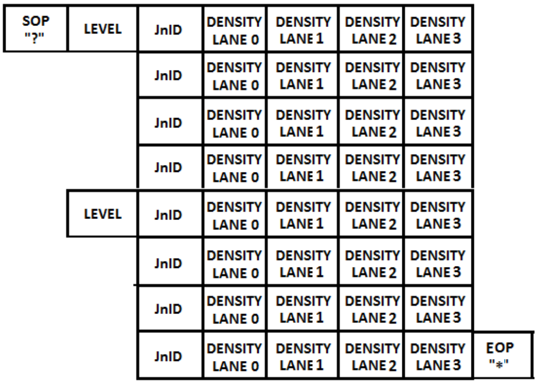


Fig. 5. Levels of Synchronization
The Inner circle represents level-1 Peer junctions And the outer Circle, the level 2 peer junctions

The Junction forming the nearest neighbor circle can be called as the level – 1 of synchronization and the junctions falling in the next outer circle is referred as level – 2 of synchronization. The response packet would be as in figure 6.



SOP "?"	LEVEL	JnID	DENSITY LANE 0	DENSITY LANE 1	DENSITY LANE 2	DENSITY LANE 3	
		JnID	DENSITY LANE 0	DENSITY LANE 1	DENSITY LANE 2	DENSITY LANE 3	
		JnID	DENSITY LANE 0	DENSITY LANE 1	DENSITY LANE 2	DENSITY LANE 3	
		JnID	DENSITY LANE 0	DENSITY LANE 1	DENSITY LANE 2	DENSITY LANE 3	
	LEVEL	JnID	DENSITY LANE 0	DENSITY LANE 1	DENSITY LANE 2	DENSITY LANE 3	
		JnID	DENSITY LANE 0	DENSITY LANE 1	DENSITY LANE 2	DENSITY LANE 3	
		JnID	DENSITY LANE 0	DENSITY LANE 1	DENSITY LANE 2	DENSITY LANE 3	
		JnID	DENSITY LANE 0	DENSITY LANE 1	DENSITY LANE 2	DENSITY LANE 3	EOP "*"

Fig. 6. Response packet
Level – Level of peer junctions

3) Central Map Display

The proposed system also provides a map display at the police control room where the authorized police personals can have access to the Real-Time traffic scenario of the entire city and traffic densities at the corresponding junctions. This will help the police to control the traffic flow in and out of the city and also provide with alternate routes when any traffic congesting situations arise.

5 Test Results

A few simulations were carried out at the level 1, in order to test the efficiency of the system by applying the configuration parameters of the junction [Refer Index]. One of the major questions in the simulation was how far will the system be successful in controlling the traffic providing a green timed based on the density.

The simulations pointed out, yet a bit shocking result. On keeping the green time a fixed constant value that is 1 min, the density builds to a total of 3726.6 vehicles as shown in table 1, now this is an average green time provided it usually climbs up to more than 3 min, one can guess out the number of vehicles accumulated in that case. Hence, a fixed timing cannot be followed.

Table 1. Simulation Results

Flow Rate	**3**	**2**	**1**	**0.1**	**Total**
Fixed (1 Min)	2166	1278	279	3.6	3726.6
Density based	1617	1182	407	26.9	3232.9
Density based (Max 1Min)	2241	957	171	15.1	3384.1
Density based (Max 2Min)	2058	1008	164	3.1	3233.1
Density based (Max 5 Mins)	1617	1182	407	26.9	3232.9

When we tried out with a pure density based technology, the results were favorable. The total number of vehicles accumulated in the junction was 3232.9, much better result as compared to fixed timer based. Yet the density based green times went on escalating as time passed due to increasing number of vehicles in each lane hence it had to be clamped to a level. Now the question was what the level should be.

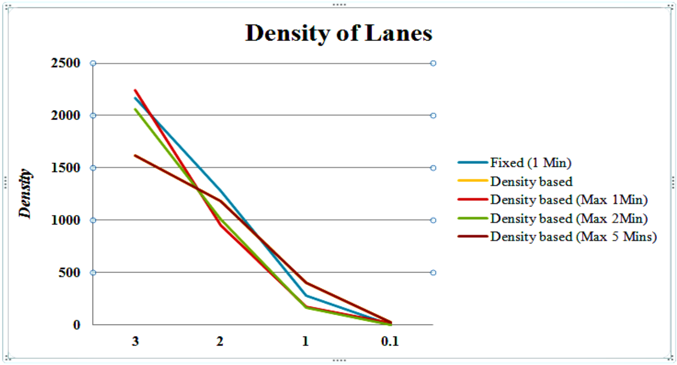


Fig. 8. Density v/s Flow Rate Curve

We randomly started with a value of 1 min, to see how better it would be as compared to the fixed green time of 1 min. And we had it, it read 3384.1 far better to 3726.6. When compared to the purely density based results, the results were satisfactory except for those of flow rate 3. Our next aim was to reduce this value. Hence we clamped the density based system to 5 minutes.

In this case, the density values followed the same curve as that of the purely density based system hence; we concluded that higher the clamping limit the curve tends to move closer to the pure density based curve.

Next, we tried with a clamp limit of 2 min, this one was better than even 1 min clamp limit value as the densities now accumulated to 3233.1 rather than 3384.1, hence a better result and it was successful in reducing the traffic accumulated in lane with flow rate 3 too. Hence, this was the best suite.

6 Advantages

The density based traffic synchronization is a much better technology in handling the traffic congestions as it offers:

1) Quick Response:

The server acting as a high speed router provides a quick response to the increasing density. Also the central map display alerts the traffic police when it senses an abnormal growth in the traffic densities, hence providing a quicker response before the situation turns worse.

2) Robust

Being designed in a mutually independent manner, the system will not fail even if the central server crashes or vice versa even if a junction PC goes down. In case of a server failure or the junction PC break down, the police authority will be notified.

3) Scalable

Newer junctions can be integrated smoothly into the system with minimum modifications into the existing system, by updating a few parameters like the distance from the nearest junctions etc. The system can be expanded to cover all the junctions in the state and finally the country too. Hence, moving towards a well organized traffic control system.

4) Integrated

The system can work well with the video sensor based technology and is independent of the method used to collect the traffic density information. This feature can be expanded to track vehicles while on road with the help of the video sensor system.

5) Hierarchical

The system is based on a hierarchical approach, by which a high level control and coordination of traffics of larger areas can be handled with ease. The city junctions may be controller by a central server for the city, which in turn would be controlled by the state server which would further be controlled by a central national server.

6) Re-Configurable

The system can be reconfigured as the synchronization factor can be formulated independently for each junction. Hence, offering better and long term sustenance to the system.

7 Future Enhancement

It is still a common site here in Kochi, with emergency vehicles too not spared by the traffic chaos. The emergency services as the ambulances and the fire Engines too end up in blocks with no way out situations, even though these are the highest priority services. By using Zigby tags or high frequency VHF tags in these emergency vehicles, the system can be enhanced to provide priority to such services by ensuring that the lanes having these vehicles get the least waiting time and the overall route of the vehicle is synchronized such that as soon as the vehicle approaches the corresponding junction signal goes green. This technology will require tracking the path and the speed of the emergency vehicle to synchronize the traffic flow of the route in the best efficient manner.

8 Conclusion

Various traffic synchronizations algorithms have been designed and a few implemented too, but the model based approach supersedes them in terms that the system at no point, can fail. If it does so, the system can still be recovered by making it able to handle the situation by working over the system manually and come up with optimized algorithms to deal the newly created situation of the traffic bottle neck. An article from the Hindu stated of implementation of traffic synchronization algorithms implemented based on video processing at Anna Salai [3], Tamil Nadu. The system implemented on an experimental basis covered a large area with a good number of traffic signals, but collapsed when ultimately the traffic at each lane turned equal.

Acknowledgements. At the very outset, we would honor the Almighty, who gave the wisdom and knowledge to complete the paper successfully. Our extreme thanks to the Principal, Prof. S. G. Iyer for providing us, the necessary facilities and the environment for the completion of this paper. We extend our sincere gratitude to Mr. R Rajaram, Head of Department (HOD) of Computer Science, for his direction and support. This paper would have been without the active and able support from Mr. Mohan kumar, The Director of Grey Technolabs, Edapally, Kerala, for helping us out with the project. We extend our heartfelt gratitude to Mr Maneesh T, Head of Department (HOD) of Information Technology, for his valuable guidance and support. Last but not the least; we are highly obliged to our parents, friends and all well wishers who had supported us directly and indirectly for making this paper.

References

1. Sehgal, V.K., Dhope, S., Goel, P., Chaudhry, J.S., Sood, P.: An Embedded Platform for Intelligent Traffic Control. IEEE (2010)
2. Image Processing Algorithms for detecting and counting vehicles waiting at a traffic light. J. Electron Imaging 19, 043025 (December 21, 2010), `http://dx.doi.org/10.1117/1.3528465`
3. Sreevatsan, A.: Intelligent traffic signals on Anna Salai prove ineffective. The Hindu (June 25, 2011), `http://www.thehindu.com/news/cities/Chennai/article2132620.ece`

Partition Sort versus Quick Sort: A Comparative Average Case Analysis with Special Emphasis on Parameterized Complexity

Niraj Kumar Singh[1] and Soubhik Chakraborty[2,*]

[1] Department of Computer Science & Engineering, B.I.T. Mesra, Ranchi-835215, India
[2] Department of Applied Mathematics, B.I.T. Mesra, Ranchi-835215, India
{niraj_2027,soubhikc}@yahoo.co.in

Abstract. In our previous work we introduced Partition sort and found it to be more robust compared to the Quick sort in average case. This paper does a more comprehensive comparative study of the relative performance of these two algorithms with focus on parameterized complexity analysis. The empirical results revealed that Partition sort is the better choice for discrete distribution inputs, whereas Quick sort was found to have a clear edge for continuous data sets.

Keywords: Partition Sort, Quick sort, average case, parameterized complexity, robustness.

1 Introduction

The true color of an algorithm cannot be explored completely until it is subjected to the parameterized complexity analysis. This paper is a continuation of our previously published work [8] which dealt with design and subsequent analysis (both mathematical and empirical) of Partition Sort algorithm. Our objective there was the development of a divide and conquer based fast as well as a robust sorting algorithm compared to the popular Quick Sort. Here, in this paper our intention is to perform an empirical study of the relative average case performance of Partition and Quick Sorts through their parameterized complexity analysis for various probability distribution data (both uniform and non uniform). In [8] we have theoretically proved that Partition Sort has better worst case complexity ($n\log_2^2 n$).

It is also evident from our experimentation that the idea of comparing two algorithms on the very same fixed parameters(s) value(s) is not a sufficient mean in itself. In this light the relative performance order, for various distribution data, obtained in [8] needs to be reviewed. We have carried out this reviewing task through a comprehensive parameterized complexity analysis. The experimental results exhibited that with its robust characteristics the Partition sort has an edge over Quick sort for discrete distribution (especially non uniform) data sets. However, for continuous data sets it is the Quick sort which performed better. Although inferior to

---

[*] Corresponding author.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 107–113.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Quick sort for Continuous data sets, the robustness of Partition sort remained unchallenged. It is important to notice that Quick sort fails to exhibit the average case robustness property for major discrete distribution inputs [9].

2 Parameterized Complexity Analysis

Parameterized complexity is a branch of computational complexity theory in computer science that focuses on classifying computational problems according to their inherent difficulty with respect to multiple parameters of the input. The complexity of a problem is then measured as a function in those parameters [10]. The parameterized complexity analysis is done for average case performance of candidate algorithms. Average complexity is explained best by the weight based statistical bounds and their empirical estimates, the so called empirical O (also denoted by the symbol O_{emp}). The statistical bounds are different from mathematical bounds in the sense that here the operations are weighed. Weighing permits collective consideration of all operations for determining the bound. Also such a bound is guaranteed to be realistic [1, 2].

This section includes the experimental results of parameterized complexity analysis performed over Partition and Quick sort algorithms. Each time data (in second) is taken for input size N=50000, and is averaged over 100 readings. Average case analysis was done by directly working on program run time to estimate the weight based statistical bound over a finite range by running computer experiments [5, 7]. The terms T(QS) and T(PS) represents the mean time for Quick and Partition sorts respectively.

System Specification: All the computer experiments were carried out using PENTIUM 1600 MHZ processor and 512 MB RAM.

2.1 Empirical Study of Binomial Distribution Inputs

The Binomial distribution has two distinct parameters, namely m and p. As the first case study, Binomial distribution inputs were taken with parameter m fixed at 1000 and p varied in the range [0.1 - 0.9]. The experimental results for both Quick and Partition sorts are shown in Table 1 and Fig. 1. The result exhibits that Partition Sort has a clear performance edge over the quick sort over the range of p values (0.1 to 0.9 in our case). It is interesting to notice that the response for Quick sort follows the bath tub pattern, whereas for Partition sort it seems to follow an inverted bath tub pattern!

Table 1. Binomial distribution for N=50000, m=1000 fixed and p varying

P	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
T(QS)	0.1943	0.14632	0.13108	0.12126	0.1194	0.1206	0.1294	0.1472	0.19586
T(PS)	0.09084	0.09662	0.09884	0.10198	0.10186	0.10034	0.0989	0.09884	0.09096

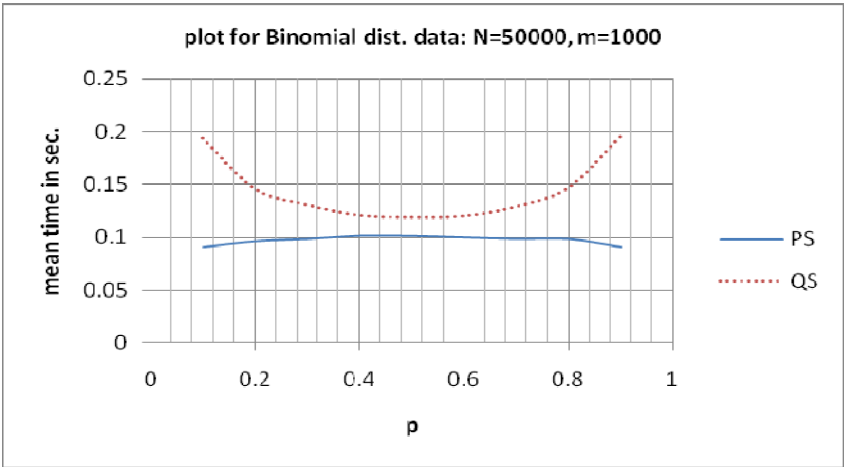


Fig. 1. Binomial distribution input curve

Table 2. Binomial distribution for p=0.5 fixed and m varying

M	100	300	500	700	900	1100	1300	1500
T(QS)	0.35874	0.21028	0.16554	0.14094	0.125	0.11604	0.10596	0.0969
T(PS)	0.07968	0.09066	0.09586	0.09968	0.10154	0.10438	0.10282	0.10618

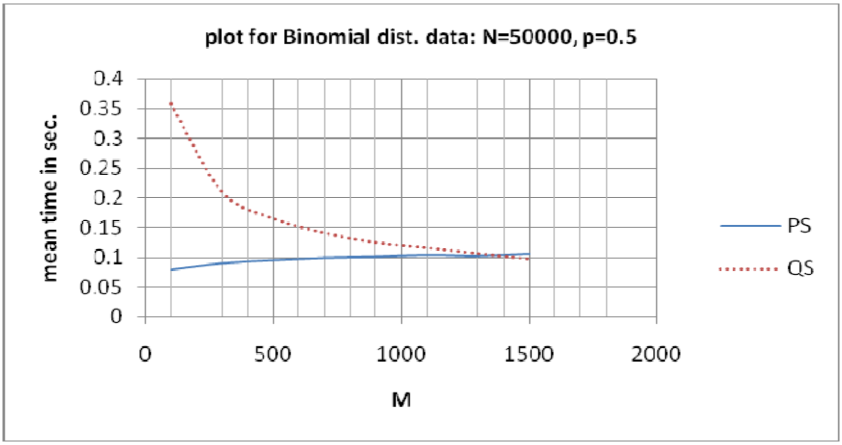


Fig. 2. Binomial distribution input curve

Next we obtained the Binomial inputs for p fixed at 0.5 and m varying in the interval 100 through 1500. The corresponding empirical result is given through Table 2 and Fig. 2. The result revealed that for fixed p value the running time is a decreasing function on m for Quick sort but an increasing function for the Partition sort algorithm. The relative plot in Fig. 2 suggests that Partition sort behaves smoothly over the considered range of m values compared to Quick sort algorithm.

For Binomial distribution inputs, Partition sort has average complexity $Y_{avg}(n, m, p) = O_{emp}(n\log_2 n)$ [1], whereas Quick sort exhibits $Y_{avg}(n, m, p) = O_{emp}(n^2)$ complexity[2].

2.2 Empirical Study of Discrete Uniform Distribution Data

The Discrete Uniform distribution U[1, 2…K] with probability 1/K for each variate value evidently depends on the parameter K. For this distribution inputs were taken for varying K values in the range [10–10000]. The corresponding experimental result is given in Table 3 and Fig. 3. The experimental result is suggesting that the Partition sort's performance is superior for smaller K values. However, for larger K values it's the Quick sort which ultimately wins the race. The intersection points of these two curves lie somewhat in between 50-100.

Table 3. Discrete Uniform distribution for N=50000

K	10	25	50	100	200	300	400	500	1000	10000
T(QS)	0.7365	0.3615	0.24376	0.14104	0.08442	0.06214	0.05174	0.04672	0.0364	0.0304
T(PS)	0.13576	0.16854	0.18926	0.20372	0.22786	0.22618	0.23312	0.24374	0.24682	0.26306

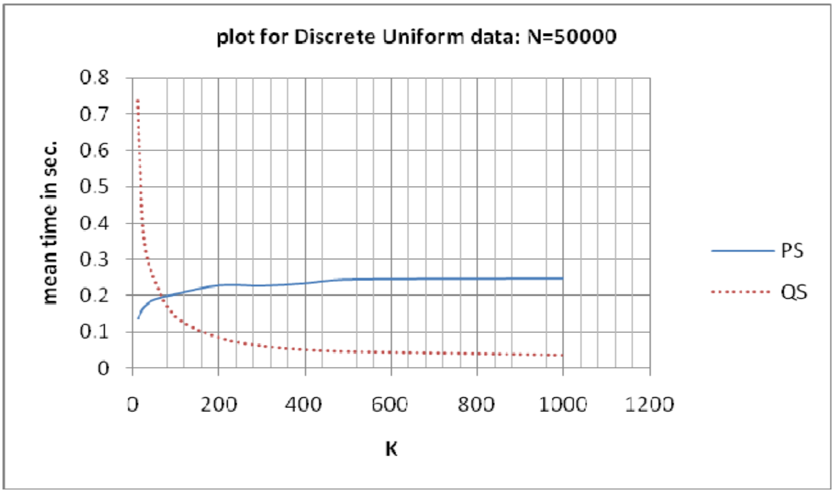


Fig. 3. Discrete Uniform distribution input curve

For Discrete Uniform distribution inputs, Partition sort has average complexity $Y_{avg}(n, K) = O_{emp}(n\log_2 n)$ [8], whereas Quick sort exhibits $Y_{avg}(n, K) = O_{emp}(n^2)$ complexity[9].

2.3 Empirical Study of Poisson Distribution Data

The Poisson distribution $\exp(-\lambda)\,\lambda^x/x!$, x=0, 1, 2… depends on the parameter λ, which is both the mean and the variance of the distribution. Lambda (λ) should not be large as it is associated to rare events. In our experiment we have taken λ value in the range of 1 to 5 in the interval of 1. The corresponding experimental result is as shown in

Table 4 and Fig. 4. The Partition sort does a clean sweep when it comes to Poisson distribution inputs. So we strongly recommend Partition sort to Quick sort if we have reasons to believe that sorting elements can be approximated by a Poisson model (which can be statistically tested using, e.g., Chi-Square goodness of fit test).

For Poisson distribution inputs, Partition sort has average complexity $Y_{avg}(n, \lambda) = O_{emp}(n\log_2 n)$ [8], whereas Quick sort exhibits $Y_{avg}(n, \lambda) = O_{emp}(n^2)$ complexity[9].

Table 4. Poisson distribution for N=50000

. λ	1	2	3	4	5
T(QS)	2.0573	1.41938	1.15706	1.0021	0.90826
T(PS)	0.08654	0.10788	0.1167	0.11904	0.12224

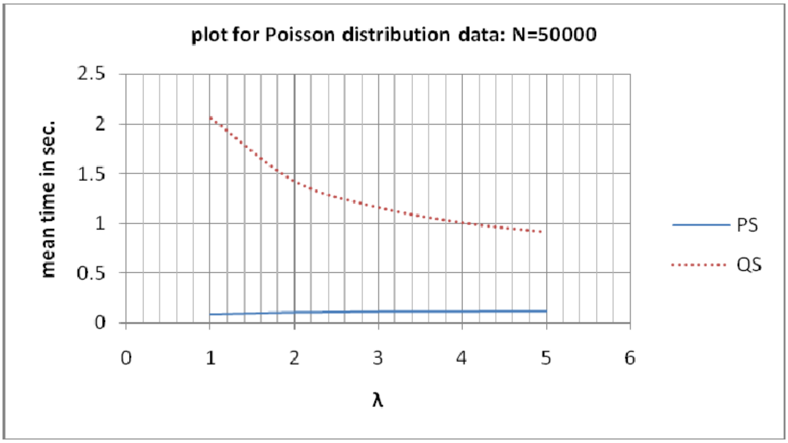


Fig. 4. Poisson distribution input curve

2.4 Empirical Study of Continuous Uniform Distribution Data

The Continuous uniform distribution depends on the parameter of [a, b] where a is the minimum value of parameter and b is maximum value parameter. Here we have taken a=0 and b=1 and simulated a U [0, 1] variate and multiplied it with the positive real theta (θ) to generate U [0, θ] variate. The corresponding experimental result is as given in Table 5 and Fig. 5. This result revealed that although both the curves seemed to be independent of the parameter θ, the performance of Quick sort is superior to that of partition sort algorithm.

For Continuous distribution inputs, both Partition and Quick sorts exhibit $Y_{avg}(n, \theta) = O_{emp}(n\log_2 n)$ complexity[8, 9].

Table 5. Continuous Uniform distribution for N=50000

Θ	1	10	100	1000	10000
T(QS)	0.0325	0.0322	0.0316	0.03188	0.0325
T(PS)	0.27258	0.26666	0.27886	0.27472	0.28078

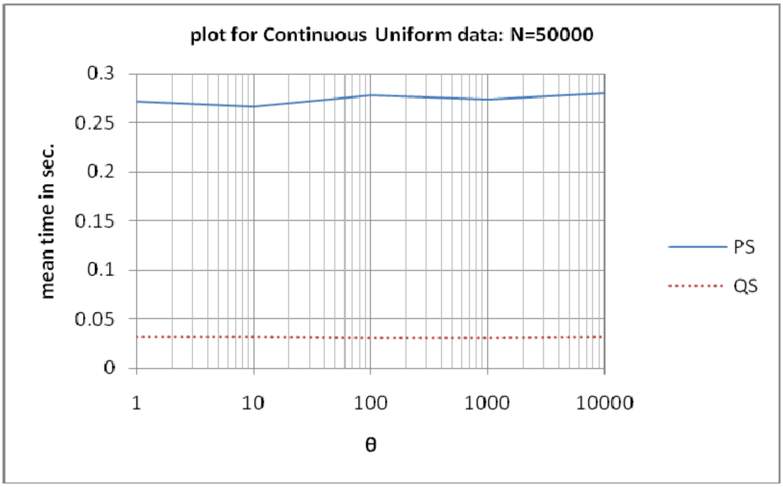


Fig. 5. Continuous Uniform distribution input curve

3 Conclusions

The empirical study performed over Partition and Quick sorts once again revealed that the parameterized complexity analysis is an essential part of statistical method of algorithmic analysis. Chakraborty et al. [3], in their work strongly advocated as to why for certain algorithms like sorting, the parameters of the input distribution should be taken into account for explaining the complexity, not just the parameter characterizing the size of the input. For other works advocating the necessity of parameterized complexity [4, 6] can be referred. The experimental results exhibited that with its robust characteristics the Partition sort has an edge over Quick sort for discrete distribution (especially non uniforms) data sets. However, for continuous data sets it is the Quick sort which performed better. Although inferior to Quick sort for Continuous data sets, the claim for robustness of Partition sort remained unchallenged [8]. It is important to notice that Quick sort fails to exhibit the average case robustness property for major discrete distribution inputs [9].

References

1. Chakraborty, S., Sourabh, S.K.: A Computer Experiment Oriented Approach to Algorithmic Complexity. Lambert Academic Publishing (2010)
2. Chakraborty, S., Modi, D.N., Panigrahi, S.: Will the Weight-based Statistical Bounds Revolutionize the IT? International Journal of Computational Cognition 7(3), 16–22 (2009)
3. Chakraborty, S., Sourabh, S.K., Bose, M., Sushant, K.: Replacement sort revisited: The "gold standard" unearthed! Applied Mathematics and Computation 189(2), 384–394 (2007)
4. Downey, R.G., Fellows, M.R.: Parameterized Complexity. Springer (1999)

5. Fang, K.T., Li, R., Sudjianto, A.: Design and Modeling of Computer Experiments. Chapman and Hall (2006)
6. Mahmoud, H.: Sorting: A Distribution Theory. John Wiley and Sons (2000)
7. Sacks, J., Weltch, W., Mitchel, T., Wynn, H.: Design and Analysis of Computer Experiments. Statistical Science 4(4) (1989)
8. Singh, N.K., Chakraborty, S.: Partition Sort and Its Empirical Analysis. In: Das, V.V. (ed.) CIIT 2011. CCIS, vol. 250, pp. 340–346. Springer, Heidelberg (2011)
9. Sourabh, S.K., Chakraborty, S.: How robust is quicksort average complexity? arXiv:0811.4376v1 (cs.DS)
10. Wikipedia, `http://en.wikipedia.org/wiki/Parameterized_complexity`

Removal of Inconsistent Training Data in Electronic Nose Using Rough Set

Anil Kumar Bag[1], Bipan Tudu[2], Nabarun Bhattacharyya[3],
and Rajib Bandyopadhyay[2]

[1] Department of Applied Electronics and Instrumentation Engineering,
Future Institute of Engineering and Management, Kolkata-700 150, India
[2] Department of Instrumentation and Electronics Engineering, Jadavpur University,
Salt Lake Campus, Sector III, Block LB, Plot No. 8, Kolkata-700 098, India
[3] Centre for Development of Advanced Computing(C-DAC), E-2/1, Block – GP,
Sector – V, Salt Lake, Kolkata-700 091, West Bengal, India
`anilkumarbag@gmail.com, {bt,rb}@iee.jusl.ac.in,`
`nabarun.bhattacharya@cdac.in`

Abstract. Inconsistency in the electronic nose data set may appear due to noise that originates from various sources like electrical equipments, measuring instruments and some times the process itself. The presence of high noise leads to produce data that are of conflicting decision and thus encounters misleading or biased results. Also the performance of the electronic nose depends upon the number of relevant, irredundant features present in the data set. In an electronic nose the features correspond to the sensor array. While deploying an electronic nose for a specific application, it is observed that some of the features (sensors response) may not be required rather than only a subset of the sensor array contributes to the decision, which implies the optimization of sensor array is also important. To obtain a consistent precise data set both the conflicting data and irrelevant features must be removed. The rough set theory proposed by Z. Pawlak, is capable of dealing with such an imprecise, inconsistent data set and in this paper, the rough-set based algorithm has been applied to remove the conflicting training patterns and optimize the sensor array in an electronic nose instrument used for sensing aroma of black tea samples.

Keywords: Electronic nose, feature selection, reduct, rough set, sensor array.

1 Introduction

For the activity and performance evaluation of different application specific instrument, it is necessary to create and maintain huge databases with activity specific information. Productive analysis on these databases is important for the strategic solution making which solely depends on the data consistency, irredundancy of feature information. The electronic nose [1] is an example of such an application specific instrument now a day extensively used in many food and agro industries for quality estimation of the products.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 115–123.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

The electronic nose comprises of a sensor array and its associated electronic circuitry. The knowledge base in electronic nose are feature information in the form of electrical sensual response produced by each individual sensor because of combined multidimensional effect of different innumerable attributes present in the food or agro products. These sensual responses in terms of numerical data pattern contain the signature, which is related to the quality of the exposed substance. Unfortunately, the limitations in data collecting, the high complexity of sensory inputs, transient effects, and equipment failure restrict the classifier to be trained by the data that have the desired characteristics in a statistically sufficient way. Thus, the data pattern produced may contain a number of irrelevant, redundant features and some decision conflicting data patterns leading poor granularity of representation of the information. Such a data set not only increases time complexity, also degrades classification accuracy. As the effective information for classification often lies within a lower dimensional feature space, the feature extraction or dimensionality reduction has proven to be a crucial step in all analytical methods or applications. The aim of this work is to develop a strategy based on rough set theory that addresses discovery of relevant features or attributes and filtering of process signals in the presence of outliers/noise.

Rough set [2] theory (RST) was proposed by Z. Pawlak in the early 1980s and has received more and more attention in the domain of artificial intelligence and cognitive sciences, especially in the spheres of machine learning, knowledge acquisition, knowledge discovery from databases, decision analysis, expert systems, inductive reasoning, data mining and pattern recognition. The philosophy behind the rough set theory was the observation that the presence of uncertainty and imprecision in knowledge base induces vague decision and vagueness may be caused by granularity of representation of the information. The tight granularity of representation of information in information system insists more objects to be in each equivalence class and thus leads to a consistent rule base. Thus, it is important to filter data in knowledge base in order to remove unnecessary granularity while keeping essential information.

Few research articles have been reported to filter out the inconsistent and imprecise data. Duntsch et al. [3] developed a simple data filtering procedure that is compatible with the rough set approach in which the main tool is ‘binary information system’. The binary information system always is of larger dimension than the original information system as binarization process splits each attribute to its sub-attribute based on the variation in attribute vales of the information system. Thus, the method generally results in an increased complexity. Another approach of data filtering based on rough set is reported by Yin et al. [4]. Here the improvement of granularity of information in each attribute is attempted using rough set but the newly formed information system is of same dimension as earlier. Brunner et al. [5] apply spatial filtering for motor imagery EEG data using independent component analysis (ICA) algorithms (Infomax, FastICA and SOBI). Some of these ICA algorithms use PCA for dimension reduction leading to information loss up to a certain extent and the performance of the algorithms depends on the tuning of different parameters in it. Wavelet-based robust filtering of process data hasbeen reported by Doymaz et al. [6] where, the strategy is based on the use of moving median (MM) filter in tandem with the coefficient de-noising approach. The limitation of this strategy lies in the choice of proper window size for MM filter, which depends upon the knowledge about the outliers present in the data and the proper threshold value for de-noising.

In this paper, we use rough set theory to remove redundant attributes as well as the inconsistent data or the data that have conflicting decisions. At the first step, the granularity of information is increased by discretization of the real valued attribute data. Then the irredundant attributes are selected using the concept of decision relative reduct [7], which is the set of most relevant and non-redundant features having highest significance in decision-making. This attribute reduced discretized data set forms the rule base. Finally, the objects corresponding to contradicting rules are removed from the rule base to form the consistent rules. Thus, the inconsistent data and superfluous features are removed from the large size inconsistent database.

2 Rough Set Theory

Z. Pawlak introduces the concept of rough set theory in the early 1980s. It is an excellent mathematical tool for the analysis of a vague description of objects. The data in rough set theory is stored in a table called a decision table or sometimes the information system IS . As rough set theory requires discrete features, the real valued attribute features are discretized before to be present for rough set analysis. The rows of IS correspond to the objects and the columns correspond to features. The last column of IS , known as decision attribute indicates the class to which each example belongs. Formally, an information system is a quadruple

$$IS = \left(U, A \cup \{d\}, V, f\right) \tag{1}$$

where U is a non-empty finite set of objects, A is a non-empty finite set of attributes, V is the union of attribute domains (i.e., $V = \bigcup_{a \in A} V_a$, where V_a denotes the domain of attribute a) and f is a function such that for any $u \in U$ and $a \in A$, $f\left(u,a\right) \in V_a$ while d is called decision attribute.

For each possible subset of attributes $B \subseteq A$, a decision table generates an equivalence relation called an indiscernibility relation $IND(B)$, where two objects $\left(u_i,u_j\right)$ are members of the same equivalence class if and only if they cannot be discerned from each other on the basis of the set of attributes B . The equivalence classes of the B -Indiscernibility relation are denoted $\left[u\right]_B$. Indiscernibility relation is defined as

$$IND\left(B\right) = \left\{\left(u_i,u_j\right) \in |U| \times |U| : \forall a \in B, f(u_i,a) = f(u_j,a)\right\} \tag{2}$$

which induces a partitioning of the universe U according to the attribute set B .

The discernibility knowledge of the information system is commonly recorded in a symmetric $|U| \times |U|$ matrix called the discernibility matrix [8]. Thus any set $X \subseteq U$ can be approximated solely on the basis of information in $B \subseteq A$ by constructing a B *-lower approximation* and B *-upper approximation* of X .

The B -lower approximation of X is defined as the union of all the elementary sets which are certainly in X i.e.

$$\underline{B}X = \{x : [x]_B \subseteq X\}. \tag{3}$$

The B-upper approximation of X is defined as the union of the elementary sets, which have a non-empty intersection with X i.e.

$$\overline{B}X = \{x : [x]_B \cap X \neq \phi\}. \tag{4}$$

3 Feature Selection and Inconsistent Data Removal

In rough set based data analysis, the decision table or the information system (IS) is first formed from the experimental data following equation (1). However, rough set theory cannot deal with continuous attributes although the real datasets include continuous values. The solution lies in partitioning of the numerical values into a number of intervals and treating each interval as a category. This process of partitioning to different category is termed as discretization [9], [10]. Feature selection may be considered as the process of finding an optimal subset from the original set of pattern features according to some specified criterion. Rough set theory defines the reduct as a minimal set of attributes that describes all the variations in the data set. Therefore, the idea of feature selection can be explored by using the concept of reduct. The algorithms for discretization and reduct generation can find in the paper published by the authors Bag et al. [11]. Thus, the newly formed information system consists of reduct set attributes with their discretized object values. Note that this newly formed information system is a preliminary rule base, which may contain some contradicting rules. Presence of such contradicting rules in the rule base will degrade the classification accuracy. In addition, it may be concluded that the objects corresponding to contradicting rules i.e. conflicting decisions are imprecise, inconsistent

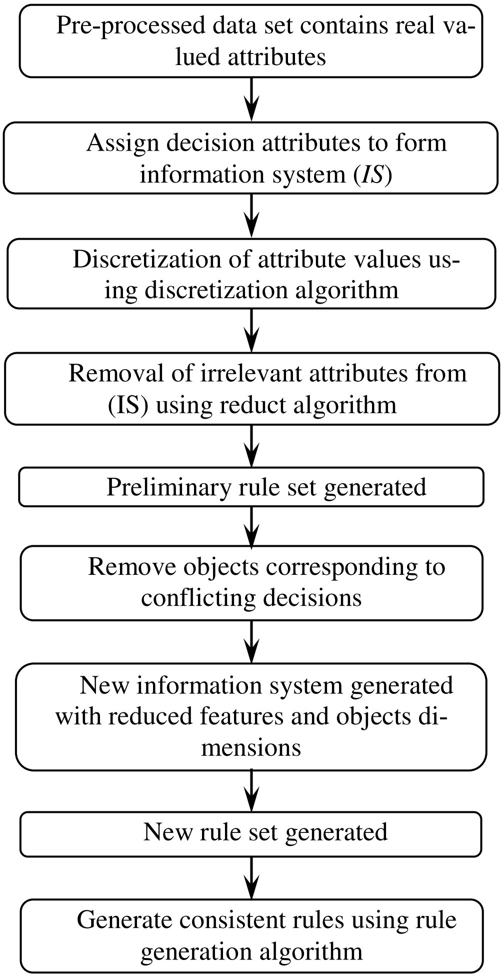

Pre-processed data set contains real valued attributes
Assign decision attributes to form information system (*IS*)
Discretization of attribute values using discretization algorithm
Removal of irrelevant attributes from (IS) using reduct algorithm
Preliminary rule set generated
Remove objects corresponding to conflicting decisions
New information system generated with reduced features and objects dimensions
New rule set generated
Generate consistent rules using rule generation algorithm

Fig. 1. Operational flow chart

information about the features. Hence, the next step is removal of those conflicting objects from the information system. Now this reduced information system is used to get the optimized rules using rule generation algorithm stated in Bag et al. [11]. The operational flow chart for the whole process is shown in Fig.1.The sensors corresponding to the attributes in the reduct set have been considered for the classification of black tea and the results have been compared between raw data set, data set with optimized sensors and final reduced data set.

Removal of the objects with conflicting decisions is done in the following way

- Objects with their attributes having same values are grouped together.
- Search for the objects having different decision within each group and delete them from the information system.

4 Experimental

Electronic noses employ an array of chemical gas sensors. Extensive range of applications achieved by the use of several pattern recognition techniques that provides a higher degree of selectivity and reversibility to the systems. Degrading of these properties of the systems is due to the presence of redundant attributes as well as the inconsistent data rather the data that have conflicting decisions. Such problems successfully are counteracted by the use of rough set theory. In practice, tea samples are tested by human sensory panel called "Tea Tasters", who assign quality scores in the scale of 1 to 10 according to the quality of tea. This method is highly subjective, and the scores vary from taster to taster. In the pursuit of objective estimation of black tea quality by experimental means, co-relation of sensor array [12] output signature with Tea Tasters' scores have already been accomplished by the authors in [13]with good accuracy, where a number of conventional neural network topologies have been utilized. In the present study, rough set theory is used to optimize the features as well as to remove imprecise, inconsistent information about the features and generate the rules for the co-relation between sensor array signature and tea tasters' scores.

4.1 Customized Electronic Nose Setup for Black Tea

A customized electronic nose setup has been developed for quality evaluation of tea aroma, the details of which are presented in [13]. Eight gas sensors from Figaro, Japan – TGS 2610, TGS 2620, TGS 2611, TGS 2600, TGS 816, TGS 831, TGS 832, and TGS 823 constitute the sensor array for the setup. The outputs of the sensors are acquired in the PC through peripheral component interconnect (PCI) data acquisition cards (PCI 6035E from National Instruments). The MOS sensors are conductometric in nature, and their resistance decreases when subjected to the odor vapor molecules. The change in resistance with respect to their original values $\left(\Delta R/R0\right)$ is converted into voltage and is then taken to the PC through analog-to-digital converter cards for subsequent analysis in the computational model.

The experimental conditions of the electronic nose for classification of black tea aroma are given as follows:

- Amount of black tea sample = 50 grams,
- Temperature $= 60^0 C \pm 3^0 C$,
- Headspace generation time = 30s,
- Data collection time = 100s,
- Purging time = 100s,
- Airflow rate = 5 ml/s.

4.2 Sample Details with Tea Taster's Score

Experiments have been carried out on 102 different tea samples from Tocklai TRA, Assam in India. One experienced tea taster has been assigned to evaluate the samples and assign a score against the aroma for each tea sample. These aroma scores given by the taster have been considered for the correlation study with the computational model. The tea of samples and their scores as given by the tea taster are given in Table 1.

Table 1. Sample Details

Sample serial number (Set of objects :U)	Number of samples	Aroma score
01-12	12	6
13 - 42	30	7.5
43-58	16	8.5
59-102	44	10

5 Results and Discussion

Tea samples collected form the gardens were presented to the electronic nose and the knowledge base in terms of feature information in the form of electrical response produced by each individual sensor in the array because of combined multidimensional effect of different innumerable attributes present in tea. These sensual responses in terms of numerical data pattern contain the signature, which is related to the quality of the exposed substance. The quality measures of the tea are the aroma scores (as indicated in Table 1) given by the experienced taster and these scores are assigned with the data patterns. The numerical data patterns thus produced are arranged in the form of an information system following the tabular form of OBJECT→ATTRIBUTE VALUE relationship. As rough set theory cannot deal with continuous attributes although the real datasets include continuous values, so the attribute values are discretized. The reduct finding algorithm selects the optimized features i.e. the sensors. In our experiment it is seen that for the tea samples considered, only four sensors were selected. Table 2 shows the sensors before and after the optimization process. The presence of outliers in the dataset degrades the classification accuracy. Therefore the objects with conflicting decisions are removed to get a consistent data set. The data dimension before and after sensor optimization with consistent data is shown in

Table 2. Sensor Array Before and After Optimization

Sensor array in E-nose	Optimized sensor array
TGS 2610, TGS 2620, TGS 2611, TGS 2600, TGS 816, TGS 831, TGS 832 and TGS 823	TGS 2610, TGS 2600, TGS 832and TGS 823

Table 3. Data Dimension Before and After Data Reduction

Original dimension	Data dimension after sensor optimization	Data dimension after sensor optimization and removing inconsistent data
102 × 8	102 × 4	52 × 4

Table 3. The next few subsections show that the optimized set of sensors and consistent data set shows better results than the inconsistent data with non-optimized features and hence the efficacy of the rough set-based approach is demonstrated.

5.1 Performance Analysis with Class Separability Index

Separability index [14] is defined as the fraction of a set of data points whose classification labels are the same as those of their nearest neighbors. Thus it is a measure of the degree to which inputs associated with the same output tend to cluster together. This measure is defined by the ratio of the trace of the 'between class scatter matrix' (S_B) to that of the 'within class scatter matrix' (S_W), and the expressions are given below:

$$S_B = \sum_{i=1}^{c} n_i \left(m_i - m\right)\left(m_i - m\right)^T \tag{8}$$

$$S_w = \sum_{i=1}^{c} \left(\sum_{j=1}^{n_i} (x_{i,j} - m_i)(x_{i,j} - m_i)^T \right) \tag{9}$$

where, c is the number of classes, n_i denotes the number of samples in the i^{th} class, and $x_{i,\,j}$ denotes the j^{th} sample in the i^{th} class. m_i mean vectors of the samples in the i^{th} class and m denotes the mean vector of the samples. The results with class separability index are shown in Table 4.

It is seen that after sensor optimization the separability index improved and is further improved when the decision conflicting objects were filtered from the feature optimized data set. The improvement in separability index due to feature optimization and removal of decision conflicting data shows the utility of the method.

Table 4. Separability Index for Different Set of Sensors

Data dimensions	Separability Index
102 × 8	0.1868
102 × 4	0.1900
52 × 4	1.6169

5.2 Classification Accuracy Using BPMLP Algorithm and Rule Based Rough Set Classifier

The back propagation multilayer perceptron (BPMLP) algorithm [15] has been used to compare the accuracy of prediction with the original data set and the final reduced data set. The number of input nodes is equal to the number of sensors and the number of output nodes is four corresponding to the scores given by the taster. Only one hidden layer has been considered with 8 nodes in the model of the neural network. We follow here the 10-fold cross validation [16], [17] method to determine the accuracy of classification. Here the training dataset comprises of 90% data samples and the classification accuracy has been calculated by testing on remaining 10% samples. The results are shown below in Table 5.

The rough set classifier uses the rule set for classification. The data size is reduced from 102×8 to 52×4. As the final data set does not contain any decision conflicting objects, consistent rule base is generated and so classification accuracy is 100%.

Table 5. Classification Accuracy using BPMLP and Rough Set

Data dimensions	Classification accuracy	
	BPMLP	Rough Set
102 × 8	63.63 %	--
52 × 4	83.33 %	100 %

6 Conclusion

The theory of rough set has been developed to deal with vagueness and inconsistent data. In gas sensor based electronic noses, there are lot of uncertainties in the data from the sensors. Moreover, the optimum sensor set is not known for a particular application beforehand. The presence of non-contributing sensors and the conflicting decisions in the training data make the rough set based classifier an appropriate one for such applications. In this paper, it has been demonstrated that rough set based classifier can optimize the sensor set and it can remove the conflicting training patterns before classification. The method has been tested at two different tea gardens in India and the results presented in the paper prove the efficacy of the method.

References

[1] Natale, C.D., Macagnano, A., Mantini, A., Davide, F., D'Amnico, A., Paolesse, R., Boschi, T., Faccio, M., Ferri, G.: Advances in food analysis by electronic nose. In: Proc. of the IEEE Int. Symp. on Industrial Electronics, vol. 1, pp. 122–127 (1997)

[2] Pawlak, Z.: Rough set theory and its applications. J. Telecom. Inform. Techno. 3, 7–10 (2002)

[3] Duntsch, I., Gediga, G.: Simple data filtering in rough set systems. Int. J. of Approximate Reasoning 18, 93–106 (1998)

[4] Yin, X.-R., Zhou, Z.-H., Li, N., Chen, S.-F.: An Approach for Data Filtering Based on Rough Set Theory. In: Wang, X.S., Yu, G., Lu, H. (eds.) WAIM 2001. LNCS, vol. 2118, pp. 367–374. Springer, Heidelberg (2001)

[5] Brunner, C., Naeem, M., Leeb, R., Graimann, B., Pfrtscheller, G.: Spatial filtering and selection of optimized components in four class motor imagery EEG data using independent components analysis. Pattern Recognition Letters 28, 957–964 (2007)

[6] Doymaz, F., Bakhtazad, A., Romagnoli, J.A., Palazoglu, A.: Wavelet-based robust filtering of process data. Computers and Chemical Engineering 25, 1549–1559 (2001)

[7] Pawlak, Z.: Some Issues on Rough Sets. In: Peters, J.F., Skowron, A., Grzymała-Busse, J.W., Kostek, B.z., Świniarski, R.W., Szczuka, M.S. (eds.) Transactions on Rough Sets I. LNCS, vol. 3100, pp. 1–58. Springer, Heidelberg (2004)

[8] Yang, P., Li, J., Huang, Y.: An attribute reduction algorithm by rough set based on binary discernibility matrix. In: Proc. of the Fifth In. Conf. Fuzzy Systems and Knowledge Discovery, vol. 2, pp. 276–280 (October 2000)

[9] Nguyen, S.H., Skowron, A.: Quantization of real value attributes, Rough set and boolean reasoning approach. In: Proc. of the Second Joint Annual Conf. on Information Science, Wrightsville Beach, North Carolina, pp. 34–37 (1995)

[10] Dai, J.-H., Li, Y.-X.: Study on discretization based on rough set theory. In: Proc. of the First Int. Conf. on Machine Learning and Cybernetics, Beijing, pp. 1371–1373 (November 2002)

[11] Bag, A.K., Tudu, B., Roy, J., Bhattacharyya, N., Bandyopadhyay, R.: Optimization of sensor array in electronic nose: a rough set-based approach. IEEE Sensors Journal 11, 3000–3008 (2011)

[12] Dutta, R., Hines, E.L., Gardner, J.W., Kashwan, K.R., Bhuyan, M.: Tea quality prediction using a tin oxide-based electronic nose: An artificial intelligence approach. Sens. Actuators B: Chem. 94, 228–237 (2003)

[13] Bhattacharyya, N., Bandyopadhyay, R., Bhuyan, M., Tudu, B., Ghosh, D., Jana, A.: Electronic nose for black tea classification and correlation of measurements with "Tea Taster" marks. IEEE Trans. Instrum. Meas. 57, 1313–1321 (2008)

[14] Duda, R.O., Stork, D.G., Hart, P.E.: Pattern classification, 2nd edn., p. 115. John Wiley and Sons (2001)

[15] Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Pearson Education, Asia (2001)

[16] Rodriguez, J.D., Perez, A., Lozano, J.A.: Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Trans. Pattern Anal. Mach. Intel. 32(3), 569–575 (2010)

[17] Singh, S., Hines, E.L., Gardner, J.W.: Fuzzy neural computing of coffee and tainted-water data from an electronic nose. Sens. Actuators B 30(3), 185–190 (1996)

A Decentralised Routing Algorithm for Time Critical Applications Using Wireless Sensor Networks

R.A. Roseline[1] and P. Sumathi[2]

[1] Postgraduate and Research Department of Computer Science, Government Arts College,
Coimbatore, Tamilnadu, India
`roselinera@yahoo.com`
[2] Department of Computer Science, Chikkanna Government Arts College,
Tiruppur, Tamilnadu, India
`sumathirajes@hotmail.com`

Abstract. Wireless Sensor Networks (WSN) are presently creating scenarios of decentralised architectures where application intelligence is distributed among devices. Decentralised architectures are composed of networks that contain sensors and actuators. Actuators base their action on the data gathered by sensors. In this paper, a decentralised routing algorithm called DRATC for time critical applications like fire monitoring and extinguishing is proposed that makes use of the Decentralised Threshold Sensitive routing algorithm. The sensing environment consists of many Monitoring Nodes that sense fire and report the data to the Cluster Head. The Cluster Head directs the Extinguishing Node to extinguish the fire before sending the data to the Base Station.

Keywords: Wireless Sensor Networks, Decentralised routing algorithm, Clusters, Cluster Head, sensors, actuator.

1 Introduction

Wireless Sensor Networks(WSN) consists of small ,low powered sensing devices equipped with programmable computing , multiple parameter sensing and wireless communication capability . WSNs offer information about environment, habitat, remote structures, military applications, healthcare and inhospitable terrains. The sensors should react immediately to drastic changes in the environment, for example in time critical applications like fire detection. The end user should be aware of the ground situation with minimum delay by making use of the limited wireless bandwidth.

Firefighting is life threatening event and even though some systems exist to provide information about the fire, the most important that are required during firefighting are proximity of the firefighters to the danger, health status of the firefighters, better radio communication, and proper information of the building floor plans. They also face sudden dangers like ignition of room, explosions occurring due to sudden oxygen entry in oxygen starved fire locations, hidden fires in walls and release of toxic gases[4]. Wireless Sensor Networks could be of great importance in such

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 125–133.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

applications where sensors are used to detect fires and actuators are used to extinguish the fire.

WSN routing algorithms pay much attention to energy savings as it is impossible to replace or recharge batteries of sensor nodes. The operating states of a sensor node can be categorised as transmitting, receiving and idle or sleep states. A sensor node in transmitting state consumes the most energy while in receiving or idle states consumes a little less energy. The energy consumption for data transmission is directly proportional to the square of a wireless transmission distance. A WSN therefore uses routing protocols that are, capable of data aggregation, distribution of energy dissipation evenly and energy efficient in order to increase the network lifetime.

This paper makes use of the Local Clustering and Threshold Sensitive routing algorithm [6] for threshold sensing. But the data transmission is done using Schedule Channel Polling(SCP) since SCP is proved to be more energy efficient than TDMA for event based reporting like fire sensing. Mini-slot structure works fine in short range wireless transmission environment however it cannot work in a Wireless Long Distance Environment (WILD)[5].

The contributions of this paper are described as follows.

(1) A solution for data gathering from the environment based on a certain threshold like finding all places where temperature level is greater than say T. This is done by Monitoring Nodes (MN) and they report the data to the Cluster Head(CH) in a single hop.
(2) The Extinguishing Node(EN) is the actuator and takes care of extinguishing the fire in case the CH orders it to extinguish fire in a certain direction based on the intensity of the fire.
(3) If many Monitoring Nodes(MNs) sense higher temperature then the Extinguishing Node(EN) is informed of a high intensity fire and appropriate extinguishing takes place.

2 Related Work

In this section, related routing protocols, with a focus on clustering sensor nodes in WSNs are discussed.

In TEEN[2], at every cluster change time, in addition the attributes ,the CH broadcasts to its members,

Hard Threshold (H_T): This is a threshold value for the sensed attribute. It is the absolute value of the attribute beyond which, the node sensing this value must switch on its transmitter and report to its CH.

Soft Threshold (S_T): This is a small change in the value of the sensed attribute which triggers the node to switch on its transmitter and transmit.

The H_T tries to reduce the number of transmission by allowing the nodes to transmit only when the sensed attribute is in the range of interest. The S_T further reduces the number of transmissions by eliminating all the transmissions which have otherwise occurred when there is little or no change in the sensed attribute once the H_T.

But the main drawback of this algorithm is that if the thresholds are not reached, the nodes will not communicate, the user will not get any data from the network, and will not come to know even if the nodes die. Therefore this scheme is not suited for applications where it is necessary to get data on a regular basis. Another problem with this algorithm is that there should not be any collisions in the cluster. So a TDMA or CDMA schedule is necessary to solve this problem.

APTEEN[1] is a variation of TEEN, designed as a hybrid protocol that changes the periodicity or threshold values used to provide a periodic state view of the network. It uses combination of proactive and reactive network's features. The CH selection in APTEEN is based on the mechanism used in LEACH-C. The cluster exists for a period called the cluster period, and the BS regroups the clusters, at a time called the cluster change time. APTEEN uses modified TDMA, where each node in the cluster is assigned a transmission slot, to avoid collisions. For query responses, APTEEN uses node pairs. This implies adjacent nodes that sense similar data, but only one of them responds to a query; the other can go to sleep. These two nodes can take the role of handling queries alternately, which helps them saving resources.

Power-Efficient Gathering in Sensor Information Systems(PEGASIS) [7] is an extension of the LEACH protocol, which chains from sensor nodes so that each node transmits and receives from a neighbour and only one node is selected from that chain to transmit to the Base Station(sink). The data is gathered and moves from node to node, aggregated and eventually sent to Base Station (BS). The chain construction is done in a greedy way. Unlike LEACH, PEGASIS avoids cluster formation and uses only one node in a chain to transmit to the BS instead using multiple nodes. A sensor transmits to its local neighbours in the data fusion phase instead of sending directly to its CH as in the case of LEACH.

In [3] the optimal planning of sensor's states in cluster-based sensor networks is discussed. Typically any sensor can be turned on, turned off, or promoted cluster head and a different power consumption level is associated with each of these states. An energy-optimal topology that maximizes network lifetime ensuring simultaneous full area coverage and sensor connectivity to cluster heads which are constrained to form a spanning tree is used as a routing topology.

3 Reference Network Model

As mentioned in the introduction, this paper focuses on how to gather information from the environment based on a certain threshold, the locations where the temperature is higher than the threshold. Accordingly the following assumptions of the WSN are made.

- The network consists of many Monitoring Nodes(MN) that sense the environment and form static clusters; one actuator or Extinguishing Node(EN) in every cluster and one MN that acts as a Cluster Head(CH) in every cluster.
- All Monitoring Nodes(MN) are homogeneous and have the same initial energy supply;
- All the MNs can directly communicate with the Cluster Head(CH) in their region;
- The CH can order the actuator or EN to start or stop extinguishing based on the intensity of the fire.

- The radio channel is symmetric, i.e., the energy consumption for transmitting a message from one node to another is the same as on the reverse direction; and
- Energy consumption for a data transmission only depends on
  (1) the size of the data packet
  (2) the distance between the sender and receiver

Figure 1 illustrates the architectural model of such a WSN with MNs and Extinguisher Nodes (ENs).

The clusters dynamically change later depending on the available energy of the other nodes and CH are elected based on rotation. The network is assumed as a 50 x 50 m network of Sensor Nodes as in Figure 1.

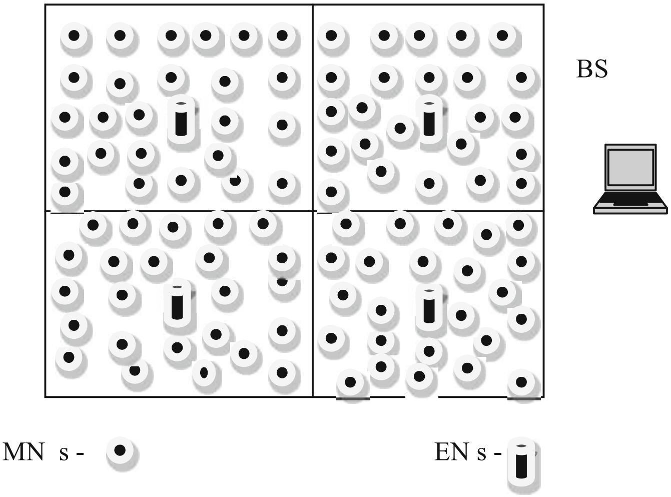


Fig. 1. Architectural model of WSN of 100 nodes showing Monitoring Nodes(CN) in clusters and one actuator or Extinguishing Node(EN) for every cluster

For energy analysis the first order radio model is adopted. Energy consumption in the circuitry for running the transmitter or receiver and in radio amplifier for wireless communication are $E_{ciruitry}$ = 50 nJ/bit and $E_{amplifier}$ = 100 pJ/bit/m2 respectively. The value of $E_{amplifier}$ is directly proportional to the square of transmission distance.

Therefore the energy for transmitting a packet where k is the size of the transmitted packets, and d is the distance between a transmitter and receiver is

$$E_{tansmit}(k,d) = E_{ciruitry} \text{ x } k + E_{amplifier} \text{ x } k \text{ x } d^2 \tag{1}$$

The energy for receiving a packet is

$$E_{receive}(d) = E_{ciruitry} \text{ x } k \tag{2}$$

An efficient routing algorithm aims at reducing the energy required for transmission and receiving and so DRATC is aimed at energy efficiency as described in the section below.

4 Decentralised Routing Algorithm for Time Critical (Dratc) Applications

DRATC works in the following phases:

4.1 The Clustering and Initial Cluster Head Set-Up Phase

The main activities of this phase are creation of clusters and selection of initial CH by the Base Station.

The network contains some Monitoring Nodes (MN) that form clusters in a region. The clusters are static and every cluster has one Extinguisher Node (EN) as shown in figure 1. The Cluster Heads for every cluster are created based on the decision taken by the Base Station (BS). Since all the nodes have the same energy initially, the BS decides Cluster Heads from the MNs in a cluster based on their locations .Every region has a midpoint where the Extinguisher Node(EN) is placed and the node that is closest to the EN is selected as CH initially. The CH thus selected by the BS will not be CH again until all other nodes with higher energy level is selected as CH since being a CH drains the battery of the node .Thus the clusters are static but Cluster heads are dynamic within each cluster.

4.2 Threshold Sensitive Data Transmission Using the Slotted Protocol Schedule Channel Polling (SCP)

Data Transmission is done by the MNs if the sensed value is greater than the Threshold as is in our previous work[6].The SCP protocol [8] is used for data gathering whenever the a fire is sensed. This protocol is proved to be more energy efficient than most of the slotted MAC protocols as seen in the performance evaluation of the algorithm.The MNs contend for the channel in case the temperature crosses the Threshold and then the CH check if it has data to receive. If there is no MN that crosses the Threshold then, all MNs will observe a clear channel and go to sleep immediately. Figure 2 illustrates this process

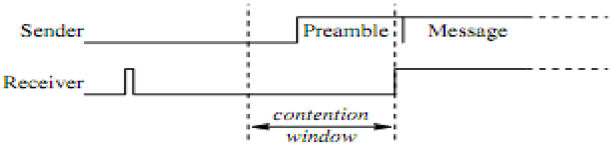


Fig. 2. SCP-sender only contention resolution by means of stretched preamble

Each slot starts with a contention window. At that moment, if a MN wants to send data, it chooses a random moment within this window. If the channel is clear, the MN switches on its radio and starts sending a preamble .The preamble acts as a busy tone and continues until the end of the contention window and thereby locks out any potential senders. Right after the contention window the CH wakes up and performs a carrier sense to see if there is a preamble followed by a message. Without any traffic, SCP-MAC thus only needs to perform one carrier sense per slot making it the most efficient protocol of its class.

Using SCP-MAC schedule as described above, each sensor transmits the sensed information to the CH if the sensed information is above the Hard Threshold(H_T).The sensed value is stored in an internal variable called *sensedvalue (SV).*The nodes will send again the value of SV only if it differs from SV by an amount equal to or greater than a Soft Threshold(S_T).Whenever a node transmits the data, SV is set equal to the current value of the sensed attribute. Thus, the H_T tries to reduce the number of transmissions by allowing the nodes to transmit only when the sensed attribute is in the range of interest. The S_T further reduces the number of transmissions by eliminating all the transmissions which might have otherwise occurred when there is little or no change in the sensed attribute once the hard threshold as in algorithm below.

Algorithm for threshold sensing.

```
If      ( (newsensordata > HT)  or  (( newsensordata  - SV )   > =  ST) ) then
        let  SV = newsensordata;
            Send SV along with rem_energy_level  of  Sensor Node
            to Cluster Head;
else    send only rem_energy_level of SensorNode   to the Cluster Head.
```

During this phase, for each time slot, the sensor nodes will sense the environment and let its value be newsensordata and the energy available at each node be rem_energy_level.

After the clusters are formed, the ENs keep sensing the environment and if the temperature exceeds the Threshold(T), it sends data immediately to the CH. The CH receives the sensed data and sends the extinguishing instruction to the EN. If more than one MN sends data exceeding the Threshold(T), then the data is aggregated and sent to the EN. The EN extinguishes the fire based on the data send and the direction of the MNs and the CH sends the aggregated data to the BS for the user.

4.3 Cluster Head Change after a Round of Fire Detection and Extinguishing

The CH changes after every round of detecting temperature greater than the Threshold. The MNs that do not sense the fire send their remaining energy levels and the next CH is determined as given below.

Algorithm for Cluster Head Selection in every cluster after every round of Threshold sensing

1. For every cluster, select a Monitoring Node(MN) with maximum residual energy as Cluster Head
2. In case of ties, calculate the Euclidean distance between the EN of that cluster and the MNs of equal residual energy.
3. Select the MN with the least distance to the EN found in step 2 as the next CH

In case more than one node has the same energy level, then the node that is closer to the EN is chosen as the CH. This ensures that the CH can communicate with the EN to extinguish the fire faster since the lower the distance between the nodes , faster the data transmission. The CH keeps their transmitters on during this phase to listen to the MNs.

Assuming that there are N nodes in a cluster, and the time for each frame is t_f and the channel bandwidth is $B_{w,}$ Each node will get $t_s = t_f/N$ seconds in which to transmit data. Assuming a 1 bit/sec/Hz signalling scheme, each node can transmit

$$B_w\, t_s = B_w\, t_f/N \text{ bits per frame} \tag{3}$$

$$\text{or } R_b = B_w/N \text{ bps.} \tag{4}$$

Once data from all the MNs have been received, in case of sensing a fire, the CH performs data fusion and reduces the amount of raw data that have to be sent to the Base Station. The compressed data, along with the information required by the BS to properly identify and decode the cluster data, are routed back to the BS by CH-CH routing path created by the BS.

Radio interference is another issue in neighbouring clusters. Code Division Multiple Access (CDMA)codes are used to counteract this problem. Each cluster is assigned a spreading code that the nodes in the cluster use to distinguish from those nodes in neighbouring clusters. Once the data gathering process is complete, The CH uses the same spreading code assigned to the cluster to route data back to the BS. After this phase the BS uses the data sent by the CH regarding the energy levels of the nodes to determine the next CH for the static clusters.

5 Performance Evaluation

The performance of DRATC was assessed by simulation using NS2. Performance is measured by quantitative metrics like average energy consumption and number of nodes alive. A random network of size 100 nodes where each node has an initial energy of 2J was considered. Further the number of data frames transmitted for each round is set at 50 and the data message size is fixed at 100 bytes of which 25 bytes represent the length of the packet header, 75 bytes for the sensed data. The simulation is done for the network where all nodes are assigned an initial energy of 2J.

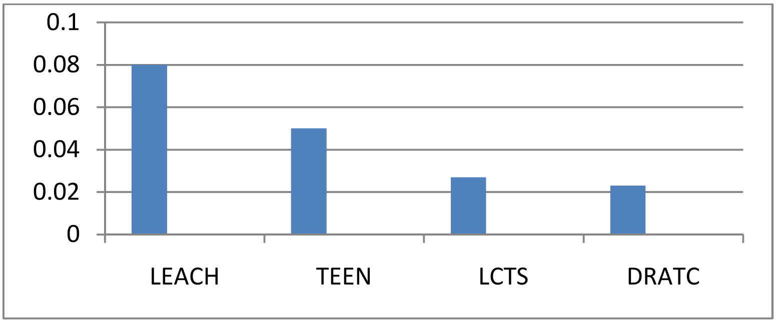


Fig. 3(a). Average energy consumption during a round in a 50 x 50 m network

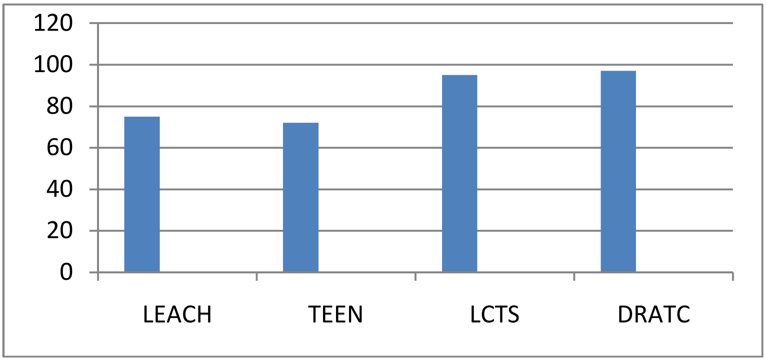


Fig. 3(b) Number of nodes alive after 100 rounds in a network of 100 nodes

The 75 bytes used for sensed data are not always transmitted as this is relevant only to the nodes that cross the threshold. On an average if we assume that only 50% of the nodes send their data we found that k (the number of bits transmitted) is reduced by N/2 * 75*100 bits for every round of transmission. This significantly saves energy of the transmitting sensor nodes and the receiving CH as shown in Equation (1) and (2).

Further, assuming that there are only 50% of the nodes in a network that transmit sensed data because of thresholds, the number of bits transmitted are reduced and thus energy can be saved as seen in equation (3) and (4). The transmission distance determines the energy consumption and so in DRATC the transmission distance is reduced because the data is transmitted to the actuator instead of the BS.

In DRATC 97 nodes were alive after 100 rounds, while only 75 nodes are alive in LEACH,72 were alive in TEEN and 95 in LCTS. This proves that DRATC is more energy efficient than its comparatives.

6 Conclusion

This algorithm is proposed with an aim to provide a solution to time critical applications like fire extinguishing. SCP-MAC is used for data gathering since TDMA is not suitable for time critical applications. This protocol helps in reducing the energy of the transmitter and receiver and therefore increases the network lifetime.

Performance of the proposed DRATC routing algorithm is assessed by simulation and compared with other clustering protocols like LEACH, LCTS and TEEN. The simulation results show that DRATC outperforms its comparatives by using a decentralised approach using actuators to take care of extinguishing in the clusters itself instead of data being sent to the BS for decision making. This reduces the transmission distance and helps in energy savings and also ascertains faster actions taken by CH to order the EN in case of fire. Further the nodes that do not meet the threshold are selected as cluster heads since they have more energy than the other nodes because they are not involved in sensing. Therefore DRATC provides an energy efficient routing scheme for vast range of time critical applications like temperature sensing and fire monitoring and extinguishing.

References

1. Manjeshwar, A., Agarwal, D.P.: APTEEN: A Hybrid Protocol for Efficient Routing and Comprehensive Information Retrieval in Wireless in Wireless Sensor Networks. In: The Proceedings of the 2nd International Workshop of Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, San Francisco, CA (April 2001)
2. Manjeshwar, A., Agarwal, D.P.: TEEN: A Protocol for Enhanced Efficiency in Wireless Sensor Networks. In: The Proceedings of the 1st International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, San Francisco, CA (April 2001)
3. Chamam, A., Pierre, S.: On the Planning of Wireless Sensor Networks: Energy–Efficient Clustering under the Joint Routing and Coverage Constraint. IEEE Transactions on Mobile Computing 8(8) (August 2009)
4. Steingart, D., et al.: Augmented Cognition For Fire Emergency Response: An Iterative User Study. In: Proceeding of the 1st International Conference on Augmented Cognition, Las Vegas (July 2005)
5. Wang, H., et al.: On the Flow Classification Thresholds of FD-MAC Protocol. In: IEEE ICC Proceedings (2011)
6. Roseline, R.A., Sumathi, P.: Local Clustering and Threshold Sensitive routing algorithm for Wireless Sensor Networks. In: The IEEE Sponsored International Conference on Devices Circuits and Systems, ICDCS 2012 (March 2012)
7. Lindsey, S., Raghavendra, C.S.: PEGASIS: Power-efficient Gathering in Sensor Information System. In: Proceedings IEEE Aerospace Conference, Big Sky, MT, vol. 3, pp. 1125–1130 (March 2002)
8. Ye, W., Silva, F., Heidemann, J.: Ultra-low duty cycle mac with scheduled channel polling. In: The 4th ACM Conference on Embedded Networked Sensor Systems, Boulder, CO (November 2006)

Application Security in Mobile Devices Using Unified Communications

Ramakrishna Josyula

Mobile Solution Architect, HTSC, HiTech ISU
Tata Consultancy Services Limited
Synergy Park, Gachibowli
Hyderabad, India
`j.ramakrishna@tcs.com`

Abstract. Unified communications is evolving with each passing day, by making its presence felt in several new user scenarios. It enables the consumer to avail information through various channels, thus ensuring a faster message delivery. Therefore, the time taken to take a decision based on the information or the time taken to act on the information is reduced, resulting in an enterprise becoming more agile, and meeting the ever increasing demand of its customers (internal or external).

Unified communications comprises many user scenarios, such as display of presence information of email recipients, exchange of information using Instant Messenger, video chat over Instant Messenger and so on. Till date, solutions related to unified communications have been predominantly deployed on desktop / laptop computers and only a small subset of use cases are enabled on the mobile platforms. The server component for desktop / laptop and mobile devices is the same to keep the information delivered through various media the same. One of the most important developments is the enablement of enterprise applications / processes / workflows on the mobile devices for faster decisions, capitalizing on the ubiquitous presence of mobile devices around the globe. Once, enterprise applications are deployed over mobile platforms / networks, we need to be mindful about:

- How to prevent data theft / data corruption when data flows through the network?
- **How to prevent data theft when the mobile device is lost / misplaced?**

The first problem can be handled through the use of various digital security mechanisms such as Cryptography APIs, Hashing, and Asymmetric Key Encryption. This paper explores a unique way of providing a solution to the second problem by extending the concept of Unified Communications to embrace enterprise applications / workflows / processes. The solution described in this paper integrates the enterprise application server with an IVR based approach to minimize enterprise data theft if a mobile device is lost.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 135–143.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

1 Introduction

Usage of mobile devices is rising at an exponential pace in the end user market. Enhanced usage for purposes of work is reflected by the fact that nearly seven out of 10 organizations rely more on mobile devices than they did 12 months ago[4]. From a survey conducted of senior IT decision-makers and end users worldwide by Vanson Bourne on behalf of Carnegie Mellon University and McAfee on the mobile security and consumerization of IT,48 percent of respondents informed that they use their smart phones for work[4]. Development of enterprise applications on mobile devices is also on the rise, although not at the same pace as development of end user applications. The two types of applications for mobile platforms are *Rich Client Applications* that get installed on the mobile device and *Mobile Web Applications* that are accessed through the browser on mobile device. The biggest advantage of Mobile Web Applications is that they can be accessed on different mobile platforms and different browsers with single source code base. They had the disadvantage of not providing compelling user experience and lack of local storage features, compared to rich clients. But HTML5 overcomes these problems with new specifications which enhance user experience and provides local storage.

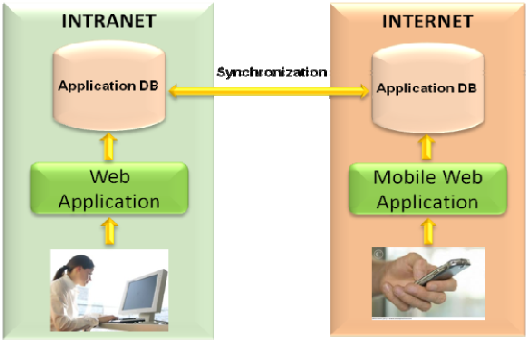


Fig. 1. As-Is Architecture for an Enterprise Application

With many mobile browsers supporting HTML 5, **mobile web solutions** built using HTML 5 are an attractive option for enterprises to adopt, if the aim is to make the enterprise applications available on mobile platforms. The advantage of **local storage** available in rich client applications is also available in HTML 5. Hence, the inclination towards a **mobile web application** is justified for enterprises. The architecture to enable an enterprise application on mobile platform is depicted in Figure 1:

- **Accessibility:** Enterprise users access the enterprise application from their desktops / laptops through a web application when they are logically located within the intranet of their enterprise.
- **Multi-Portal Support:** A web site, which is specifically designed for the smaller form factor of a mobile device, is deployed on the Internet

- **Database Co-location:** Web site for mobile devices interacts with a distinct application database physically located on Internet.
- **Mobile Access:** Enterprise users access the mobile web site from the browsers in their mobile devices.
- **Data Synchronization:** Two way synchronization happens between the two databases located in Intranet and Internet respectively

But one of the serious problems encountered by mobile device users is**loss of the mobile device**. According to phone interviews with the lost property offices of 15 UK airports, more than 5100 mobile phones had been left behind during the holiday season in 2010 - 2011, the bulk of which have not been claimed[5]. Four in 10 organizations say some of their mobile devices have been lost or stolen, half of which housed business-critical information, according to the survey by VansonBourne[4]. Many portable devices often **do not even have a password** on them to protect the unit's data. Surveys have also concluded that more than one-third of lost mobile device cases resulted in financial loss to the organization. The total cost of the loss is north of $2 billion[6]. Gartner estimates that the commercial cost to a business when an employee loses a mobile phone or a PDA is between USD 2000 and USD 3000.In such cases where mobile device is lost and security of the enterprise data is at stake, the enterprise user chooses one of the following actions to prevent data access from mobile device:

- Enterprise user accesses an Internet site [1]to wipe off / reset device to factory settings
- Enterprise user calls the telecom operator to block voice and data traffic from the lost mobile device.
- Enterprise user informs an administrator (either by calling or by sending an mail) to prevent access to enterprise applications [2]from his / her mobile device
- Enterprise user goes to a de provisioning portal and de provisions[3] the mobile device from accessing enterprise applications.

Figure 2 depicts the preventive actions taken by the enterprise user when security of the enterprise application is at stake.

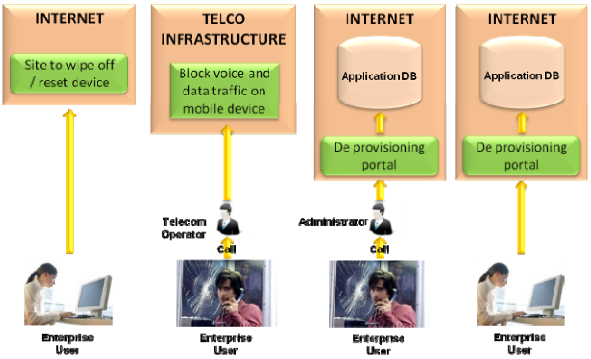


Fig. 2. De provisioning Methods

Some of the solutions available in market to lock out lost devices or erase data on lost devices, and their associated challenges are shown below:

Solution Provider	**Some Salient Features**	**Challenges**
YouGetItBack.com[7]	• SMS the **lock** code from another mobile device to lock lost device • **Lock** lost device from their web site	• Cannot lock from a normal landline phone • Accessing their website can take time in the situation where mobile device is lost
Microsoft[8]	• Lock & Erase using eMail account hosted on Exchange Server	• Accessing eMail account can take time in the situation where mobile device is lost • Does not work if eMails not configured on Exchange Server (for eg. Configured on Lotus Notes)
Lookout Mobile Security[9]	• Lock & Erase data on mobile devices	• Works only for iOS& Android devices. Doesn't work on other mobile platforms

2 Drawbacks of Existing Approach

Following are the drawbacks of the existing approaches to prevent access to the confidential enterprise data:

- It is time consuming for the user to access a desktop / laptop to visit the Internet site to wipe off / reset a device to factory settings. It provides enough time for the miscreant to steal enterprise data.
- It is time consuming to block the mobile device's voice and data traffic,even if the telecom operator is informed immediately. This also provides a very large window for data theft.
- Enterprises will have to bear the cost of anadministrator being employed 24 X 7 to take up calls from enterprise users to prevent access to enterprise applications from their lost mobile devices.
- It is difficult to access the de-provisioning portal if the device is lost, and the user does not have a desktop/laptop.
- Existing solutions in market work only on some mobile platforms or only in constrained scenarios.

3 Suggested Solution

The solution that is provided in this paper overcomes the drawbacks mentioned in section 2 by providing an easy way to prevent access to enterprise applications from lost / misplaced mobile devices, by integrating unified communications with workflows / processes.

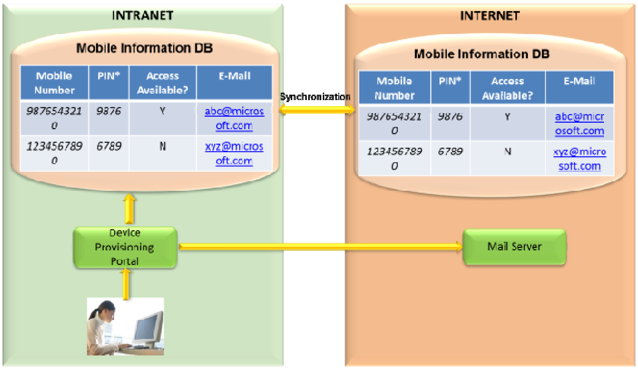
The solution provided is achieved without the presence of an intermediate human element. It also minimizes the data theft window. Hence, it is the most optimum solution for security of enterprise data from lost / misplaced mobile devices.

The solution enables the user to make a call to a preconfigured telephone number (could be toll free) when the mobile device is lost. The user will be authenticated to prevent misuse of the system. An IVR system is used to connect the telephone network to server software during authentication, thus adding a new channel / dimension to unified communication. Any system that uses an authentication to trigger an action, must first allow the user to provide unique credentials to the system. This is the first step in the entire procedure and is called **provisioning a mobile device** for using enterprise applications. It will then use information about whether a mobile device is provisioned or not before serving requests from a mobile device for enterprise applications. Lastly, the system provides a mechanism to **de provision a mobile** device in case a mobile device is lost / misplaced.

3.1 Provisioning a Mobile Device

Here a user provides the **mobile device identification number** (could be mobile number OR IMEI number) and a secret **PIN** to the system on an intranet portal. The mobile device identification number and the PIN are later used by the system to authenticate a user during the process of de provisioning the lost / misplaced mobile device. The architecture for provisioning is depicted in Figure 3.

The PINs are stored in the database in an encrypted format. The **Access Available** column indicates whether the mobile device can access enterprise applications or not. Information about newly provisioned mobile devices is synchronized with the database located in Internet at periodic intervals. In order to ensure that only required



Mobile Number	PIN*	Access Available?	E-Mail
9876543210	9876	Y	abc@microsoft.com
1234567890	6789	N	xyz@microsoft.com

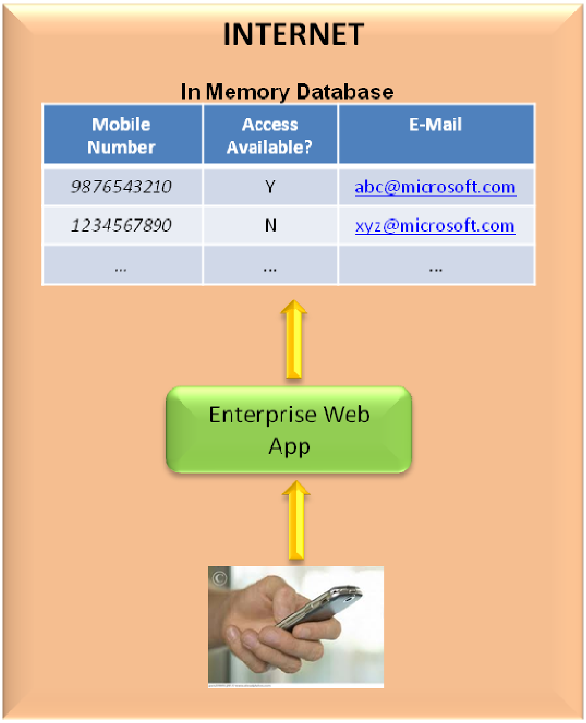
Mobile Number	PIN*	Access Available?	E-Mail
9876543210	9876	Y	abc@microsoft.com
1234567890	6789	N	xyz@microsoft.com

Fig. 3. Solution for Device provisioning

users have access to enterprise applications from mobile device, the process of supervisor approval for access to enterprise applications from mobile device can also be added before the mobile device is finally provisioned.

3.2 Serving Mobile Requests

When a web request comes from a mobile device, the mobile web application checks an in memory database if the mobile device is allowed access to the enterprise web application, as depicted in Figure 4.



Mobile Number	Access Available?	E-Mail
9876543210	Y	abc@microsoft.com
1234567890	N	xyz@microsoft.com
...	...	...

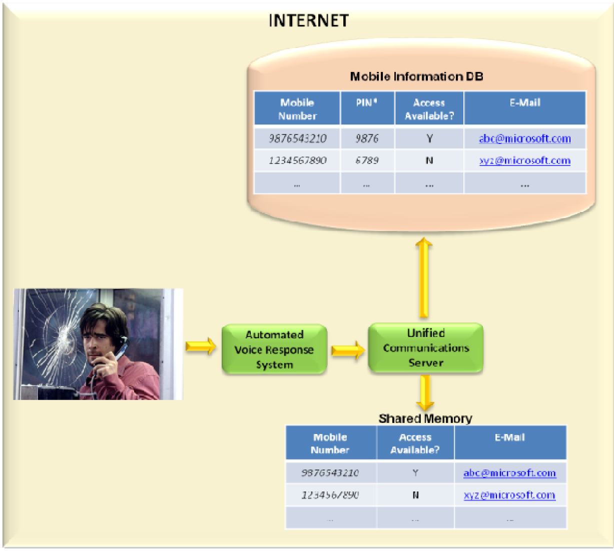
Fig. 4. Serving Mobile Requests

Figure 4 explains the concept that can be extended to all the enterprise mobile web applications. Allthe applications can refer to the same in memory database to find out if a mobile device is allowed access to enterprise applications. The in memory database is a snapshot of the data store containing information about each provisioned mobile device. This is done in order to speed up retrieval of provisioning information of mobile devices, because this information has to be retrieved for each mobile request.

3.3 De Provision Mobile Device

In case a mobile device is lost, the access to the enterprise web applications from the lost device should be blocked / prevented as soon as possible. This is done by calling up a preconfigured telephone number and authenticating oneself in the ensuing voice call. Once, authentication is successful, the mobile device is prevented from accessing enterprise applications. The architecture for de provisioning is depicted in Figure 5.

When a user loses the mobile device, providing a solution which involves calling up a preconfigured helpline to de provision a mobile device, is the best way to minimize data theft, rather than providing a web based solution. Web based solution requires the user to find a laptop / desktop when the mobile device is lost. Finding another phone (either public phone booth or borrowing a mobile device from another person for de provisioning) is much easier than finding a desktop / laptop. Hence a telephone network based solution is suggested to de provision a mobile device.


INTERNET
Mobile Information DB
Automated Voice Response System
Unified Communications Server
Shared Memory

Mobile Number	PIN*	Access Available?	E-Mail
9876543210	9876	Y	abc@microsoft.com
1234567890	6789	N	xyz@microsoft.com
...	...	...	...

Mobile Number	Access Available?	E-Mail
9876543210	Y	abc@microsoft.com
1234567890	N	xyz@microsoft.com
...	...	...

Fig. 5. De Provisioning Architecture

The telephone network integrates with Unified Communications Server to provide an automated voice response system. It begins with the authentication of the user where the user provides the mobile device identification number and PIN. The Unified Communications Server validates the authenticity of the PIN accessing the mobile information database. Once authentication is completed, the database is updated to state that web requests from the corresponding mobile device will be de provisioned. The in memory database is also updated to reflect the de provisioned status of mobile device. Once de provisioning is completed in database and memory, subsequent web requests from mobile device are redirected to an authentication page.

4 Advantages with Suggested Solution

The advantages of the solution are highlighted in sections 4.1 and 4.2.

4.1 Reduction in Data Theft Window

Data theft window is defined as the phase between the enterprise user losing the device and the enterprise user calling the automated voice response helpline to block the device. This scenario will occur when the miscreant finds the misplaced mobile device before it gets locked. This is depicted in Figure 6 with all the activities carried out, shown along a horizontal time line:

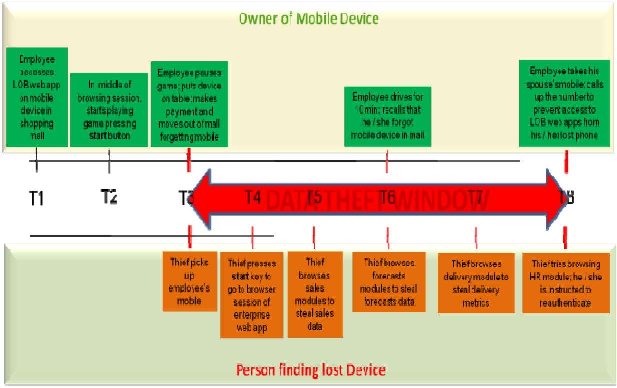

Owner of Mobile Device
Employee accesses LOB web app on mobile device in shopping mall
In middle of browsing session, starts playing game pressing start button
Employee pauses game; puts device on table; makes payment and moves out of mall forgetting mobile
Employee drives for 10 min; recalls that he / she forgot mobile device in mall
Employee takes his spouse's mobile; calls up the number to prevent access to LOB web apps from his / her lost phone
T1
T2
T3
T4
T5
T6
T7
T8
DATA THEFT WINDOW
Thief picks up employee's mobile
Thief presses start key to go to browser session of enterprise web app
Thief browses sales modules to steal sales data
Thief browses forecasts modules to steal forecasts data
Thief browses delivery module to steal delivery metrics
Thief tries browsing HR module; he / she is instructed to reauthenticate
Person finding lost Device

Fig. 6. Data Theft Window

The owner's activity at T8 will be delayed if an Internet site is to be accessed to reset / wipe off a mobile device. This demonstrates the reduction in data theft window with the solution presented in this paper.

4.2 Decrease in Operational Cost

The solution suggested in this paper does not need manual intervention during de provisioning, resulting in **lesser operational cost to the enterprise**, since process between the enterprise user and the system is automated.

5 Conclusion

- This paper has focussed on very important user scenario of securing enterprising applications in case a mobile device is lost.
- This paper has looked at the solutions already available today. Existing solutions result in a considerable time lag before device is blocked,and data can be stolenduring this period. This is unacceptable to enterprises.
- This paper presents an innovative solution that has the merits of minimizing data theft in case a mobile device is stolen, and also reducing the operational costs to the enterprise.
- This solution ensures that access to all enterprise applications from the misplaced / stolen mobile device is blocked in one go through a telephone call.

- The integration between the telephone network and the enterprise application database is achieved through **Unified Communication Server**.

This paper would suggest enhancing this idea to achieve the same end result using Short Messaging Services to prevent access to enterprise applications from lost / stolen mobile devices.

Acknowledgements. This paper would not have been written without the constant push by my manager **JitendraMaan**. He was unwavering in his opinion that this idea should materialize into a white paper. I would like to thank him for the birth of this paper. I would like to thank **Bala Prasad** for reviewing this paper from a technical standpoint and providing me valuable suggestions to improve this from a technical perspective.

Finally, I would like to thank **Sravani** who provided valuable inputs related to improving the format of this paper, which improved the readability of this paper.

References

[1] Remote wipe a mobile device, `http://www.google.com/support/a/bin/answer.py?answer=173390`
[2] Remote wipe using MobileMe, `http://support.apple.com/kb/TS2734`
[3] Mobile Device Management in iOS, `http://www.apple.com/iphone/business/integration/`
[4] Mobile device security, `http://www.darkreading.com/cloud-security/167901092/security/news/229625511/half-of-lost-or-stolen-mobile-devices-store-sensitive-company-data.html`
[5] Enterprise data security, `http://www.infosecurity-magazine.com/view/14865/thousands-of-mobile-devices-set-to-go-missing-over-the-holidays/`
[6] Cost of lost mobile device, `http://theemf.org/2010/12/03/whats-the-cost-of-a-lost-mobile-device/`
[7] Mobile security solution from YouGetItBack, `http://www.yougetitback.com/mobile_superhero`
[8] Mobile security solution from Microsoft, `http://www.microsoft.com/online/help/en-us/helphowto/0d03ee46-50da-43c1-837b-a868b208a764. htm`
[9] Mobile security solution fromMyLookout, `https://www.mylookout.com/features/missing-device/`

Iterative Image Fusion Using Fuzzy Logic with Applications

Srinivasa Rao Dammavalam[1], Seetha Maddala[2], and M.H.M. Krishna Prasad[3]

[1] Department of Information Technology, VNRVJIET, Hyderabad, India
[2] Department of CSE, GNITS, Hyderabad, India
[3] Department of CSE, JNTU College of Engineering, Vizianagaram, India
dammavalam2@gmail.com, smaddala2000@yahoo.com,
krishnaprasad.mhm@gmail.com

Abstract. Image fusion is the process of reducing uncertainty and minimizing redundancy while extracting all the useful information from the source images. Image fusion process is required for different applications like medical imaging, remote sensing, machine vision, biometrics and military applications. In this paper, an iterative fuzzy logic approach utilized to fuse images from different sensors, in order to enhance visualization. The proposed work further explores comparison between fuzzy based image fusion and iterative fuzzy fusion technique along with quality evaluation indices for image fusion like image quality index, mutual information measure, root mean square error, peak signal to noise ratio, entropy and correlation coefficient. Experimental results obtained from fusion process prove that the use of the proposed iterative fuzzy fusion can efficiently preserve the spectral information while improving the spatial resolution of the remote sensing images and medical imaging.

Keywords: image fusion, panchromatic, multispectral, fuzzy logic, image quality index, mutual information measure, entropy, correlation coefficient.

1 Introduction

Image fusion is a technique to combine information from two or more images of a scene into a single composite image that is more informative and is more suitable for visual perception or computer processing. Image fusion approach has been used in great many disciplines like medical imaging, remote sensing, navigation aid, machine vision, automatic change detection, biometrics and military applications etc. Multisensor image fusion for surveillance systems is proposed in which fuzzy logic approach utilized to fuse images from different sensors, in order to enhance visualization for surveillance [1]**.** In [2] urban remote image fusion using fuzzy rules approach utilized to refine the resolution of urban multi-spectral images using the corresponding high-resolution panchromatic images**.** After the decomposition of two input images by wavelet transform three texture features are extracted and then a fuzzy fusion rule is used to merge wavelet coefficients from the two images according to the extracted features. In [3] image fusion algorithm based on fuzzy logic and wavelet, aimed at the visible and infrared image fusion and address an algorithm

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 145–152.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

based on the discrete wavelet transform and fuzzy logic. In [3] the technique created two fuzzy relations, and estimated the importance of every wavelet coefficient with fuzzy reasoning. In [4] an Iterative Fuzzy and Neuro Fuzzy approach proposed for fusing medical images and remote sensing images and found that the technique very useful in medical imaging and other areas, where quality of image is more important than the real time application. In [5] a new method is proposed for Pixel-Level Multisensor image fusion based on Fuzzy Logic in which the membership function and fuzzy rules of the new algorithm is defined using the Fuzzy Inference System. A fuzzy radial basis function neural networks is used to perform auto-adaptive image fusion and in experiment multimodal medical image fusion based on gradient pyramid is performed for comparison [6]. In [7] a novel method is proposed using combine framework of wavelet transform and fuzzy logic and it provides novel tradeoff solution between the spectral and spatial fidelity and preserves more detail spectral and spatial information. Pixel & Feature Level Multi-Resolution Image Fusion based on Fuzzy logic in which images are first segmented into regions with fuzzy clustering and are then fed into a fusion system, based on fuzzy if-then rules [8].

2 Basic Elements of Fuzzy Approach

Fuzzy based image fusion requires basic elements to be introduced and exploited.

2.1 Fuzzy Logic

The necessity of fuzzy logic derives from the fact that most modes of human reasoning and especially common sense reasoning are approximate in nature.

The essential characteristics of fuzzy logic as founded by Zader Lotfi are as follows

- In fuzzy logic, exact reasoning is viewed as a limiting case of approximate reasoning
- In fuzzy logic everything is a matter of degree
- Any logical system can be fuzzified as
- In fuzzy logic, knowledge is interpreted as a collection of elastic or, equivalently, fuzzy constraint on a collection of variables
- Inference is viewed as a process of propagation of elastic constraints

The goal of this approach is to describe powerful features of fuzzy sets when used specially for image information processing.

It has been chosen to focus on the following points:

- Fuzzy sets are to represent spatial information in images along with its imprecision
- Operations recently generalized to fuzzy sets in order to manage spatial information and
- Information fusion using fuzzy combination operators
- Fast computation using fuzzy number operations.

2.2 Fuzzy Sets

In [9] Zadeh proposed that fuzzy set is a class of objects with a continuum of grades of membership. Fuzzy set is characterized by a membership function which assigns to each object a grade of membership ranging between zero and one. It was introduced as a mean to model the vagueness and ambiguity in complex systems. The idea of fuzzy sets is simple and natural.

2.3 Membership Functions

The membership function is a graphical representation of the magnitude of participation of each input in the input space. Input space is often referred as the universe of discourse or universal set, which contain all the possible elements of concern in each particular application. It associates a weighting with each of the inputs that are processed, define functional overlap between inputs, and ultimately determines an output response. The rules use the input membership values as weighting factors to determine their influence on the fuzzy output sets of the final output conclusion. Once the functions are inferred, scaled, and combined, they are defuzzified into a crisp output, which drives the system. There are different membership functions associated with each input and output response [10].

2.4 Fuzzy Rules

Human beings make decisions based on rules. Fuzzy machines, which always tend to mimic the behavior of man, work the same way. However, the decision and the means of choosing that decision are replaced by fuzzy sets and the rules are replaced by fuzzy rules. Fuzzy rules also operate using a series of if-then statements. For instance, if X then A, if Y then B, where A and B are all sets of X and Y. Fuzzy rules define fuzzy patches, which is the key idea in fuzzy logic.

3 Fuzzy Logic Based Image Fusion

Fuzzy image processing is not a unique theory. Fuzzy image processing is the collection of all approaches that understand, represent and process the images, their segments and features as fuzzy sets. The representation and processing depend on the selected fuzzy technique and on the problem to be solved.

3.1 Fuzzy Logic in Image Processing

Fuzzy image processing has three main stages: Image fuzzification, Modification of membership values and Image defuzzification. The coding of image data (fuzzification) and decoding of the results (defuzzification) are steps that make possible to process images with fuzzy techniques. The main power of fuzzy image processing is in the middle step (modification of membership values). After the image data are transformed from gray-level plane to the membership plane (fuzzification), appropriate fuzzy techniques modify the membership values [11].

3.2 Steps Involved in Image Fusion Using Fuzzy Logic

The original image in the gray level plane is subjected to fuzzification and the modification of membership functions is carried out in the membership plane. The result is the output image obtained after the defuzzification process [12].

3.3 Steps Involved in Iterative Image Fusion Using Fuzzy Logic

The input image in the gray level plane is subjected to fuzzification and the modification of membership functions is carried out in the membership plane. The result is the output image obtained after the defuzzification process.

Algorithm steps for iterative image fusion using fuzzy logic approach [13].

- Read first image in variable I1 and find its size (rows: rl, columns: c1).
- Read second image in variable I2 and find its size (rows:`r2,` columns: c2).
- Variables I1 and I2 are images in matrix form where each pixel value is in the range from 0-255. Use Gray Colormap.
- Compare rows and columns of both input images. If the two images are not of the same size, select the portion, which are of same size.
- Convert the images in column form which has C= rl*cl entries.
- Make a fis (Fuzzy) file, which has two input images.
- Decide number and type of membership functions for both the input images by tuning the membership functions. Input images in antecedent are resolved to a degree of membership ranging 0 to 255
- Make rules for input images, which resolve the two antecedents to a single number from 0 to 255.
- For num=l to C in steps of one, apply fuzzification using the rules developed above on the corresponding pixel values of the input images which gives a fuzzy set represented by a membership function and results in output image in column format
- continue the fusion process with two inputs, in which one of the inputs is the latest output and second is the required input image.
- Convert the column form to matrix form and display the fused image.

Membership functions and rules considered in the fuzzy system

1. if (input1 is mf1) and (input2 is mf1) then (output1 is mf1)
2. if (input1 is mf2) and (input2 is mf1) then (output1 is mf2)
3. if (input1 is mf2) and (input2 is mf2) then (output1 is mf2)
4. if (input1 is mf3) or (input2 is mf2) then (output1 is mf3)
5. if (input1 is mf1) and (input2 is mf3) then (output1 is mf1)
6. if (input1 is mf3) or (input2 is mf3) then (output1 is mf2)

4 Evaluation Indices of Image Fusion

Evaluation measures are used to evaluate the quality of the fused image. The fused images are evaluated, taking the following parameters into consideration.

4.1 Image Quality Index

Image quality index (IQI) measures the similarity between two images (I1 & I2) and its value ranges from -1 to 1. IQI is equal to 1 if both images are identical. IQI measure is given by [14]

$$IQI = \frac{m_{ab}\; 2\,xy\; 2\,m_a m_b}{m_a m_b\, x^2 + y^2 m_a^2 + m_b^2} \tag{1}$$

Where x and y denote the mean values of images I1 and I2 and m_a^2, m_b^2 and m_{ab} denotes the variance of I1, I2 and covariance of I1 and I2.

4.2 Mutual Information Measure

Mutual information measure (MIM) furnishes the amount of information of one image in another. This gives the guidelines for selecting the best fusion method. Given two images *M (i, j) and N (i, j)* and MIM between them is defined as:

$$I_{MN} = \sum_{x,y} P_{MN}(x,y) \log \frac{P_{MN}(x,y)}{P_M(x) P_N(y)} \tag{2}$$

Where, P_M (x) and P_N (y) are the probability density functions in the individual images, and $P_{MN}(x, y)$ is joint probability density function.

4.3 Root Mean Square Error

The root mean square error (RMSE) measures the amount of change per pixel due to the processing. The RMSE between a reference image R and the fused image F is given by

$$RMSE = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (R(i,j) - F(i,j))} \tag{3}$$

4.4 Peak Signal to Noise Ratio

Peak signal to noise ratio (PSNR) can be calculated by using the formula

$$PSNR = 20 \log_{10} \left[\frac{L^2}{MSE} \right] \tag{4}$$

Where MSE is the mean square error and L is the number of gray levels in the image.

4.5 Entropy

The entropy of an image is a measure of information content. It is the average number of bits needed to quantize the intensities in the image. It is defined as:

$$E = -\sum (p * \log_2(p)) \tag{5}$$

Where p contains the histogram counts returned from imhist.

4.6 Correlation Coefficient

The Correlation Coefficient (CC) method is used to determine how closely the input and output images co-vary. Correlation coefficient is widely used for comparing images. It is widely used in statistical analysis, pattern recognition, and image processing.

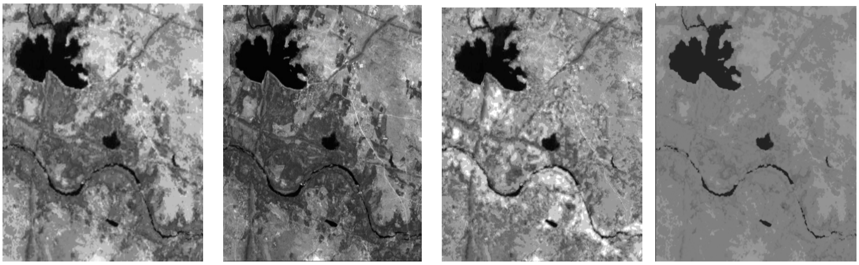
$$CC = \frac{\sum_{i=1}^{n}(Xi - X)(Yi - Y)}{\sqrt{\sum_{i=1}^{n}(Xi - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Yi - \overline{Y})^2}}$$

Where, X_i is the intensity of the i^{th} pixel in image1, Y_i is the intensity of the i^{th} pixel in image2, X is the mean intensity of image1 and Y is the mean intensity of image2.

5 Experimental Results and Analysis

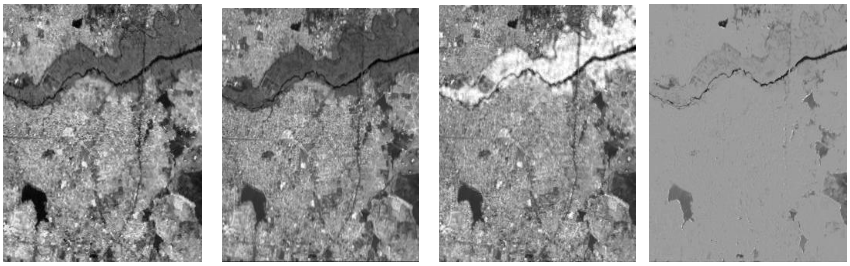
In this section, we performed fusion between a panchromatic (PAN) image and multispectral (MS) image with proposed algorithm. Case 1, Case 2 Panchromatic and Multispectral images of the Hyderabad city, AP, INDIA are acquired from the IRS 1D LISS III sensor at 05:40:44, Case 3 images are MRI and CT scan images of brain.

The proposed algorithm has been implemented using Matlab 10.0. It can be seen from the above table and the image results that the iterative fuzzy logic approach are having much better results when compared with the fuzzy logic approach. Table 1 show that proposed iterative fuzzy logic approach gives comparatively better IQI, MIM and PSNR and Correlation Coefficient through preserving more spectral and spatial information. Considerable differences in evaluation indices are obtained through iterative fuzzy logic with lower RMSE values. Iterative image fusion using

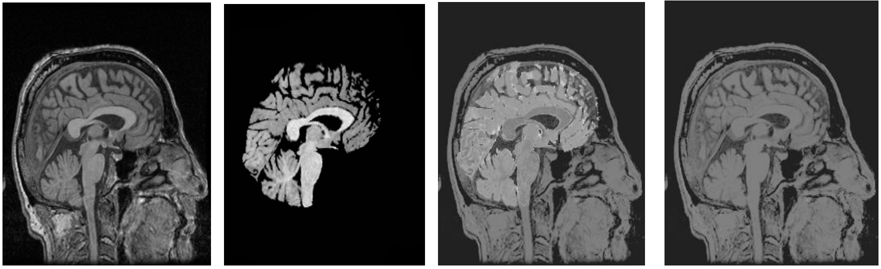


Case 1: (a) (b) (c) (d)

Fig. 1. Some example images (a), (b), (e), (f), (i) and (j): original input images; (c), (g) and (k): fused images by fuzzy logic and (d), (h) and (l): fused images by iterarative fuzzy logic.



Case 2: (e) (f) (g) (h)



Case 3: (i) (j) (k) (l)

Fig. 1. (*continued*)

Table 1. The evaluation indices of image fusion based on fuzzy and iterative fuzzy logic

Method	IQI	MIM	RMSE	PSNR	Entropy	Correlation Coefficient with MS/MRI	Correlation Coefficient with PAN/CT
Fuzzy Logic							
(Case 1)	0.9491	1.4662	37.2497	16.7084	5.1603	0.7589	0.7127
(Case 2)	0.9758	0.4628	44.8448	15.0966	4.4881	0.5833	0.5808
(Case 3)	0.7696	1.4684	41.8933	15.6879	5.7515	0.8941	0.6470
Iterative Fuzzy Logic							
(Case 1)	0.9680	2.9638	30.7766	18.3664	4.4974	0.8555	0.7495
(Case 2)	0.9883	1.1439	29.5840	18.7097	5.5088	0.8163	0.6824
(Case 3)	0.8624	1.6859	30.7798	18.3655	5.2903	0.9658	0.4747

fuzzy logic improves visualization and information from the source images which are important in many different applications. Therefore it is ascertained from experimental results that iterative fuzzy logic based image fusion schemes perform better results in remote sensing and medical imaging applications.

6 Conclusions

In this paper, iterative image fusion using fuzzy logic approach for remote sensing and medical imaging has been explored. In order to evaluate the outputs and compare the methods, the assessment criteria, evaluation indices of image fusion are employed. The experimental results clearly show that the introduction of the iterative image fusion using fuzzy logic gives a considerable improvement on the performance of the fusion process. The iterative fuzzy technique can be further extended to real time images and to integrate valid evaluation metric of image fusion schemes. Image fusion using neuro fuzzy logic, iterative neuro fuzzy, automatic determinations of the percentage of overlapping among fuzzy sets, membership functions and determination of fuzzy rules are also worthy of research. The work could be extended to video image processing for real time processing.

Acknowledgments. This paper is partially supported by the All India Council for Technical Education, New Delhi, India under Research Promotion Scheme (RPS) Grant No. 8023/RID/RPS-80/2010-11.

References

1. Yi, Z., Ping, Z.: Multisensor Image Fusion Using Fuzzy Logic for Surveillance Systems. In: IEEE Seventh International Conference on Fuzzy Systems and Discovery, Shanghai, pp. 588–592 (2010)
2. Yang, X.H., Huang, F.Z., Liu, G.: Urban Remote Image Fusion Using Fuzzy Rules. In: IEEE Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, pp. 101–109 (2009)
3. Mengyu, Z., Yuliang, Y.: A New image Fusion Algorithm Based on Fuzzy Logic. In: IEEE International Conference on Intelligent Computation Technology and Automation, Changsha, pp. 83–86 (2008)
4. Ranjan, R., Singh, H., Meitzler, T., Gerhart, G.R.: Iterative Image Fusion technique using Fuzzy and Neuro fuzzy Logic and Applications. In: IEEE Fuzzy Information Processing Society, Detroit, USA, pp. 706–710 (2005)
5. Zhao, L., Xu, B., Tang, W., Chen, Z.: A Pixel-Level Multisensor Image Fusion Algorithm Based on Fuzzy Logic. In: Wang, L., Jin, Y. (eds.) FSKD 2005, Part I. LNCS (LNAI), vol. 3613, pp. 717–720. Springer, Heidelberg (2005)
6. Wang, Y.P., Dang, J.W., Li, Q., Li, S.: Multimodal Medical Image fusion using Fuzzy Radial Basis function Neural Networks. In: IEEE International Conference on Wavelet Analysis and Pattern Recognition, Beijing, pp. 778–782 (2007)
7. Tanish, Z., Ishit, M., Mukesh, Z.: Novel hybrid Multispectral Image Fusion Method using Fuzzy Logic. I. J. Computer Information Systems and Industrial Management Applications, 096–103 (2010)
8. Bushra, N.K., Anwar, M.M., Haroon, I.: Pixel & Feature Level Multi-Resolution Image Fusion based on Fuzzy Logic. In: ACM Proc. of the 6th WSEAS International Conference on Wavelet analysis & Multirate Systems, Romania, pp. 88–91 (2006)
9. Zadeh, L.A.: Fuzzy Sets. J. Information and Control 8, 338–353 (1965)
10. Praveena, S.M.: Multiresolution Optimization of Image Fusion. In: National Conference on Recent Trends in Communication and Signal Processing, Coimbatore, pp. 111–118 (2009)
11. Maruthi, R., Sankarasubramanian, K.: Pixel Level Multifocus Image Fusion Based on Fuzzy Logic Approach. J. Information Technology 7(4), 168–171 (2008)
12. Dammavalam, S.R., Maddala, S., Krishna Prasad, M.H.M.: Quality Evaluation Measures of Pixel – Level Image Fusion Using Fuzzy Logic, pp. 485–493 (2011)
13. Thomas, M., David, B., Sohn, E.J., Kimberly, L., Darryl, B., Gulshecn, K., Harpreet, S., Samuel, E., Grmgory, S., Yelena, R., James, R.: Fuzzy Logic bascd Image Fusion Aerosense, Orlando (2002)
14. Mumtaz, A., Masjid, A.: Genetic Algorithms and its Applicatio to Image Fusion. In: IEEE International Conference on Emerging Technologies, Rawalpindi, pp. 6–10 (2008)

A Review on Clustering of Web Search Result

Mansaf Alam and Kishwar Sadaf

Department of Computer Science, Jamia Millia Islamia
New Delhi, India
{mansaf_alam2002,kishwarsadaf}@gmail.com

Abstract. The over abundance of information on the web, makes information retrieval a difficult process. Today's search engines give too many results out of which only few are relevant. A user has to browse through the result pages to get the desired result. Web search result clustering is the clustering of results returned by the search engines into meaningful groups. This paper throws light and categorizes various clustering techniques that have been applied on the web search result.

Keywords: Information Retrieval, document-clustering, web search result.

1 Introduction

The information available on the web is unstructured, disorganized, dynamic and heterogeneous in nature and enormously large. Moreover the process of retrieval is highly affected by the vague query put up by the average user. Today's search engines return too many results which are not necessarily relevant to the user's need. Usually user has to traverse several search result pages to get to the desired result. A way of assisting users in finding what they are looking for quickly is to group the search results by topic. The user does not have to reformulate the query, but can merely click on the topic most accurately describing his or her specific information need. This grouping of result is called Clustering. More specifically, it is a process of grouping similar documents into clusters so that documents of one cluster are different from the documents of other clusters. There are many web clustering engines available on the web (Carrot2, Vivisimo, SnakeT, Grouper etc) which give the search results in forms of clusters. A web clustering engine takes the result, returned by the search engine as input and performs clustering and labelling on that result. This process is usually seen as complementary rather than alternative and different to the search engine [1]. The main use for web search result clustering is not to improve the actual ranking, but to give the user a quick overview of the results. Having divided the result set into clusters, the user can quickly narrow down his search further by selecting a cluster. This resembles query refinement, but avoids the need to query the search engine for each step. Web search result clustering has been the focus of IR community since the emergence of web search engine. Therefore numerous works has been done in this area. The Scatter/Gather system by [2] is held as the predecessor and conceptual father of all web search result clustering. Web Search engine is the most commonly used

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 153–159.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

tool for information retrieval on the web; however, its current status is far from satisfaction for several possible reasons [3]:

- Information on the Web multiply continuously;
- Different users have different requirements and expectations for search results;
- Users want whole picture of their search result on the first page of the search engine.
- Sometimes search request cannot be expressed clearly just in several keywords;
- Synonymous and polysemous words make searching more complicated;
- Users may be just interested in "most qualified" information or small part of information returned while thousands of pages are returned from search engine;
- Many returned pages are useless or irrelevant;
- Many useful information/pages are not returned for some reasons.

2 Traditional Clustering Techniques

Clustering in IR context can be classified as pre-retrieval and post-retrieval. In preretrieval clustering approach, all the documents that contain the query terms are retrieved and a clustering is done using some similarity function. The result is then presented to the user. While in postretrieval clustering approach, clustering is applied on the documents that are returned by the search engine. Clustering whether preretrieval or post retrieval can be classified into main two categories: Hierarchical clustering and Flat clustering. Although there are numerous clustering techniques but these clustering methods form the basis for other clustering techniques. Hierarchical clustering methods group the documents into a hierarchical tree structure by *Agglomerative (*bottom-up) approach or *Divisive* (top-down) approach. [4] [5]. Hierarchical methods are widely adopted, but its time complexity of $O(n^2)$ struggle to meet the speed requirements of the web. The K-Means algorithm is the most common flat clustering and comes in many flavors [Steinbach]. Although it has O(n) time complexity, it produces a fixed number (k) of flat clusters and a "bad choice" in the random selection of initial clusters can severely degrade performance.

Above mentioned clustering techniques use the vector based representation of the document where documents are grouped only when they share exact common individual words separately. Frequent itemset clustering technique is characterized by focusing on grouping documents that share sets of frequently occurring phrases**.** In [6] Fung et al propose using the data mining notion of frequent itemsets to cluster documents. Frequent itemsets originate from association rule mining. The idea is that documents that share a set of words i.e. itemsets that appear frequently are related, and this is used to cluster documents.

The traditional clustering techniques can be applied on web search result. In case of hierarchical approach, there is tradeoff between quick result and good quality result. Since web search result clustering is an online process, time can't be traded. Usually operating on document vectors with a time complexity of $O(n^2)$ or more, clustering more than a few hundred snippets is often unfeasible. Another problem is that if two clusters are incorrectly merged in an early state there is no way of fixing this later in the process. Finding the best halting criterion that works well with all queries can also be very difficult. In flat clustering approach, the number of clusters should be known prior to clustering. The search engine returns thousands of documents for a simple query. It is difficult to know in advance that how many clusters will be formed from the numerous documents. Several problems exist with this approach: It can only produce a fixed number of clusters (k). It performs optimally when the clusters are spherical but we have no reason to assume that documents clusters are spherical. Finally, a "bad choice" in the random selection of initial clusters can severely degrade performance.

3 Search Result Clustering

Clustering of web search results has been studied in the area of Information Retrieval (IR). The goal of clustering search result is to give user an idea of what the result contains. This idea is in the form of clusters. Clustering in context of web search result means organizing query result pages into groups based on their similarity between each other. Vivisimo, Carrot2, Kartoo etc are some of common commercial clustering engines available. Search result clustering techniques specific to the search engine result can be broadly classified as content-based and topology-based clustering. Document snippet clustering can be classified as the content-based clustering. Graph based clustering can be categorized as topology-based clustering.

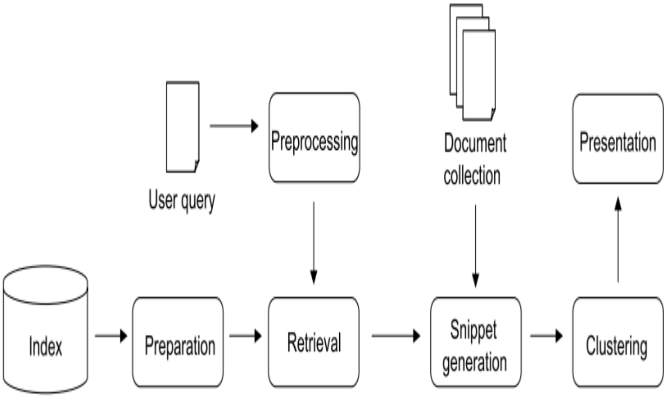


Fig. 1. A generic web search result clustering system using snippets

3.1 Document Snippet Clustering

A common technique used by clustering engines is to cluster so-called document snippets rather than entire documents. Snippets are the small paragraphs often displayed along with web search results to give the user a suggestion of the document contents. Snippets are considerably smaller than the documents (typically only

100-200 characters), thereby drastically reducing the computational cost of the clustering. This is very important since scalability and performance are major challenges for most clustering engines. When building clusters based only on short extracts from the documents, the quality of the snippets returned by the search engine naturally becomes very important. Snippet generation approaches vary from naive (e.g. first words in the document) to more sophisticated (e.g. display the passage containing the most words from the query or multiple passages containing all or most of the query keywords).

Clustering algorithms differ in their sensitivity to document length, but generally the effect of using snippets as opposed to entire documents is surprisingly small as demonstrated by [7]. Only about 15% average loss of precision for the clusters was found when using snippets rather than entire documents. The article suggests that this is caused by the search engines efforts to extract meaningful snippets concerning the user query, which reduces the noise present in the original document so much that the results do not deteriorate significantly. This further emphasizes the importance of high quality snippet extraction for snippet clustering approaches. In [8], Yao et al put forward a token-based web-snippet clustering. Direct probability graph is used to represent the snippet features. The documents which share the same features are grouped into one cluster.

An important snippet-based clustering, Suffix Tree Clustering (STC), is based on the Suffix Tree Document (STD) model which was proposed by Zamir et al [7]. The STC algorithm was used in their meta-searching engine to cluster the document snippets returned from other search engine in realtime. The similarity between documents is based on matching phrases rather than on single words only. A phrase in this context is an ordered sequence of one or more words. The STC algorithm focuses on clustering document snippets returned by the search engine, faster than standard data mining approaches. Its time complexity is linear to the number of snippets, making it attractive when clustering a large number of documents. There are numerous works available, which are derived from STC algorithm [9] [10]. In [11], authors propose an online clustering method using the STC algorithm. This algorithm groups web search results through a hierarchical, semantic and online clustering approach and named as SHOC. It consists of three steps-data collection and cleaning, feature extraction and identifying and organizing clusters. The problem with STC is the use of continuous phrases as the only features measuring similarity between documents. It can cause certain problems in languages where the positional order of parts of speech in a sentence may change. In [12], Osinski proposes a method where first, labels for clusters are defined using the input document snippets and then documents are assigned to these clusters according to their similarity with the labels.

In [13], Mecca et al use Singular Value Decomposition (SVD) on documents returned by the search engine as a whole instead of document snippets. Their algorithm has been integrated with Noodles search engine.

3.2 Graph-Based Search Result Clustering

The documents returned by the search engine in answer of a query can be looked as a subgraph of the whole web graph. The documents to be clustered can be viewed as a set of nodes and the edges between the nodes represent the relationship between them.

The edges bare a weight, which denotes the strength of that relationship. Graph based algorithms rely on graph partitioning, that is, they identify the clusters by cutting edges from the graph such that the edge-cut, i.e. the sum of the weights of the edges that are cut, is minimized. Since each edge in the graph represents the similarity between the documents, by cutting the edges with the minimum sum of weights the algorithm minimizes the similarity between documents in different clusters. The basic idea is that the weights of the edges in the same cluster will be greater than the weights of the edges across clusters. Hence, the resulting cluster will contain highly related documents. Sha et al [14] propose a web search result clustering based on lexical graph. Authors show that lexical graph structure is suitable in finding the word relationship and synonyms. The web search result is structured as a graph. They assert that their method performs better than STC and k-means. Navigelli et al [15] use graph-based clustering approach to cluster web search results. They first use graph clustering for word sense disambiguation and then cluster the results based on their semantic similarity.

Search engine like Google uses the hyperlink structure of the web to retrieve query results. This hyperlink structure is basically a directed graph, where a node represents a page and a link is characterize d by a directed edge. The pioneer works in the field of link-based web search are [16] and [17]. They have inspired many other works. Applying clustering on the hyperlink structure of web documents is an evolving area in IR research. Wang et al in [18], propose a web search result clustering which makes use of the hyperlinks between the pages and employs the HITS [16] algorithm and k-means clustering. Authorities are pages that are recognized as providing significant, trustworthy, and useful information on a topic. Hubs are index pages that provide lots of useful links to relevant content pages. PageRank uses an alternative link-analysis method. It ranks pages just by authority. Its applied to the entire web rather than a local neighborhood of pages surrounding the results of a query. In [19], Bradic uses the graph structure of the document that is preserved in the search result. Then this subgraph is partitioned to form topic related clusters.

3.3 Rank-Based and Hybrid Search Result Clustering

Clustering can be applied on the ranked result returned by the search engine or ranking can be done within each clusters formed. Leuski et al [20] propose a method where ranking and clustering are combined. The approach first traverses through the ranked list returned by the search engine until a relevant document is found. This document is then used as a cluster seed and clustering is performed on unexamined documents. Duhan et al [21] combine the power of ranking and clustering. First they cluster the documents in accordance with the query and then apply ranking within each cluster. Combining the topology and contents of the documents on the web, search result clustering can perform proficiently. Wang et al [18] propose a web search result clustering which makes use of the hyperlinks between the pages and employs the HITS algorithm and k-means clustering. Bekkerman et al [22] propose a multiagent, and bidirectional based heuristic search in the web graph to form clusters. They apply beam search in the search result graph in parallel to traditional topical clustering method on the clusters so formed. In [23], authors propose an approach based on the topology i.e. hyperlink and contents of the documents returned by the

Table 1. Search Result Clustering

Clustering Type	Input Data	General Clustering Methods
Snippet-based	Document Snippets returned by the Search Engine	STC, SHOC, SVD and other Hierarchical and flat clustering methods
Graph-based	Underlying Web graph of the search result	Graph Clustering Methods
Hybrid	Underlying web graph and the content of the documents of the search result	Combination of graph and semantic or lexical based clustering methods
Rank-based	Documents returned in the ranked search result	Various Hierarchical and Flat clustering methods

search engine. They first apply heuristic search on the web search result graph to form cluster and then perform Latent Semantic Indexing process in each cluster to derive semantic similarity between documents.

4 Discussion and Future Work

Users want a complete depiction of their search result at once. Clustering is the best possible solution for this problem. Clustering in a data mining setting has been researched for decades. Lately, document clustering used to cluster web search engine results has received much attention. Although commercial clustering engines exist, clustering is yet to be deployed on major search engines like Google. As the primary aim of a search results clustering is to decrease the effort required to find relevant information, user experience of clustering-based search result is of crucial importance. Part of this experience is the speed at which the results are delivered to the user. Ideally, clustering should not introduce a noticeable delay to normal query processing. This is presumably because of the computational overhead caused by data mining methods. It should have a low response time. Another issue related to search result clustering is labelling the clusters. The labels should be such that they must define the clusters. Unfortunately, regardless of how good the document grouping is, users are not likely to click on clusters if the labels are ill-defined. Defining accurate labels for cluster is an interesting and important area of research in the field of IR. Clustering performance is also a major issue, because web users expect fast response times.

References

[1] Carpenito, C., Osinski, S., Romano, G., Weiss, D.: A Survey of Web Clustering Engines II. ACM Computing Surveys 41(3), Article 17 (2009)

[2] Cutting, D.R., Kager, D.R., Pedersen, J.O.: Tukey JW Scatter/gather: a cluster-based approach to browsing large document collections. In: The 15th Annual International ACM Sigir Conference on Research and Development in Information Retrieval (1992)

[3] Wang, Y., Kitsuregawa, M.: Link Based Clustering of Web Search Results. In: Wang, X.S., Yu, G., Lu, H. (eds.) WAIM 2001. LNCS, vol. 2118, pp. 225–236. Springer, Heidelberg (2001)

[4] Han, J., Kamber, M.: Data Mining -Concepts and Techniques. Academic Press (2001)

[5] Steinbach, M., Karypis, G., Kumar, M.: A Comparison of Document Clustering Techniques II. In: KDD Workshop on Text Mining (2000)

[6] Fung, B.C.M., Wang, K., Ester, M.: Hierarchical Document Clustering (2003)

[7] Zamir, O., Etzioni, O.: Web Document Clustering: A Feasibility Demonstration. In: Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 46–54 (1998)

[8] Yao, T., Li, J.: A Token-based Online Web-Snippet Clustering Approach based on Directed Probability Graph. Journal of Computational Information Systems 5(3), 1235–1244 (2009)

[9] Branson, S., Greenberg, A.: Clustering Web Search Results Using Suffix Tree Methods. Stanford University (2009)

[10] Janruang, J., Guha, S.: Semantic Suffix Tree Clustering. In: First IRAST International Conference on Data Engineering and Internet Technology, DEIT (2011)

[11] Zhang, D., Dong, Y.: Semantic, Hierarchical, Online Clustering of Web Search Results. In: Yu, J.X., Lin, X., Lu, H., Zhang, Y. (eds.) APWeb 2004. LNCS, vol. 3007, pp. 69–78. Springer, Heidelberg (2004)

[12] Osinski, S.: A Concept-Driven Algorithm for Clustering Search Results. IEEE Intelligent Systems 20(3), 48–54 (2005)

[13] Mecca, G., Raunich, S., Pappalardo, A.: A New Algorithm for Clustering Search Result. Journal of Data & Knowledge Engineering 62(3) (2007)

[14] Sha, Y., Zhang, G.: Web Search Result Clustering Algorithm based on Lexical Graph. Journal of Computational Information Systems 5(1) (2009)

[15] Navigli, R., Crisafulli, G.: Inducing Word Senses to Improve Web Search Result Clustering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2010)

[16] Kleinberg, J.: Authoritative Sources In A Hyperlinked Environment. In: Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, SODA (1998)

[17] Page, L., Brin, S.: Web document clustering: A feasibility demonstration. In: Proceedings of SIGIR 1998, Melbourne, Australia (1998)

[18] Bradic, A.: Search Result Clustering via Randomized Partitioning of Query-Induced Subgraphs. Telfor Journal 1(1) (2009)

[19] Leuski, A., Allan, J.: Improving Interactive Retrieval by Combining Ranked Lists and Clustering. In: Proceeding of RIAO (2000)

[20] Duhan, N., Sharma, A.K.: A Novel Approach for Organizing Web Search Results using Ranking and Clustering. International Journal of Computer Applications 5(10) (2010)

[21] Wang, Y., Kitsuregawa, M.: Link Based Clustering of Web Search Results. In: Wang, X.S., Yu, G., Lu, H. (eds.) WAIM 2001. LNCS, vol. 2118, pp. 225–236. Springer, Heidelberg (2001)

[22] Bekkerman, R., Zilbersteinn, S., Allan, J.: Web Page Clustering using Heuristic Search in the Web Graph. In: Proceedings of IJCAI 2007, the 20th International Joint Conference on Artificial Intelligence (2007)

[23] Alam, M., Sadaf, K.: Web Search Result Clustering using Heuristic Search and Latent Semantic Indexing. International Journal of Computer Applications 44(15) (2012)

Application of Information Theory for Understanding of HLA Gene Regulation in Leukemia

Durjoy Majumder

Department of Physiology, West Bengal State University,
Berunanpukuria, Malikapur, Barasat, Kolkata 700 126
durjoy@rocketmail.com, durjoym@in.com

Abstract. The classical concept of information entropy can be useful in analyzing data pertaining to bioinformatics. In the present work, this has been utilized in understanding of the regulation of HLA gene expression by the inducible promoter region binding transcription factors (TFs). Human HLA surface expression data acquired through flow cytometry and corresponding different TFs expression data acquired through semi-quantitative PCR have been used in this work. The gene regulation phenomenon is considered as an information propagation channel with an amount of distortion. Information entropies computed for the source, receiver and computation of channel equivocation and mutual information are used to characterize the phenomenon of HLA gene regulation. The results obtained in the current exercise reveals that the state of leukemia alters the role of each TF, which tally with the current hypotheses about HLA gene regulation in different leukemias. Hence, this work shows the applicability of information theory in understanding of HLA gene regulation derived from human data.

Keywords: MHC expression, Information entropy, channel entropy.

1. Introduction

Human leukocytic antigen (HLA) class I (HLA – ABC) molecules in conjunction with β2-microglobulin (β2M) molecule are present in almost all the nucleated cell surface of the human. HLA molecules class II (HLA – DP, DQ and DR) is present only on the cell surface of B-cell, T – helper cells, macrophage and dendritic cells of the immune system. It has been reported that in cancer cells HLA class I gene expression is frequently down-regulated that may enable them to escape from immune attack. It is noted that HLA downregulation is also evident in leukamia cells both at the transcriptional and at the translational level [1-3]. Therefore understanding the mechanism of HLA gene regulation would be of interest in cancer/leukemia immunology. In this connection it would be interesting to note that in cancer no mutation has been identified in HLA gene so far [2].

Transfection experiments with different combinations of construct of the HLA upstream sequence together with marker gene reveal several regulatory sequence elements present in the HLA promoter region. They are named as enhancer (Enh) A

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 161–173.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

and B. TF NF-κB (Rel) binds to Enh A region. Enh B region (also known as MARM, MHC antigen regulatory module) is composed of X1, site α (in class I)/X2 (in class II), CREB-1 (cyclic AMP responsive element) binding site and inverted CCAAT box. Another sequence element called IRE (Interferon responsive element) or IRSE (Interferon responsive sequence element) is also present that partially overlaps with Enh A and is responsible for binding with different interferon regulatory transcription factors (IRF). Several TFs, namely RFX5 (regulatory factor – X5), RFXB (or RFXANK), RFXAP and CIITA (HLA class II transactivator) bind to site α. Though there is sequence homology between Enh A and Enh B sequence, however, Enh A together with IRE element is absent in HLA class II promoter sequence element [4].

Enh A is responsible for the constitutive expression of HLA class I gene, whereas Enh B is responsible for interferon (IFN) induced expression via CIITA promoter [5-9]. Alteration in binding of NF-κB to EnhA causes downregulation in the constitutive expression of HLA class I [10]. Aberrant NF-κB activity, mutations and rearrangements of NF-κB/IκB have been described in various types of leukemia and lymphomas [11-12]. IRFs have also been correlated to the response to IFN therapy in leukemia [13-14].

It has been shown that CIITA has an intrinsic histone acetyl transferases activity [15]. A cytokine IK has been identified that has been shown to downregulate the activity of CIITA even in the presence of IFN- stimulated B cells [8-9]. However, in most of the AML and B-ALL and 5-17% T-ALL showed HLA-DR expression on the blasts' cell surface and lack of HLA class II in T-cell malignancy is due to loss of CIITA expression [1, 16-18]. IK over-expression is also documented in cutaneous T cell lymphoma [19].

For understanding gene regulation, conventionally experimental biologists perturb the system either over-expressing the gene of interest (GI) (say, TF) within a cell line deficient to that gene or silencing the GI followed by estimation of the effect on the downstream target gene. For human cancer cases, information regarding inherited mutation of HLA is not known and aberrant expression or binding of TF to Enh A is already reported. Therefore, it would be interesting to find out the possibility (role) of switching over/synergistic activation of other Enh region. Moreover, in such understanding, particularly in human disease cases, that sort of straight-forward experimental approaches may not be suitable. Therefore it has been suggested that Enh B has a tissue specific function and plays a significant role in pathogenic transformation [20]. In view of this, a recent attempt has been made to find out the relative importance of different TFs in HLA class I regulation in leukemia by using non-parametric statistical analysis [21].

Moreover, with the recent perspectives of Systems Biology, development of powerful analytical tool is more desirable in understanding of gene regulation of a system without any artificial forceful perturbation. In recent times, information theoretic tools (like entropy analysis, mutual information) have been used in the development of sets of over- and/or under-expressed genes (clustering) from microarray profile of different gene expression [22, 23]. Very recently, mutual information has been utilized to reveal the genetic and epigenetic regulation of target genes by TFs [24]. However in the earlier approaches data are obtained from cell line based experimentation. Contrary to those earlier approaches, we have utilized the information theory to reveal the relative importance of different TFs involved with the HLA expression under the condition of human hematological malignancies. Hence we have chosen the gene expression profile of diagnosed *de novo* leukemia cases.

2 Conceptual Frameworks

The classical concept involves a source of information that emanates certain symbols according to a probability distribution. These symbols pass through a channel and are received at the other end. The received symbol probabilities are different from the source to an extent depending upon the distortion properties of the channel.

This concept can be extended to cover data points of a TF attribute which acts as the source and an attribute surface expression (SE) as the receiver with the phenomenon of gene regulation as the underlying channel. Using measures of average uncertainty for the source, receiver and the channel, one can throw light on the phenomenon of gene regulation. Information entropy function provides us with this important metric. Information entropies computed for the source, receiver and computation of channel equivocation and mutual information could be useful to characterize the phenomenon of gene regulation. Below we provide detailed theoretical background on the concepts used in this work.

Information Entropy Background. Given the n data points pertaining to a variable, the range is sub-divided into q intervals and if f_i is the number of data points occurring in the i^{th} interval, then $p_i = f_i / n$ defines a probability distribution for the variable over the chosen q intervals. Entropy (H) = $\sum_{i=1}^{q} p_i \times \log (1/p_i)$ gives a measure of surprise associated with this probability distribution of the variable. In general for r-based logarithm,

$$I(E) = -\log_r (1/p_E) \text{ r-ary units.}$$

In natural logarithms (base e) the units are nats. In our calculation we have calculated all values to 10 based logarithm. That means here $r = 10$.

Some of the properties of entropy function is listed here that would be useful for the present work [25-26].

i) H is symmetric and continuous. This ensures that any choice of sub-interval changes can bring out the required uncertainty measure.

ii) $H_{n+1} (p_1, p_2, p_3,\ldots, p_{n-1}, 0) = H_n (p_1, p_2, p_3,\ldots p_n)$ i.e., if an interval is empty, it does not affect entropy. This means extending the range to some global (max, min) does not affect the sample data point based calculation. Due to this property, all the different groups (normal, disease) can be governed by the same sub-interval choice without affecting the desired metric.

iii) $H_n (p_1, p_2, p_3,\ldots p_n) \leq H_n (1/n, 1/n, 1/n,\ldots 1/n)$. This means that if the data is uniformly distributed the entropy will be maximum while the same falls down when the data is clustered more in a certain interval. This allows an upper bound on the chosen metric and thereby facilitates comparison.

Joint entropy is the amount of average information provided by the two attributes jointly. The joint entropy approaches the summation of the individual entropies when the two taken attributes are independent. This allows any arbitrary sub-ranging of the two-dimensional array involving TF and SE pair while finding the metric.

$$H(X, Y) = \sum_{X} \sum_{Y} p(X, Y) \times \log \frac{1}{p(X,Y)}$$ where X and Y are two random variables.

The concept of information entropy has been widely applied in various fields [27]. In bioinformatics too, for characterization of gene sequence, this concept is used [28-29].

3 Materials and Methods

3.1 Collection of Data

All gene expression data has been collected from the Ph.D. thesis of Jadavpur University, India, 2006 [21]. Primarily we have data of two attributes – HLA surface expression (HLA-ABC and HLA-DR) and transcriptional data (β2M, IRF-1, RFX5, RFXB, CIITA, CREB-1) of 10 normal volunteers (NV) and different leukemic patients [18 AML (acute myelogenous leukemia), 14 ALL (acute lymphocytic leukemia), 12 CML (chronic myelogenous leukemia) and 6 CLL (chronic lymphocytic leukemia)]. The demographic description of the patients is same as mentioned in the early work [3]. The TFs gene expression data and SE data were acquired through semi-quantitative reverse transcription polymerase chain reaction (RT-PCR) and by flow cytometric method respectively. The characteristics of the collected data are shown in Table 1 and 2 [21].

Table 1. Cell surface HLA-ABC, β2-microglobulin and HLA-DR expression. Data are presented as mean ± SD; Mdn, Max and Min stands for median, maximum and minimum value obtained in the population.

Sample	HLA-ABC	β2-microglobulin	HLA-DR
NV	57.23±21.97 Mdn 48.6 Max 107.37 Min 38.53	83.705±26.91 Mdn 75.84 Max 145.32 Min 60.26	36.793±13.78 Mdn 32.48 Max 58.33 Min 21.03
AML	29.035 ± 17.325 Mdn 28.02 Max 59.81 Min 1.12	36.519±22.45 Mdn 38.01 Max 86.97 Min 2.66	51.508±46.29 Mdn 42.96 Max 165.01 Min 1.19
ALL and CLL	32.721 ± 23.44 Mdn 25.19 Max 81.42 Min 9.2	54.083±36.05 Mdn 44.56 Max 145.1 Min 11.24	195.909±192.43 Mdn 106.82 Max 626.36 Min 32.91

* Note: In CML cases, identification of malignant cell diagnosis is not possible through flow cytometry, hence investigation on HLA surface expression is not done. Statistical test of significance is available in Ref. 3.

Table 2. Transcriptional expression of different TFs in leukemic and normal individuals. Data are presented as mean ± SD.

	IRF1	RFX5	RFXB	CIITA	IK	CREB1
NV	1.049 ±0.632	1.102 ±0.376	0.711 ±0.392	0.412 ±0.353	2.14 ±1.429	0.0
AML	1.37 ±1.451 NS	0.984 ±0.597 NS	1.83 ±0.588 $P<0.005$	0.801 ±0.742 NS	1.986 ±1.98 NS	0.801 ±0.76 $P<0.001$
ALL	0.831 ±0.978 NS	1.181 ±0.5 NS	1.84 ±0.905 $P\leq0.02$	0.735 ±0.486 $P\leq0.02$	1.551 ±1.453 NS	0.533 ±0.286 $P<0.001$
CML	0.842 ±1.419 $P<0.02$	1.15 ±0.66 NS	1.528 ±1.1 $P<0.05$	0.717 ±0.668 NS	1.134 ±1.504 $P<0.02$	0.228 ±0.218 $P<0.001$
CLL	1.83 ±2.16 NS	1.234 ±0.411 NS	2.253 ±1.403 $P\leq0.05$	0.795 ±0.491 NS	3.032 ±2.617 NS	0.155 ±0.158 $P<0.001$

NS: Not statistically significant; P means the level of statistical significance through Mann-Whitney U test.

3.2 Proposed Scheme

Five attributes pertaining to TF and two attributes of HLA surface expression (SE) have been considered (Fig. 1A). For each pair (altogether $5\times2 = 10$ pairs) of TF and SE, we consider the existence of an informational channel through which the concerned TF manifest into the corresponding SE. Our aim is to examine these channels. For the chosen attributes we have collected data of individuals from normal population and some individuals with different leukemic conditions. We separately examine the channels in different disease groups and compare them with normal group. This gives an insight into the phenomenon through which a TF regulates the SE. The results are not in absolute terms but based on the comparative analysis.

The TF and SE data collected is divided into a certain number of intervals. The number of intervals is analogous to the number of symbols in classical information theory. Calculation of the frequency distribution from data is analogous to the symbol probabilities at the source and the receiver side respectively.

Now we have considered the joint probabilities of the symbols of the source (TF) and receiver (SE). In this way we convert the TF-SE pair into a source-receiver pair. Now we calculate the information entropies at the source $H(X)$ and receiver $H(Y)$. We also calculate the joint entropy of source and receiver pair $H(X,Y)$ by considering the joint probabilities. These provide the measure of uncertainty and from these we derive the channel equivocation $H(X|Y)$ or $H(Y|X)$ i.e., the average conditional entropy of the source given the receiver symbol or vice versa.

$$H(Y|X) = H(X, Y) - H(X) \text{ and}$$
$$H(X|Y) = H(X, Y) - H(Y).$$

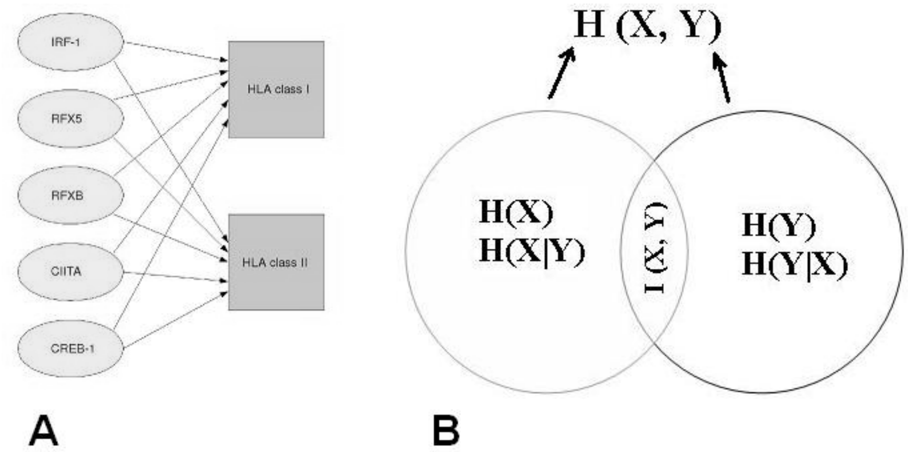


Fig. 1. Analogy between transcriptional regulation and information channel (A) and Venn Diagram showing relation between different entropies (B)

Mutual information I(X;Y) can now be calculated as:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

The channel equivocation is an important metric that provides the information about the nature of the channel, i.e., how the channel contributes to the uncertainty propagation from source to receiver. In analogy, the metric chosen by us could provide the uncertainty with which the TFs' express themselves into the SE. In other words, it gives an idea about the contribution of the channel in the propagation of uncertainty of TF into the uncertainty of SE. Venn diagram (Fig. 1B) shows how the different entropies are related to one another. Results indicate how differently the channel behaves from normal to disease cases. Also, the relative importance of propagation of a TF to SE would be manifested.

3.3 Grading of Independence

We sum the individual entropies of attributes already computed and we compare them with their joint entropies. If sum comes close to joint entropy values we can say that the two considered attributes are independent.

So after finding the joint entropy we have given the grading to them that denotes the degree of independence between those two attributes. Say for a joint distribution of X, Y we obtain $H_{X,Y} = P$ units of entropy and Q be the individual sum of entropies, then, $\dfrac{Q-P}{Q} \times 100$ is taken as a measure to find the grade which is expressed as a range of percentage. Thus grading system indicates that the lesser the percentage or grade, the two attributes are more independent to each other; whereas higher is the grade, higher is the dependency.

4 Results

4.1 Analysis of Single Attribute

Table 3 indicates the extent of deterministic behavior of different attributes in different populations. In Table 3, if any attribute value is closer to the maximum value, the more is its information entropy or uncertainty. As discussed in entropy function properties, this also means that uncertainty will maximize if all the intervals are equally likely i.e., data points are scattered equally over the entire range. Here deterministic behavior implies that most of the attribute values fall within a few sub-ranges while randomized behavior means that the attribute values are scattered over the entire range.

It is observed that in acute leukemia (AML, ALL) the HLA-ABC expression is more randomized, whereas in CLL it is more deterministic compared to normal. Interestingly HLA-DR expression is fully deterministic for NV. The behavior is randomized in disease cases with ALL being most prominent.

β2M expression is more randomized in diseased samples compared to NV. IRF1 expression in CML sample is more deterministic, however, in other disease samples the values are randomized compared to normal population. Table 3 also indicates that RFX5 expression in all the diseased population is more randomized, most strikingly in CML cases. Similarly, CIITA is randomized in all the disease samples, most strikingly in ALL cases.

Table 3. Entropy of different attributes

	HLA-ABC	HLA-DR	β2M	IRF1	RFX5	CIITA	IK
NV	0.6988	0.0000	0.4301	0.5301	0.4729	0.4582	0.5536
AML	0.7270	0.2725	0.7346	0.6552	0.7348	0.5856	0.5022
ALL	0.7465	0.6373	0.6382	0.6671	0.6242	0.7268	0.5353
CML	ND	ND	0.7400	0.4769	0.7903	0.4269	0.4167
Total	0.8930	0.4732	0.8327	0.7338	0.7338	0.8261	0.6318
Maximum	1.0414	0.9542	0.3030	1.0414	0.8451	0.8451	0.8451

ND: not done due to inadequate data. Data from CLL group has not been considered due to inadequate sample size.

4.2 Analysis of Combinations of Different Leukemias

Next we have the conjunction of different disease conditions and performed entropy analysis to find out the behavior of individual parameters in different states and types of leukemia. This has been compared with the behavior of normal population (Table 4). From Table 4 we observe that HLA-ABC in chronic leukemia cases (CML + CLL) and HLA-DR in myeloid leukemia (AML +CML) behave more deterministically.

Again, β2M is more scattered in different disease combination. However, CIITA, IK and IRF1 do not show any significant difference in behavior from NV.

4.3 Joint Entropy Analysis

Joint entropy analysis provides the dependency between two attributes and a comparison between disease and normal reflects the alteration in transcriptional efficiency. The joint entropy of different combinations and its comparison with the summation of their individual entropies have been tabulated and a grade has been provided as per the grading rule mentioned in the Methods section. Example cases are tabulated in Table 5. The table shows that in normal samples, CIITA is more potent in induction of HLA-DR (more dependency) compared to HLA-ABC. Generally, in disease cases, HLA is independent of CIITA with some minor dependency in case of lymphoid leukemia. The detailed results for all TFs can be derived from mutual information analysis.

Table 4. Entropy of each attribute of different leukemia combinations

Type of combination	Malignancy	HLA-ABC	HLA-DR	β2M	IRF1	CIITA	IK
Myeloid combination	AML+CML	0.7270	0.2725	0.7429	0.6662	0.5855	0.5067
Lymphoid combination	ALL + CLL	0.7972	0.6926	0.8263	0.7290	0.7108	0.5067
Acute combination	AML+ ALL	0.7768	0.5164	0.7392	0.6761	0.7269	0.5506
Chronic combination	CML+ CLL	0.4771	0.4771	0.8412	0.6650	0.5058	0.6405
All cases of Leukemia	Total – NV	0.8189	0.8146	0.7999	0.7192	0.6820	0.6072
	Total	0.8930	0.4332	0.8327	0.7338	0.8261	0.6318
	Maximums	1.0414	0.9542	0.9030	1.0414	0.8451	0.8451
Normal volunteers	NV	0.6988	0.9031	0.4301	0.5301	0.4582	0.5536

Table 5. HLA-ABC vs. CIITA and HLA-DR vs. CIITA. DS: total disease samples

	HLA-ABC vs. CIITA			HLA-DR vs. CIITA		
Type	Joint entropy	Sum of entropies	Grading*	Joint entropy	Sum of entropies	Grading*
NV	0.6796	1.1570	C	0.4582	0.4582	A
AML	0.9451	1.3126	B	0.6543	0.8581	B
ALL	1.0414	1.4733	B	0.9319	1.3641	C
Total (DS)	1.3149	1.7191	B	1.5133	1.2933	B
Maximum	1.8865			1.7993		

*A: 0-15%, B: 15-30%, C: 30-45%, D: above 45%.

4.4 Mutual Information Analysis

Quantitative mutual information analysis may reveal the relative importance of an attribute (TF) on the regulation of SE (HLA-ABC or HLA-DR) in normal and leukemic state (Table 6). If mutual information decreases, it indicates that the association becomes more independent and the channel (gene regulation) distorts the passage of information from TF to SE.

From the analysis it is revealed that mutual information for RFXB is high in general. This indicates that both HLA class I and II (HLA-DR) are dependent on this TF with an indication that dependency of HLA class I is more than HLA-DR. This dependency becomes more pronounced in leukemic condition. Though under normal condition CIITA dependency of HLA-DR is more but under the condition of malignancy this dependency decreases for AML but increased in ALL cases. However, HLA-ABC is less dependent on RFX5 and CIITA in normal and myeloid leukemic cases. For lymphoid leukemia dependency is almost unaltered.

Results imply that in induction of HLA-ABC, CREB1 has no role both in normal and leukemia, however, in lymphoid leukemia it plays a role. Similarly, in HLA-DR expression CREB1 has no role in normal and leukemia in general excepting lymphoid leukemia. The overall observation is that different TF plays different role in different type of leukemia (cell). IRF-1 is excluded from this study as this TF becomes functional through Enh A region and/or influence the Enh B region through CIITA. Similarly, IK is a cytokine, becomes active through its inverse effect on CIITA; hence is also excluded for this analysis.

Table 6. Mutual information analysis

HLA gene	Type	$I(X;Y) = H(X) - H(X \mid Y)$			
		CIITA	RFX5	RFXB	CREB1
HLA-ABC	NV	0.4774	0.4751	0.6988	0.4304
	AML	0.3675	0.3857	0.7270	0.2289
	ALL	0.4319	0.3841	0.7465	0.3965
HLA-DR	NV	0.6392	0.6392	0.5941	0
	AML	0.2038	0.112	0.5401	0.1331
	ALL	0.4322	0.405	0.6373	0.3728

5 Discussion

Conventionally in biology, classification and distinction between disease and normal is made from gene expression data obtained by microarray method followed by analysis through artificial intelligence (AI). However, criticism to such approach is that microarray only provides a range and AI is not truly a mechanistic way of analysis in understanding the biological mechanism [30]. Moreover, as for few genes particularly for HLA binding TFs together with HLA expression, no microarray chip data is available. So we depend on the PCR (Polymerase Chain Reaction) based gene expression data.

The available biological information is that disruption of the constitutive region binding transcription factor (TF) downregulates the HLA surface expression and information from in vitro experimental data suggest that each of the inducible region binding transcription factor (which are become activated only in emergency situation like infection) is enough to transcribe the gene expression, so multiple regression may not the ideal for analyzing the data of the present situation.

From our analysis it is revealed that channel behaves differently from normal to disease cases. Also, the relative importance of propagation of a TF to SE is also manifested. This indicates that the immune escape mechanism by the reduction of HLA class I expression is due to unavailability of both the constitutive [10] and inducible region binding TFs. The approach of induction of HLA class I by inducible binding TFs [5-6] requires both CIITA and CREB1 factors in the myeloid leukemia cases whereas CREB1 in single could be sufficient for lymphoid leukemia cases. This finding tallies with the findings in [21].

The parametric variation in population (signal) is a major concern of any biological investigation. However, the measurement uncertainty (noise) is another important aspect. So a high signal to noise ratio bears some significance. To reduce the noise and for quantitative accuracy, presently, experimental molecular biologists rely mainly on the data collected by Real-time PCR based method. If the measurement variability is high then every relation would appear as independent, however, in this exercise, relationship between some attributes become dependent on each other. This signifies that the measurement noise is less.

In this connection it is necessary to mention that the coefficient of variation of the PCR reaction has been determined for the source data (and mentioned in Methodology section), together with the relative abundance of each of the gene expression with half log dilution and where possible PCR reaction was performed in triplicate and corroborate the semi-quantitative findings (RNA level data) with the flow cytometric (protein level data) for all the genes. Moreover in the source data, semi-quantitaive findings of different loci of HLA genes have been validated with the Real time PCR [3, 21].

It is to be noted here that a lot of criticisms are also available for Real Time PCR methodology, namely, threshold settings, amount of input RNA and size of the target sequence in the PCR reaction etc. [31]. The only advantage of real time PCR based method over conventional PCR is that it can identify the differences in addition to PCR master mix in the PCR reaction (experimental error). It is worthwhile to mention here that both PCR based methods actually quantify the relative abundance of mRNA with respect to a house-keeping gene (RQ) in a sample.

So we expect that the improved analytical tools could able to validate the low cost semi-quantitative PCR methodology for understanding the gene expression regulation in terms of them. Such data variability between samples and uncertainty associated with the measurement method could be tested by different conventional analytical tool rather than looking for instrumental sophistication. Thus in conjunction with the conventional analytical tools for association analysis (like nonparametric statistical test – χ^2 test as in [21]), information theory may help in comprehensively narrowing down the experimental dimensionality (without any artificial perturbation of the system) as well as experimental cost.

Previously several biological conclusions have been drawn from the gene expression data with the aid of information theory. However, to the best of our knowledge, none of them utilized the concept of channel equivocation in understanding of complex gene regulation of HLA from human disease data. It is worthwhile to mention here that HLA gene regulation is complex in the sense that it has an inducible promoter region and tissue as well as pathogenic diversification. So, with this work we hope that information theory would be utilized in revealing that context specific regulation. Moreover, this analytical tool could be accepted by the experimental biological community in drawing the conclusion in understanding of gene regulatory mechanism in terms of transcriptional efficiency from the population based gene expression data specially from human disease cases.

Acknowledgement. Author acknowledges West Bengal State University for the help to attend the ACITY 2012 and Ravi Babu Arangi, M.Tech student (2008) of Computer Science & Technology at Bengal Engineering & Science University for his technical help during the initial phase of the work.

References

[1] Wetzler, M., McElwain, B.K., Stewart, C.C., Blumenson, L., Mortazavi, A., Ford, L.A., Slack, J.L., Barcos, M., Ferrone, S., Baer, M.R.: HLA-DR antigen-negative acute myeloid leukemia. Leukemia 17, 707–715 (2003)

[2] Demanet, C., Mulder, A., Deneys, V., Worsham, M.J., Maes, P., Claas, F.H., Ferrone, S.: Down-regulation of HLA-A and HLA-Bw6, but not HLA-Bw4, allospecificities in leukemic cells: an escape mechanism from CTL and NK attack? Blood 103, 3122–3130 (2004)

[3] Majumder, D., Bandyopadhyay, D., Chandra, S., Mukhopadhayay, A., Mukherjee, N., Bandyopadhyay, S.K., Banerjee, S.: Analysis of HLA class Ia transcripts in human leukaemias. Immunogenet 57, 579–589 (2005)

[4] van den Elsen, P.J., Holling, T.M., Kuipers, H.F., van der Stoep, N.: Transcriptional regulation of antigen presentation. Curr. Opin. Immunol. 16, 67–75 (2004)

[5] Martin, B.K., Chin, K.C., Olsen, J.C., Skinner, C.A., Dey, A., Ozato, K., Ting, J.P.: Induction of MHC class I expression by the MHC class II transactivator CIITA. Immunity 6, 591–600 (1997)

[6] Gobin, S.J.P., Peijnenburg, A., Keijsers, V., van den Elsen, P.J.: Site alpha is crucial for two routes of IFN gamma-induced MHC class I transactivation: the ISRE-mediated route and a novel pathway involving CIITA. Immunity 6, 601–611 (1997)

[7] Kushida, M.M., Dey, A., Zhang, X.L., Campbell, J., Heeney, M., Carlyle, J., Ganguly, S., Ozato, K., Vasavada, H., Chamberlain, J.W.: A 150-base pair 5’ region of the MHC class I HLA-B7 gene is sufficient to direct tissue-specific expression and locus control region activity: the alpha site determines efficient expression and in vivo occupancy at multiple cis-active sites throughout this region. J. Immunol. 159, 4913–4929 (1997)

[8] Muhlethaler-Mottet, A., Otten, L.A., Steimle, V., Mach, B.: Expression of MHC class II molecules in different cellular and functional compartments is controlled by differential usage of multiple promoters of the transactivator CIITA. EMBO J. 16, 2851–2860 (1997)

[9] Piskurich, J.F., Linhoff, M.W., Wang, Y., Ting, J.P.: Two distinct gamma interferon-inducible promoters of the major histocompatibility complex class II transactivator gene are differentially regulated by STAT1, interferon regulatory factor 1, and transforming growth factor beta. Mol. Cell Biol. 19, 431–440 (1999)

[10] Girdlestone, J.: Transcriptional regulation of MHC class I genes. Eur. J. Immunogenet 23, 395–413 (1996)

[11] Cabannes, E., Khan, G., Aollet, F., Jarrett, R.F., Hay, R.T.: Mutations in the IkBa gene in Hodgkin's disease suggest a tumour suppressor role for IkappaBalpha. Oncogene 18, 3063–3070 (1999)

[12] Mori, N., Fujii, M., Ikeda, S., Yamada, Y., Tomonaga, M., Ballard, D.W., Yamamoto, N.: Constitutive activation of NF-kappaB in primary adult T-cell leukemia cells. Blood 93, 2360–2368 (1999)

[13] Hochhaus, A., Yan, X.H., Willer, A., Hehlmann, R., Gordon, M.Y., Goldman, J.M., Melo, J.V.: Expression of interferon regulatory factor (IRF) genes and response to interferon-alpha in chronic myeloid leukaemia. Leukemia 11, 933–939 (1997)

[14] Schmidt, M., Nagel, S., Proba, J., Thiede, C., Ritter, M., Waring, J.F., Rosenbauer, F., Huhn, D., Wittig, B., Horak, I., Neubauer, A.: Lack of interferon consensus sequence binding protein (ICSBP) transcripts in human myeloid leukemias. Blood 91, 22–29 (1998)

[15] Nekrep, N., Fontes, J.D., Geyer, M., Peterlin, B.M.: When the lymphocyte loses its clothes. Immunity 18, 453–457 (2003)

[16] Miller, K.B., Daoust, P.R.: Clinical manifestions of acute myeloid leukemias. In: Hoffman, R., Benz Jr., E.J., Shantil, S.J., Furie, B., Cohen, H.J., Silberstein, L.E., McGlave, P. (eds.) Hematology: Basic Principles and Practice, 3rd edn., Churchill Livingstone, New York, London, Philadelphia, pp. 999–1024 (2000)

[17] Hoelzer, D.: Acute lymphocytic leukemia in adults. In: Hoffman, R., Benz Jr., E.J., Shantil, S.J., Furie, B., Cohen, H.J., Silberstein, L.E., McGlave, P. (eds.) Hematology: Basic Principles and Practice, 3rd edn., Churchill Livingstone, New York, London, Philadelphia, pp. 1089–1105 (2000)

[18] van den Elsen, P.J., Holling, T.M., van der Stoep, N., Boss, J.M.: DNA methylation and expression of major histocompatibility complex class I and class II transactivator genes in human developmental tumor cells and in T cell malignancies. Clin. Immunol. 109, 46–52 (2003)

[19] Willers, J., Haffner, A., Zepter, K., Storz, M., Urosevic, M., Burg, G., Dummer, R.: The interferon inhibiting cytokine IK is overexpressed in cutaneous T cell lymphoma derived tumor cells that fail to upregulate major histocompatibility complex class II upon interferon-gamma stimulation. J. Invest. Dermatol. 116, 874–879 (2001)

[20] Girdlestone, J.: Regulation of HLA class I loci by CIITA. Blood 97, 1520 (2001)

[21] Majumder, D.: Transcriptional regulation of immune recognition in hematological malignancies. Ph. D. Thesis, Jadavpur University, India (2006)

[22] Margolin, A.M., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D., Califano, A.: ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinfor. 7(suppl. 1), S7 (2006), doi:10.1186/1471-2105-7-S1-S7

[23] Priness, I., Maimon, O., Ben-Gal, I.: Evaluation of gene-expression clustering via mutual information distance measure. BMC Bioinfor. 8, 111 (2007), doi:10.1186/1471-2105-8-111

[24] Wang, K., Saito, M., Bisikirska, B.C., Alvarez, M.J., Lim, W.K., Rajbhandari, P.R., Shen, Q., Nemenman, I., Basso, K., Margolin, A.A., Klein, U., Dalla-Favera, R., Califano, A.: Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nature Biotech. 27(9), 830–837 (2009)
[25] Abramson, N.: Information theory and coding. McGraw-Hill, New York (1963)
[26] Hamming, R.W.: Coding and Information theory. Prentice Hall Inc., Englewood Cliffs (1980)
[27] Cover, T.M., Thomas, J.A.: Elements of Information theory. John Wiley & Sons, Inc., New Delhi (1999)
[28] Troyanskaya, O.G., Arbell, O., Koren, Y., Landau, G.M., Bolshoy, A.: Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. Bioinfor. 18, 679–688 (2002)
[29] Adami, C.: Sequence complexity in Darwinian evolution. Complexity 8, 49–56 (2002)
[30] Weston, A.D., Hood, L.: Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J. Proteome Res. 3, 179–196 (2004)
[31] Bustin, S.A., Nolan, T.: Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. J. Biomol. Techq. 15, 155–166 (2004)

Improving Security in Digital Images through Watermarking Using Enhanced Histogram Modification

Vidya Hari[1] and A. Neela Madheswari[2]

Assistant Professor, Department of Information Technology
Associate Professor & Head, Department of Information Technology.
KMEA Engineering College, Aluva
`{vidya.brijesh,neela.madheswari}@gmail.com`

Abstract. Image transmission plays an important role in recent studies of engineering and scientific research fields. While transmitting the image, they have to be secured. Many approaches are available for secure transmission of images. This method focuses towards the use of invisible watermarking for encryption purpose. The attacker cannot able to find the difference when watermarking is used. There are two algorithms used for making the watermarking object. The performance evaluation is done by introducing various attacks to the watermarked object and it is done using Matlab 7.11.

Keywords: Digital image, watermarking, histogram, enhanced histogram modification for watermarking.

1 Introduction

The evolution of World Wide Web and Internet plays a great role in data transmission. The data can be in any form i.e. text, image or object. Many advanced researches such as medical diagnostics etc are in need of transferring the image through networks or Internet for fast diagnostics. Many fields in nowadays are focusing towards digital image transmission for various purposes. The general methods used for secure transmission of image are cryptography and information hiding. The information hiding can be broadly classified as steganography and watermarking. Although cryptographic methods have long been applied in digital content security, the decrypted content requires further protection. This work focuses towards the use of invisible watermarking for secure image transmission. Digital watermarks can provide extra protection to the decrypted content since it is embedded into the content. Some of the main applications of watermarking are owner identification, copy protection, broadcast monitoring, and medical applications and data authentication.

2 Motivation

With digital multimedia distribution over World Wide Web, IPR (Intellectual Property Rights) are more threatened than ever due to the possibility of unlimited copying [1-4]. Stephen wolthusen [5] has attempted to present the challenges faced by digital

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 175–180.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

watermarking techniques when entering the application domains envisioned for it. Frank enhances the hierarchical watermarking security by using the genetic algorithm to embed watermarks into the frequency domain of a host image. The algorithm not only detects vector quantization attacks, but also provides a fundamental platform for other fragile watermarking techniques [6].

Hsu and Wu proposed an image authentication technique by embedding digital watermarks into images. They embed the watermarks with visually recognizable patterns into the images by selectively modifying the middle frequency parts of the image [7].

Tang and Hang [8] introduced a robust digital image watermarking scheme that combines image feature extraction and image normalization. The goal is to resist both geometric distortion and signal processing attacks. A feature extraction method called Mexican Hat wavelet scale interaction is used. The extracted feature points can survive a variety of attacks and be used as reference points for both watermark embedding and detection.

Invisible watermarking techniques with their limitations, attacks and implications are described [9]. This work helps to prove the importance of invisible watermarking and the problems faced by visible watermarking.

3 Digital Watermarking Process

The section 3.1 gives the short description of the generic watermarking process. the section 3.2 describes the attacks on watermarks and the section 3.3 describes the system model considered for this work and the various kinds of attacks introduced in the system for performance evaluation.

3.1 Generic Watermarking Process

Digital image watermarking schemes can be modeled as a communication process involving an embedder and a detector. A watermark signal is embedded into a cover image to produce a stego image. No extra space is required to store the signal. The stego image is then transmitted to the consumer. Distortions due to unintentional modification, malicious attacks could occur during this process. Finally watermark detector is applied to determine whether the watermark exists in a possibly distorted image. The entire process is shown in fig. 1.

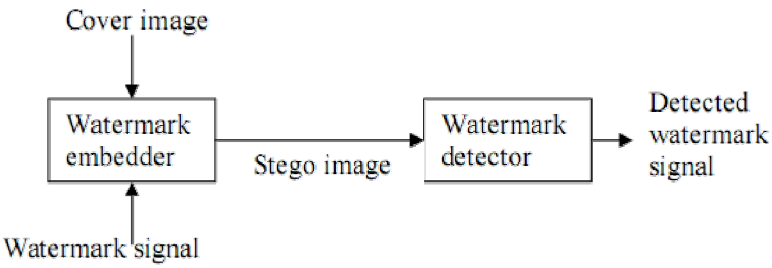


Fig. 1. A generic watermarking system

In this work, the real image is considered and histogram is applied and the final image is the watermarked image. The actual embedder makes use of the original image and the image obtained after applying histogram and then transmitted. While transmission some forms of distortions can be occurred.

3.2 Attacks on Watermarks

A watermarked image to be likely subjected to certain manipulations. Some of them may be unintentional such as usage of compression techniques or some of them may be intentional such as cropping, re-watermarking, etc. Some of the forms of attacks are:

a. Lossy compression: Many compression schemes like jpeg and mpeg can potentially degrade the data's quality through irretrievable loss of data.

b. Geometric distortions: They are specific to images, videos and include operations such as rotations, translation, scaling and cropping.

c. Common signal processing operations: Some of the common operations are D/A conversion, A/D conversion, re-sampling, filtering, color reduction, addition of constant offset values to the pixel values and local exchange of pixels.

d. Other intentional attacks: Some intentional attacks include printing and rescanning and re-watermarking.

3.3 System Model

For experimentation, a sample Lena image of 256 x 256 pixel of jpeg type is considered for original image. For the watermarked object, the same original image after performing enhanced histogram modification is considered. The entire process is carried out using the image simulator, MatLab 7.0 and the performance is evaluated by introducing various attacks in the original image. The attacks used in this work are Median filters, Cropping, Rotation and Compression.

4 Enhanced Histogram Modification for Watermarking

This algorithm is used for getting the watermarked image, which is useful for encryption process. The histogram modification is done in two steps. They are: i) using R and G components of image, which is given in Algorithm1, ii) using B component of the image, the steps to be involved in this algorithm are given in Algorithm2.

Algorithm 1

```
1. Decompose the original image into R, G and B components. The segment is called HRG histogram. From the R and G component a 2D Histogram of R and G component is made.

2. Generate a watermark w which is a pseudo random binary pattern of size L*L. $L=2^n$ $1<L<256$
```

3. Segment the histogram HRG into blocks of size 16x16.

4. Find the blocks which are having nonzero values > m, where m=28, to embed the watermark sequence.

5. The modification is done according to the conditions:

   a. If W(j,k)=0 then BHrg(j,k)=0

   b. If W(j,k)=1 then BHrg(j,k)=non zero value.

To Enforce BHrg(j,k) to be zero the nonzero bin value of BHrg(j,k) is distributed uniformly between its four neighbours. To make BHrg(j,k)to be non zero ,the non-zero neighbours bin value is transferred.

6. After the modification the watermarked components Rw and Gw are restored.

Algorithm 2

This algorithm is done for the second modification on the histogram based on B component and features extracted from the image.

1. To extract the image feature the Rw component is used.

2. JPEG compression is used which eliminates the high frequency components of the image.Rf(x,y)is calculated according to

$$Rf(x,y)=\sum \sum Rw(x+(!x))(y+(!y))/N$$

3. Divide Rf and B component into Pi regions of size 16*16.

4. For each Pi, a 2D histogram composed by Rf and B component made.

5. Each 2D histogram is partitioned into four blocks.

6. The watermark bit is embedded into the 2D histogram modifying the bin values of B components. Some bins are moved from one block to another modifying the bin values. If wmk(i)==1, then some bins in BB and BC are moved into BA and BD and if wmk(i)==0 then some pixels in BA and BD are moved in to BB and BC.

5 Detection Process

The watermark is extracted from the watermarked image using the detection algorithm. Algorithm3 is used for the detection process.

Algorithm 3

```
1.    Obtain the different watermarked components from the watermarked image.

2.    Obtain the watermark component with all zeros.

3.    Extract the watermark from the blocks having non-zero values greater than 128.Here 128 is set as a threshold value, to determine in how many blocks the watermark is being embedded.

4. Then we compare the original watermark with extracted one.
```

6 Experimental Results

In [10], the algorithm for histogram modification is specified. But the algorithm will first find a block which has some nonzero pixel values. Only that single block is considered and made the histogram updating process. But the algorithms specified in this work will find the number of blocks which has n number of nonzero pixel values. If m number of blocks is available with n nonzero pixel values, then m blocks are considered for histogram modification and finally obtained the watermarked image for the embedding process. The output is shown in fig. 2 and fig. 3. The original image is given in fig.2 and the watermarked image after embedding is shown in fig. 3. There is no much difference between fig.2 and fig.3. After introducing the attacks, the picture is withstanding its original quality.

Fig. 2. Original Image

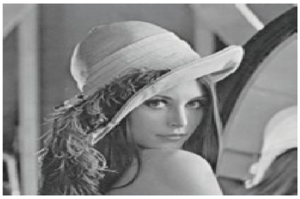


Fig. 3. Watermarked Image

The above enhanced watermarking algorithm based on histogram modification show better performance against most of the common geometric attacks like median filter, rotation, cropping and compression. The number of correct watermarked bits extracted from the watermarked image is taken as the parameter for measuring robustness and is given in the table 1.

Table 1. Robustness against attacks based on number of correct bits extracted

Attacks	Number of bits
Median filter	76%
Rotation	92%
Cropping	78%
Compression	72%

7 Conclusion

An enhanced watermarking algorithm based on histogram modification is proposed in this paper. It is based on two modifications in the two dimensional histogram. The modifications are done on different components of the two dimensional histogram. The 2D histogram is partitioned into different blocks and the watermark pattern is embedded into the different blocks based on a threshold value T. This threshold value is based on the number of non zero values in the different blocks. If the number of blocks satisfying the threshold value is more then watermark pattern is embedded into all those blocks which help to improve the robustness. The experimental result shows better performance against most of the common geometric attacks. Then performance comparison is done when watermark is embedded in only one block having maximum non zero values.

References

[1] Niu, X.W.J.W.P.: A new digital image watermarking algorithm resilient to Desynchronization attacks. IEEE Transactions on Information Forensics and Security (2007)

[2] O'Runuanaidh, J.J.K., Pun, T.: Rotation Scale and Translation invariant spread spectrum digital image watermarking. Spread Spectrum (1998)

[3] Dong, P., Brankov, J.G., Galatsanos, N.P., Yang, Y., Davoine, F.: Digital watermarking robust to geometric distortions. IEEE Transactions on Image Processing (2005)

[4] Kim, H.S., Lee, H.K.: Invariant image watermark using Zernike moments. IEEE Transactions on Circuits and Systems for Video Technology (2003)

[5] Wolthusen, S.: On the limitations of digital Watermark: A cautionary note (1998)

[6] Shih, F.Y., Wu, Y.T.: A novel fragile Watermarking technique. In: IEEE International Conference on Multimedia and Expo (2004)

[7] Hsu, C.T., Wu, J.L.: Hidden digital Watermarks in images. IEEE Transactions on Image Processing (1999)

[8] Tang, C.W., Hang, H.M.: A feature-based robust digital image watermarking scheme. IEEE Transactions on Signal Processing (2003)

[9] Craver, S., Memon, N., Yeo, B.L., Yeung, M.M.: Resolving rightful ownerships with invisible watermarking techniques: limitations, attacks and implications. IEEE Journal on Selected areas in Communications (1998)

[10] Hernandez, M.C., Miyatake, M.N., Meana, H.P.: Robust Watermarking based on histogram modification. In: IEEE International Conference on Multimedia and Expo (2009)

Survey on Co-operative P2P Information Exchange in Large P2P Networks

S. Nithya[1,*] and K. Palanivel[2]

[1] Department of Computer Science, School of Engineering,
Pondicherry University, Puducherry, India
`nithyasugumaar@gmail.com`
[2] Department of Computer Center, Pondicherry University, Puducherry, India
`Kpalani@yahoo.com`

Abstract. Peer to Peer Information Exchange (PIE) is a technique to improve the data availability using data present within the peers in the network. Network Coding (NC) which gives a set of combined data dissemination procedure to improves the transmission efficiency of the network. These two techniques are merged in such a way that the PIE can be performed in much improved way. But immense problem concerned with the technique is that the performance degrades gradually with increase in the number of peers in the network. Clustering approach is proved to be more efficient for solving the scalability issues in large networks. Thus in this paper, we have presented a detailed study on the techniques like PIE, NC and clustering from the view of their requirement, merits and demerits and performance improvement on their combine usage. In this paper, we have also provided a distant view of Co-operative P2P Information Exchange (cPIE) which incorporates clustering technique in the existing PIE with NC technique to eradicate its scalability and performance bottleneck issues.

Keywords: Peer to Peer Information Exchange, Network Coding, Clustering, Co-operative.

1 Introduction

The traditional wired communication uses the node to node communication for transmission of information. But the wireless network uses two distinct types of communication patterns named upstream communication and local communication. The upstream communication is the communication between the node and the base stations where as the local communication is the exchange of information among the peers. Information exchange is the exchange of information among the nodes is the

---

[*] No academic titles or descriptions of academic positions should be included in the addresses. Either this information should be omitted altogether (preferably), or it should be included in a footnote at the end of the first page. Information of this nature, given in the addresses, will be deleted by our typesetters.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 181–193.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

basic aspect of wireless networks because of which several operations like cluster head election and data aggregation can be made effective.

The traditional communication uses the flooding or forwarding approach for the information exchange. This exchange becomes an ineffective process for many applications. So in order to make it efficient, the network coding technique has been proposed to enhance the performance.

Network coding was first proposed by Ahlswede, et al. and it is a novel method for transmitting data, gaining lots of importance recently. When compared to traditional routing method, network coding offers many throughput benefits. And it is achieved by combining packets from different incoming data streams among the intermediate nodes instead of simply forwarding them. Network coding grants numerous advantageous not only for multicast flows, but also for other traffic patterns, such as information exchange. Earlier the nodes when receiving the information, sends the data in sequence in the normal approach. But when incorporating the network coding in the PIE, the node broadcasts the information instead of sending it in series. By this approach the number of transmissions is reduced. Therefore, the transmission energy cost and time consumption are also minimized.

The above mentioned techniques are applicable for smaller networks. When considering larger network. It is difficult to maintain the broadcasting mechanisms due to its larger size and more number of nodes. When the network size increases the performance of the network decreases proportionally to it. So in order to overcome this problem, the larger networks are grouped into smaller size called clusters. These clusters improve the performance by enhancing the throughput, fault tolerance and load balancing of large P2P networks.

This paper has been organized as follows, section 2 describes about the works published that concerned with the study made and section 3 and 4 provides a detailed study about the Peer to Peer Information Exchange (PIE and Network Coding (NC) techniques and their contribution to the P2P environment. Section 5 and 6 gives a description on technique to use PIE and NC merged and their importance and scalability issues concerned with it. Section 6 explains about the vital contribution of clustering technique on scalability issues and section 7 elucidates a brief method to achieve the PIE with NC using clustering approach. Finally section 8 concludes the paper with future work.

2 Related Work

P2P Information Exchange and Retrieval (PIER) [2] is a P2P query engine used for query processing in Internet scale distributed systems. PIER offers a method for possible scalable sharing and querying of finger print information, which is used in network monitoring applications including intrusion detection. PIER in its design uses four guiding principles. First, it grants relaxed consistency semantics - best effort results, as achieving ACID properties maybe difficult in Internet scale systems [3]. Second, it assumes organic scaling, meaning that there are no data centres/warehouses and machines can be added in typical P2P fashion. Third, the query engine imagines data is available in resident file systems and need not necessarily be loaded into local databases. The fourth principle is that instead of waiting for breakthroughs on semantic technologies for data integration, PIER tries to combine local and reporting mechanisms into a global monitoring facility. PIER is realized over CAN, the hypercube based P2P system [4].

For sharing information in networks of autonomous sources a number of different frameworks and systems have also been planned and studied [5, 6, 7, 8, 9]. In mobile peer-to-peer networks information exchange has been reviewed, for example, by Buchholz et al. [10], Kurhinen et al. [11] and Kurhinen & Vuori [12]. The process has been categorized by the number of links included in a data transfer process (single-hop, multiple-hop) and by the message delivery method (proactive, reactive) [11]. In the field of P2P file-sharing systems the majority of the research spotlights on improving efficient search, replication and security techniques. In addition, there subsist various significant research areas for information exchange systems such as resource management issues that include fairness and administrative ease [13].Caching improves information exchange performance in mobile peer-to-peer system [14].Since caching can able to save computing power and bandwidth as long as you have enough memory space. A P2P network can demonstrate a power-law topology [15] such that it can propagate queries rapidly and, if executed efficiently [16], it can locate objects in log n time, where n is the number of nodes in the network. On the other hand, there are remaining problems in the P2P information exchange theory which complicate its operation. Free-riding and the misfortune of the commons are two main problems. Distributed hash table (DHT) systems such as CAN (content-addressable network) [11] achieve furnish excellent scalability and deterministic guarantee, however these methods only suggest a simple interface for storing and retrieving (key, value) pairs. Directly applying them to information exchange would entail users to indicate accurate document IDs (keys) for retrieval, an impractical assumption in an environment where content is produced by millions of organizations and individuals, independently.

It has been noted that the problem observed in the COPE scheme [21] is different from the peer scheduling problem which highlights from the perspective of a single peer regarding how to opportunistically snooping neighbour states and cleverly XOR-ing these blocks, whereas the peer scheduling problem is how to increase the wireless coding gain from the point of view of all peers for cleverly scheduling the sending sequence of peers in a wireless network as a complete. The majority existing research concentrations are on block scheduling problems. Furthermore opportunistic snooping neighbour states, the COPE scheme eminently handles the block scheduling problem by cleverly XOR-ing packets.

In traditional networks according to [22] a peer scheduling problem is confirmed to be NP-hard; with network coding, due to the coded packets in wireless networks a peer scheduling problem becomes exacerbated and also it supposed as NP-hard. Several theoretical results followed, showing that optimal throughput achievement, which is NP-hard with routing, is possible with network coding [23]. In wireless ad hoc net-works for broadcasting with reference to network coding distributed probabilistic broadcast algorithms and deterministic broadcast algorithms have been proposed by Fragouli et al. [24,25] and Li et al. [26] , proportionately, resulting in a considerable energy saving.[27] Discusses the advantage of coding in lossy networks. When intra-flow network coding is used it gives the hopeful unicast throughput gains and also it has been shown in [28] via experiments. In the wired domain for multicast even though the majority outcomes on network coding have been given, for retrieving the advantages of network coding the broadcast nature of a lossy wireless medium turns out to be very helpful for unicast as well. Overall a solitary wireless transmission is frequently received by more than one node. Might overhear transmissions when nodes are located beyond than a one-hop distance and serve relay packets for preceding hops. Such opportunistic overhearing/listening has been expansively learnt in conjunction with inter-flow network coding2 in [29]. Over conventional routing significant unicast throughput profits for single-channel single-radio wireless mesh networks have been accounted [29] via wide-ranging experiments.

Multiple transmissions are permitted by multiple access techniques like FDMA [30], spatial reuse [30] and CDMA [31], at the same time. On the other hand these proposals signify to avoid intervention in frequencies, codes, or space for separating channel facility among multiple users. In variation the capacity of the network is enlarged by using ancient network coding. By utilizing space-time coding techniques Co-operative diversity [32], analog forwarding [33] and MIMO systems [30] grants multiple synchronized transmissions. A few of this work presumes antenna arrays and coherent combining at the receiver, which we do not guess. More vitally, these techniques vary from ancient network coding since they do not make use of the receiver's knowledge of one of the interfering signals to increase the capacity of the network.

Structure on network coding [34] across multiple generations of video packets has been examined, where one generation is identified at the transport layer despite application layer GOP structures. [35] The application of Markov decision process [36] to network coding has been discussed in which the network coding scheduling and optimization are centralized at the base station.

The type of linear NC is allowed by the nearly well-known topology called the butterfly topology [37] and in many situations it has been exposed that the butterfly can be globalized to sustain pair-wise linear NC [38] and these type topologies can be seen in a distributed manner [39]. Additionally to show the benefits of network coding compared with routing, the well-known butterfly network was proposed in papers [40] [41]. Also, relay networks that have multicast capability in the downstream, like wireless meshes or Passive Optical Networks (PONs), can be map to the butterfly topology [42]. Ho et al [43] proposed the random linear network coding and beside provided numerous upper bounds on the failure probabilities of random linear network coding. Balli, Yan, and Zhang [44] progressed on these bounds and reviewed the maximum behaviour of the failure probability as the field size goes to infinity. Koetter and M´edard [45] offered an algebraic characterization of network coding.

A different significant model for network coding using clustering is provided by the work in [46]. Latest work made an effort to mutually optimize video streaming and network coding. In order to combat internet bandwidth vacillation for both CDN and P2P networks [47] employed the hierarchical network coding design. Yeung [48] presents that network coding accomplishes the optimal delay performance in a time-synchronized model for any transmission schedules in P2P networks and in addition when compare and contrast to a original successive dissemination, a shorter broadcast delay of k blocks can be arrived by network coding in a complete graphs within a time-synchronized model demonstrated by Deb et al. [49]. Further[50] in arbitrary graphs the broadcast delay using network coding in arbitrary graphs has been investigated using network coding and illustrates its correlation with the spectral characteristics of the graph.

3 Peer to Peer Information Exchange

In most recent years, information exchanges in Peer to peer networks have become very familiar. Due to the hasty increase of decentralized and structured or unstructured peer-to-peer (P2P) networks, this handles immense potential for efficient information exchange in the Internet. A peer-to-peer, or “P2P,” information exchange means provide you access to a prosperity of information and allows the user to share computer files through the Internet. These exchanges are set up to permit users to search for and download files to their computers, and to facilitate users to make files available for others to download from their computers.

This Peer to peer information exchange generates a network of linked users since they are extremely decentralized one. In order to discover the needed file this permits a user to search through the files of all of the linked computers. Hence to use one of these services, a user must download the appropriate software from the Internet and install and configure it. The main intention of information exchange is to recognize information share among systems. Numerous different frameworks for sharing information between sovereign stores have been formulated and investigated in depth. Information exchanges one of the conceptually simpler, yet technically challenging, such frameworks [1]. In an information exchange background, data from a source schema are transformed to data over a target schema according to specifications given by source-to-target constraints. This framework models a situation in which the target passively receives data from the source, as long as the source-to-target constraints are satisfied.

The major benefit of PIE is there is no central exchange. PIE mainly support applications like which present file sharing and content exchange like music, movies, etc. The idea of PIE has also been successfully utilized for distributing computing and Internet-based telephony. The most important advantage of P2P information exchange is that these systems got significant efficiency gains which are completely scalable so there are no bounds on the membership and the network capacity hence every extra peer gets further capability to the system so ease of expansion and set up. A P2P environment can grow and use all the existing computers connected with a peering portal. A peer can act together as a client and a server so the network always functions as long as there are peers connected to the network. Instead of building complex and ex -pensive networking infrastructures, information systems can be integrated with a P2P program, or peering portal. The participating peers mark at least part of their resources as ‘exchanged’, allowing other contributing peers to access these resources. Thus, if peer 1 publishes something and peer 2 downloads it, then when peer 3 asks for the similar information, it can access it from either peer 1 or peer 2. As a result, as new users access a particular file, the system’s capability to supply that file increases [17].

4 Network Coding

Network coding is a new and elegant transmission paradigm that proved its strength in optimizing the usage of network resources introduced at the turn of the millennium to develop network performance. The emergence of network coding has brought about a metamorphosis in thinking about network communication, with its easy but important principle that in communication networks, we can permit nodes to not only forward but also process the incoming autonomous information flows. In common network coding is achieved by encoding and decoding several packets either from the same client or from dissimilar users. The former is called intra-session network coding [18, 19]. Network coding has been extensively appeared as a prospective approach to the operation of communication networks, particularly wireless networks [18]. Sanders et al [20] expressed illustrations where the space between the throughput using network coding to that without using coding was Ω (log n), where n is the number of receivers.

The NC scheme refines the accurate flow of data in a network by transmitting combined digital messages from source to recipient. Aforementioned to network coding, within a network the only work of intermediate nodes (i.e., routers and switches) has to forward data packets towards the destinations. NC principles support that, in addition to forwarding packets, intelligent mixing of packets (from different sources before forwarding) increases the network throughput. Hence, in network coding, instead of using routers and switches by replacing coders it allow encoding the incoming messages by the intermediate nodes and then forward these messages to other nodes. Number of bottlenecks has been reduced by increase in effective capacity of networks by incorporating this encoding and forwarding methods.

Network coding technique presents numerous advantages over the traditional store-and forward routing approach such as an enhancement in reliability and robustness, increase in throughput [51], energy efficiency [51] and an improvement of delay minimization [51].Even though the benefits of network coding in local area networks have been examined, its application in wider area networks such as WiMAX [52], LTE, and LTE-advanced is mostly unfamiliar.

Applying network coding to robust video transmission which takes place in the circumstance of wireless networks is a one more big challenge. In real-time video conferencing Video quality, communication bandwidth, and stringent delay necessities all concealment dreadful challenges through error-responsive wireless networks which endure from dynamic channel variations and intervention in a shared medium. These issues are individually addressed by applying NC elimination protection over the uplink, downlink, and overhearing channels [53, 54, 55, 56, 57] in video multicast, broadcast and conferencing scenarios.

Major applications of network coding techniques are tactical communications in military networks [58], multimedia streaming [59] in peer-to-peer (P2P) overlay networks, information delivery in wireless networks [60]. File distribution and multimedia streaming on P2P networks [60], data persistence and data transmission in sensor networks [58], resilient to network attacks like snooping, eavesdropping or replay attacks, Bidirectional low energy transmission in wireless sensor networks, security, Decentralized Network Operation, Multiple Unicast Sessions and decrease the number of packet retransmission for a single-hop wireless multicast transmission, and hence improve network bandwidth.

5 Combination of PIE and Network Coding

In wireless networks multicast routing i.e. the distribution of information from a source peer to a large number of destination peers , has recently attracted a lot of attention for more than a decade (e.g. native IP multicast, CDNs, and, recently, peer-to-peer networks). An essential problem in large scale distribution is the optimal scheduling of the data streams. Recently, network coding proposed a new significant solution to the scheduling problem by encouraging the network nodes to mix the transmitted information and can boost multicast throughput and transmission reliability to be the minimum of the min-cut from the source to the multicast receivers. Combination of network coding and peer to peer information exchange can be used to increase network capacity by reducing the number of transmissions required to exchange data over wireless media. By defining a cooperative peer to peer information exchange using network coding as one in which peers exchange information to coordinate efforts and maximize application-related performance.

Network Coding permits intermediate nodes to combine packets by taking their exclusive-OR (XOR) bit-by-bit to reduce number of transmissions, which decreases energy utilization and helps in throughput improvement. Sending maximum number of hops for a packet to arrive at a receiver node helps in delay minimization. It also raises enormous transmission efficiency and decreases computational overhead. In a wireless networks several network coding techniques are utilized to increase the bandwidth efficiency of consistent exchange which reduce the number of broadcast transmission from one sender to multiple receivers. And also it combines different lost packets from different receivers in such a manner that multiple receivers are able to recover their lost packets by means of single exchange by the source. Thus it has been identified that the information exchange in wireless networks is a further application scenario where network coding shows exclusive gains over conventional routing.

Information exchange using network coding finds many useful applications. These include voice conversations, Videoconferencing between two participants, and instant messaging. In fact, the scope of information exchange goes much further beyond the generic two-way end-to-end communications listed above applications. Even though many networks are usually utilized for information exchange between peers, they have either protected peers anonymity, or required transacting peers to trust each other implicitly. These two methods are vulnerable to attacks by malicious peers who can misuse the P2P system to spread viruses, incorrect, or damaging information. In order to exchange securely network coding is used since it employs encoding mechanisms. And also some secure network coding is designed by combining information theoretic approaches with cryptographic approaches. With information-theoretic approaches, it is proved that complicated modification detections are done at sink nodes. In this scheme, random network coding is used by incorporating a polynomial hash value in each packet. By this way, the computing complexity is much less and also efficiently prevents the propagation of malicious attacks.

PIE can not only fully exploit the broadcast nature of wireless channels, but also take advantage of cooperative peer-to-peer information exchange. So far PIE using network coding is done for small network and at a time only one exchange takes place. When the number of nodes in the network increases, the network performance decreases. It is evident from the theoretical analysis that even under the optimal circumstances, the throughput of each host decreases towards zero rapidly. Despite of the various solutions available, cluster formation seems to be more efficient for solving the scalability problem in adhoc networks. Thus clustering will help us to get scalability and it increases the network performance. Clustering algorithms designed at producing the minimum number of clusters that maximize the network lifespan and fault tolerance and provide load balancing and data throughput.

From the above trace, we can justify that the cooperative approach among the peers improves the performance of cPIE application in wireless environment. But the throughput of the performance of the network degrades with increase in the number of nodes in the network. Thus in this paper, we propose an efficient Co-operative Peer-to-peer Information Exchange (cPIE) technique with an effective network coding using clustering approach.

6 Clustering

Network clustering provides an approach to partition a network topology into groups such that nodes in the same cluster are highly connected and nodes between clusters are sparsely connected. And these nodes communicate with each other and employ toward a common goal. But there is no formal hierarchy for how information passes between the nodes. Clustering is one of the fundamental approaches for scheming energy-efficient, robust and highly scalable wireless P2P networks. By using clustering in networks, it reduces the communication overhead, thereby decreasing the energy consumption and interference among the nodes.

Generally there are three types of nodes in clustering networks .They are cluster heads, cluster members (CM) and gateway nodes. In all clusters, one node is selected as Cluster Leader (CL) to operate as a local organizer and these nodes are vested with the responsibility for routing node messages within each cluster and managing power control and synchronization. The size of the cluster (the number of nodes in the cluster) will be based on the transmission range of the nodes in single hop cluster and the number of hops made by the cluster in multi-hop clusters. The cluster members send or relay data to the CL which transmits the collected packets to the next hop. The gateway node, belonging to more than one cluster, bridges the CLs in those clusters. The communications between two adjacent clusters are conducted through the gateway nodes. Both gateways and member nodes are managed by their cluster heads. However these CLs and gateway nodes form the backbone network, but the presence of gateway node is not compulsory in the clustering network.

There is no physical backbone architecture available in wireless P2P networks for routing of the message, a node depends on other nodes to relay packets if they do not have direct links. Wireless backbone architecture can be used to support efficient communications between nodes [61]. To support backbone architecture, the cluster heads should be a part of the backbone and the fewer the number of backbone nodes the better. Fewer nodes in the backbone can reduce the quality of messages exchanged by backbone nodes [61].

Some benefits of clustering are,

- Reducing the number of messages sent to each BS from each node, channel access, power control and bandwidth control.
- Clustering makes the topology more stable even though there is change in the nodes since it affects only the part of the topology.
- Only CLs or gateway nodes necessitate sustaining the route information.
- Only the CLs and gateway nodes produce the backbone network, results in much simpler topology, less overhead, flooding and collision.

7 Incorporating cPIE and Networking in a Large Network Using Clustering

In this chapter, we will discuss in detail about how to carry out PIE with NC in a large network using clustering approach. In this phenomenon, we can assume two facts, Fact 1: Each packet sent by the tower should be received by at least one peer. Fact 2: No peer shall receive all the packets that are sent by the tower. These facts imply that it is possible to find each packet sent by the tower within the network instead of requesting to broadcast all the packets again which will be tedious. By this technique, the peer can receive the packets directly from other peers which received it errorless rather than getting the missing packets from the tower. Here we elucidate the process to effectively achieve cPIE technique using the clusters which are already formed in the network in three stages.

- Stage 1 – Sharing packets within single cluster
- Stage 2 – Sharing packets among the CLs
- Stage 3– Sharing new packets within the cluster which are received from other CLs

Stage – 1:

In this stage, the packets received correctly by each CM nodes are shared among its cluster through their corresponding CLs. In this stage NC being used by the CLs to control the packet flow and performance improvement.

Stage – 2:

After the completion of Stage 1, each CLs share the packets they have with each other. This is because some packets are received by CMs of any one cluster, to make it available to all other clusters stage 2 is performed. At the success of stage 2, each CL has all the packets, without any error, broadcasted by the tower.

Stage – 3:

At this stage each CLs share the new packets, which they received during stage 2, with their CMs. This makes sure that each CMs of each cluster has all the packets that are broadcasted by the tower without any error. At the successful completion of all these three stages, each peer in the network holds equal and all the packets broadcasted by the tower without asking for re-broadcasting. Thus the Co-operative

Peer-to-Peer Information Exchange with Network Coding can be achieved using the Clustering approach to improve the efficiency of the system.

8 Conclusion

In this paper, we have conducted a detailed study on Peer-to-Peer Information Exchange (PIE), Network Coding (NC) and the advantages of merging PIE and NC in performance improvement factor. It is also studied the scalability issues that arise on the PIE with NC and the capability of clustering which handles the scalability issues. Based the study, we presented a overview method to incorporate clustering technique on PIE with NC to improves the performance than the existing PIE with NC technique in network with large number of peers. As a future work, we are working on to devise a comprehensive methodology to justify the work presented in the study.

References

[1] Fagin, R., Kolaitis, P.G., Miller, R.J., Popa, L.: Data Exchange: Semantics and Query Answering. In: Calvanese, D., Lenzerini, M., Motwani, R. (eds.) ICDT 2003. LNCS, vol. 2572, pp. 207–224. Springer, Heidelberg (2002)

[2] Huebsch, R., Hellerstein, J.M., Lanham, N., Loo, B.T., Shenker, S., Stoica, I.: Querying the Internet with PIER. In Johann Christoph Freytag. In: Freytag, J.C., Lockemann, P.C., Abiteboul, S., Carey, M.J., Selinger, P.G., Heuer, A. (eds.) Proceedings of 29th International Conference on Very Large Data Bases, VLDB 2003, September 9-12, pp. 321–332. Morgan Kaufmann (2004)

[3] Gilbert, S., Lynch, N.: Brewer's Conjecture and the Feasibility of Consistent, Available, Partition Tolerant Web Services. SIGACT News 33(2), 51–59 (2002)

[4] Ratnasamy, S., Francis, P., Handley, M., Karp, R., Schenker, S.: A Scalable Content Addressable Network. In: SIGCOMM 2001: Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, pp. 161–172. ACM Press, New York (2001)

[5] Bernstein, P., Giunchiglia, F., Kementsietsidis, A., Mylopoulos, J., Serafini, L., Zaihrayeu, I.: Data management for Peer-to-Peer computing: A vision. In: WebDB, pp. 89–94 (2002)

[6] Li, C.: Raccoon: A peer-based system for data integration and sharing. In: ICDE, p. 852 (2004), System Demonstration

[7] Calvanese, D., Giacomo, G.D., Lenzerini, M., Rosati, R.: Logical foundations of peer-to-peer data integration. In: PODS, pp. 241–251 (2004)

[8] Franconi, E., Kuper, G., Lopatenko, A., Serafini, L.: A robust logical and computational characterisation of peer-to- peer database systems. In: VLDB Workshop on Databases, Information Systems and Peer-to-Peer Computing (2003)

[9] Franconi, E., Kuper, G., Lopatenko, A., Zaihrayeu, I.: The coDB robust peer-to-peer database system. In: Symposium on Advanced Database Systems, pp. 382–393 (2004)

[10] Buchholz, T., Hochstatter, I., Treu, G.: Profile-based Data Diffusion in Mobile Environments. In: 2004 IEEE International Conference on Mobile Ad-hoc and Sensor Systems (2004)

[11] Kurhinen, J., Korhonen, V., Vapa, M., Weber, M.: Modelling Mobile Encounter Networks. In: Proceedings of IEEE PIMRC 2006 (2006)

[12] Kurhinen, J., Vuori, J.: Information Diffusion in a Single-Hop Mobile Peer-to-Peer Network. In: Proceedings of the 10th IEEE Symposium on Computers and Communications, ISCC (2005)

[13] Daswani, N., Garcia-Molina, H., Yang, B.: Open Problems in Data-Sharing Peer-to-Peer Systems, `http://www-db.stanford.edu`

[14] Shen, H., Kumar, M., Das, S.K., Wang, Z.: Energy-Efficient Caching and Prefetching with Data Consistency in Mobile Distributed Systems. In: Proc. of IEEE International Parallel and Distributed Processing symposium (IPDPS), Santa Fe, NM (April 2004)

[15] Ripeanu, M.: Peer-to-peer architecture case study: Gnutella network. Computer Science Dept., University of Chicago (2001)

[16] Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: scalable peer-to-peer look up service for internet applications. In: ACM SIGCOMM (2001)

[17] Tewari, S.: Performance Study of Peer -to-Peer File Sharing, Ph. D Thesis, University of California, Los Angeles (2007)

[18] Ahlswede, R., Cai, N., Li, S.-Y.R., Yeung, R.W.: Network information flow. IEEE Trans. Inf. Theory 46(4), 1204–1216 (2000)

[19] Chen, L., Ho, T., Low, S., Chiang, M., Doyle, J.: Optimization based rate control for multicast with network coding. In: Proc. of IEEE INFOCOM, Anchorage, AK (May 2007)

[20] Sanders, P., Egner, S., Tolhuizen, L.: Polynomial Time Algorithms for Network Information Flow. In: Proc. 15th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), pp. 286–294 (2003)

[21] Katti, S., Rahul, H., Hu, W., Katabi, D., Medard, M., Crowcroft, J.: XORs in the Air: Practical Wireless Network Coding. IEEE/ACM Trans. Networking 16(3), 497–510 (2008)

[22] Cheung, G., Li, D., Chuah, C.-N.: On the Complexity of Cooperative Peer-to-Peer Repair for Wireless Broadcasting. IEEE Communications Letters 10(11), 742–744 (2006)

[23] Li, Z., Li, B.: Network coding in undirected networks (2004)

[24] Fragouli, C., Widmer, J., Boudec, J.-Y.L.: A network coding approach to energy efficient broadcasting: from theory to practice. In: IEEE INFOCOM (2006)

[25] Fragouli, C., Widmer, J., Boudec, J.-Y.L.: Efficient broadcasting using network coding. IEEE/ACM Transactions on Networking 16(2), 450–463 (2008)

[26] Li, L., Ramjee, R., Buddhikot, M., Miller, S.: Network coding-based broadcast in mobile ad hoc networks. In: IEEE INFOCOM (2007)

[27] Lun, D.S., Médard, M., Koetter, R.: Efficient operation of wireless packet networks using network co ding. In: International Workshop on Convergent Technologies, IWCT (2005)

[28] Chachulski, S., Katti, S.: Trading structure for randomness in wireless opportunistic routing. In: Proc. the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, New York, NY, USA, pp. 169–180 (2007)

[29] Katti, S., Rahul, H., Hu, W., Katabi, D., Médard, M., Crowcroft, J.: XORs in the air: Practical wireless network coding. IEEE/ACM Trans. on Networking 16(3), 497–510 (2008)

[30] Tse, D., Vishwanath, P.: Fundamentals of Wireless Communications. Cambridge University Press (2005)

[31] Pickholtz, R.L., Milstein, L.B., Schilling, D.L.: Spread spectrum for mobile communications. IEEE Trans Veh. Technology 40, 313–322 (1991)

[32] Laneman, J.N., Tse, D.N.C., Wornell, G.W.: Cooperative diversity in wireless networks: Efficient protocols and outage behavior. IEEE Trans. on Inform. Theory 50(12), 3062–3080 (2004)

[33] Ramanathan, R.: Challenges: A Radically New Architecture for Next Generation Mobile Ad Hoc Networks. In: ACM MOBICOM (2005)

[34] Halloush, M., Radha, H.: Network coding with multi-generation mixing: Analysis and Applications for video communication. In: IEEE International Conference on Communications (May 2008)

[35] Nguyen, D., Nguyen, T., Yang, X.: Multimedia wireless transmission with network coding. In: IEEE 16th International Packet Video Workshop, Lausanne, Switzerland, pp. 326–335 (November 2007)

[36] Chou, P., Miao, Z.: Rate-distortion optimized streaming of packetized media. IEEE Transactions on Multimedia 8(2), 390–404 (2006)

[37] Ahlswede, R., Cai, N., Li, S.-Y.R., Yeung, R.W.: Network information flow. IEEE Transactions on Information Theory 46, 1204–1216 (2000)

[38] Wang, C.-C., Shroff, N.B.: Beyond the butterfly a graph-theoretic characterization of the feasibility of network coding with two simple unicast sessions. In: Proc. IEEE International Symposium on Information Theory (June 2007)

[39] Wang, C.-C., Shroff, N.B.: Intersession network coding for two simple multicast sessions. In: 45th Allerton Conference on Communication, Control and Computing (2007)

[40] Ahlswede, R., Cai, N., Li, S.-Y.R., Yeung, R.W.: Network Information Flow. IEEE Transactions on Information Theory 46(4), 1204–1216 (2000)

[41] Li, S.-Y.R., Yeung, R.W., Cai, N.: Linear Network Coding. IEEE Transactions on Information Theory 49(2), 371–381 (2003)

[42] Biermann, T., Polgar, Z.A., Karl, H.: Cooperation and coding framework. In: Proc. International Workshop on the Network of the Future (Future-N et) (June 2009)

[43] Ho, T., Koetter, R., Médard, M., Effors, M., Shi, J., Karger, D.: A Random Network Coding Approach to Multicast. IEEE Transactions on Information Theory 52(10), 4413–4430 (2006)

[44] Balli, H., Yan, X., Zhang, Z.: On Randomized Linear Network Codes and Their Error Correction Capabilities. IEEE Transactions on Information Theory 55(7), 3148–3160 (2009)

[45] Koetter, R., Médard, M.: An Algebraic Approach to Network Coding. IEEE/ACM Transactions on Networking 11(5), 782–795 (2003)

[46] On the Viability of a Cooperative-Network Coding Protocol in Clustered Networks. In: IEEE MILCOM, San Diego, CA, USA, November 17-19 (2008)

[47] Nguyen, K., Nguyen, T., Cheung, S.-C.: Video streaming with network coding. The Springer Journal of Signal Processing Systems Special Issue: ICME 2007 (February 2008)

[48] Yeung, R.W.: Avalanche: A Network Coding Analysis. Communications in Information and Systems 7(4), 353–358 (2007)

[49] Deb, S., Médard, M., Choute, C.: Algebraic Gossip: A Network Coding Approach to Optimal Multiple Rumor Mongering. IEEE Transactions on Information Theory 52(6), 2486–2507 (2006)

[50] Mosk-Aoyama, D., Shah, D.: Information Dissemination via Network Coding. In: Proc. of IEEE International Symposium on Information Theory (ISIT 2006), Seattle, WA (October 2006)

[51] Ahlswede, R., Cai, N., Li, S.-Y.R., Yeung, R.W.: Network information flow. IEEE Transactions on Information Theory 46, 1204–1216 (2000)

[52] Wang, C.-C., Shroff, N.B.: Beyond the butterfly – a graph-theoretic characterization of the feasibility of network coding with two simple unicast sessions. In: Proc. IEEE International Symposium on Information Theory (June 2007)

[53] Karande., S., Misra, K., Radha, H.: Clix: Network coding and cross layer information exchange of wireless video. In: Proc. of 2006 IEEE International Conference on Image Processing (October 2006)

[54] Nguyen, D., Nguyen, T., Yang, X.: Multimedia wireless transmission with network coding. In: Proc. of Packet Video (November 2007)

[55] Katti, S., Rahul, H., Hu, W., Katabi, D., Medard, M., Crowcroft, J.: XORs in the Air: Practical Wireless Network Coding. IEEE/ACM Trans. Networking 16(3), 497–510 (2008)

[56] Seferoglu, H., Markopoulou, A.: Opportunistic network coding for video streaming over wireless. In: Proc. of Packet Video (November 2007)

[57] Hui Wang, R.Y.C., Kuo, C.-C.J.: Wireless multi-party video conferencing with network coding. In: Proc. of ICME (2009)

[58] Kamra, A., Misra, V., Feldman, J., Rubenstein, D.: Growth codes: maximizing sensor network data persistence. In: Proc. ACM SIGCOMM 2006, pp. 255–266 (2006)

[59] Gkantsidis, C., Miller, J., Rodriguez, P.: Comprehensive view of a live network coding P2P system. In: Proc. 6th ACM SIGCOMM Conf. Internet Measurement, pp. 177–188 (October 2006)

[60] Feng, C., Li, B.: On large-scale peer-to-peer streaming systems with network coding. In: Proc. ACM Multimedia, Vancouver, BC, Canada, pp. 269–278 (2008)

[61] Baker, D.J., Ephremides, A.: The Architectural Organisation of a Mobile Radio Network via a Distributed algorithm. IEEE Trans. Commun. 29(11), 1694–1701 (1981)

SNetRS: Social Networking in Recommendation System

Jyoti Pareek[1], Maitri Jhaveri[2], Abbas Kapasi[2], and Malhar Trivedi[2]

[1] Department of Computer Science,
Gujarat University,
Ahmedabad-380009, Gujarat, India
drjyotipareek@yahoo.com
[2] GLS-Institute of Computer Technology,
Law Garden, Gujarat Technological University, Ahmedabad, Gujarat, India
jmaitri@glsict.org, mr_abbas_kapasi@yahoo.com, tmalhar@gmail.com

Abstract. With the proliferation of electronic commerce and knowledge economy environment both organizations and individuals generate and consume a large amount of online information. With the huge availability of product information on website, many times it becomes difficult for a consumer to locate item he wants to buy. Recommendation Systems [RS] provide a solution to this. Many websites such as YouTube, e-Bay, Amazon have come up with their own versions of Recommendation Systems. However Issues like lack of data, changing data, changing user preferences and unpredictable items are faced by these recommendation systems. In this paper we propose a model of Recommendation systems in e-commerce domain which will address issues of cold start problem and change in user preference problem. Our work proposes a novel recommendation system which incorporates user profile parameters obtained from Social Networking website. Our proposed model SNetRS is a collaborative filtering based algorithm, which focuses on user preferences obtained from FaceBook. We have taken domain of books to illustrate our model.

Keywords: User preferences, social networking, Recommendation System (RS), Collaborative Filtering (CF).

1 Introduction

As time passes, World Wide Web (WWW) goes on growing. Lots of information is available on WWW. All the information which we get is not relevant, only few of them are relevant. When a user tries to search something on WWW s/he lands up with thousands of result. As a result, s/he will mess up with huge information. Hence fetching the actually required details becomes cumbersome and time consuming. This gives rise to data filtering system. In early days, for data filtering, Information Filtering (IF) was used. IF was basically developed for filtering documentation, articles, news etc. Looking to our era, e-commerce is growing explosively. Whenever a user makes a search for particular item on internet to buy, s/he will get many options. Looking at the options user gets confuse what to buy, and will not able to sort the

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 195–206.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

item that is suitable to him/her. This problem gave rise to Recommendation System [RS]. A recommender system is a personalization system that helps users to find items of interest based on their preferences. Recommender systems are efficient tools that overcome the information overload problem by providing users with the most relevant contents [8]. The importance of contextual information has been recognized by researchers and practitioners in many disciplines including E-commerce, personalized IR, ubiquitous and mobile computing, data mining, marketing and management. There are many existing e-commerce websites which have implemented recommendation systems successfully. We will discuss few website in our coming section that provides recommendation. Items are suggested by looking at the behavior of like-minded-users. Groups are formed of such users, and items preferred by such groups are recommended to the user, whose liking and behavior is similar to the group. In our model we have incorporated user preferences obtained from Social Networking Site. Social Networking sites are used intensively from last decade. According to the current survey, Social Networking sites have the largest data set of users. Each social networking site notes/records each and every activity of user (like: what user likes? what user is doing? what is user's hobby? Etc.). Social Networking site will prove to be largest domain in understanding the user behaviour. One of the best examples of social networking is FACEBOOK. According to current news FACEBOOK is trying to develop algorithm, to understand user behaviour. Social Networking sites can help us in getting important information of user's, such as age, gender, location, language, actives, likes etc. our model takes into account these parameters of the user to recommend books.

2 Literature Review

Study of few recommendation pattern used by websites**:** Amazon recommendations change regularly based on a number of factors. These factors include time and day of purchase, rate or like a new item, as well as changes in the interests of other customers. Because your recommendations will fluctuate, Amazon suggests you add items that interest you to your Wish List or Shopping Cart. **E-Bay** recommends product on bases of features of items. **You Tube** recommends items based on like/dislikes concept. **In.com** recommends the songs that are popular, songs from the same movie, similar actor-actress, artist, director etc. RS is used to filter the item/product according to the user interest [1,2] and looking at the like-minded-users [3]. There are many popular recommendation algorithms based on collaborative filtering [3,4]. Collaborative Filtering creates a group of users with similar behaviour, and finds the items preferred by this group. Ratings from user will be taken from user in two ways explicit rating and implicit rating [5]. CF algorithms are divided into two types, memory-based algorithm and model based algorithm. Memory-Based algorithm simply stores all the user ratings into memory. There are two variants of memory-based recommendation and both are based on the k-Nearest Neighbour algorithm: user-based filtering and item-based filtering. In User - Based Filtering, Rating matrix is used to find neighbouring users for the active user. This is done by using cosine or Pearson's correlation matrix. After knowing the neighbouring user for active user, items preferred by neighbouring users will be sorted on frequency and rating of items. Items that are not known to active user will be recommended. Item – Based Filtering finds the most similar items. Items are considered to be similar when the same set of users has purchased them or rated them highly. For each item of an active user, the neighbourhood of most similar items is identified. Collaborative filtering techniques can be expanded to other algorithms such as tag based and attribute aware and trust aware recommender systems. A diffusion-based recommendation algorithm is proposed [9] which consider the personal vocabulary. A hybrid user profiling strategy is proposed [10] that take advantage of both content-based profiles describing long-term information interests that a recommender system can acquired along time and interests revealed through tagging activities, with the goal of enhancing the interaction of users with a collaborative tagging system. Trip Tip system is proposed [11] to help negotiate traveller's way through the immense amount of information that is often available by recommending a set of choices. Trip Tip recommends to the users the next place, which they would most likely want to visit given their preference in previous choices. To generate this information, tags that are attached on a given place by users give the characteristics of a place and the reasons for visiting the place. Attribute-aware method proposed [12] takes into account item attributes, which are defined by domain experts. In addition, content-based algorithms can provide very accurate recommendations [13]. Collaborative tagging systems (CTSes), allow users to freely assign tags to their collections, provide promising possibility to better address the above issues. A generic method [14] was proposed that allows tags to be incorporated to the standard collaborative filtering, via reducing the ternary correlations to three binary correlations and then applying a fusion method to re-associate these correlations. Some diffusion-based algorithms are recently proposed for personalized recommendations. A spreading ACTtion based collaborative filtering [15] was proposed which is essentially an iterative diffusion process. A diffusion-based [16] top-k collaborative filtering, performs better than pure top-k CF and pure diffusion-based algorithm. Besides recommender systems, research on context-aware computing seems promising. Context-awareness allows software applications to use information beyond those directly provided as input by users [17]. More recently, there were attempts [18] to define architectures for context-aware recommender. However, authors don't give details about the deployment of such architectures. An algorithm is proposed [19] which adopt item-based algorithms in the early stage of the cold-start period and eventually switching to SVD-based algorithms. A collaborative filtering recommendation algorithm based on the implicit information of the new users and multi-attribute rating matrix is proposed [20] to solve the cold start problem.

3 Our Approach

We propose the architecture of SNetRS as shown in the following figure 1.

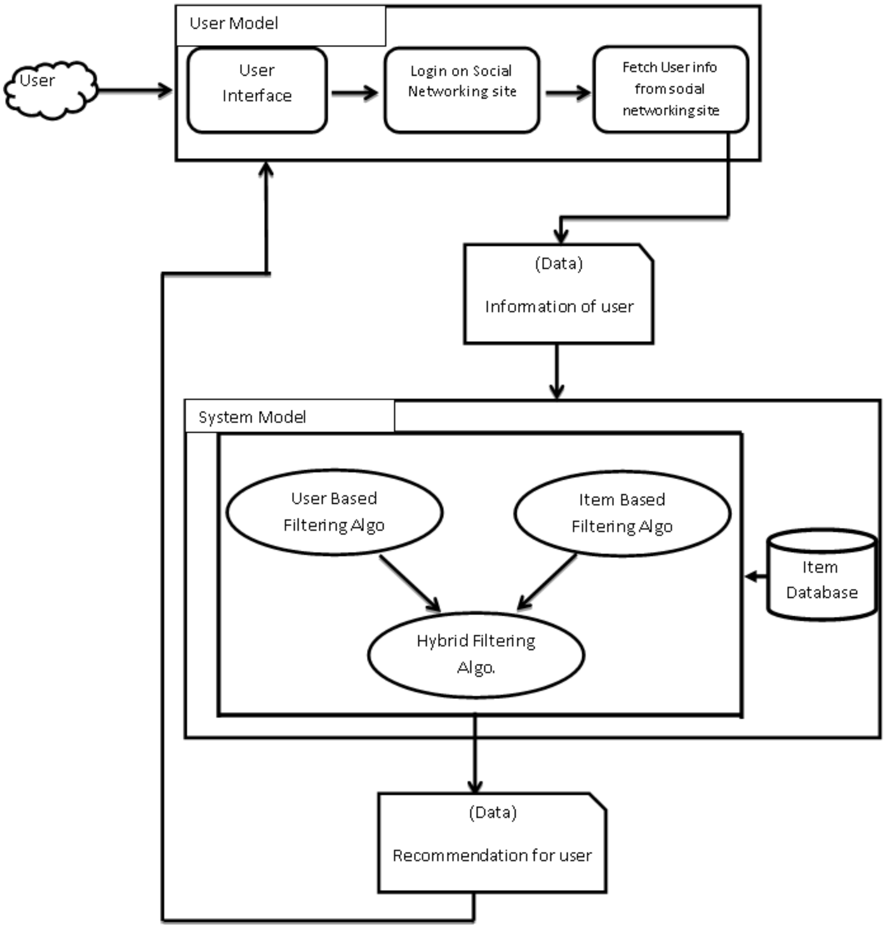


Fig. 1. Architecture of SNetRS

It is divided in two models. User model and System model. User model gives the information of the user which is then utilized by the system model which ultimately gives the recommendations. User model uses FaceBook as a source of fetching user details such as user own interests and interest of his/her friends. Each social networking site gives API, which can be used to fetch information from the user profile. Each social networking site gives there plugins and SDK [7] in different platforms, which will help to include their service to our site. We have used the API of the highest used and famous social networking site “FACEBOOK”. Facebook provides Graph API [8] through which we can access the user information. The Graph API presents a simple, consistent view of the Facebook social graph, uniformly representing objects in the graph (e.g., people, photos, events, and pages) and the connections between them (e.g., friend relationships, shared content, and photo tags). System model takes as input the information of the user to whom the item is to be recommended. This model is a combination of item based filtering and user based filtering.

Experimental Setup

The experimental data contains 8 two wheeler vehicles and 7 movies. We have taken into consideration the location and gender as parameters for the two wheeler vehicles (Table 2) and language and age group as parameters for the movies(Table 1). The aim is to find the likes of devang (Table 3 and 4) in the experimental products based on his rating given for the products in the training dataset (table 5 and 6).

Based on the training data of table 5 and 6, we aim to obtain the ratings for experimental data which should match with the data of table 3 and 4.

Experimental data includes finding and matching likes of devang for two wheeler KARI and movie ROC. Table 7 and 8 shows rating(1 to 5) of different products obtained from a survey of 50 users. Category of users who participated in the survey was students, accountants, house wives and professors. The rating of the each product is obtained by applying SVD++ [4] algorithm on the ratings obtained from the 50 users who participated in the above survey.

Implementation

Item Based Algorithm

Step 1: following is the information of user "devang" obtained from the user model.

Age:	24
Gender:	Male
Location:	Gujarat, India
Language Known:	Gujarat, Hindi.
Activities:	Tennis, Guitar, Cooking

Step 2: Find Satisfaction rate for each products and add ratings of each product. Table 9 shows the ratings of each product obtained from the survey and the satisfaction rate of each product. The satisfaction rate of each product is obtained by the satisfaction of location and genderparameter of devangl.. Table 10 shows satisfaction rate of each movie. The satisfaction rate of HP remains zero irrespective of the age group because language is the key parameter.

User Based Algorithm

Step 1: We obtain set of items liked by user "devang" from the user model created from facebook.. Ref Table 11

Step 2: We obtain the users with similar likes as that of user "devang" and Find number of common likes of other user for user "devang". See table 12

Step 3: Find the other likes of the users set. Set the priority of user, based on from Table 13 and 14 ehavior count.

.Step 5: Remove the items that are already liked by user "devang".

Step 6: Find frequency of product that are common between users. See table 15 and 16

Step 7: Find the final priority for recommendation,

Summation of: Priority + Frequency + Ratings. See table 16 and 17

Hybrid Algorithm

Step 1: Combine both result of Item based filtering and user based filtering.
Step 2: Sort in descending order on final priority bases.
Step 3: If there is new duplicate item then place its final priority index as highest see table 18 and 19

4 Conclusion and Future Work

We conclude from our research and analysis that, scope of recommendation is much in e-commerce domain. Recommendation using social networking information will really help in recommending the best product suitable to the user. Social networking is the best means of knowing user behaviour. We are going to have further research on the same topic. We plan to implement this model and to add time factor and cross-domain filtering. Time factor model will help in knowing the rating gaps base on time. Cross – domain filtering will help to know the purpose of user, visiting our site. From cross-domain filtering system will get an idea, about the product user is looking for.

References

1. Bruke, R.: Hybrid Recommender System: Survey and Experiments. User Modeling and User- Adapted Interaction 12(4), 331–370 (2001)
2. Montaner, M., Lopez, B., de la Rosa, J.L.: A taxonomy of recommender agents on the internet. Artificial Intelligent Review 19(4), 285–330 (2003)
3. Bogers, A.M.: Recommender Systems for Social Bookmarking ISBN 978-90-8559-582-3
4. Koren, Y., Bell, R.: Advances in Collaborative Filtering
5. Oard, D.W., Kim, J.: Implicit Feedback for Recommender Systems. In: Proc. 5th DELOS Workshop on Filtering and Collaborative Filtering, pp. 31–36 (1998)
6. `http://developers.facebook.com/docs/reference/api/`
7. `http://developers.facebook.com/docs/sdks/`
8. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, pp. 43–52 (1998)
9. Shang, M.S., Zhang, Z.K.: Diffusion-based recommendation in collaborative tagging systems. Chinese Physics Letters 26(11) (2009)
10. Godoy, D., Amandi, A.: Hybrid content and tag-based profiles for recommendation in collaborative tagging systems. In: La-Web (Latin American Web Conference), pp. 58–65 (2008)
11. Kim, J., Kim, H., Ryu, J.H.: TripTip: A trip planning service with tag-based recommendation. Extended Abstracts on Human Factors in Computing Systems, 3467–3472 (2009)
12. Tso, K., Schmidt-Thieme, L.: Attribute-aware collaborative filtering. In: Proceedings of 29th AnnualConference of the German Classification Society (2005)
13. Pazzani, M.J., Billsus, D.: Content-Based Recommendation Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 325–341. Springer, Heidelberg (2007)

14. Tso-Sutter, K.H.L., Marinho, L.B., Schmidt-Thieme, L.: Tag-aware recommender systems by fusion of collaborative filtering algorithms. In: Proceedings of the ACM Symposium on Applied Computing, pp. 1995–1999 (2008)
15. Huang, Z., Chen, H., Zeng, D.: Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Transactions on Information Systems 22(1), 116–142 (2004)
16. Liu, J.G., Wang, B.H., Guo, Q.: Improved collaborative filtering algorithm via information transformation. International Journal of Modern Physics C 20(2), 285–293 (2009)
17. Dey, A.K., Abowd, G.D., Salber, D.: A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Human-Computer Interaction Journal 16, 97–166 (2001)
18. Baltrunas, L.: Exploiting contextual information in recommender systems. In: ACM RecSys, pp. 295–298 (2008)
19. Cremonesi, P., Turrin, R.: Analysis of cold-start recommendations in iptv systems. In: Proceedings of the Third ACM Conference on Recommender Systems, pp. 233–236 (October 2009)
20. Yin, H., Chang, G., Wang, X.: A cold-start recommendation algorithm based on new user's implicit information and multi-attribute rating matrix. In: Proceedings of the Ninth International Confer-ence on Hybrid Intelligent Systems, vol. 2, pp. 353–358 (2009)

Appendix -1

Two Wheeler Vehicles								
Items	Discover	TwiSter	Pulzar	Karishma	CBZ	Splendor	Scooty Pap	Activa
Short-form	DIS	TWI	PUL	KARI	CBZ	SPLEN	SCOOP	ACT
Movies								
Items	Rock Star	Chiller party	KAHANI	House full 2	Andaz Apna Apna	Love Aaj Kal	Harry Potter	
Short-Form	ROC	CHI	KAH	Hf2	AZAA	LAK	HP	

Tables

Table 1. Survey of movies among different age group (Language is the key parameter for recommendation)

Sr. no.	1	2	3	4	5	6	7
Movie	ROC	CHI	KAH	HF2	AZAA	LAK	HP
Language	Hindi	Hindi	Hindi	Hindi	Hindi	Hindi	English
Preferred Age group	20-35	5-18	20-90	10-40	10-90	20-40	5-50

Table 2. Survey of vehicles among people of different gender (Location is the key parameter for recommendation)

Sr. no.	1	2	3	4	5	6	7	8
Vehicle	DIS	TWI	PUL	KARI	CBZ	SPLEN	SCOOP	ACT
Location	India	India	India	India	India	India	India	India
Preferred gender	male	Male	male	Male	Male	male	female	Male/female

Table 3. Likes of devang for two wheelers, taken from Facebook

Two Wheeler Vehicle's Like							
DIS	TWI	PUL	KARI	CBZ	SPLEN	SCOOP	ACT
-	-	like	Like	like	-	-	Like

Table 4. Likes of devang for movies, taken from Facebook

Movies Like						
ROC	CHI	KAH	HF2	AZAA	LAK	HP
Like	-	Like	like	-	-	-

Based on the training data of table 5 and 6, we aim to obtain the ratings for experimental data which should match with the data of table 3 and 4.

Table 5. Training data for two wheeler

DIS	TWI	PUL	KARI	CBZ	SPLEN	SCOOP	ACT
-	-	Like	-	Like	-	-	Like

Table 6. Training data for movies

ROC	CHI	KAH	HF2	AZAA	LAK	HP
-	-	Like	like	-	-	-

Table 7. Ratings (1 to 5) of two wheeler vehicle taken by survey

Ratings Of Two Wheeler Vehicle								
Items	DIS	TWI	PUL	KARI	CBZ	SPLEN	SCOOP	ACT
Ratings	3	3	4	4	3	3.5	3	3

Table 8. Ratings (1 to 5) of movies taken by survey

Ratings of Movies						
ROC	CHI	KAH	HF2	AZAA	LAK	HP
4	3.5	3.5	3	3	3.5	4

Table 9. Satisfaction rate of each two wheeler. Note: gray colored products are already liked by user.

	PUL	KARI	SPLEN	CBZ	TWI	ACT	DIS	SCOOP
Ratings	4	4	3.5	3	3	3	3	3
Satisfaction rate	2	2	2	2	2	2	2	1
Final ratings	6	6	5.5	5	5	5	5	4

Table 10. Satisfaction rate of each movie. Note: gray colored items are already liked by user.

	ROC	KAH	LAK	HF2	AZAA	CHI	HP
Ratings	4	3.5	3.5	3	3	3.5	4
Satisfaction rate	2	2	2	2	2	1	0
Final rating	6	5.5	5.5	5	5	4.5	4

Table 11. Set of items liked by user “devang”

Two Wheeler Vehicle				Movies		
PUL	CBZ	ACT		ROC	KAH	HF2
like	like	Like		Like	like	like

Table 12. Users with similar likes as that of user “devang”

PUL	Sandip	Kandarp	Jagdish	Ravi	Malhar	Abbas	Ekta	-	-
CBZ	Sandip	Kandarp	Jagdish	Ravi	Malhar	Abbas	Ekta	-	-
ACT	-	Kandarp	-	Ravi	Malhar	Abbas	Ekta	Dhara	Chinmayee
KAH	Sandip	Kandarp	-	-	Malhar	Abbas	Ekta	Dhara	Chinmayee
HF2	-	-	Jagdish	Ravi	Malhar	Abbas	Ekta	Dhara	Chinmayee
Common behavior count	3	4	3	4	5	5	5	3	3

Table 13. Other likes of the users set with priority for two wheelers

Priority	Users	Two Wheeler vehicle					
9	Abbas	-	-	PUL	KARI	CBZ	SPLEN
8	Malhar	-	TWI	PUL	KARI	CBZ	SPLEN
7	Ekta	-	-	PUL	KARI	CBZ	-
6	Kandarp	-	TWI	PUL	-	CBZ	SPLEN
5	Ravi	-	TWI	PUL	KARI	CBZ	SPLEN
4	Sandip	DIS	TWI	PUL	KARI	CBZ	SPLEN
3	Jagdish	-	TWI	PUL	-	CBZ	SPLEN
2	Dhara	-	-	-	-	-	-
1	Chinmayee	-	-	-	-	-	-

Table 14. Other likes of the users set with priority for movies

Priority	Users	Movies						
9	Abbas	ROC	CHI	KAH	HF2	AZAA	-	HP
8	Malhar	ROC	CHI	KAH	HF2	AZAA	LAK	-
7	Ekta	ROC	-	KAH	HF2	AZAA	LAK	HP
6	Kandarp	ROC	-	KAH	-	AZAA	LAK	HP
5	Ravi	ROC	CHI	-	HF2	AZAA	LAK	HP
4	Sandip	ROC	CHI	KAH	-	AZAA	LAK	HP
3	Jagdish	ROC	CHI	-	HF2	AZAA	LAK	HP
2	Dhara	ROC	-	KAH	HF2	AZAA	LAK	HP
1	Chinmayee	ROC	-	KAH	HF2	AZAA	LAK	HP

Table 15. Is the set for two wheeler vehicles with frequency count

User's	Two wheeler vehicle				
Abbas	-	-	KARI	SPLEN	-
Malhar	-	TWI	KARI	SPLEN	-
Ekta	-	-	KARI	-	SCOOP
Kandarp	-	TWI	-	SPLEN	-
Ravi	-	TWI	KARI	SPLEN	-
Sandip	DIS	TWI	KARI	SPLEN	-
Jagdish	-	TWI	-	SPLEN	-
Dhara	-	-	-	-	SCOOP
Chinmayee	-	-	-	-	SCOOP
Frequency	1	5	5	6	3

Table 16. Is the set for movies with frequency count

Abbas	ROC	CHI	AZAA	-	HP
Malhar	ROC	CHI	AZAA	LAK	-
Ekta	ROC	-	AZAA	LAK	HP
Kandarp	ROC	-	AZAA	LAK	HP
Ravi	ROC	CHI	AZAA	LAK	HP
Sandip	ROC	CHI	AZAA	LAK	HP
Jagdish	ROC	CHI	AZAA	LAK	HP
Dhara	ROC	-	AZAA	LAK	HP
Chinmayee	ROC	-	AZAA	LAK	HP
Frequency	9	5	9	8	8

Table 17. Is set for two wheeler vehicle based on final priority for recommendation

Product Name	KARI	SPLEN	TWI	SCOOP	DIS
Priority	9	9	8	7	4
Frequency	5	6	5	3	1
Rating	4	3.5	3	3	3
Final ratings	18	18.5	16	13	8

Table 18. Is set for movies based on final priority for recommendation

	AZAA	ROC	CHI	HP	LAK
Prioirty	9	9	9	9	8
Frequency	9	9	5	8	8
Ratings	3	4	3.5	4	3.5
	21	22	17.5	21	19.5

Table 19. Is the final recommendation for movies. HP is removed from the recommendation as its satisfaction rate is "zero" as per item base algorithm.

	ROC	AZAA	LAK	CHI
User based rating	22	21	19.5	17.5
Item based Rating	6	5	5.5	4.5
Final ratings	31	29	28	25

Table 20. Is the final recommendation for two wheeler vehicle

	SPLEN	KARI	TWI	SCOOP	DIS
User based rating	18.5	18	16	13	8
Item based Rating	5.5	6	5	4	5
Final ratings	27	27	24	20	16

Linguistic Conversion of Syntactic to Semantic Web Page

G. Nagarajan[1] and K.K. Thyagarajan[2]

[1] Research Scholar, Sathyabama University, Chennai, Tamil Nadu, India
[2] Professor, Dept. of Information & Technology, RMK College of Engineering & Technology, Chennai, Tamil Nadu, India

Abstract. Information is knowledge. In earlier days one has to find a resource person or resource library to acquire knowledge. But today just by typing a keyword on a search engine all kind of resources are available to us. Due to this mere advancement there are trillions of information available on net. So, in this era we are in need of search engine which also search with us by understanding the semantics of given query by the user. One such design is only possible only if we provide semantic to our ordinary HTML web page. In this paper we have explained the concept of converting an HTML page to RDFS/OWL page. This technique is incorporated along with natural language technology as we have to provide the Hyponym and Meronym of the given HTML pages.

Keywords: Ontology, OWL, RDFS, Name entity recognition, Probability Reasoner.

1 Introduction

Information is the main source of intelligent. In today's revolutionary era the amount of information available on web is enormous. Just by typing any keyword on any search engine will provide you with millions of information, but the amount of relevant information would be very few and the user again has to search manually on all those million result. So, this kind of search engine is not user friendly. The primary approach of this paper is to design a Intelligent search engine which has to provide only the needed relevant information regarding the given query.

Semantic search engine is the only key answer for this kind of search. As said in [2][3] here both the machine and the user tries to search some information on web. There are many research papers regarding the design of a new semantic search engine. In [6] even listed top five to ten Semantic Search Engine. The main drawback of Semantic Search Engine is that the available Semantic Web page on web is very few. As the concept of Semantic Web had started on around 2000, we have very few Semantic Web page. As the creation and design of Syntactic Web page is ease of work also we have lot of in-built software for it still people are interested in creating simple Syntactic web page with general XHTML,XML,PHP,ect. coding instead of construct ontology for it.

The main focus of this paper is to design an Web Intelligent Framework for converting a Syntactic web page to Semantic Web page. Thus if a framework is available we can convert all those web pages and make it available for Semantic Search Engine thus we can provide a way for efficient search engine where both the User and Machine search an information from Web.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 207–213.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

This paper is design in such a way that first we listed some of the related work, then we start with providing the design of Web Intelligent Framework, then in next session we explained the concept of converting HTML document to XML document, next the major conversion of XML to RDFS/OWL is explained then we concluded our work with future scope.

2 Related Work

In [10] discuss the way of converting the HTML to OWL using table. They consider the TABLE tag of HTML page and tried to convert to OWL. This won't produce any semantic to the ontology. In [11] they tried to convert the HTML to OWL using the FRAME set tags they also tried to incorporate UML to identify the class and subclasses. In [12] the conversion is done by first annotating the web page. The annotation they consider is the semantic annotation thus they tried to provide semantic of the page. They use the tool called GATE to analyze the semantic through natural language processing. In [13] the conversion is take place using the tag and they used GRDDL tool for conversion.

3 Web Intelligent Frame Work

As we search for intelligent information we need to design an intelligent system for this Syntactic to Semantic conversion. Fig.1 shows the proposed framework, where the collection of Syntactic web pages are collected via a Web crawler as the output of an web crawler is list of URL we required a genius system to filter out the unwanted URL. Then with the available list of URL of input the XML is created with that entity concern XML a conversion of XML to OWL is implemented and the crated ontology is collected in repository which can be used by an of the Semantic Web search services. The conversion of HTML to XML and the XML to OWL is explained in detail in the forth coming session. The concept of Web Crawler is already explained in [1] we are not specifying in this paper.

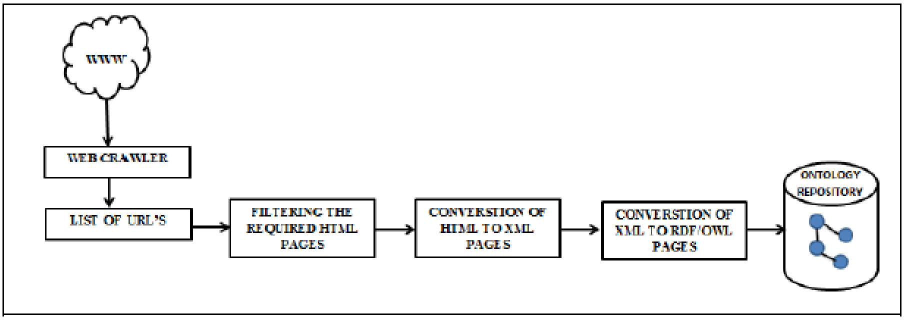

WWW
WEB CRAWLER
LIST OF URL'S
FILTERING THE REQUIRED HTML PAGES
CONVERSION OF HTML TO XML PAGES
CONVERSION OF XML TO RDF/OWL PAGES
ONTOLOGY REPOSITORY

Fig. 1. Web Intelligent Framework

4 HTML to XML Conversion

The first phase of this Web Intelligent Framework is the conversation of all the web page collected from a web crawler to a standard XML files with name entity as the main entity. Name Entity Recognition is a concept of Natural Language Processing. In short it is called as NER. The main technology used here are patterns and Lexicons. For the given Corpus text NER classifies the entity as Person Name, Organization Name, Location and Miscellaneous (Date, Time, Number, Percentage, Monetary expression, Number expression and Measurement expression.

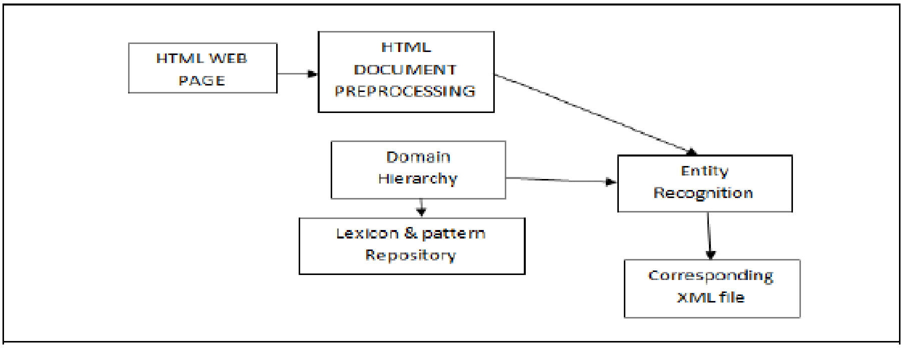


Fig. 2. HTML to XML Conversion

Figure 2 shows the general framework for converting HTML document to XML using Name Entity concept this technique is derived from [5].

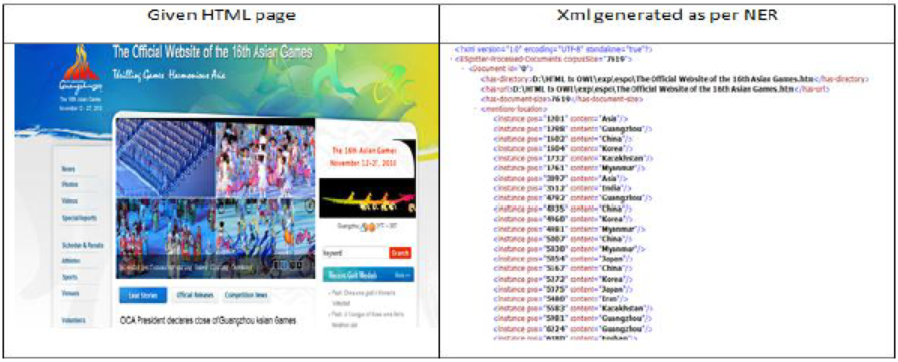


Fig. 3. Output of the conversion

Figure 3 shown the output of XML creation of the given website which is relevant to Asian games. The concern entity relation XML for Organiztion is given below:

```
<mentions-organization>
      <instance content="Guangzhou Online News Centre" pos="5954" />
      <instance  content="Guangzhou  Asian  Games  Organising  Committee"
pos="7257" />
      <instance content="Spectator Services" pos="393" />
      <instance content="Media Services" pos="412" />
      <instance content="Olympic Council" pos="1198" />
      <instance content="Press Conferences" pos="3112" />
      <instance content="The Radio Management" pos="5807" />
      <instance content="News Coverage Tour" pos="5977" />
      <instance content="Media Friends'" pos="5983" />
   </mentions-organization>
```

5 XML to RDFS/OWL Conversion

The main work on this phase of the project is the conversion of XML to OWL/RDFS. Thus through this conversion we provide semantic to the web page. As the converstion is take over automatically we need same format of well formatted XML file, that's the reason we use the generalized NER technique for XML conversion which provide same entity name tag. With this how we can convert is the main focus of this work. The syntactic rule of XML is converted to its concern Semantic via two main technique. One is Syntatic Analysis, where without seeing what inside with the use of XML's XSD files we can add the RDFS of that page [4]. The another approach is Semantic Analysis where we process the XML using Natural language processing and create the Ontology and Schema of the web page [7].

5.1 Syntactic Analysis

Here we generally map the XSD element [4] and convert to OWL element for the mapping the strategy shown in Table 1 is used.

Table 1. XSD to OWL

XSD	OWL
Xsd:elements,containing other elements or having at least one attribute	Owl:class,coupled with owl:ObjectProperties
Xsd:elements,with neither sub-elements nor attributes	Owl:DatatypeProperties
Named xsd:complexType	Owl:class
Named xsd:SimpleType	Owl:DatatypeProperties
Xsd:minOccurs,xsd:maxOccurs	Owl:minCardinality,owl:maxCardinality
Xsd:sequence,xsd:all	Owl:intersectionOf
Xsd:choice	Combination of owl:intersectionOf, owl:unionOf and owl:complementOF

An Ontologies main element are the owl classes, Object property, Data Property and all those constrain and cardinality element. Here as shown in Table 1 the XML element is converted. The main drawback of this approach is that some time an irrelevant data element would be tagged in OWL may produce irrelevant output.

5.2 Semantic Analysis

In this analysis the RDFS/OWL is generated using Natural Language Processing techniques[9]. For a Semantic Web page we have to create both RDFS and OWL. RDFS which Resource Descriptor Framework is a kind a rules and logic regarding the content on the page. A human identifies and analysis any intelligent information only via logical reasoning. Likewise RDF produce logic to the web page. OWL, Web ontology language used to produce the ontology of the web page with this only we will have the whole conceptual idea of any general concept we can give accurate results.

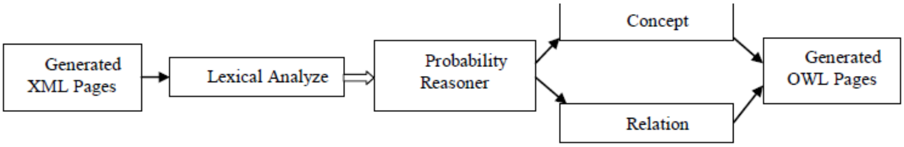


Fig. 4. OWL Generator

Figure 4 show a general framework for generated OWL from the generated XML. To analyze the content of XML we uses Lexical Analyzer which analyze each entity XML tag and represent whether they are noun or verb or any other verbal notation. Once analyzed we use the concept of Probability Reasoner to determine the concept and relationship between them as ontology is of considering the concept and their relationship between them. We use Probability as they used to handle uncertainty with the help of deductive logic i.e using a set of hypothesis for reasoning. The primary relationship between the concept is to be identified are “is-a” and “part-of” relation. To identify this kind of relationship [8] to create a full structured ontology we have to determine the Hyponym and Meronym of the identified tag. There is an automatic way to determine and extract the Meronym with the following linguistic pattern

1) `Such NP as NP,*(or|and) NP`
2) `NP, NP* or other NP`
3) `NP, including NP, or|and NP`

With this concept we can create an Ontology. The RDFS can be created using the Lexical Analyzer the frame work is shown in Figure 5, For the mapping the concept shown in Table 2 can be referred

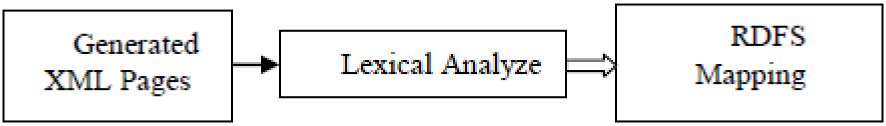


Fig. 5. RDFS Generator

Table 2. XML to RDF

XML'S POS	RDF COMPONENTS
Verb(phrase)	Predicate
Noun(phrase) denoted at first	Subject
Noun(Phrase) denoted after predicate	Object
Noun with adjective (phrase)	Subclass of the original noun

6 Conclusion

If we have knowledge about what we are searching for, we can easy retrieve the desire information. The main drawback in information retrieval procedure in web technology is that the technology doesn't know the semantic and syntax of what the user searching for. This gives birth to the Semantic Web Technology. In this paper we deal with the reusability technique of using converting the available HTML pages to an Ontology enriched Semantic web page. With the successful of this work we would like to design a Semantic search Engine [14][15][16] for efficient information retrieval.

References

1. Nagarajan, G., Thyagharajan, K.K.: A Novel Image Retrieval Approach for Semantic Web. International Journal of Computer Applications (January 2012)
2. Nagarajan, G., Thyagharajan, K.K.: A Survey on the Ethical Implications of Semantic Web Technology. Journal of Advanced Reasearch in Computer Engineering 4(1) (June 2010)
3. Minu, R.I., Thyagharajan, K.K.: Evolution of Semantic Web and Its Ontology. In: Second Conference on Digital Convergence (2009)
4. Bohring, H., Aure, S.: Mapping XML to OWL ontologies (2004)
5. Zhu, J., Uren, V., Motta Espotter, E.: Adaptive named Recognition for web browsing (2004)
6. Esmaili, K.S., Abolhassani, H.: A Categorization scheme for semantic web search engines (2005)
7. Hassanzadeh, H., Keyvanpour, M.R.: A Machine learning based analytical framework for semantic annotaion requirements. International Journal of Web and Semantic Technology (2011)
8. Abeyruwan, S.W.: Prontolearn: unsupervised lexico semantic ontology generation using probabilistic methods. These of universtiy of miami (2010)

9. Lenci, A., et al.: NLP based ontology learning from legal texts. A case study (2006)
10. Tijerino, Y.A., et al.: Towards ontology generation from tables. Kluwer academic publishers (2004)
11. Benslimane, S., et al.: Towards ontology extration from data intensive web sites: An html forms based reverse engineering approach. International Arab Journal of Information Tecnology (2006)
12. Mukhopadhyay, D., et al.: A New semantic web services to translate HTML pages to RDF. In: Int. Conference of IT (2007)
13. Hwangbo, H., et al.: Reusing of information constructed in HTML document: a conversion of HTML to OWL. In: Int. Conference on Control, Automation and Systems (2008)
14. Minu, R.I., Thyagharajan, K.K.: Automatic image classification using SVM Classifier. CiiT International Journal of Data Mining Knowledge Engineering (July 2011)
15. Minu, R.I., Thyagharajan, K.K.: Scrutinizing Video and Video Retrieval Concept. International Journal of Soft Computing & Engineering 1(5), 270–275 (2011)
16. Nagarajan, G., Thyagharajan, K.K.: A Machine learning technique for Semantic Search Engine. In: ICMOC NI University will publish in Elsevier Procedio (April 2012)

Two-Stage Rejection Algorithm to Reduce Search Space for Character Recognition in OCR

Srivardhini Mandipati, Gottumukkala Asisha, S. Preethi Raj, and S. Chitrakala

Department of Computer Science and Engineering, Easwari Engineering College,
Chennai, India

Abstract. Optical Character Recognition converts text in images into a form that the computer can manipulate. The need for faster OCRs stems from the abundance of such text. This paper presents a Two-Stage Rejection Algorithm for reducing the search space of an OCR. It is tacit that the reduction in search space expedites an OCR. Preprocessing operations are applied on the input and features are extracted from them. These feature vectors are clustered and the Two-Stage Rejection Algorithm is applied for character recognition. With about the same character recognition rate as other OCRs, an OCR reinforced with the Two-Stage Rejection Algorithm is considerably faster.

Keywords: Optical Character Recognition, Feature Extraction, K-means.

1 Introduction

Optical character recognition has been an active area of research for many decades. The fact that OCRs have the potential to simplify data entry in the future adds value to research in this area. OCRs use various pattern matching techniques for character recognition. Most OCRs typically use classifiers like SVM or neural networks for character recognition. The training process for these classifiers is time consuming. Moreover, with an increase in the number of classes, the comparisons made increases and consequently the time taken for character recognition increases. Hence, they cannot be easily extended to recognize characters from additional languages. The proposed system uses a structural approach as opposed to statistical approach for feature extraction. The strength of the structural method over the statistical one is its representation of a pattern that is similar to the way human perceive it. The structural features help retain the local shape description of the characters. Like all other OCRs, any image undergoes preprocessing. Additionally, the dataset is clustered and a Two-stage Rejection Algorithm is applied to it to reduce the search-space for character recognition. A considerable increase in the performance was observed during the experimentation.

2 Related Works

Numerous works have been carried out in the field of OCR. When an OCR is being extended to recognize characters from multiple languages, the dataset increases which

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 215–223.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

will considerably increase the number of comparisons required to recognize a character. This is all the more true when a single document contains characters from different languages. In our paper, we focus on the reduction of the search space for character recognition. This is done by clustering the training dataset and reordering the clustering.

Weijie Su and Xin Jin [1] propose a hidden Markov model with parameter-optimized K-means clustering for handwritten character recognition. Here, they improve K-means clustering by considering the influence of neighboring pixels and different weights of pixels in different places. This model aims at improving the average accuracy of HMM with K-means clustering for handwriting characters recognition.

Karthik Sheshadri et al. [2] address the problem of Kannada character recognition, and propose a recognition mechanism based on K-means clustering. Here they propose a segmentation technique to decompose each character into components from 3 base classes, thus reducing the magnitude of the problem. They have also used probabilistic and geometric seeding as heuristics to ensure uniformity of centroids from the extracted character with the centroids in the training database.

Mu-King Tsay, Keh-Hwashyu, Pao-Chung Chang [3] designed a feature transformation module to extract discriminative features from the input scanned document to enhance the recognition performance. The initial feature transformation matrix is obtained by using the Fisher's linear discriminant function. Template matching with minimum distance criterion recognizer is used and each character is represented by one reference template. These reference templates and the elements of the feature transformation matrix are trained by using a generalized learning vector quantization algorithm.

B. Vijay Kumar, A. G. Ramakrishnan[4], a neural network in which the only hidden layer in the network applies non-linear transformation from input space into hidden space, called a Radial Basis Function Network is trained with wavelet features using K-means. It is used along with the subspace projection method, which represents higher dimensional data in lower dimensional space, to recognize printed Kannada characters.

Premnath Dubey, Wasin Sinthupinyo [5], the normalized image is split into zones. The histogram of contour direction in each of these zones is used to represent a feature vector. These feature vectors are clustered to create a feature template and template matching is used for character recognition.

Igor Kleiner et al.[6] proposed a rejection based approach which follows the paradigm of property testing which quickly determines whether an input image satisfies a predefined property. Assuming that most of the input images are far from the satisfying condition and rejecting them considerably reduced the search space.

OCRs in general have multiple templates for the recognition of a single character. This leads to a reasonably large search space for character recognition, hence slowing down the process.

3 System Description

The proposed system minimizes the use of templates and also uses clustering to narrow down the search space. This results in considerably lesser comparisons for character recognition.

The architecture of the proposed OCR system is shown in Fig.1.

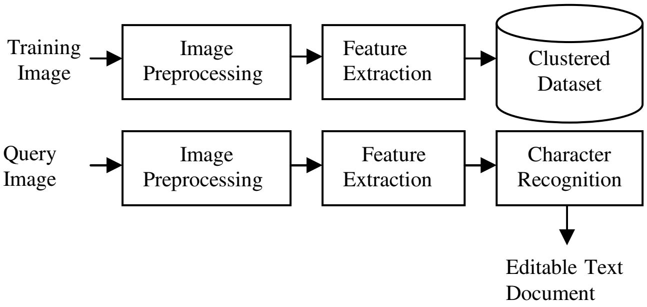


Fig. 1. Functional architecture of the system

First, we acquire a scanned image of a printed document. We assume the input image to be devoid of skews. The input image then undergoes image preprocessing stage which consists of a series of operations namely binarization, noise removal, thinning, line and word segmentation and finally, character segmentation. The result of this stage is a segmented character upon which further processing is done. The next step is feature extraction. Here, desired features are extracted from the character and are represented in the form of feature vectors. These feature vectors are then clustered and this makes up the dataset for character recognition. During character recognition, a two stage rejection algorithm is applied to obtain the closest match for the query image. The character hence recognized is displayed on an editor.

4 Image Preprocessing

A scanned paper containing printed text is fed as input to the OCR. The input is assumed to have no skews and to contain only printed text without images. The desired features of the input are first enhanced to support further processing, most importantly feature extraction. This is done by a series of low-level operations called image preprocessing. The sequence of preprocessing operations is shown in Fig. 2. The operations are, in order, binarization, noise removal, thinning and segmentation.

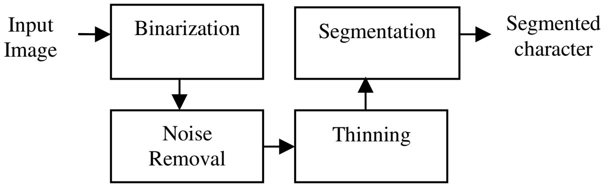
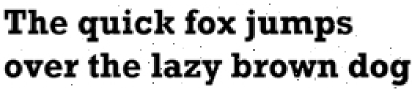


Fig. 2. Image preprocessing

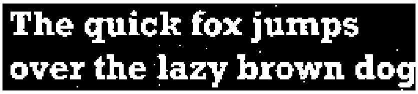
The transformation of an input image when it undergoes preprocessing is shown from Fig.3 to Fig.8.



The quick fox jumps
over the lazy brown dog

Fig. 3. Input image

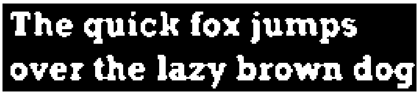
The first preprocessing step is image binarization. Here, the input image, which is either a color or a grayscale image, is converted to its binary equivalent. Otsu's algorithm is the most commonly used algorithm for binarization and is employed here as well. After binarization, the image is inverted. This is done to aid with thinning. However, this can be done at any later stage but it must be done before thinning.



The quick fox jumps
over the lazy brown dog

Fig. 4. After binarization and inversion

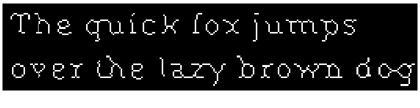
The next step is noise removal. Here, unwanted noise is removed. Noise is introduced into the input through a variety of ways like the quality of the scanner and the age of the paper. The most common noise present here is pixel noise. Pixel noise is characterized by the presence of random pixels throughout the image. A median filter is employed to remove this noise.



The quick fox jumps
over the lazy brown dog

Fig. 5. After noise removal

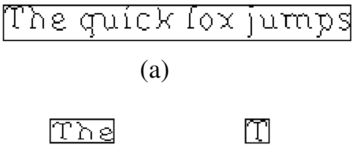
Noise removal is followed by thinning. Here, the image is represented using the lowest possible bits without affecting the general shape of the image. The image is inverted after thinning.



The quick fox jumps
over the lazy brown dog

Fig. 6. After thinning

The last preprocessing step that an image goes through before feature extraction is segmentation. The image goes through three kinds of segmentation in a sequence, namely line segmentation, word segmentation and character segmentation. In line segmentation, the lines of text in the image are separated and the image is represented as a series of lines. This is done by horizontal projection profile.



(a)

(b) (c)

Fig. 7. (a) Line (b) Word (c) Character Segmentation

Line segmentation is followed by word segmentation. Here, the words in the lines are separated and the lines are represented as a series of words. This is done by using vertical projection profile. Here, the threshold is set for the gap between two consecutive words. The final segmentation step is character segmentation in which words are represented as a series of characters. This is also done by using vertical projection profile but with a smaller threshold value. Further processing is done on these individual segmented characters.

5 Feature Extraction

The features from character images are extracted through an approach which is based on a structural feature model. A data structure referred to as a feature template is created from the feature vector extracted. The feature that is extracted is the contour of the character which is represented by the feature template. The contour direction of the character is obtained by a combination of edge detection algorithms such as the x gradient and y gradient from Sobel operator (1).

$$G_x = z_7 + 2z_8 + z_3$$

$$G_y = z_3 + 2z_6 + z_9$$

$$\alpha(x,y) = \arctan\left(G_x / G_y\right) \tag{1}$$

$\alpha(x,y)$is the direction of the gradient on the character image contour. Since the input images may vary in size, they are initially normalized to a 32 x 32 pixel image. The normalized image is then divided into 4 x 4 grid or 16 zones. The contour direction helps to preserve some of the local information about the shape of the character. The character image is further divided into smaller regions or zones to obtain higher level of description. The contour direction in each zone is divided equally into 6groups and is then collected in a histogram. This 6 bin histogram is the feature vector which will be used as a descriptor of the shape in a particular zone. The output of feature extraction gives a sequence of 16 zone labels for each character. This method is generalized and can be used for a new set of characters.

6 Clustering

The extracted feature vectors are clustered using K-means. K-means is a flat clustering algorithm that groups a given data set into a predefined number of clusters. The objective of this clustering is to minimize the average squared distance of the cluster objects to the cluster centers.

The feature vectors of the characters of the training dataset are clustered. Euclidean distance is the measure used for clustering. The Euclidean distance, D(p, q) between two features vectors q and p is defined in (2)

$$D(p,q)=\sqrt{\sum_{i=0}^{n}(q_i - p_i)} \tag{2}$$

where p = (p_1, p_2,...,p_n) and q = (q_1, q_2,..., q_n) are n-dimensional feature vectors

The optimal number of clusters was found out to be 4 through trial and error. However, this number changes with a change in the number of characters in the dataset.

After clustering the dataset, the order of the clusters is modified. The order of the elements in a cluster is denotes when an object was placed in the cluster and not by how close an object is to the cluster center. Hence, the order of the cluster is explicitly changed to denote the position of the object with respect to its cluster center. The rejection algorithm is applied after reordering the clusters.

7 Rejection Algorithm and Character Recognition

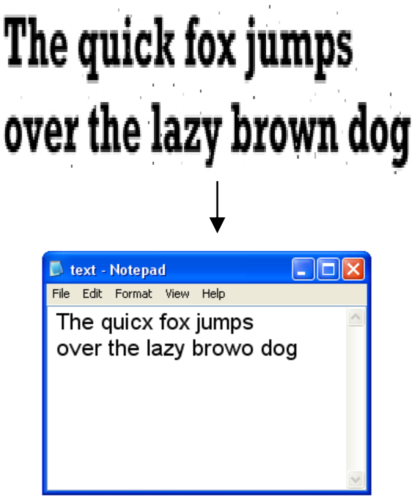
A two-stage rejection algorithm is proposed, that aids with character recognition. In the first stage, the feature vector of the query input image and all the cluster centers are compared. The clusters whose centers are far from the input feature vector are rejected. At this stage only the cluster whose center is relatively close to the input feature vector is selected and the remaining clusters are ignored. The second stage performs rejection within the identified cluster. The clusters are reordered or sorted according to the position of the cluster member from the cluster center. In other words, the clusters are sorted in ascending order of the distances from the center. The query feature vector is initially compared with both, the cluster center and the cluster member farthest from the cluster center. If the query feature vector is closer to the cluster center, it is compared with the remaining cluster objects starting from the cluster center to the next closest member in ascending order till it reaches its closest match. On the other hand, if the query feature vector is closer to the farthest cluster member, then it is compared with the remaining cluster objects starting from this farthest member to the next closest member in descending order till it reaches its closest match. The cluster member hence recognized corresponds to the desired character and this character hence recognized will be displayed on a text editor.

Algorithm: Two-stage Cluster Rejection algorithm

```
Input:
        fv: feature vector of the input image
        idx: clustering of the dataset
        ctr: array of centroids in idx
        k: number of clusters
        last:index of last centroid in the cluster
Output:
        res: feature vector closest to the feature vector of the input image

Initialize s, which represents the cluster number of the cluster containing the
centroid that is closest to the feature vector of the input image, to 1

FOR i = 2 to k
        IF D(fv, ctr[i])<D(fv, ctr[s]) THEN
                s = i
        ENDIF
ENDFOR

IF D(fv, idx[s][last])>D(fv, ctr[s]) THEN
        i = 1 and res = ctr[s]
        WHILE D(fv, res)>D(fv, idx[s][i])
                res = idx[s][i++]
        ENDWHILE
ELSE
        i = last and res = idx[s][last]
        WHILE D(fv, res)>D(fv, idx[s][i])
                res = idx[s][i--]
        ENDWHILE
ENDIF
```

8 Results and Discussion

In order to accurately evaluate the performance of the proposed system, feature vectors were extracted from 52 characters of the English alphabet (uppercase and lowercase). After training, these feature vectors were clustered. Testing was done using images of text in 5different fonts. The fonts used for testing the OCR were Arial, Times New Roman, Verdana, Comic Sans MS, and Courier New. The recognition rate varied between fonts and was the highest for Arial.



The quick fox jumps
over the lazy brown dog

The quicx fox jumps
over the lazy browo dog

Fig. 8. Input image and the corresponding output text

The recognition rate was computed using the formula below and the results are tabulated in Table 1.

$$recognitionRate = \frac{P}{N} \times 100 \tag{3}$$

where P denotes the number of characters recognized correctly and N denotes the total number of characters in the tested image.

Table 1. Recognition Rates For Tested Fonts

Font Style	Recognition Rate
Arial	99.05%
Times New Roman	98.21%
Verdana	98.31%
Comic Sans MS	97.52%
Courier New	97.49%

The average character recognition rate was 98.11%. Recognition was tested on an estimate of about 2000 split between different images. Only images which contain a single font are considered for experimental analysis.

Furthermore, the time taken to convert the text images to its equivalent editable version is lesser than other OCRs.

9 Conclusion and Future Work

In this paper, a Two-Stage Rejection Algorithm is proposed to expedite an OCR. Instead of using classifiers, this algorithm is used along with clustering to reduce the search space for faster character recognition. Encouraging experimental results have been observed. Future work will address extending Two-Stage Rejection Algorithm for a multilingual OCR.

References

[1] Su, G., Jin, X.: Hidden Markov Model with Parameter-Optimized K-means Clustering for Handwriting Recognition. In: International Conference on Internet Computing and Information Services, pp. 435–438 (2011)

[2] Sheshadri, K., Ambekar, P.K.T., Prasad, D.P., Kumar, R.P.: An OCR system for Printed Kannada using K-means clustering. In: International Conference on Industrial Technology, pp. 183–187 (2010)

[3] Tsay, M.-K., Keh-Hwashyu, Chang, P.-C.: Feature Transformation with Generalized Learning Vector Quantization for Hand-Written Chinese Character Recognition. IEICE Transactions on Information & System E82-D (1992)

[4] Vijay Kumar, B., Ramakrishnan, A.G.: Radial Basis Function And Subspace Approach For Printed Kannada Text Recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, p. V-321-4 (2004)

[5] Dubey, P., Sinthupinyo, W.: New Approach on Structural Feature Extraction for Character Recognition. In: International Symposium on Communications and Information Technologies, pp. 946–949 (2010)

[6] Kleiner, I., Keren, D., Newman, L., Ben-Zwi, O.: Applying property testing to an image partitioning problem. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(2) (2011)

[7] Mohanty, S., Dasbebartta, H.N., Behera, T.K.: An Efficient Bilingual Optical Character Recognition (English-Oriya) System for Printed Documents. In: Seventh International Conference on Advances in Pattern Recognition, pp. 398–401 (2009)

[8] Trier, O.D., Jain, A.K., Taxt, T.: Feature Extraction Methods For Character Recognition–A Survey. Pattern Recognition 29, 641–662 (1995)

[9] Vuori, V., Laaksonen, J.: A Comparison of Techniques for Automatic Clustering of Handwritten Characters. In: 16th International Conference on Pattern Recognition, vol. 3, pp. 168–171 (2002)

Transforming Election Polling from Electronic Voting to Cloud as a Software Service in India

P. Vidhya

School of Management and Computer Applications, SSN College of Engineering,
Chennai, Tamil Nadu, India
carolinevidhya@gmail.com

Abstract. Today election polls have been re-engineered using electronic voting system in India. The system does come with several drawbacks and huge expenses including cost for security of the machine, transport, etc. In-fact customizing the electronic firmware to booth requirements is still a manual procedure and takes more effort and cost than modifying the same in a software application. In this paper we propose a model to transform the electronic polling to a cloud polling system to bring less cost, scalability, save time, easier modification, centralized control, speedy results, voter's convenience and running a smooth poll. We also discuss the issues and possible solutions in moving towards the cloud polling system.

Keywords: Electronic Voting System, Cloud Polling System.

1 Introduction

Cloud computing is a recent concept that is receiving an increasing attention in both the public and the private sectors. Cloud Computing is changing the traditional way of how hardware and software providers sell their products to their customers. This is another kind of business in which the product is not sold as such, but a service is sold. Mainly, this platform allows sharing resources as a service. These resources include infrastructure (IaaS), software (SaaS), platform (PaaS), and data storage (dSaaS). Cloud computing has the potential to transform the use of government services from being manual procedure to innovative technology solutions that are integrated and efficient. Gartner, defines cloud computing as "a style of computing where massively scalable IT-related capabilities as a service using internet technologies to connect multiple external customers" [1]. Infact NASSCOM report conveys that in the next few years Indian IT companies are to venture into major cloud projects and is providing datacenters [2].

One of the major government electronic initiatives in India is the polling system enabled through electronic voting machines introduced in 2004 general elections. In this paper we propose to overcome the drawbacks of electronic polling system by re-engineering into cloud polling system. India being predicted to be one of the good cloud hubs in the next few years makes it feasible for the cloud polling system to be introduced seamlessly.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 225–232.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

2 Literature Survey

To automate government services with current technology is a challenging task. Be it the procurement of hardware, vendor selection or storage capacity, all of these things can present grim bureaucratic hurdles for public administrators. Vivek Kundra, the U.S. Federal Chief Information Officer, said that “by using cloud computing services, the Federal Government can gain access to powerful technology resources faster and at lower costs. Ultimately, this will allow the Government to better serve the American people and focus on mission-critical tasks instead of on purchasing, configuring and maintaining redundant infrastructure” [3].

Currently there are several government initiatives to move government services on cloud. United Kingdom’s government based cloud computing initiative is called G - cloud that spans all the units and departments within the government [4]. The goal of G-cloud is to migrate from government’s services to the web. Open Cirrus open cloud project is a research cloud initiative of InfoCommunication Development Authority (IDA), Singapore [5]. In Saudi Arabia, AWAL cloud is being introduced through the telecommunication companies aiming at providing IaaS services [6,7]. The US Army is using cloud computing as a means to enhance it recruitment processes [8]. The national government of Japan has taken an initiative to create the Kasumigaseki private cloud for the purpose of hosting all the government’s computing resources and services [9]. The Kasumigaseki cloud facilitates sharing of information and promoting higher levels of standardization and control over the Japanese government’s technology resources. The private cloud is expected to promote and revive Japenese economy after the recent disaster due to natural calamities. The Canadian Government is to move Government Services on Cloud [10]. Government organizations in City of Miami, City of Edmonton, Federal Government and NASA have adopted cloud computing solutions to improve their service delivery and to reduce their IT costs [11].

To implement cloud solutions in Indian government, a tri-partite agreement was signed, in June 2010, by Kerala State IT Mission (KSITM), Indian Institute of Information Technology and Management-Kerala (IIITM-K), Chennai (C-DAC) [12]. In the year June 2010, the state of Jammu & Kashmir successfully utilized computing services offered by the state of Madhya Pradesh (MP) to move citizen services online within 60 days at zero initial cost [13]. The technology was successfully completed, and the state of MP received revenues from Jammu on a per transaction basis.

3 Existing System - Electronic Polling in India

Currently the electronic voting machines in India are designed by Election Commission of India in collaboration with Bharat Electronics Limited, Bangalore and Electronics Corporation of India Limited, Hyderabad.

An Indian voting machine has a Control Unit which is with the Polling Officer and a Balloting Unit which is placed inside the voting compartment.

The benefits of Electronic Voting Machines include [14]:

- The cost per EVM is Rs.5500.
- It will be easier to transport the EVMs compared to ballot boxes.
- Vote-counting is faster using EVM and the results are declared within 2 to 3 hours as compared to 30–40 hours in a ballot-paper system.
- In India, where illiteracy rate is high, voters can cast votes using EVM in a user friendly manner than using a ballot paper system.
- Bogus voting can be greatly reduced as an EVM is programmed to record only five votes in a minute.
- In case an EVM goes out-of-order then, the Election Officer, in-charge of the polling booth, can replace a spare EVM.
- Votes are recorded memory of the Control Unit so that even if the EVM crashes polling data can be recovered avoiding the necessity of a re-polling.
- The Control Unit can store the result in its memory for 10 years and even more.
- Invalid votes can be avoided by use of EVMs.
- Since EVMs work on a 6-volt battery, there is absolutely no risk of any voter getting an electric shock.

The drawbacks of Electronic Voting Machines include [14]:

- The number of candidates per EVM is limited to 64.
- With the increase in population more cost must be spent for the EVM.
- Cost of reconfiguring the EVM is more.
- Transporting EVM in rugged environment and bad climatic conditions is difficult.
- The life of any electronic gadget is short.
- Security threats like Clip-on memory manipulator attack and dishonest display attack are possible [15].
- The number of voters that can be maintained in an EVM is 3840 only. In such cases paper ballots are the only alternatives.
- Configuring the candidate list is a manual process and often confuses illiterate voters when excess of free options are there.
- Malpractices are possible during EVM randomization.
- Handling physical problems in EVM during voting and power supply solutions are an issue.
- Protection is needed during transport of EVM.
- After the polls protection must be given to EVM from natural calamities and deliberate intruders till the counting day.
- Scalability of EVM to the growing population is a major drawback.
- There is no logging of votes to track possible fraud.
- Though booth capturing can be prevented destruction of EVM can disrupt peaceful poll.

- On the day of the election the physical presence of the voter at the exact polling booth becomes mandatory which leads to plenty of votes wasted without casting by voters who cannot make up to the poll booth.
- There are more security demands.
- EVM being standalone makes integration of polling results time consuming and a herculean task.
- A wrong announcement based on polling results of other constituencies is possible.
- A huge segment of votes are not cast due to voters not really propelled to going to the polling booth by laziness.

4 Proposed System – Cloud Polling System

In this paper we propose a system to migrate from standalone EVM to cloud polling solution. Though the polling system has been currently re-engineered with electronic voting system the electronic expenses are at a higher rate. Moreover achieving scalability and customization for the growing population becomes difficult. In fact in the 2011 polls there was a 60 per cent hike to the existing poll expense limits in certain States [16]. Poll expense limits are not uniform across the country and vary according to the size, demographics and other factors of the States and Union Territories.

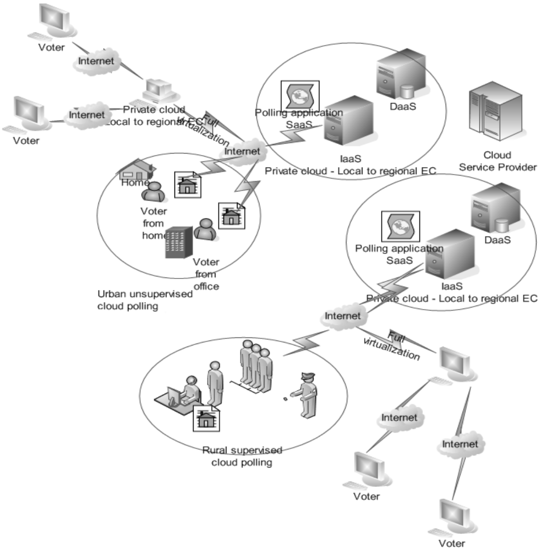


Fig. 1. Cloud Polling System

Replacing polling application as a software service would reduce the cost greatly. The polling service can run on state wide datacenters and virtualized to constituencies reducing cost expenses drastically. The infrastructure services (IaaS) like server and database (DaaS) can be deployed as private cloud.

Voters can access the polling service on the day of the election irrespective of their current location via internet. To support voters below the digital divide index supervised polling is proposed in special limited booths while other voters can go on an unsupervised poll from home or office or any internet cafes.

Issues to be handled:

- Bandwidth
- ISP reliability
- Security and key management
- Upstream and downstream rates
- Lack of Cloud computing and government legislative norms
- Lack of Cloud computing standards

The key challenge in CPS is how the password is distributed. For every registered voter (credentials verified) whose name appears in the voter list a password is delivered in person by post only if the receiver has voter id proof. The voter has to sign for the receipt of the password only if the password slip is not tampered. The passwords have to be changed by the voter just one day before the poll by initial sign-in to the poll application using voter id as account name and password sent by post. The changing of the password cannot be done on the day of the poll or any day other than the day prior to the polling date. During the poll period the voter has to sign-in with voter id, other credentials including Adarsh-id or ration card number and the newly altered password. For those opting supervised polling a fingerprint authorization is required instead of password at the polling booth.

The best way to access cloud is through internet. Using basic public internet to access the cloud polling application provides easy accessibility and highly fault tolerant. Internet provides SSL based HTTP over secure socket layer (https) encrypted access cloud poll application with confidentiality. Anyhow internet lacks end to end QOS and there is a probability of poor response over high latency connections. Therefore we suggest an optimized internet overlay that provides enhancement on the provider's cloud where the voters access the cloud via public internet. Enhancements at POP provide optimized real time routing.

To handle the issues of reliability and uptime redundancy virtualization should be done by the service provider. The virtualized polling cloud will be a good solution to reduce capex and opex of Election Commission of India.

5 Evaluation of Cloud Polling System

Cloud polling solution offers the following benefits:

- CPS offers scalability to the expanding voting population
- The drawbacks limited candidate list and voter strength per EVM can be overcome using CPS.
- Unlike EVM which is standalone CPS is networked and reduces the infrastructure and cost requirements greatly. CPS can be easily upgraded.
- To fit the candidate requirement no manual intervention is needed instead a simple software service provides the solution.
- To audit the transparency of polls logs are good solution not available in EVM.
- Centralized monitoring of transparent poll is possible by ECI than in a distributed standalone solution.
- Does not require huge security in every local ward. Anyhow security is needed to monitor the internet service provider side.
- Security for the EVM in the centers until counting is not needed.
- The voter need not be physically present at the polling booth.
- Polling day need not be declared as a holiday by easy access to polling service can be done from workspace of home using internet.
- The resources can be virtualized to reduce cost.
- Counting can be done fast.
- Reliability is achieved by redundant resources so that in case of failure a redundant resource continues the polling than postponing the poll to some other day.
- Booth capture and other violent incidents can be prevented.
- In this solution we propose that users with internet awareness go on an unsupervised poll while for the rest there is a supervised poll in poll booths running polling service over optimized internet overlay.
- CPS will also improve the percentage of voting population through ease of access.

Cloud polling solution potential risks:

- All the virtual machines of any virtualized resource are at equal risk if being hacked.
- There is a lack of cloud regulations and standards making cloud adoption risky.
- Ownership of data must be strategically regulated.
- Cloud polling system should be immune to web security threats including phishing, denial of service, etc.
- For cloud computing to efficiently materialize services should be deployed over an appropriate platform and have large data centers to support large scale services.
- There are other big issues like poor electricity supply; internet reliability and low internet penetration [17, 18] that are to be dealt in the transformation process.

6 Conclusion

Adoption of Cloud polling system would transform the entire outlook of Indian polling from a poll-from-booth scenario to a poll-from-anywhere scenario. A centralized monitoring of smooth polls is a convenience of the new system. Though cloud is at its infancy today major Indian IT giants are competing to migrate government services over the cloud. The total cloud computing market in India is expected to reach $1,084 million by 2015. Cloud Polling System would definitely out perform the current electronic voting system in India in terms of expense and transparency.

References

1. Smyth, P.: Cloud computing – A strategy guide for broad level executives. Kynetix Technology Group report (2009)
2. NASSCOM Perspective 2020: Transform business, Transform India, Knowledge partner McKinsey and Company. NASSCOM, Magnum custom publishing (April 2009)
3. Kundra, V.: Official website of U.S. government, July 1 (2010), `http://www.cio.gov/pages.cfm/page/Vivek-Kundra-Testimony-on-Cloud-Computing`
4. Suffolk, J.: Data Centre Strategy, G-Cloud & Applications Store for Government (ASG) Programme – draft Cabinet office (January 2010)
5. Campbell, R., Gupta, I., Heath, M., et al.: Open CirrusTM Cloud Computing Testbed: Federated Data Centers for Open Source Systems and Services Research. In: Proceeding of HotCloud 2009, July 15. USENIX Association, Berkeley (2009) ACM Digital Library
6. Saudi Gazette (2011), `http://www.saudigazette.com.sa/index.cfm?method=home.regcon&contentID=20111204113143` (December 4, 2011)
7. IT report 2010 On the Internet Ecosystem in Saudi Arabia, Communications and Information Technology Commission (June 18, 2011)
8. Kundra, V.: Federal cloud computing strategy, February 8 (2011)
9. Amrhein, D., Anderson, P., de Andrade, A., et al.: Cloud computing Use Cases, Licensed under Creative Commons Attribution-Share Alike 3.0 Unported License, October 30 (2009)
10. McEvoy, N.: Canada Cloud 3.0 - Building Canada's Digital Economy advantage through Cloud Computing. Wordpress (March 2011)
11. Tsaravas, C., Themistocleous, M.: Cloud Computing and Egovernment: A Literature Review. In: European, Mediterranean & Middle Eastern Conference on Information Systems 2011, Athens, Greece, May 30-31 (2011)
12. Economic Review, Kerala State Planning Board, ch. 21 (2010)
13. Julka, H.: Economic Times Bureau, June 24 (2010)
14. `http://en.wikipedia.org/wiki/Indian_voting_machines`
15. Prasad, H.K., Alex Halderman, J., Gonggrijp, R., Wolchok, S., Wustrow, E., Kankipati, A., Sakhamuri, S.K., Yagati, V.: Security Analysis of India's Electronic Voting Machines. In: Proceedings of 17th ACM Conference on Computer and Communications Security, CCS 2010, pp. 1–14. ACM, New York (2010) ISBN: 978-1-4503-0245-6

16. Sainath, P.: The Hindu, Chennai (2011), `http://www.thehindu.com/news/national/article1487696.ece` (February 25, 2011)
17. Measuring the Information Society 2011, International Telecommunication Union, Geneva (2011) ISBN 92-61-13801-2
18. Shrivastava, B., Abhichandani, T., Biswas, A., Thakare, M.: Group Report on Internet in India (I-Cube). Internet & Mobile Association of India, IAMAI (2011)

Area Efficient Architecture for Frequency Domain Multi Channel Digital Down Conversion for Randomly Spaced Signals

Latha Sahukar[1] and M. Madhavi Latha[2]

[1] Associate Professor,
Aurora's Technological and Research Institute (ATRI),
Parvatapur, Uppal, Hyderabad, AP- 500039, India
latha.sahukar4@gmail.com
[2] Professor,
HOD ECE Dept.,
JNTU College of Engineering, Hyderabad, India
mmadhavilatha@jntuh.ac.in

Abstract. A complete frequency domain processing based digital down conversion architecture is presented in this paper. The conventional complex NCO multiplication is achieved with direct spectrum rotation and various possibilities for frequency domain filtering are discussed. An FFT-IFFT based architecture is implemented in Xilinx Virtex-6 family XC6VLX240T FPGA platform and synthesis is verified. The overlap and add method at the output of IFFT is employed to avoid time domain overlapping. The results demonstrate highly optimized area implementation with respect to conventional DDC architectures. The synthesis results show that the developed core can work upto clock rates of 250 MHz while occupying only 10% of the FPGA slices.

Keywords: Frequency Domain Filtering (FDF), Digital down conversion (DDC), Sample rate conversion, non-cooperative communication, dynamic decimation.

1 Introduction

The software defined radio based receivers, which are designed to work in non-cooperative communication systems aim to intercept unknown signals and perform analysis on them. As the challenges in homeland security are growing rapidly and wireless devices being used for various types of anti-social activities, monitoring the whole spectrum and analyzing signals has become very important issue.

The challenge in dealing with such communication signals in a portable receiver architecture requires the usage of software radio principles. The basic architecture of such receiver is shown in below figure.

The wideband antenna system consisting one or multiple set of antennas is used to receive the signals of entire band of interest. The RF front end process the signal to provide sufficient gain and filtering. The super heterodyne receiver technique is used to convert the signal with band of interest to the fixed IF band. The sweeping synthesizer is used to sweep the entire band of interest in steps of IF bandwidth. The signal

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 233–241.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

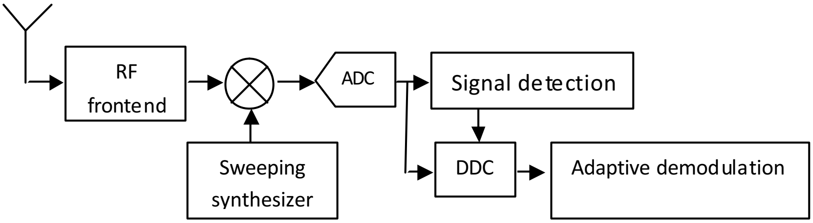


Fig. 1. Architecture of signal monitoring receiver

is digitized at IF stage using suitable high dynamic range ADC. The digital processing must realize the required functionality with the ADC output.

The DDC block shown in figure.1 performs the frequency shifting and sample rate conversion.

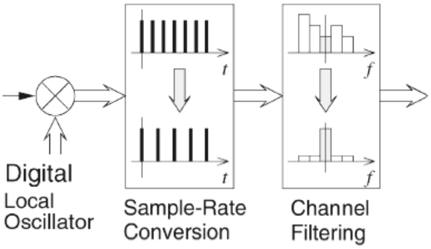


Fig. 2. Digital down conversion with channel filtering

The DDC requirement in monitoring application of non cooperative communication requires the following capabilities and architecture wise can be divided as shown in figure 2

(a) Must handle simultaneous signals which are spectrally apart.
(b) The signal detection to monitoring time must be minimized in order to capture the LPI (Frequency hopping) signals.
(c) Must be realized in smaller FPGAs (to result in low power) so that man portable systems can be realized.

The section 3 explains how the proposed architecture is designed to achieve all the above requirements.

The challenge in monitoring unknown signals over a wideband requires fast identification of presence of signal in the selected band and tuning the DDC dynamically. Several architectures [1][2][3][4] are proposed to achieve this requirement. The digital front end common to this kind of applications is discussed in [1] and proposes channelizer based architecture. The filter bank based channelizer architecture divides signal into multiple small bands with FFT and further allows decimation after complex NCO multiplication and half band filtering(HB). The architecture proposed doesn't result in full frequency domain implementation, rather it is mixture of both time and frequency domain. The channelizer based architecture is overhead for the signal monitoring application as described above.

The architecture given in [2] explains the modified architecture for handling arbitrary sample rate conversion. However this still implements the filter in time domain and requires filter coefficient storage and high speed implementation of MAC based on input sampling rates. This architecture could be suitable at the adaptive demodulation stage level, but at the digital front end this becomes inefficient.

Sampling rate conversion with fraction FFT is explained in [3], however this is very complex architecture to implement in FPGAs. The article given at [4] presents the basic idea of sample rate conversion in frequency domain. The approach presented in this paper extends this concept and evolves a full dynamic tuning based low area multi channel DDC. There are few FPGA IP cores used in defense systems [5][6] which are targeted for similar applications. The work is motivated to realize such architectures in low area implementation style.

Rest of the paper is organized in 4 sections. The section 2 elaborates the motivating factors for frequency domain DDC. The section 3 explains the frequency domain filtering operation. The section 4 presents the proposed architecture. The section 5 discusses on obtained results.

2 Motivation for Frequency Domain DDC

In this section the frequency domain implementation of all the DDC blocks are explained.

Multiplication with digital Local oscillator

The complex digital local oscillator shifts the signal in frequency domain[7]. Let a signal x(t) with spectrum X(f) is digitized, which results in sampled digital signal x[n] with spectrum $X_a(f)$

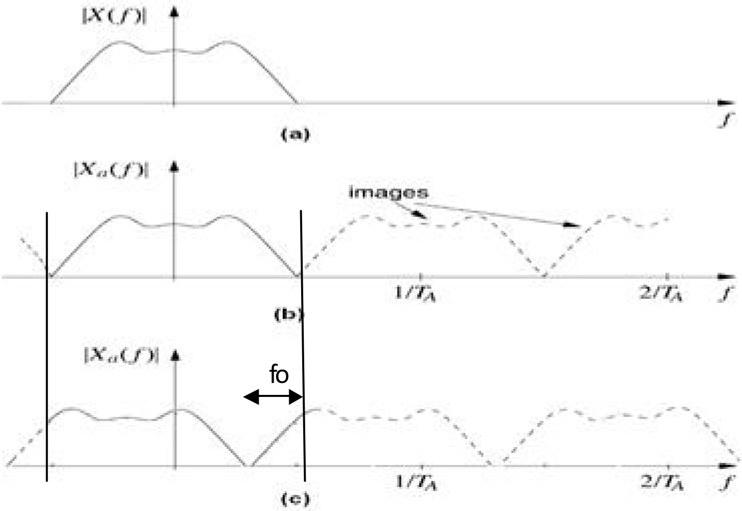


Fig. 3. Multiplying signal with $e^{-j2\pi fon}$

The figure 3 shows the effect of multiplying signal with negative complex exponential. This can be achieved by simply rotating the spectrum left. This avoids area consuming complex oscillator realization on FPGA. Similarly the multiplication with positive complex exponential signal can be achieved with right rotating the whole spectrum. The implementation level details of proposed rotation based complex oscillator multiplication are given in next section.

3 Proposed Dynamic Frequency Domain DDC

Frequency domain filtering

The change in sampling rate can cause either aliasing (in case of decimation) or imaging (in case of interpolation). To avoid that we need to perform the filtering operation over the band of interest and then do sample rate conversion.

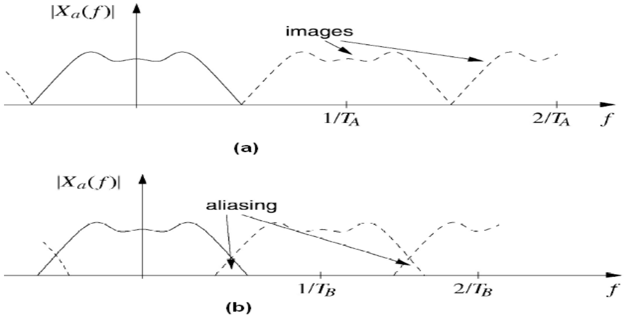


Fig. 4. Aliasing due to change in sampling rate

Filtering is usually performed with FIR filters, with area efficient polyphase architectures. The input signal is convolved with the filter coefficients and then sample rate conversion is performed. Doing same in frequency domain can be achieved with multiplication of input signal and filter coefficients in frequency domain. The inverse Fourier transform of this product can give the time domain signal[8].

Because the FFT provides the means to reduce the computational complexity of the DFT from order (N^2) to order $(N \log_2(N))$, it is often desirable to do FFT-based processing for DSP systems. As the FFT is most fundamental element in digital signal processing, lot of architectures were evolved to provide high speed and low area implementation. Even the computational cost of doing both FFT and IFFT in FPGAs is less than conventional methods

The DFT is a sampled version of the Fourier transform, so multiplying DFTs corresponds to circular convolution. Circular convolution can result in time-domain aliasing. As we want linear convolution, we must ensure time-limited input signals to avoid time-domain aliasing similar to band limiting to avoid frequency-domain aliasing.

Linear convolution with circular convolution

Consider a unit sample response h[n] with finite length P, and a signal x[n] of length L. The linear convolution h*x has length L+P-1. To avoid time-domain aliasing, we zero pad both sequences to at least length L+P-1, do FFT, multiply the transforms, then IFFT to get L+P-1 result. Let us consider filter response of length P, but assume input signal is of streaming type coming from ADC.

The signal is windowed into consecutive blocks of length L, pad each with zeros to length L+P-1, and FFT can be performed. Next add L-1 zeros to the P number of filter coefficients and perform FFT. Both the FFT outputs are multiplied bin by bin. The product is given to IFFT block to compute the time domain signal back.

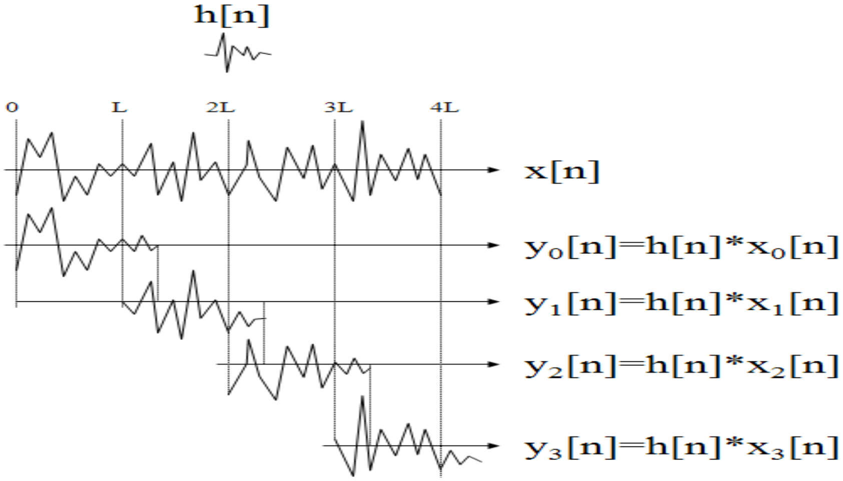


Fig. 5. Overlapp and add method

It is to be noted that the last P-1 output samples will overlap the start of the next block, and the overlapping points must be added to get the proper response. This is known as the overlap-add algorithm. The figure 5, explains the same.

4 Proposed Architecture

For the purpose of digital down conversion (DDC) either we can apply rectangle window or any selected window function as frequency domain low pass filter. The results shall be similar to window based FIR filter design. Obviously the rectangular window based implementation is low area solution as it results in only multiplying with 0 or 1. Even in other function case for all the channels only one memory set is used for storing the window function.

The proposed architecture for complete frequency domain digital down conversion is shown in below figure.

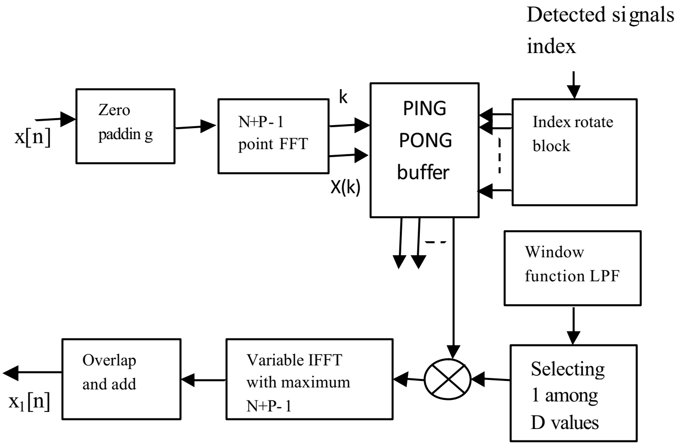


Fig. 6. Proposed architecture

The input x[n] is zero padded with P-1 zeros and FFT is performed with N+P-1 point FFT core. For simplicity the filter length is considered as N+1 hence the FFT size is 2N

$$X[k]=\sum_{n=0}^{2N-1} x[n]e^{-j\frac{2\pi nk}{2N}} \tag{1}$$

The output of the FFT is stored in dual buffer memory. This is also called ping-pong buffer, as when the FFT output is written in PING memory block the pong memory is used to read the FFT output with necessary shift values.

This dual buffer memory system is required as the FFT output reading can start at any address value (k value) depending on the required FFT index rotate which depends on the detected signal frequency value.

The FFT rotate can be very easily achieved by changing the address offset and performing modulus 2N addition with FFT output index k.

$$k_i = MOD\_SUM(k + Detected\_k_i) \tag{2}$$

The k_i is used as read address to read from the PING-PONG memory.

Let the outputs of i-th channel after necessary rotation be $X_i(k)$. After multiplying with window function the $X_{if}(k)$ is obtained.

$$X_{if}(k) = X_i(k) \cdot W(k) \tag{3}$$

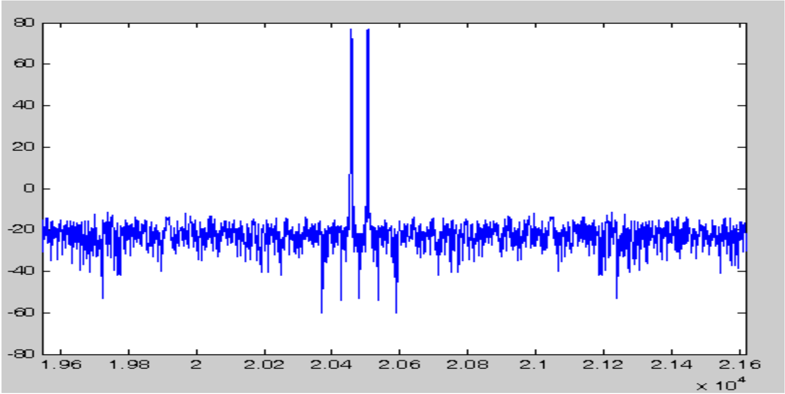
The required decimation is achieved by selecting only the (2N/D) bins around the mid (zero frequency) FFT bin. These bins are streamed to the IFFT core which is configured for computing the IFFT with (2N/D) length. Where D is the decimation factor which is provided by signal detection block.

The IFFT output also results in 2N/D time domain samples. Overlap and add algorithm is performed to result in N/D time domain samples $x_1[n]$ shown in figure 6 for first output channel. Since for one frame of N samples at input N/D output samples are coming this achieves the decimation effect by D factor.

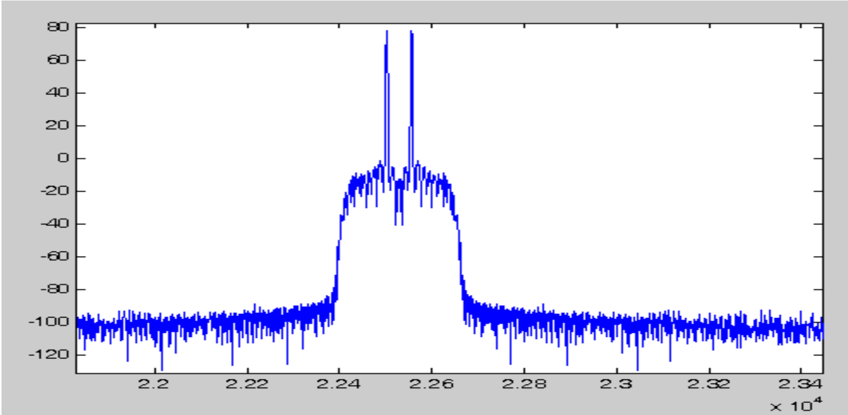
This limits the decimation to only 2^i , for i= 1,2,4,8 .. . etc, because the FFT also exists in only these steps. However DFT cores can be used for lower bandwidth signals (i.e. higher D factor), which do not reflect in much area penalty.

5 Results

The design is simulated both in MATLAB and VHDL and the results are verified. The MATLAB simulation are given in below figures.



(a) Spectrum of simulated input signal.



(b) Spectrum of the output filtered and decimated signal.

Fig. 7. MATLAB simulation results

The Figure 7 shows the zoomed input and output spectrum which shows clearly the stop band attenuation of 80dB. The spectrum is also shifted in frequency domain before filtering and decimation. After completing the conceptual verification the architecture given in figure 6 is coded in VHDL and synthesized for Virtex-6 family XC6VLX240T FPGA and observed to occupy only 10% of the slices and reports 285 MHz of operating frequency. The design also is simulated using Modelsim Xilinx Edition (MXE) 6.2c and the result screen shot is shown in below figure. The results show for one channel with rectangular window and output IFFT size of 64. The IFFT input and output are shown in below figure. Further simulation for higher number of channels is in progress.

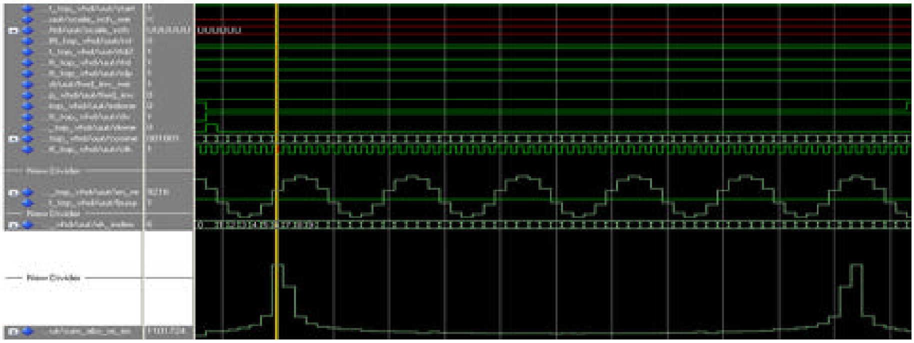


Fig. 8. VHDL simulation results

6 Conclusions and Future Scope

The presented work proposes FPGA based complete frequency domain DDC addressing the requirements of dynamic decimation and multi channel support. The complex digital oscillator is achieved by rotating the FFT output with appropriate value. The architecture uses one 1 FFT core and 2 buffers to store FFT outputs (PING-PONG) and M number of IFFT cores to process M simultaneous signals. The IFFT core can be chosen from Xilinx core gen with runtime FFT length selection option.

The work is aimed to be continued in the direction of realizing full architecture on FPGA to support multiple channels and profiling for selected FPGA device.

Acknowledgment. The Authors acknowledge the support given by ECE dept of ATRI college, Hyderabad, for providing the FPGA development tools and boards. The authors also would like to acknowledge the support given by JNTU ECE dept for regular review and suggestions on the presented work.

References

[1] Han, Y.: A Flexible and Compact Digital Front-End Design for Wideband Software Radio Receivers. In: Science and Technology on Communication Information Security Control Laboratory Zhejiang, China. IEEE (2011)

[2] Xu, Y.-J., Wang, H.-Y., Shen, Z.: Modified Polyphase Filter for Arbitrary Sampling Rate Conversion. Huazhong University of Science and Technology Wuhan, China. IEEE (2010)
[3] RanTao, Senior Member, IEEE, BingDeng, Zhang, W.: Student Member, IEEE, Yue-Wang: Sampling and Sampling Rate Conversion of Band Limited Signals in the Fractional Fourier Transform Domain. IEEE Transactions on Signal Processing 56(1) (January 2008)
[4] Bi, G., Mitra, S.K.: Sampling Rate Conversion in the Frequency Domain [DSP Tips and Tricks]. IEEE Signal Processing Magazine 140 (May 2011)
[5] Channelizer cores from RFEL, `http://www.rfel.com/channeliser-cores.aspx`
[6] Signal detection IP core from NSS Communications, `http://nsscomm.com/spectrum_search_and_report.html`
[7] Sample Rate Conversion in Software Configurable Radios, Hentschel, Artech house
[8] Lecture notes ECEN4002/5002: Digital Signal ProcessingLab, `http://ecee.colorado.edu/~ecen4002/index.html`

New Distances for Improving Progressive Alignment Algorithm

Ahmed Mokaddem and Mourad Elloumi

Laboratory of Technologies of Information and Communication and Electrical Engineering
(LaTICE), Higher School of Sciences and Technologies of Tunis (HSSTT),
University of Tunis, Tunisia
`moka.ahmed@yahoo.fr, Mourad.Elloumi@fsegt.rnu.tn`

Abstract. Distance computation between sequences is an important method to compare between biological sequences. In fact, we attribute a value to the sequences in order to estimate a percentage of similarity that can help to extract structural or functional information. Distance computation is also more important in the progressive multiple alignment algorithm. Indeed, it can influence the branching order of the sequences alignment and then the final multiple alignment. In this paper, we present new methods for distance computation in order to improve the progressive multiple alignment approach. The main difference between our distances and the other existed methods consists in the use of all the sequences of the set in the pair-wise comparison. We tested our distances on BALIBASE benchmarks and we compared with other typical distances. We obtained very good results.

1 State of the Art of the Pairwise Comparison

Progressive approach is the most used and the most effective approach to resolve the problem of Multiple Sequences Alignment (MSA) it operates in three steps:

1. Pairwise comparisons: it consists in pairwise comparison between every pair of sequences.
2. Sequences clustering: in this step, we define the branching order of the sequences by construction a *guide tree*.
3. Sequences integration: this step allows aligning the sequences using the branching order.

The goal of the pairwise comparison step is to estimate the similarity between pairs of sequences in order to distinguish the sequences that are the first to be aligned. The distances computed between all pairs of sequences are stored in a symmetric diagonal matrix, called *distance matrix*. Several methods are used to compute distance, we group the distances computation method in two approaches:

1. *Direct computation*, i.e., using pairwise alignment.
2. *Indirect computation*, i.e., without pairwise alignment.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 243–251.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

1.1 Direct Computation

The first approach for the distances computation is the *direct computation*. Using this approach, we compute a distance between two sequences without constructing a pairwise alignment, thus allows reducing the time computation and the memory space required. Several distances have been defined. Among these distances we mention:

1. *The k-mer distance*: a *k-mer* is a string of length k. The *k-mer distance* is based on finding the common substring having a fixed length k.

   $$D(x,y) = 1 - \sum_T \frac{min\{nx(T), ny(T)\}}{min\{Lx, Ly\} - k + 1} \qquad (1)$$

   Where $nx(T)$: number of occurrence of the k-mer T in x ; $ny(T)$: number of occurrence of the k-mer T in y ; Lx is the length of x, Ly is the length of y. This distance is used by MUMMALS [14], MAFFT [5] and MUSCLE [1]. The distance D is used to filter the k-mer in excess. The drawback of this distance is that it is rarely to find multiple occurrences of a k-mer in the same sequences. A variant of the distance D can be defined to verify only the existence of a shared k-mer between sequences. We obtain the following formula:

   $$D(x,y) = 1 - \sum_T \frac{\lambda_{x,y(T)}}{min\{Lx, Ly\} - k + 1} \qquad (2)$$

   Where: $\lambda_{x,y(T)}$ = 1 if the k-mer is a common substring in the two sequences, otherwise $\lambda_{x,y(T)}$= 0. This distance is used by the MUSCLE [1]. The k-mer distance is very simple, fast to compute but still imprecise and inefficient for the estimation of similarity between closest sequences. Indeed, this distance restricts the comparison between the two sequences to a fixed length of substrings and can not find all fragments and residues repeated in the sequence. There are many other distances based on the k-mer [19].
2. *The edit distance*: The edit distance is used by the string matching algorithms to compute the minimum number of operations, i.e., substitutions, deletions, insertions, needed to transform one string to another. The progressive alignment algorithm Kalign [8] and Kalign2 [7] use respectively the Wu and Manber algorithm [21] and Muth and Manber algorithm [12] to compute the edit distance.
3. *Distance based on grammar dictionary*: This distance is based on sets of fragments called *grammar dictionaries* [15]. Indeed, for each sequence i of the initial set of sequences, we build a dictionary $E_{i,i}$ made up of fragments called *rules*, this dictionary is used to generate the sequences i by concatenating the several fragments of the dictionary.

   $$D(x,y) = 2 * \frac{|E_{x,y}| - |E_{x,x}| + |E_{y,x}| - |E_{y,y}|}{|E_{x,y}| + |E_{y,x}|} \qquad (3)$$

   Where $|E_{i,j}|$ represents the number of rules in $E_{i,j}$

   This distance is used by the GRAMALIGN [15] algorithm. There are many other distances such as the adaptation of the Fourier transform used by MAFFT [5] and the FDOD distance used by MSAID [10].

1.2 Indirect Computation

The *indirect computation* uses a pair-wise alignment. The pair-wise alignment is constructed using a standard dynamic programming algorithm. We classify these distances into two groups:

1. Those based on the alignment score: the distances based on the alignment score use the optimal score obtained by the pair-wise alignment. Among these distances we cite:
    a. The NS (*Normalized Score*): Represents the ratio of the optimal alignment score and the length of the alignment, i.e., number of column in the alignment.

    $$NS(x,y) = \frac{SPx,y}{L} \tag{4}$$

    Where SPx,y represents the scores of the optimal alignment of the two sequences x et y, L: represents the length of the alignment.

    This distance is used by the algorithms MAP2 [9] and OPAL [20].

    b.

    $$D(x,y) = SP_{max} - SP_{x,y} + 1 \tag{5}$$

    Where $SPmax$ represents the highest scores of the all pair-wise alignment, $SP_{x,y}$ the score of the pair-wise alignment of the sequences x et y. This distance is used by TSPMSA [22]
2. Those based on the pair-wise alignment: the distance is computing using the correspondence between the residues or regions obtained by the pair-wise alignment. Among these distances, we mention:
    a. The percent of identity: this distance is computed from the alignment by counting the number of identical residues that appear in the same column in the pair-wise alignment excluding gaps.

    $$P(A) = 1 - \frac{NI}{NC} \tag{6}$$

    Where NI: the number of identical aligned residues obtained and NC: number of compared residues. This distance is used by CLUSTALW [16], PLASMA [2] and PRRP [4]. This distance is restrictive. Indeed, we don't compute substitutions. Thus, several variants have been defined by allowing possible substitutions between residues.

    b. The percentage of similarity that allows computing substitutions in a compressed alphabet [15,6].

    c. Distance-based on probability: the stochastic and probabilistic methods uses Markov models (HMM) to create pair-wise alignments, then build a distance based on the probability of having the residue of each sequence in the correct optimal alignment.

    $$E(x,y) = 1 - \frac{1}{min\{|x|,|y|\}} \sum P\left\{x_i\tilde{}y_j \in a^*\right\} \tag{7}$$

    Where $|x|$: the length of sequence x, $P(x_i\tilde{}y_j)$: represent the probability that the residue x_i is aligned to the residue y_j in the alignment a^*.

    This distance is used by PROBCONS [3].

The indirect computation method is more accurate and can estimate adequately the similarity between sequences, however it is slower in the case of large number of long sequences, since it requires the construction of an optimal alignment for each pair of sequences.

The rest of this paper is organized as follows. In section 2, we present some notations and definitions. In section 3, we present our new distances. In section 4, we present the experimental results realized using our distances in a standard progressive alignment algorithm and we benchmarked this algorithm using the BALIBASE benchmark and we compare the obtained results with other distances. Finally, in section 5, we present our conclusion and give some perspectives.

2 Notation and Definitions

Let A be a finite alphabet, a sequence is an element of $A*$, it is a concatention of elements of A. The length of a sequence w, denoted by $|w|$, is the number of the characters that constitute this sequence. A portion of w beginning at the position i and ending at the position j, $0 \prec i \prec j \leq |w|$, is called *subsequence*.

Let f be a set of sequences, aligning the sequences of f consists in optimizing the number of matches between the characters occurring in the same order in each sequence. When $|f| > 2$, aligning the sequences of f is called *Multiple Sequence Alignment* (MSA).

Let $f = \{w_1, w_2, \dots, w_N\}$ be a family of sequences, A *profile* is a sequence that represents an alignment. The profile is constructed by selecting for each column of the multiple alignments the residue that has the maximum occurrences in this column.

We say that a *subsequence* x is a *motif* if it is extracted from a profile and not forming by gaps.

A *bloc* is a pair of motif extracted from a pairwise alignment.

3 New Distances

Classical approaches assign a value to a pair of sequence independently of all the sequences of the initial set. However, the goal of the first step of the progressive alignment algorithm is to select the appropriate sequences in the progressive process. Thus, we propose a new approach that allows to all the sequences to participate in pair-wise comparison, thereby we can more estimate the future alignment in the progressive process. In fact, our method uses a general comparison that aims to quantify the impact of each pair-wise alignment on the set of sequences. This comparison approach is totally different from conventional approaches and is beneficial because it includes all the sequences in the pair-wise comparison and allows as selecting at each step of the progressive alignment the appropriate sequence in order to obtain the optimal alignment. Our approach sheds light the interaction between each pair of sequences and other sequences selected. By using our approach we compute the distances at each step of progressive alignment as follows:

1. First, we aligned each pair of sequences using the dynamic programming algorithms of pair-wise alignment [13].
2. Then, we compared the alignment obtained with other sequences to all sequence of the set in order to compute a distance. We used the sequence alignment and the profile in this comparison.

3.1 Distance Based on Motif

We compute the distance based on motif using the following formula:

$$d_{motif}(x,y) = \frac{\sum_i^{nb} \sum_j occ(i,j) * |i|}{|profil| * N} \tag{8}$$

Where $occ(i,\ j)$=1 if the motif i appears in the sequence j else $occ(i,\ j)$=0 ; $profile$ represents the profile sequence obtained by the pair-wise alignment of the sequences x and y; N is the number of sequences and nb represents the number of motif.

The distance based on motif maximizes the weight attributed to the alignment that preserves more motifs in other sequences. Indeed, we compute the number of motifs extracted from profiles that occur in other sequences. In fact, the alignment is more interesting if it conserve longer motif in other sequence. Thus, we integrate the length of motif in the distance. Our method operates as follows:

- In the first step, we construct pair-wise alignment using a standard dynamic programming algorithm then we construct the sequence profile.
- In the second step, we extract motif from the sequence profile.
- In the third step, we compute the number of occurrences of each motif in each sequence.

Example: Let f the following set of sequences formed by 4 sequences
w_1: *tyimreaqyesaq*; w_2: *tcivmreaye*; w_3: *yimqevqqer*; w_4: *wryiamreqyes*.

We compute the distance based on motif between w_2 and w_4 as follows:

- First, we align the two sequences and we construct the profile below.

  $profile$: - - - I - M R E - Y E -

- In the second step, we extract the list of motifs and we find the occurrences of the motifs extracted in the other sequences of the set f.

Thus, we obtained the following results:

- The score corresponding to the motif 'I' is s =2*1.
- The score corresponding to the motif 'MRE' is s =3*1.
- The score corresponding to the motif 'YE' is s=2*1.

$d_{motif}(w_2,w_4) = 2+3+2/(4*12) = 7/48 = 0.14$.

We apply the same method for each pair of sequences; we obtained the following distances matrix

Table 1. Distances Matrix

	w_1	w_2	w_3	w_4
w_1	-	-	-	-
w_2	0.7	-	-	-
w_3	0.9	0.16	-	-
w_4	0.8	0.14	0.17	-

3.2 Distance Based on Profile

Our second distance is based on profile due to the importance of the profile in the progressive alignment. In fact, progressive alignment uses in every step the profile to construct the next alignment by the profile-profile alignment algorithm. Our distance is based on finding identical common ordered residues, different from gap character, i.e., '$-$' between pairs of sequences that appear in the other sequences in the set. Thus, our distance assigns a value to the most interesting alignment that preserves the maximum number of residue in other sequences. The distance is computing using this formula:

$$d_{profil}(x,y) = \sum_{i \in N} \frac{d}{min(|profil|, |w_i|)} \tag{9}$$

Where d the number of ordered common residue between the $profile$ and the sequence w_i.

3.3 Distance Based on Blocs

We use the blocs of the alignment instead of the profile sequence to compute the distance based on blocs. In fact, the blocs are fundamentals in progressive alignment. Indeed, we can not modify the column of the blocs in the progressive process, but we can only insert a column of gap between the columns of the blocs. Thus, we find the conserved blocs in the other sequence. We consider that a bloc is conserved in one sequence if one motif from the bloc appears in this sequence. This bloc is called *conserved bloc*. The distance is computed by the following formula:

$$d_{bloc}(x,y) = \frac{\sum_i \delta * |bi| * Nbi}{L * N} \tag{10}$$

Where bi is a conserved bloc, Nbi represent the number of sequence that the bloc bi appears, N the number of sequence, δ the number of identity occurring in the blocs and L the length of the pair-wise alignment.

4 Experimental Results

In this section, we present our experimental studies in order to assess the performance of our new distances. Our method consists to compare the relevance of our distances to those used by the progressive multiple alignment algorithms. To assess our distances, we used a standard progressive algorithm, we replaced a first step by our distance and we compared the several alignment. Our approach is defined as follows:

- We developed a standard progressive alignment algorithm.
- We modified the first step, i.e., the pair-wise comparison using our distance and the other distances.
- Thus, we compared between the different alignments obtained.

Indeed, we implemented the different algorithm for computing our distance and we integrated them in the progressive alignment algorithm. The progressive algorithm operates as follows:

- In the first stage we used one of our distances.
- In the second step, we used the UPGMA[18] method for the construction of the guide tree. alignment and we used an adaptation of the Needleman and Wunsch algorithm for profile-profile alignment.
- In the last step, we also implemented an efficient iterative approach, the one used by the MUSCLE.

We compared our distances to the popular one, i.e., the *percentage of identity*, the *percentage* of *similarity* in a compressed alphabet [15], the *Normalized Score* (NS) [20], the *k*-mer *distance* [1] and the *anchor distance* [11]. We tested the algorithm on the datasets of the BALIBASE benchmark. In order to compare alignments, we use the *Column Score* (CS) and the *Sum of Pairs Score* (SPS).

Table 2. The average of the CS on the reference of BALIBASE

Distances	REF1	REF2	REF3	REF4	REF5
percent of identity	68.49	23.38	16.16	35.11	56.09
percentage of similarity	68.50	26.05	16.16	34.90	59.30
k-mer	68.87	28.86	26.20	40.40	55.58
normalized score	67.77	25.49	19.16	40.05	58.27
Anchor Distance	69.61	29.50	33.00	38.70	55.72
Distance based on Profile	66.67	26.66	22.66	40.94	57.45
Distance based on motif	67.50	24.77	19.66	37.40	59.41
Distance based on bloc	65.63	20.94	12.33	38.60	57.75

These tables give respectively the average of the SPS and the average of the CS for every sets of reference datasets of BALIBASE[17] benchmark. These tables show that the CS and SPS scores obtained with our distances are significantly higher than those obtained by the *percent of identity*, the *percentage of similarity* in a compressed alphabet [15], the *k*-mer *distance* and the *normalized score*. Thus prove that our distances improve significantly the final alignment obtained by the progressive algorithm. In addition, the results obtained by the distances computing using our new method, i.e., *distance based on profile*, *distance based on motif* and *distance based on bloc* are higher than the *anchor distance*. In fact, our approach allows all the sequences of the set contribute in distance computation for each pair of sequences. Thus the result

Table 3. The average of the SPS on the reference of BALIBASE

Distances	REF1	REF2	REF3	REF4	REF5
percent of identity	77.85	79.06	59.83	65.18	76.69
percentage of similarity	78.03	77.42	59.83	64.95	78.28
k-mer	78.14	76.04	54.66	69.32	75.40
normalized score	77.74	76.50	63.33	64.64	78.27
Anchor Distance	79.21	79.96	69.18	69.33	76.08
Distance based on Profile	76.66	76.81	62.23	70.32	78.37
Distance based on motif	77.90	74.33	51.15	66.90	79.23
Distance based on bloc	77.38	79.21	49.58	68.54	78.41

is more precisely and allows selecting the appropriate sequences in each step of the progressive approach. Thus, we can resolve the problem of the minimum local of the progressive method. Indeed, we estimate from the first step the pair-wise alignment that can be give the best alignment by providing all possible alignment by comparing all pair-wise alignment to the other sequences thus allow to select the two sequences that undoubtedly allow us to have the most significant alignment.

Our distances are efficient and give good results.

5 Conclusions and Perspectives

In this paper, we present a new approach to compute a distance between sequences in order to improve the progressive approach for multiple sequence alignment. Our approach consists to integrate all sequences in the pair-wise comparison. Using this approach, we have defined new distances that use several properties of the alignment such as $profiles$ and $blocs$.

Our distances are efficient. Indeed, we tested our method in a progressive algorithm and we compared to other distances. The results obtained using a BALIBASE benchmark is good. Thus, our method improve significantly the multiple alignment progressive.

In our future work, we would like to improve the other step in progressive alignment such as the cluster step and the profile-profile alignment. Thus allow us to construct a new efficient progressive algorithm.

References

1. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy high throughput. Nucleic Acids Research 32(5), 1792–1797 (2004)
2. Derrien, V., Richer, J.M., Hao, J.K.: PLasMA: un nouvel algorithme progress if pour l'alignement multiple des séquences. In: Proc. Premières Journées Francophones de Programmation par Contraintes (JFPC 2005), pp. 39–48 (2005)
3. Do, C.B., Mahabhashyam, M.S., Brudno, M., Batzoglou, S.: PROBCONS: Probabilistic consistency-based multiple sequence alignment. Genome Res. 15, 330–340 (2005)

4. Gotoh, O.: Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J. Mol. Biol. 264(4), 823–838 (1996)
5. Katoh, K., Kuma, K., Toh, H., Miyata, T.: MAFFT version 5: Improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33(2), 511–518 (2005)
6. Kimura, M.: The neutral theory of molecular evolution. Cambridge University Press (1983)
7. Lassman, T., Frings, O., Sonnhammer, L.L.: KALIGN2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Research 37(3), 858–865 (2009)
8. Lassman, T., Sonnhammer, L.L.: KALIGN: An accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6 (2005)
9. Liang Ye, Y., Huang, X.: MAP2: multiple alignments of syntenic genomic sequences. Nucleic Acids Research 33(1), 162–170 (2005)
10. Min, Z., Weiwu, F., Junhua, Z., Zhongxian, C.: MSAID: multiple sequence alignment based on a measure of information discrepancy. Computational Biology and Chemistry 29, 175–181 (2005)
11. Mokaddem, A., Elloumi, M.: PAAA: A Progressive Iterative Alignment Algorithm Based on Anchors. In: PRIB, pp. 296–305 (2011)
12. Muth, R., Manber, U.: Approximate multiple string search. In: Hirschberg, D.S., Meyers, G. (eds.) CPM 1996. LNCS, vol. 1075, pp. 75–86. Springer, Heidelberg (1996)
13. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(1), 443–453 (1970)
14. Pei, J., Grishin, N.V.: MUMMALS: Multiple sequence alignment improved by using hidden Markov models with local structural information. Nucleic Acids Res. 34(16), 4364–4374 (2006)
15. Russell, D.J., Out, H.H., Sayood, K.: Grammar-based distance in progressive multiple sequence alignment. BMC Bioinformatics 9 (2008)
16. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleid Acids Research 22(22), 4673–4680 (1994)
17. Thompson, J.D., Plewniak, F., Poch, O.: A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 27(13), 2682–2690 (1999)
18. Sneath, P., Sokal, R.: Numerical taxonomy, pp. 230–234. Freeman, San Francisco (1973)
19. Vinga, S., Almeida, J.: Alignment-free sequence comparison-a review. Bioinformatics 19(4), 513–523 (2003)
20. Wheeler, T.J., Kececioglu, J.D.: Multiple alignment by aligning alignments. Bioinformatics 23(13), 559–568 (2007)
21. Wu, S., Manber, U.: Fast Text Searching Allowing Errors. Communications of the ACM 35, 83–91 (1992)
22. Zhong, W.: Using Traveling Salesman Problem Algorithms to Determine Multiple Sequence Alignment Orders. Thesis (2002)

Service-Oriented Architecture (SOA) and Semantic Web Services for Web Portal Integration

T.G.K. Vasista[1] and Mohammed A.T. AlSudairi[2]

[1] King Saud University, Riyadh, KSA
gatapudi@ksu.edu.sas, tgkvasista@gmail.com
[2] College of Business Administration, King Saud University, Riyadh, KSA
mas@ksu.edu.sa

Abstract. Service Oriented Architecture (SOA) is gradually replacing monolithic architecture as the premier design principle for new business applications with its inherently systematic nature and capability. I T requires investment of advanced architectural thinking into definition of services before any development of services or service consumers can begin. Earlier efforts of notable styles of SOA such as CORBA and XATMI have failed to be adopted as main stream projects because of demanding design process requirement with sense-making activities and even have been residing with the modern SOA or Web services middleware. Thus this paper aims in incorporating such sense-making design activities with the proposed semantic web service based architecture. This paper tries to tackle the above problem by proposing a service-oriented architecture for web data and service integration. Firstly, it proposes a service-oriented platform independent architecture and Secondly, it presents a specific deployment of such architecture for data and service integration on the web using semantic web services implemented with the WSMO (Web Services Modeling Ontology).

1 Introduction

Service-Oriented Architectures are particularly well suited to cope up with the needs of such an ongoing incremental process optimization [6] as represented by Capability Maturity Model Integration (CMMI). CMMI in software engineering and organizational development is a process improvement approach that provides organizations with the essential elements for effective process improvement [15] with the characteristics of the maturity levels as given in Table 1.

When making architectural decisions, one must carefully analyze the advantages and disadvantages of the level of coupling [2, 3]. Generally speaking, OLTP (online transaction processing) style applications, as they are found throughout large enterprises, do not normally require a high degree of loose coupling, as these applications are tightly coupled by their nature. ERP systems are usually found suitable for such kind of single large monolithic representation and handling [6]. But the recent trend in Globalization is increasing the scale and complexity of the modern enterprises by forcing the enterprise to be represented as a global network consisting of multiple

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 253–261.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Table 1. Characteristics of CMMI Levels [15]

Maturity Level No.	Maturity level Title	Maturity Level Description
Level 1	Initial	Processes unpredictable, poorly controlled and reactive.
Level 2	Managed	Process characterized for projects and is often reactive
Level 3	Defined	Process characterized for the organisation and is proactive
Level 4	Quantitatively Managed	Process Measured and Controlled
Level 5	Optimizing	Focus on Process Improvement

units and functions [12, 13]. So when business processes are highly distributed, the different sub-processes and transactions are generally more independent of each other or more loosely coupled. Loose coupling though increases the flexibility, increases the complexity. The enterprise software architecture is the architect's most important tool at hand to confront the changes to and expansion of functionality that increase system complexity and reduce efficiency [11]. Thus in enterprise software, the architect takes on the roles as an outside influencer and controller. It is his responsibility to oversee individual software projects from the strategic point of view of the overall organisation, as well as from the tactical, goal-oriented viewpoint of the individual project. He has to balance different requirements while attempting to create an enduring order within the enterprise software landscape. By technique of refactoring the current solutions, architects constantly strive to reduce complexity and thereby the agility of the system (See Figure 1) [6].

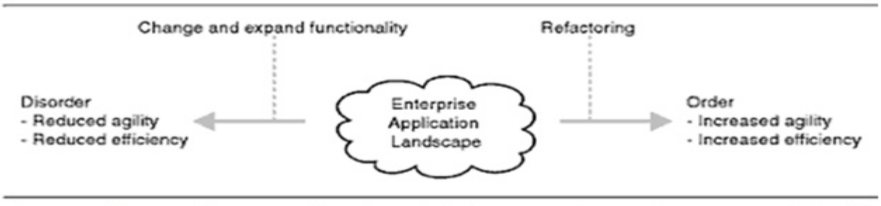


Fig. 1. Software architects use refactoring to fight the constant increase in system complexity [6]

Different tools support architectural design and test case generation. Users of these tools want them to work together to fully support the user's design and development process. Thus tool integration is intended to produce complete environments that support the CMMI model based software development life cycle of maturity levels. There are five types of tool integration issues that must be addressed. These can be termed platform integration, presentation integration, data integration, control integration and process integration [14].

Service-Oriented Computing (SOC) arises as a new paradigm for distributing applications which envisions services as fundamental elements for developing applications [9].

So X-as-a-Service kind of pattern is what SOA is suggesting in terms of cloud computing terminology by adopting the basic principle SOA's loose coupling technique where the above five types of tool integration issues can be represented as SOA based services. Where X-as-a-Service can be replaced with Platform-as-a-Service (PlaaS), Presentation-as-a-Service (PraaS1), Data-as-a-Service (DaaaS), Control-as-a-Service (CoaaS) and Process-as-a-Service (Praas2) in cloud computing terminology that are available independently as loosely coupled services.

As design and development of SOA applications are inherently systematic [7], this is of particular importance for enterprises, when given their need to become more agile in order to react as quickly as possible to changing business environments and offer new services to customers, suppliers and partners that make a difference with respect to competition. SOAs can help to significantly reduce complexity at all levels. SOAs achieve their simplicity by following features: Decomposition-SOAs decompose large systems into application frontends and services; Appropriate granularity-The granularity of services is well suited to gain a high-level understanding of the entire system; Loose coupling by SOA architectural patterns can be Decoupled using technology-SOAs can be well understood without in-depth knowledge of technology; Reuse-SOAs result in the high level reuse of existing components; Documentation – SOA based services are well documented due to service contract to provide comprehensive understanding [6].

2 Platform Independent Integration Architecture

In order to get a web portal integration architecture that can be used in different domains, it is important first to design a platform-independent architecture [1]. The key to successful integration and interoperability lies in the intelligent use and management of metadata across all applications, tools and databases. Metadata management and integration can be accomplished through the use of the OMG's core MDA standards such as CWM, UML, MOF and XMI [10], where:

CWM stands for *Common Warehouse Meta-model* – It is a meta-model of the data model representing both the business and technical metadata that is most often found in the data warehousing and business analysis domains. CWMs are intended to be highly generic, external representations of shared metadata [10]. It requires the use of generalization and abstraction technique to translate the product-specifics to generics through standard extension mechanism making it compatible to CWM format [10].

UML stands for *Unified Modelling Language* – It is used for expressing in the Unified Modelling Language, which is an OMG standard language for modelling discrete systems by Rumbaugh. UML is the notational basis for the definition of CWM. Visual UML models are capable of automatic translation to other visual or non-visual formal languages to facilitate the support for interchange of CWM models in various platform and tool independent formats (e.g., XML) as well as the construction of tool-specific metadata from CWM models (e.g. translation of a CWM relational model into SQLDDL statements that actually build the schema) [10].

MOF stands for *Meta Object Facility*- It is an OMG standard defining a common, abstract language for the specification of meta-models. MOF is an example of meta-meta-model or model of the meta-model (also called as ontology). The MOF's support for the model life cycle semantics means that MOF implementation provides an effective metadata authoring and publishing tool, when combined with support for visual modelling. For example a fully MOF-compliant repository must provide a significant number of metadata services that as: persistence, versioning and directory services [10].

XMI stands for *XML Metadata Interchange (XMI)* - It is an OMG standard that maps the MOF to the W3C's eXtensible Markup Language (XML). XMI defines how XML tags are used to represent serialized MOF-compliant models in XML. MOF based meta-models are translated to XML Document Type Definitions (DTDs) and models are translated into XML documents that are consistent with their DTDs. XMI based interchange is so important in distributed and heterogeneous environments as the communication of content is both self-described and inherently asynchronous [10].

Object Management Group's Model-Drive Architecture (MDA) is an approach to system-specific and interoperability based on the use of formal models. In MDA, platform-independent models are initially expressed in a platform-independent modelling language such as UML. The platform independent model is subsequently translated into a platform specific model by mapping platform-independent models into some implementation language (e.g. Java) or platform using formal rules. The core standards of MDA such as CWM, UML, MOF and XMI form the basis for building coherent schemes for authoring, publishing and managing models within a model-driven architecture [10].

Metadata is critical to all aspects of interoperability within heterogeneous environment. In fact Interoperability is achieved by means of metadata [10], which is being used to provide system semantic definitions and capabilities facilitated in the form of

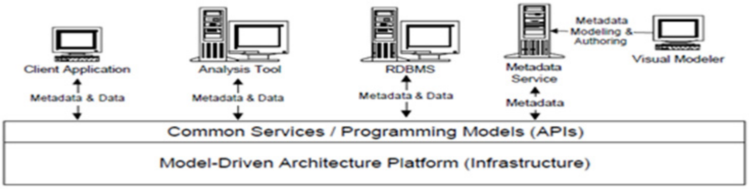

Client Application
Metadata & Data
Analysis Tool
Metadata & Data
RDBMS
Metadata & Data
Metadata Service
Metadata
Metadata Modeling & Authoring
Visual Modeler
Common Services / Programming Models (APIs)
Model-Driven Architecture Platform (Infrastructure)

Fig. 2. Example of a Realization of Model-Driven Architecture [10]

standard APIs. Any MDA based system must have the ability to store, manage and publish both application and system-level metadata including descriptions of the environment itself.

The platform-independent architecture [1] (See Figure 3) aims to offer service-based platform independent architecture for web portal integration.

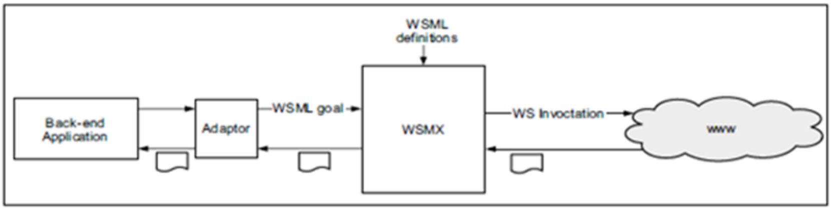


Fig. 3. External view of WSMX Architecture [1]

Web Service Execution Environment (WSMX) is a reference implementation for Web Service Modelling Ontology (WSMO) [1]. Its goal is to provide both a test bed for SMO and to demonstrate the viability of using WSMO as a means to achieve dynamic inter-operation of Semantic Web Services (SWS). The first version of WSMX provides the architecture needed for a middleware-based platform for integration, and as such it is concerned with dynamic discovery, mediation, selection and invocation and an implementation of these components.

2.1 Description of the Functional Usage of WSMX

Web Service Markup Language (WSML) descriptions of Web Services, ontologies, mediators and goals are sent to Web Service Execution Environment (WSMX) through Web Service Modelling Ontology (WSMO) editor for compilation. A back-end application creates a service requirement in a known source format and sends this to the WSMX adapter. The adapter takes the service requirement and translates it into a WSML message consisting of a goal that describes what a WSMX should execute. The goal is then sent to WSMX for execution. Before WSMX can execute the goal, WSML descriptions of the WS offering, the capability of matching the service requirement of the ontologies these WS use, and the source format ontology must have been created using User Interface (WSMO editor) are compiled to WSMX. When WSMX receives the WSML message with specific goal, it discovers the WS that best matches that goal, mediates the service requirement data following mapping rules between the source format ontology and the ontology of the discovered WS and finally invokes it, providing the data to it in the concepts and formats it expects [1].

Whereas metadata acts as control abstraction layer [5], the remaining g four generic and important layers from service based system integration perspective are Interaction, Process, Function and Data as against inclusion of Platform-as-a-Service for platform specific integration architecture (See Figure 4).

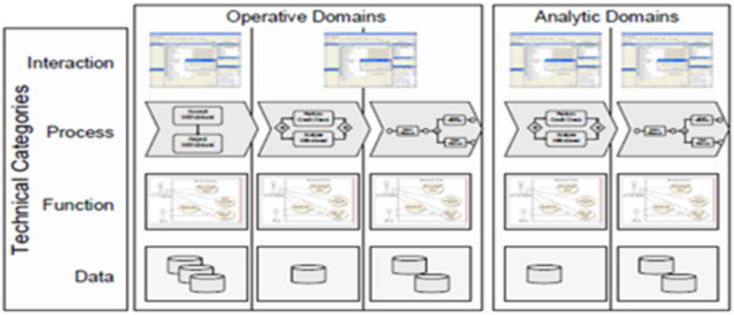


Fig. 4. Remaining Four Layers of Platform Independent Service Classification and Reference Schema [8]

3 Integrating Multi-Agent Systems (MAS) and Semantic Web Services (SWS)

The general idea of the approach to integrate MAS and SWS is given in Figures 5 & 6. For both Multi-Agent Systems (MAS) as well as Semantic Web Services (SWS), model transformations to the platform specific levels were provided by applying principles of model-driven development [4].

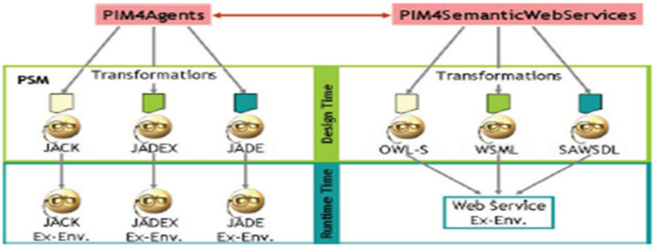


Fig. 5. Approach to Integration [4]

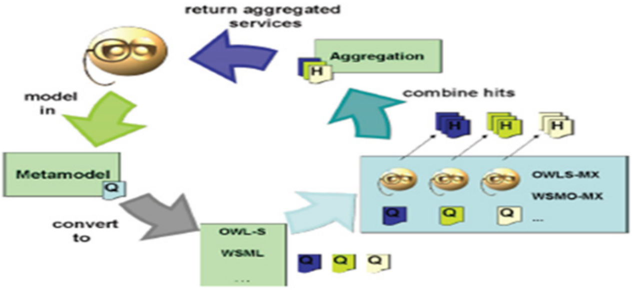


Fig. 6. Model-Driven Semantic Web Service Match Making [4]

This integration is based on a platform independent meta-model for agents and a platform independent meta-model for semantic web services.

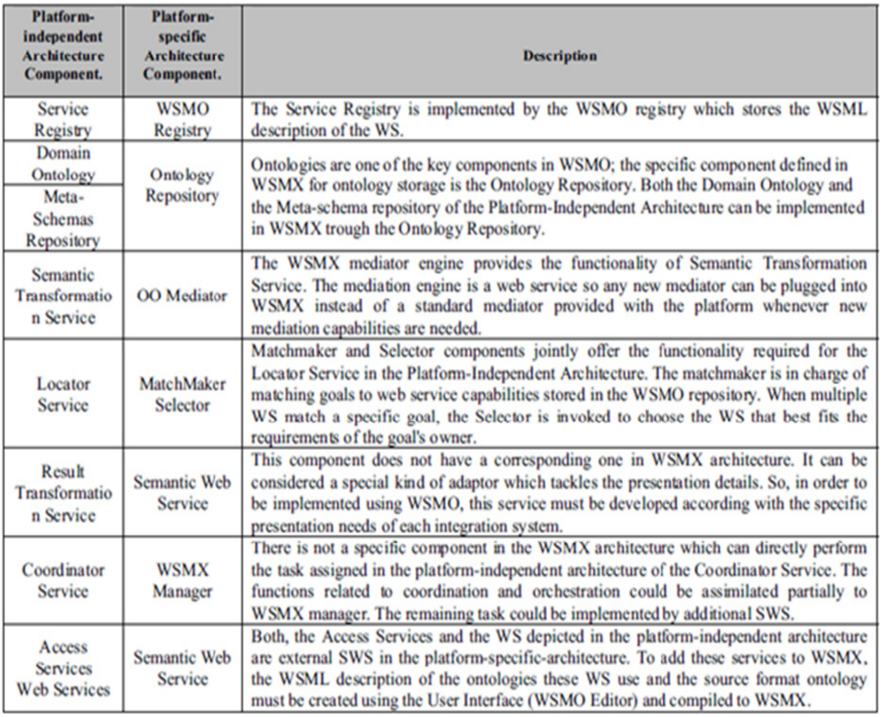
For Semantic Web Services, a model-driven semantic web services matchmaker agent is designed (See Figure 6) that discovers semantic services independent of selected description formats (OWL-S), (WSML) and (SAWSDL). The model-driven semantic service matchmaker (MDSM) performs an automatic service retrieval applying existing matchmakers (i.e. OWL-MX and WSMO-MX) for different formats and returns the most similar matches to the user [4].

4 Platform-Specific Integration Architecture

WSMO and OWL-S can be used as the specific platform to deploy the platform-independent service oriented architecture.

This section shows the specific deployment of the platform-independent architecture to a platform specific one using WSMO in the form of Table 2 that presents the transformation needed to deploy the platform independent architecture in the selected platform specific technology.

Table 2. Correspondence between platform-independent and platform specific architectures [1]



Platform-independent Architecture Component.	**Platform-specific Architecture Component.**	**Description**
Service Registry	WSMO Registry	The Service Registry is implemented by the WSMO registry which stores the WSML description of the WS.
Domain Ontology	Ontology Repository	Ontologies are one of the key components in WSMO; the specific component defined in WSMX for ontology storage is the Ontology Repository. Both the Domain Ontology and the Meta-schema repository of the Platform-Independent Architecture can be implemented in WSMX trough the Ontology Repository.
Meta-Schemas Repository		
Semantic Transformation Service	OO Mediator	The WSMX mediator engine provides the functionality of Semantic Transformation Service. The mediation engine is a web service so any new mediator can be plugged into WSMX instead of a standard mediator provided with the platform whenever new mediation capabilities are needed.
Locator Service	MatchMaker Selector	Matchmaker and Selector components jointly offer the functionality required for the Locator Service in the Platform-Independent Architecture. The matchmaker is in charge of matching goals to web service capabilities stored in the WSMO repository. When multiple WS match a specific goal, the Selector is invoked to choose the WS that best fits the requirements of the goal's owner.
Result Transformation Service	Semantic Web Service	This component does not have a corresponding one in WSMX architecture. It can be considered a special kind of adaptor which tackles the presentation details. So, in order to be implemented using WSMO, this service must be developed according with the specific presentation needs of each integration system.
Coordinator Service	WSMX Manager	There is not a specific component in the WSMX architecture which can directly perform the task assigned in the platform-independent architecture of the Coordinator Service. The functions related to coordination and orchestration could be assimilated partially to WSMX manager. The remaining task could be implemented by additional SWS.
Access Services Web Services	Semantic Web Service	Both, the Access Services and the WS depicted in the platform-independent architecture are external SWS in the platform-specific-architecture. To add these services to WSMX, the WSML description of the ontologies these WS use and the source format ontology must be created using the User Interface (WSMO Editor) and compiled to WSMX.

For Multi-Agent Systems, model transformations from the platform independent to the platform specific meta-models for JACK and JADE are defined as shown in the figure 5 [4].

It is interesting to note that there is another potential layer of service abstraction called Structure-as-a-Service (StaaS) that could be included by providing a subclass hierarchy of nested service categories, which serve as component types that can be applied to services. This structural perspective is complemented by the service connectivity. This structural perspective acts as reference architecture constraints from organisational and connectivity perspective on platform oriented runtime services (for e. g., See Figure 7).

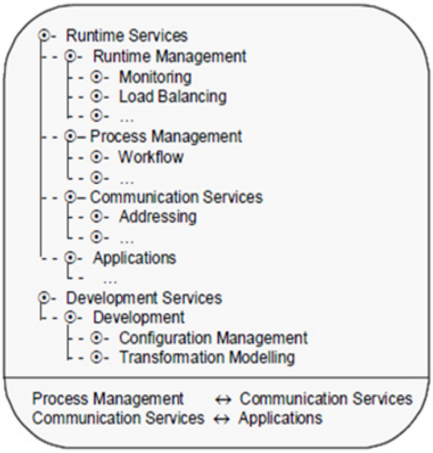


Fig. 7. Structural Perspective as a reference architecture constraint–organisational and connectivity perspective on platform oriented runtime services [8]

5 Conclusion

Since the advent of the Internet, several works have gone a long way towards resolving the web integration problem. Our effort in this regard is to propose two kinds of architecture based solutions: (1) Platform Independent architecture based on service oriented paradigm for web portal integration and (2) Platform specific architecture. It is interesting to note that the following abstraction levels-as-a-service kind of pattern has been observed to be part of the cloud computing paradigm. They are: Structure-as-a-Service, Process-as-a-Service, Control-as-a-Service, Presentation-as-a-Service, Data-as-a-Service for Platform Independent Architecture Solutions and Platform-as-a-Service get added or included for platform-specific architecture.

References

1. Acuna, C.J., Marcos, E., Gomez, J.M., Bussler, C.: Toward Web Portals Integration through Semantic Web Services. In: Proceedings of the International Conference on Next Generation Web Services Practices (NWeSP 2005). IEEE Computer Society (2005)
2. Erl, T.: Service-Oriented Architecture: A Field Guide to Integrated XML and Web Services. Prentice-Hall, NJ (2004) ISBN: 0131428985, Digital ACM Citation No. =983556
3. Erl, T.: Service-Oriented Architecture: concepts, technology and design. Prentice-Hall (2005)
4. Hahn, et al.: Integration of Multi-agent Systems and Semantic Web Services on a Platform Independent Level. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. IEEE Computer Society (2008)
5. Hedden, H.: Taxonomies and Controlled vocabularies best practices for metadata. Journal of Digital Asset Management 6, 279–284 (2010)
6. Krafzig, Banke, Slama: Enterprise SOA- Service Oriented Architecture: Best Practices. Pearson Education Inc. Publication, USA (2005)
7. Natis, Y.V.: Service Oriented Architecture Scenario. Gartner Research. ID No.: AV-19-6751, Gartner Inc, USA (2003)
8. Pahl, C., Hasselbring, W., Voss, M.: Service-Centric Integration Architecture for Enterprise Software Systems. Journal of Information Science and Engineering 25, 1321–1336 (2009)
9. Papazoglou, M.P., Georgakopoulos, D.: Service-Oriented Computing. Communications of the ACM 46(10), 25–28 (2003)
10. Poole, J.D.: Model-Driven Architecture: Vision, Standards and Emerging Technologies. In: Workshop on Metamodeling and Adaptive Object Models, ECOOP 2001 (2001)
11. Shiva Shankar, M.: The importance of Enterprise Software Architectures, `http://www.c-sharpcorner.com/uploadfile/shivamadishet/the-importance-of-enterprise-software-architectures/` (cited on April 04, 2012)
12. Vasista, T.G.K.: Adaptive Enterprise: Design and Implementation Approaches for E-Government Integration. International Journal of Conceptions on Computing and Information Technology, IJCCIT (accepted for publication, 2012)
13. Varma, V.A., Reklaitis, G.V., Blau, G.E., Pekny, J.F.: Enterprise-wide modeling & optimization—An overview of emerging research challenges and opportunities. Computers and Chemical Engineering 31, 692–711 (2007)
14. Wasserman, A.I.: Tool Integration in Software Engineering Environments. In: Long, F. (ed.) Software Engineering Environments. LNCS, vol. 467, pp. 137–149. Springer, Heidelberg (1990)
15. Wiki. Capability Maturity Model Integration (January 2012), `http://en.wikipedia.org/wiki/File:Characteristics_of_Capability_Maturity_Model.svg` (cited January 20, 2012)

Evolutionary Multi-Objective Optimization for Data-Flow Testing of Object-Oriented Programs

P. Maragathavalli and S. Kanmani

Department of Information Technology,
Pondicherry Engineering College, Puducherry, India
`{marapriya,kanmani}@pec.edu`

Abstract. This paper presents a Class-Based Elitist Genetic Algorithm (CBEGA) to generate a suite of tests for testing the object-oriented programs using evolutionary multi-objective optimization techniques. Evolutionary Algorithms (EAs) are inspired by mechanisms in biological evolution like reproduction, mutation, recombination, and selection. EA applies these mechanisms repeatedly to a set of individuals called population to obtain solution. Multi-objective optimization involves optimizing a number of objectives simultaneously. The objectives considered in this paper for optimization are maximum coverage, minimum execution time and test-suite minimization. The experiment shows that CBEGA gives 92% path coverage and simple GA gives 88% path coverage for a set of java classes.

Keywords: Evolutionary Algorithm, multi-objective, path coverage, CBEGA, immigration rate.

1 Introduction

Evolutionary algorithms (EAs) are search methods that take their inspiration from natural selection and survival of the fittest mechanisms [3]. EAs differ from traditional optimization techniques in that they involve a search from a population of solutions, not from a single point. Each iteration in an EA involves a competitive selection that filters the least favorable solutions. The solutions with high fitness are recombined with other solutions by swapping parts of a solution with another. Solutions are also mutated by making a small change to a single element of the solution. EAs are robust optimization methods used for test data generation.

Multi-objective optimization refers to the solution of problems with two or more objectives to be satisfied at the same time [1] [3]. Most real world problems have multiple objectives to achieve. This situation creates a set of problems in Multi-Objective Optimization Problems (MOOP). A MOOP has a number of objective functions, which are to be minimized or maximized [10].

Often, traditional approaches for solving multi-objective optimization problem try to change the multiple objectives into a single objective problem in which only a global optimal point is desired. The MOOP produces a set of solutions which are superior to the rest of the solutions with respect to all objective criteria but are inferior to other solutions in one or more objectives [5]. These solutions are called Pareto

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 263–271.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Optimal solutions or non-dominated solutions. A Pareto optimal set is the mathematical solution to a multi-objectives problem [10] [4]. A solution is Pareto-optimal if no other solution can improve one object function without reducing at least one of the other objectives.

Genetic Algorithm is the most popular heuristic technique to solve Multi-Objective Design and Optimization problems. In this paper, CBEGA employs both evolutionary and multi-objective optimization techniques. The objectives are maximization of coverage and minimization of execution time. The immigration rate, one of the GA parameters, is an elitist operator that controls the migration of test cases from one era to next era. Elitism is the process of selecting best individuals from a population. Elitism is important since it allows the solutions to get better over time. Elitism can speed up the performance of the GA significantly; also it helps to prevent the loss of good solutions once they have been found.

The rest of this paper is organized as follows: Section 2 describes the concept of genetic algorithms. Section 3 describes the multi-objective optimization problem. Section 4 describes the CBEGA, section 5 consists of experiment and result analysis and section 6 consists of conclusion.

2 Concept of Genetic Algorithm

Genetic Algorithms (GAs) are adaptive heuristic search techniques based on the evolutionary techniques of natural selection, recombination and mutation [7]. The principle behind GAs is that they create and maintain a population of individuals represented by chromosomes. The input to the algorithm is a set of potential solutions to that problem and a metric called a *fitness function* allows each candidate to be quantitatively evaluated. The GA then evaluates each candidate according to the fitness function. Only good individuals in the current population survive to the next generation while a bad one is eliminated from the selection process. The fitness value of each element, which could be the objective of the solution, is used to distinct good and bad individuals from the population. Reproduction selects individuals with high fitness values in the population, and through crossover and mutation of such individuals, a new population is derived in which individuals may be even better fitted to their environment. The process of crossover involves two chromosomes swapping chunks of data. Mutation introduces slight changes into a small proportion of the population and is representative of an evolutionary step. The structure of a simple genetic algorithm is given below

```
Simple Genetic Algorithm()
{ initialize population;
evaluate population;
while termination criterion not reached
{ select solutions for next population;
perform crossover and mutation;
evaluate population;}
}
```

The above algorithm will iterate until the population has evolved to form a solution to the problem, or until a maximum number of iterations have occurred.

3 Multi-objective Optimization Problem

Multi-objective optimization also known as multi-criteria or multi-attribute optimization is the process of simultaneously optimizing two or more conflicting objectives subject to certain constraints [5].

The general form for Multi-Objective Optimization Problems (MOOP) can be expressed as,

Fitness function $f_m(x)$, where m = 1, 2, 3… M

M=no. of objective functions &

x = single candidate solution.

There are two general approaches to multiple-objective optimization. One is to combine the individual objective functions into a single composite function or move all but one objective to the constraint set. In the former case, determination of a single objective is possible with methods such as utility theory, weighted sum method, etc., but the problem lies in the proper selection of the weights or utility functions to characterize the decision-makers preferences. In practice, it can be very difficult to precisely and accurately select these weights, even for someone familiar with the problem domain. The second general approach is to determine an entire Pareto optimal solution set or a representative subset [4]. A Pareto optimal set is a set of solutions that are non-dominated with respect to each other.

4 Class-Based Elitist Genetic Algorithm

In this section, CBEGA is presented to generate tests for the object-oriented programs. Fig.1 shows the interaction of evolutionary testing with CBEGA. The CBEGA parameters are initial population size, crossover probability, mutation probability, maximum no. of generations per era, maximum no. of eras and immigration rate.

4.1 Control-Flow and Data-Flow Testing

The input representation is in the form of control flow graphs (CFGs) and the output will be the generated test cases. CFG will give the dependencies between variables associated with specific nodes [2].

Data-flow testing focuses on execution of all the interactions between variables i.e. all def-use-associations for any variable ‘v’ in the program. Criteria for generating test cases consists of a triple [d, u, v] where‘d’ def_node, ‘u’ use_node & ‘v’ associated variable.

The inputs of GA are:

1) An instrumented version of the class under test, and
2) Test requirements, which satisfy a given control or data-flow test criterion.

4.2 Test Data Generation

Test generation for structural programs is the process of identifying test data, which execute the program under test and satisfy a given test criterion [8] [9]. Test generation for the object-oriented programs involves generating: 1) test cases, which are sequences of methods issue on an object of the class under test and satisfy a given test criterion, and 2) test data, which is a set of values for the arguments of the called methods. Thus test data generation in improved GA takes place in two stages. First stage generates the method sequences and the second stage generates the corresponding argument values for the methods.

A typical test case for a given [d, u, v] contains the following:

- Invocation of a constructor ‘c’
- Method m_d that causes the execution of ‘d’
- Method m_u that causes the execution of ‘u’
- Values to the parameters ‘v’

4.3 Genetic Operations

Genetic operators such as roulette wheel selection, two-point crossover, value occurrences mutation and immigration rate are experimented in CBEGA for test data generation.

4.3.1 Roulette Wheel Selection

In this type of selection, the individual is selected on the basis of fitness. The probability of an individual to be selected is calculated based on the fitness value of the individual, and thus individuals with higher fitness values have better chances of being selected.

4.3.2 Two-Point Crossover

The crossover allows us to combine individuals i.e. chromosomes that were selected for reproduction. Two-point crossover calls for two points to be selected on the parent chromosomes. Everything between the two points is swapped between the parent chromosomes, rendering two child chromosomes.

Testcase1: 1 3| 2 5 4| &7 8 3

Testcase2: 1 2| 4 3 5| &9 12 33

Offspring1: 1 3 4 3 5 &7 8 3

Offspring2: 1 2 2 5 4 &9 12 33

4.3.3 Value Occurrences Mutation

Mutation alters one or more gene values in a chromosome from its initial state. Value Occurrences mutation attempts to replace a duplicate value of an individual with a missing value to improve the individual’s fitness.

Testcase1: 1 4 4&5 9

Offspring: 1 4 2&5 9

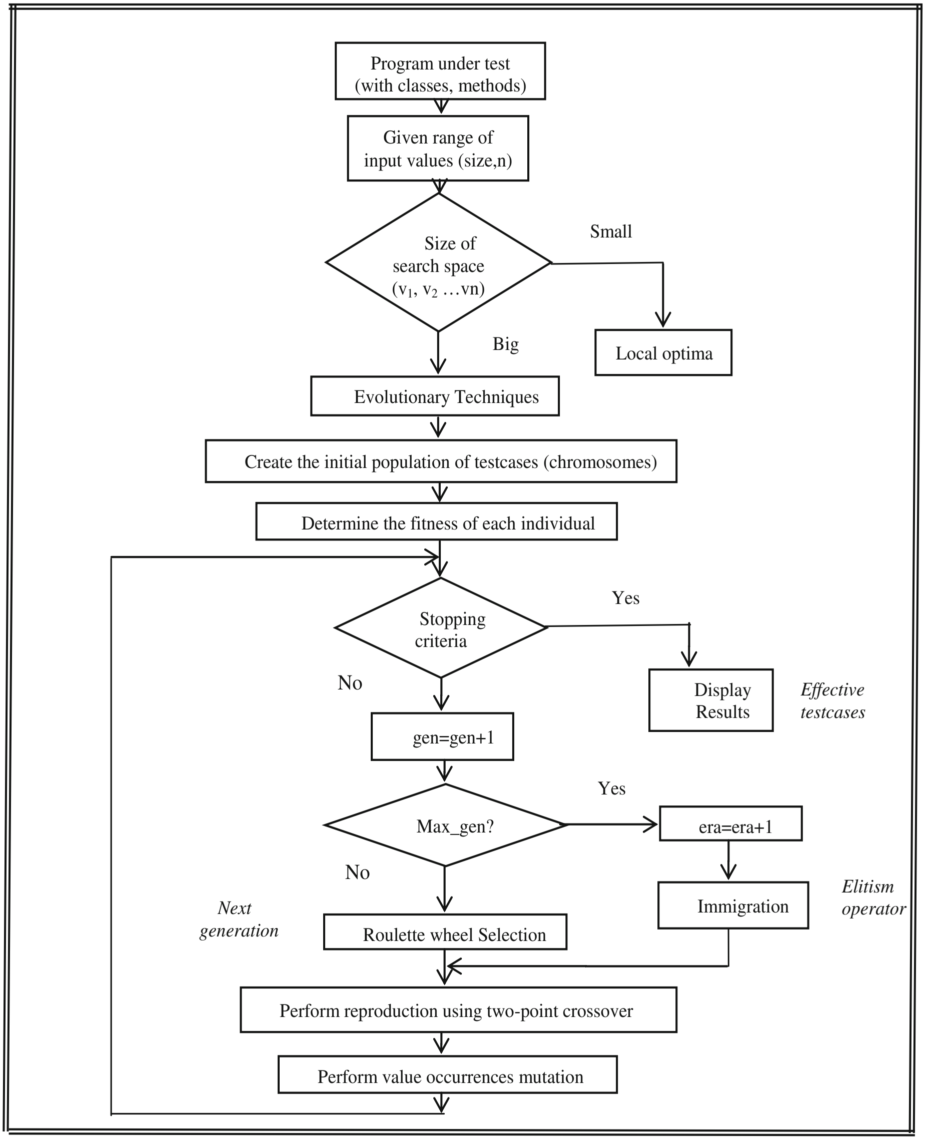


Fig. 1. Interaction of evolutionary testing with Class-Based Elitist Genetic Algorithm

4.3.4 Immigration Rate

Elitism improves the performance and convergence of the GA. Immigration is an elitist operator that controls the migration of test cases from one era to next era.

$$Immigration\ rate(I) = \frac{no\ of\ test\ cases\ reduced\ to\ next\ generation}{no\ of\ test\ cases\ in\ the\ current\ generation}$$

4.4 Fitness Function

The fitness value of the test case is calculated using two fitness functions f1 and f2 since we are considering two objectives for optimization namely maximization of coverage and minimization of execution time. The first fitness function is a maximization of coverage and the second one is a minimization function. The functions are given by

$$f1(x) = \frac{coverednodes}{totalnodes} \tag{1}$$

$$f2(x) = \frac{executiontimeofatestcase}{totalexecutiontime} \tag{2}$$

$$f(x) = \max\{f1(x)\} + \min\{f2(x)\} \tag{3}$$

The optimal test cases are chosen by their fitness values that are calculated based on the values from these two fitness functions. The best test case will have value equal to 1 for the function f1 and will have the value approximately equal to zero for the function f2.

5 Experiments and Result Analysis

The CBEGA is used for test data generation in java with two stages. The CBEGA has been tried several times with different values of population size (50, 75, 100…), mutation probability and crossover probability (merely equals 1). The effectiveness of CBEGA is studied by applying the algorithm on simple java classes.

The implementation starts with the Class Control Flow Graph (CCFG) of the class which is drawn with the help of a set of def-use associations identified from the Class under Test (CUT). CBEGA initializes each sequence by a randomly selected constructor and the methods. Stage 1 generates or updates sequences of method calls. Stage 2 performs the traditional genetic algorithm to generate the required parameters.

The execution of genetic algorithm repeats till all the test requirements is satisfied or the maximum number of eras is reached or until getting the optimal solution with fitness value nearly equals 1. The final test cases are the minimal test set obtained from the resultant test cases of GA which satisfy all the target test requirements. Thus, the minimum no. of test cases is found for testing a given class.

The sample programs taken have the no. of conditions in the range 18 to 42 with Stack having 18 conditions and Bit Set containing 42 conditions. The immigration rate is set to 0.5 and no. of generations per era is 30.The execution time increases with increasing number of conditions as shown in the Fig. 3. From Fig. 2 it is inferred that coverage is greater in CBEGA than in simple GA. Thus when compared to simple GA, CBEGA performs better in terms of efficiency in generating test cases. The results obtained for sample java programs are shown in table 1.

Table 1. The results obtained for the sample java classes comparing GA and CBEGA

Sample programs	No. of eras	No. of generations	No. of conditions	Coverage for GA%	Coverage for CBEGA%	Execution time (ms) CBEGA
Lower complexity Programs						
LinkedList	3	80	20	85	95	2100
Stack	6	170	18	83	98	2082
TreeMap	5	150	25	92	96	2445
BinomialHeap	3	95	38	92	97	3340
Bitset	3	80	42	90	95	4025
HashMap	3	100	32	90	96	2785
BinarySearchTree	4	120	24	83	91	2235
FibonocciHeap	3	90	36	88	97	2910
Disjset	3	95	40	90	95	3695
TreeSet	4	110	28	89	96	2612
Somewhat higher complexity Programs						
Car crash crisis management system	7	200	54	86	96	5015
eHealth system	6	180	48	88	95	4923
Traffic collision avoidance system	3	95	47	87	97	4265
N-queens problem	4	110	52	85	97	4420
Chess playing	4	110	58	86	96	5150

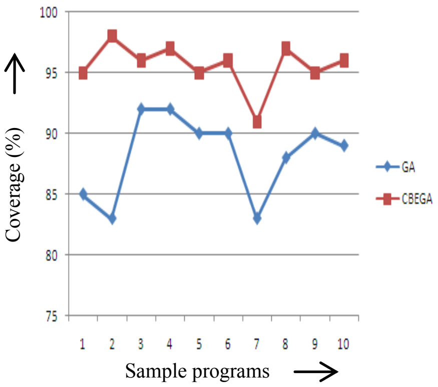


Fig. 2. Graph showing coverage obtained for GA and CBEGA

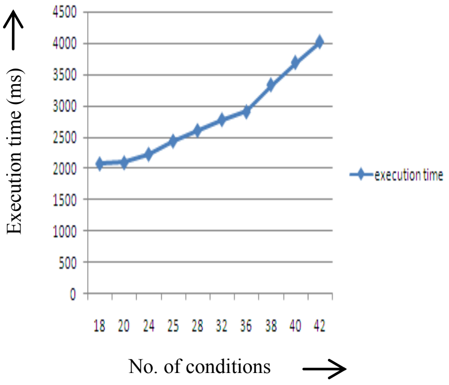


Fig. 3. Graph showing execution time for CBEGA

6 Conclusion

Thus, the class based elitist genetic algorithm has been used for automatic generation of object-oriented test cases. The fitness depends on the path coverage and execution time of the test cases. The results for sample java programs show that the efficiency of CBEGA over simple GA in terms of coverage. When the no of classes and conditions increases, the execution time will also be increased. This algorithm can be used for similar type of real-world software engineering problems.

References

1. Sukstrienwong, A.: Solving multi-objective optimization under bounds by genetic algorithms. International Journal of Computers 5(1), 18–25 (2011)
2. Ghiduk, A.S.: Automatic Generation of Object-Oriented Tests with a Multistage-Based Genetic Algorithm. Journal of Computers 5(10), 1560–1569 (2010)
3. Singh, D.P., Khare, A.: Different Aspects of Evolutionary Algorithms, Multi-Objective Optimization Algorithms and Application Domain. International Journal of Advanced Networking and Applications 2(04), 770–775 (2011)
4. Harman, M., Kiranlakhotia, McMinn, P.: A Multi-Objective Approach to Search-Based Test Data Generation. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2007, pp. 1–8 (2007)
5. Konak, A., Coit, D.W., Smith, A.E.: Multi-objective optimization using genetic algorithms: A tutorial. Reliability Engineering and System Safety, pp. 992–1007. Elsevier (2006)
6. Conway, B.A.: A Survey of Methods Available for the Numerical Optimization of Continuous Dynamic Systems. Journal of Optimization Theory and Applications (JOTA) of Springer, 1–36 (2011)
7. Malhotra, R., Garg, M.: An Adequacy Based Test Data Generation Technique Using Genetic Algorithms. Journal of Information Processing Systems 7(2), 363–384 (2011)

8. Andreou, A.S., Economides, K.A., Sofokleous, A.A.: An Automatic software test-data generation scheme based on data flow criteria and genetic algorithms. In: 7th International Conference on Computer and Information Technology, pp. 867–872 (2007)
9. Chen, Y., Zhong, Y.: Automatic Path-oriented Test Data Generation Using a Multi-population Genetic Algorithm. In: Fourth International Conference on Natural Computation, pp. 566–570 (2008)
10. Deb, K.: Single and Multi-Objective Optimization using Evolutionary Computation. KanGALRt- No. 2004002, Technical Report, pp. 1–24 (2005)
11. Zhang, Y.: Multi-Objective Search-Based Requirements Selection and Optimization. Ph.D Thesis, University of London, pp. 1–276 (2010)
12. Coello, C.A., Lamont, G.B., Van Veldhuizen, D.A.: Evolutionary Algorithms for Solving Multi-Objective Problems, 2nd edn. Springer (2007)

The Simulation of the Assembling Production Process

Róbert Pauliček, Tomáš Haluška, and Pavel Važan

Slovak University of Technology, FMST, Trnava, Slovakia
{robert.paulicek,tomas.haluska,pavel.vazan}@stuba.sk

Abstract. This contribution presented the using of a Lanner Group's Witness PWE simulator and its modeling resource PF network for creating a model of the automotive bumpers assembly system. The article also describes a solution of the problem with different types of the attribute in data loading. The subsequent simulation experiments with the model is aim to setting the minimal frequency of the input arrival times for the present orders plan and for the two estimated orders plans.

1 Introduction

Simulation is the imitation of the operation of a real-world process or system over time. Simulation involves the generation of an artificial history of the system, and the observation of that artificial history to draw inferences concerning the operating characteristics of the real system that is represented.

Simulation is an indispensable problem-solving methodology for the solution of many real-world problems. Simulation is used to describe and analyze the behavior of a system, ask "what if" questions about the real system, and aid in the design of real systems. Both existing and conceptual systems can be modeled with simulation [1].

Discrete event systems (DES) are dynamic systems which evolve in time by the occurrence of events at possibly irregular time intervals. DES abounds in real-world applications. Examples include traffic systems, flexible manufacturing systems, computer-communications systems, production lines, coherent lifetime systems, and flow networks. Most of these systems can be modeled in terms of discrete events whose occurrence causes the system to change from one state to another. In designing, analyzing and operating such complex systems, one is interested not only in performance evaluation but also in sensitivity analysis and optimization [2].

A discrete-event simulation model is defined as one in which the state variables change only at those discrete points in time at which events occur [1].

Computer simulation is the discipline of designing a model of an actual or theoretical physical system, executing the model on a digital computer, and analyzing the execution output [3].

2 Analysis of the Production System

The manufacture is oriented to the bumpers production for the passenger cars. On the input of the production is a small storage space and therefore are the bumpers

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 273–278.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

bringing into the assembly hall in the small batches and arrival intervals. The bumper is hanging at the beginning of the production process on the one of seven suspended clip of the chain conveyors and sequentially passing through the six workstations on which are on the bumper performing the operations and modifications to the required state. At the last workstation are the bumpers hanging down and transfer to a warehouse for the finished products.

In production is on each workstation operating just one bumper at the time and between two workstations on the chain conveyor is free space for one more piece. The frequency of the input arrival is one bumper every 2 minutes.

The company is producing 35 types of the bumpers. Every type has different time of the operation on each workstation. The bumpers are not producing in equal numbers for each type, but according to the received orders. Into the production are bumpers entering in the order in which they are introducing into the input store.

In the following manufacturing process is necessary to find out how busy will be the workstations, what will be a production lean time per one part and how many bumpers will be finished in one continuous operation day (1440 min.) for the different inter arrival periods of the bumpers into the production and find out the effect of the different input orders plans proportions on the production processes.

3 Building of the Simulation Model

Simulation program Witness PWE uses 4 methods of the parts transportation. It is the conveyors, paths, tracks whit the use of the vehicles and PF (power and free) networks. PF networks consisting of the PF sections which connecting the PF stations and along which are moving the PF carriers. All these PF components are assigned to the main component PF net [4].

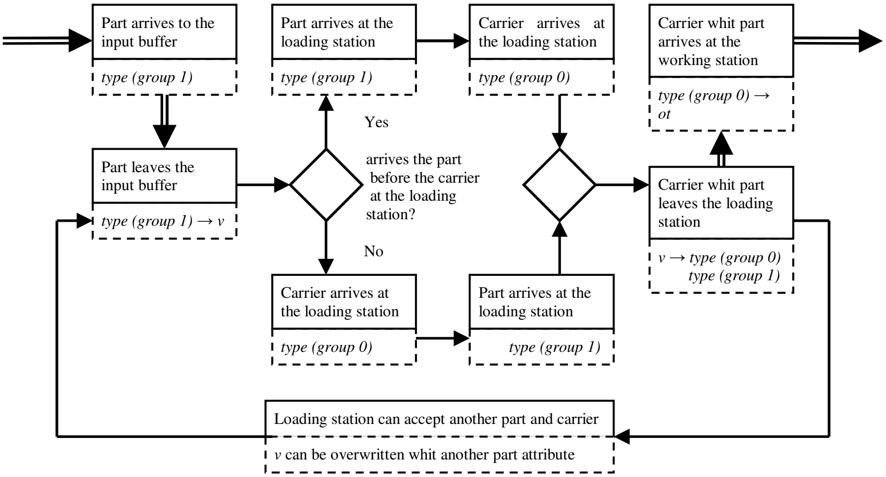
We are decided to use the PF network to solve this specified problem, for their ability to most plausible simulates the chain conveyors with workstations. PF network can also easily adjust to a limited number of the transport carriers and transport capacity of the single conveyor.

In the model is used one PF network for which is defined 6 transport sections, one carrier whit the quantity 7 and 7 workstations, from which 5 workstations was manufacturing, one loading and one manufacturing/unloading. On the input of the PF network is one buffer, into which are the parts arriving and on the output is the buffer whit quantity 35, where is for every type of the bumper reserved one buffer.

For the differentiation of the bumper type is used probability distribution, which generated the number from 1 to 35 by the defined ratios, which is subsequently added to the part as an attribute. For better overview of the bumper type is during the simulation this number added to the visual display of the part. In the modeling of the inputs we have used statistical distribution because the company is unable to declare the precise sequence of the orders and the number of the parts in these orders. Implemented input represents the statistical distribution of the long-term monitoring of the processing orders.

The times of the operations, which are on the beginning of the simulation loaded from the MS Excel, are assigning to the workstations according by the carrier attribute. The real manufacturing process have only 6 workstations, but for the simulation is needed to add one more workstation which serves only to loading parts on the carriers and have zero operation time. This is because the action "action on start" on the loading workstation is executed by the arriving of the free carrier. Otherwise it is not possible to load time of the operation, which is depending on the bumper type, when the bumper itself is not at the workstation. Therefore is preferable to use a loading workstation whit the zero time of the operation which is directly linked to the first manufacturing workstation without a PF section. This workstation is used only to assign the bumper to the carrier and the first manufacturing operation is performed on the next workstation, where is the bumper coming already included with the carrier.

At the distinguishing of the bumper type is used predefined attribute for part, which is default in attribute group 1. Throughout the simulation of the production process, which is progressing into the PF network, is possible distinguish the carriers only by attribute group 0. Therefore is in the model necessary to create another attribute (group 0) and assign it the value from the main attribute (group 1). For this attribute assignment is also necessary to create auxiliary variable (all model components use same variable group), because it is not possible to exchange the values between a different groups of the attributes. The assigning of the bumper attribute (group 1) value to the variable is taking place at the leaving of the part from the input buffer (actions on output) and this variable value is assign to the carrier attribute (group 0) at the end of the operation on the loading station (actions on finish). The auxiliary variable can not be accidentally overwriting during the assigning, because the bumper can not leave the input buffer before the carrier leave the loading station. The values transfer between different attributes groups are illustrated on the figure 1.

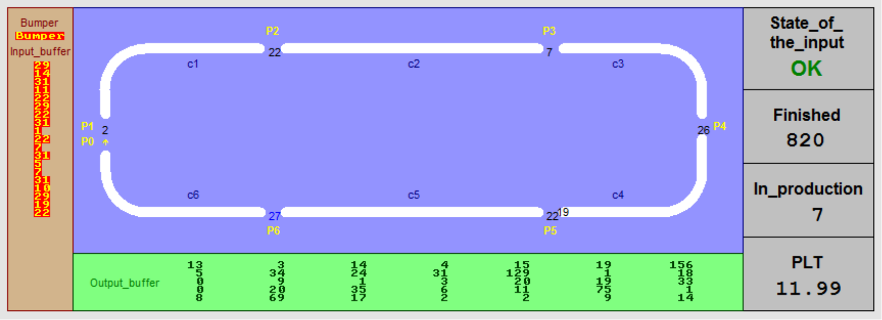

Part arrives to the input buffer
type (group 1)
Part arrives at the loading station
type (group 1)
Carrier arrives at the loading station
type (group 0)
Carrier whit part arrives at the working station
type (group 0) → ot
Yes
Part leaves the input buffer
type (group 1) → v
arrives the part before the carrier at the loading station?
Carrier whit part leaves the loading station
v → type (group 0)
type (group 1)
No
Carrier arrives at the loading station
type (group 0)
Part arrives at the loading station
type (group 1)
Loading station can accept another part and carrier
v can be overwritten whit another part attribute

type (group 1) = part attribute, *type (group 0)* = carrier attribute, *v* = auxiliary variable, *ot* = operation time on working station

Fig. 1. The scheme of the values transfer cycle between attributes

Potential overload of the input buffer capacity is notified on the right side panel of the model (State_of_the_input), where is also displayed a basic data like the number of the finished bumpers (Finished), the part in the manufacturing process (In_production) and the production lean time (PLT). When is the capacity of the input buffer overloaded, it means that the bumpers in this order will be not finished on the time.

On the figure 2 is imaged the model of the bumpers production process.



State_of_the_input
OK

Finished
820

In_production
7

PLT
11.99

Fig. 2. Model of the bumpers production process

The designed simulation model was validated according to real requirements. The preparing experiments were designed for the validation process. The simulation model was the step by step calibrated according to results experiments [4]. Finally full equality was achieved with the real production system.

4 Experiments

On the model we have tested the different input arrival intervals of the bumpers and their variants of the order plans. In the first series of the experiments was used the input orders plan based on the present statistical distribution of the bumpers in 6 months. The range of the input arrival was one bumper per 1,5 – 2 minutes. The result of the experiment is on the figure 3.

The experiment shows that reducing the input arrival interval increase the number of the produced bumpers, but at the interval 1,7 has the increase stopped. It is because the input buffer has overcrowded and the bumpers will be not possible to produce in time. Increasing of the number of the produced pieces also increase the production lean time (PLT) of one bumper. From the chart also view the decreasing usage of the loading station P0. That means that on the loading station was less free carriers (by the interval 2 (99,9% usage) has bumper waited for the carrier and by the interval 1,5 (1% usage) has carrier waited for the bumper). In the additional experiments we have tested intervals 1,6 and 1,7 with more carriers, but it had negligible effect on the total number of the produced bumpers and production lean time.

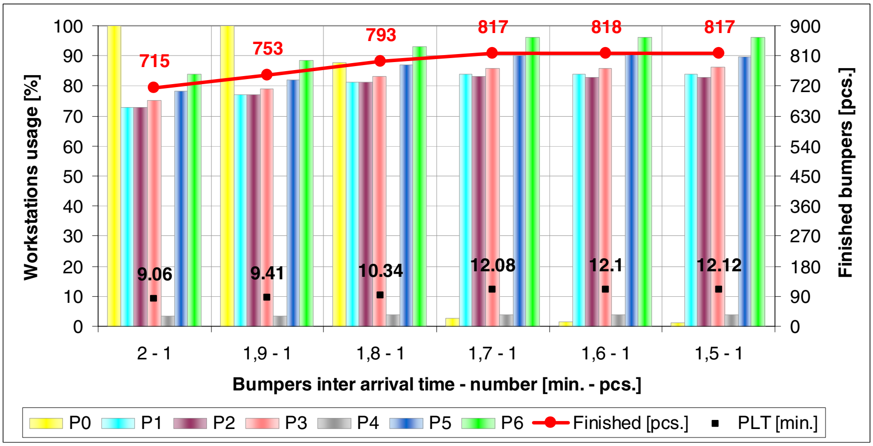


Fig. 3. Result of the experiment whit the present orders plan

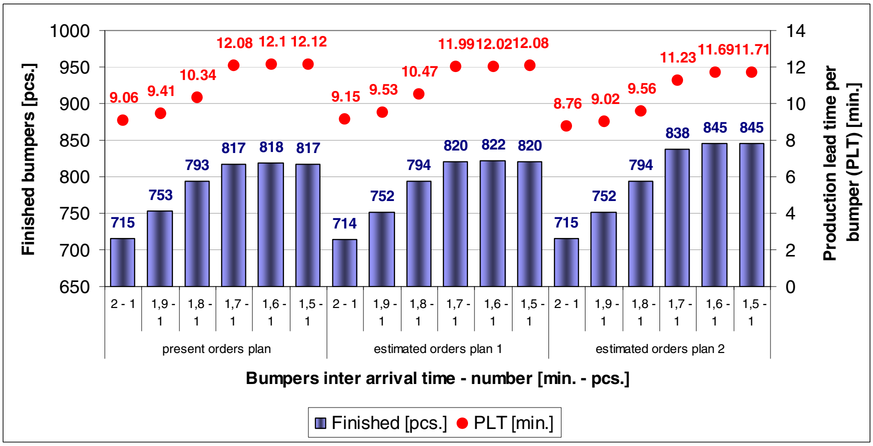


Fig. 4. Result of the experiment whit the present and estimated orders plans

On the figure 4 are the results with one present and two estimated orders plans without usage of the workstations, because the usage of two estimated orders plans have same character like a present orders plan. The estimated orders plan number 1 calculated only with the most probably changes to the current state (more then 80%). Therefore are the results very similar to the present orders plan. The estimate orders plan number 2 calculated also with the less probably changes (more then 40%). At this plan significantly increased the number of the produced bumpers and reduced the production lean time. The overloading of the input buffer occurred in all cases at the same input arrival interval 1,6.

5 Conclusion

PF network proved to be an excellent modeling tool for this production system. The model was created more easily than if we used another method of the transportation.

The results of the experiments, where the input was adjusted by the real daily sample which values were measured directly from the production, we have obtain the conformity of the simulation results with the real results with the deviation smaller than 3%. This difference is caused by our deterministic model, whereby in the real system there are small deviations of the prescribed technological operation time.

By executing of the experiments we have set the minimal input arrival interval to one bumper per 1,7 minute for all three orders plans. If it is necessary, the company is able to produce by this interval about 100 more bumpers per all-day operation without overloading of the manufacturing process.

The tested ratios of the bumpers inputs to the system and their accidental character of the inputs adversely affecting the exact production planning. If it would be possible to set up the exact deterministic inputs of the individual bumpers, the production planning in terms of the direct determination of the production lean time, thereby it would be possible to simulate the deadlines more detailed.

References

[1] Banks, J.: Discrete event simulation. In: Farrington, P.A., Nembhard, H.B., Sturrock, D.T., Evans, G.W. (eds.) The Proceedings of the 1999 Winter Simulation Conference, Phoenix (1999) ISBN 0780357817

[2] Arsham, H.: System Simulation: The Shortest Route to Applications, `http://home.ubalt.edu/ntsbarsh/simulation/sim.htm#rintroduction`

[3] Fishwick, P.A.: What is simulation?, `http://www.cise.ufl.edu/~fishwick/introsim/node1.html`

[4] Lanner Group Ltd, Witness technology for knowing: Manufacturing Performance Edition: Tutorial Manual. United Kingdom, p. 63 (2007)

[5] Masár, A., Tanuška, P., Masárová, R.: Possible particular abstract approach to validation. In: Annals of DAAAM for 2009 & Proceedings of the 20th International DAAAM Symposium "Intelligent Manufacturing & Automation: Focus on Theory, Practice and Education", vol. 20(1), pp. 175–176 (2009) ISSN 1726-9679, ISBN 978-3-901509-70-4

Parallel Performance of Numerical Algorithms on Multi-core System Using OpenMP

Sanjay Kumar Sharma and Kusum Gupta

Banasthali University, Banasthali, Rajasthan, India
skumar2.sharma@gmail.com, gupta_kusum@yahoo.com

Abstract. The current microprocessors are concentrating on the multiprocessor or multi-core system architecture. The parallel algorithms are recently focusing on multi-core system to take full utilization of multiple processors available in the system. The design of parallel algorithm and performance measurement is the major issue on today's multi-core environment. Numerical problems arise in almost every branch of science which requires fast solution .System of linear equations has applications in fusion energy, structural engineering, ocean modeling and method of moment formulation. In this paper parallel algorithms for computing the solution of system of linear equations and approximate value of π are presented. The parallel performance of numerical algorithms on multi-core system have been analyzed and presented. The experimental results reveal that the performances of parallel algorithms are better than sequential. We implemented the parallel algorithms using multithreading features of OpenMP.

Keywords: Multi-core processors, Parallelization, Parallel computation, Parallel algorithm, Performance analysis.

1 Introduction

With the invention of multi-core architecture, every laptop and desktop machine is now shared memory parallel computer. The conventional parallel computing methods focuses on multi-core architecture where multiple cores are integrated into a single CPU package [1]. In multi-core environment the sequential computing paradigm is not good and inefficient, while the usual parallel computing may be suitable. One of the most important numerical problems is solution of system of linear equations. Systems of linear equations arise in the science domain such as *fusion energy*, *structural engineering and method of moment formulation of Maxwell equation.* It is also use in mathematical modeling of numerous problems in discipline such as *applied mathematics*, and *physical* and *social sciences*. Thus it has great importance to design the parallel algorithm for some numerical problems which are frequently used in different science domain and test their performances on multi-core system.

There are some numerical problems which are large and complex; solutions of which are not efficient using sequential algorithm on a single processor machine or on multiprocessor machine. The solution of these problems can be obtained efficiently using parallel algorithm on multi-core or multiprocessor system. In this paper we

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 279–288.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

select two numerical problems. The first problem is to approximately compute the value of π using method of integration, and second is solution of system of linear equations [2]. Parallel algorithms for these numerical problems have been presented which are effective and more efficient than their corresponding sequential algorithms. We tested the performances of these parallel algorithms by measuring their execution times.

2 Multi-core Technology

Multi-core technology means having more than one core inside a single chip. This opens a way to the parallel computation, where multiple parts of the program are executed in parallel at same time [3]. The factor motivated the design of parallel algorithm for multi-core system is the performance. The performance of parallel algorithm is sensitive to number of cores available in the system, core to core latencies, memory hierarchy design, and synchronization costs. The software development tools must abstract these variations so that software performance continues to obtain the benefits of the Moore's law.

The multithreading is a technique which allow the programmer to specify the portion of code to execute in parallel with other codes, for this it requires an additional efforts in coding which are very difficult and complex. There are two threading methods available: *library based* and *compiler directed.* The library based threading (Win32 multithreading API on Windows, the Pthreads library on Linux) requires programmer to manually map concurrent task to threads. The libraries also give programmer control over the low level aspect of thread creation, thread management and synchronization. Threading an existing serial application with a library based method is a dangerous process and requires significant code modification. The directives based threading method allows the programmer to use the compiler directives to specify the region of code to execute in parallel with the other codes. The directive provides *constructs* to *create number of threads, thread synchronization* and *other constructs.* In *C/C++*, the directives are implemented as *#pragma* statement [4].

3 OpenMP

An OpenMP Application Programming Interface (API) was developed to enable shared memory parallel programming. OpenMP API is a set of compiler directives, library routines, and environment variables to specify shared-memory parallelism in *FORTRAN* and *C/C++* programs. OpenMP provides the straight forward approach to implement parallel algorithm [5]. It provides three kinds of directives: *parallel work sharing, data environment and synchronization* to exploit the multi-core, multithreaded processors. The OpenMP provides means for the programmer to: create teams of thread for parallel execution, specify how to share work among the member of team, declare both shared and private variables, and synchronize threads and enable them to perform certain operations exclusively (i.e. without interference by other threads).

A. Creation of Programs using OpenMP

OpenMP's directives let the user tell the compiler which instructions to execute in parallel and how to distribute them among the threads that will run the code. Since the parallel algorithm specify the section of code to execute in parallel. When implementing parallel algorithm, programmer has to decide which directives are used to distribute the parallel work to different processors. *OpenMP* work in conjunction with set of the programming languages C/C++. In C/C++ the directives are implemented as *#pragma* statements. The directive based parallelization approach has the advantage that it allows the same source code to be used for single and multiprocessor development, since the code will be executed serially on single core and in parallel on multi-core processors.

The OpenMP is based on *fork/join* execution model. The control structures for parallelization are embedded into to *fork/join* execution model. Thus they fork (i.e. start) new threads and execute the enclosed code block concurrently, and afterwards they join in parallel running threads to serial master thread [5]. By means of *work sharing directive* the work within code block can be divided among existing team of threads. An instance for this is *for* directive, which divides loop iterations among concurrently executing threads. This exploits the *loop level parallelism.* The required thread synchronization is done implicitly by *OpenMP* at the end of parallel region or explicitly by programmer through directive like *barrier* (wait until barrier is reached by all threads) or *critical.* The parallelization of for-loop is given as an example.

```
… /* serial code          */
#pragma omp parallel for
{
   for ( i=0; i< 100; i++)
      x[i]  +=compute( y[i] );
}
/* synchronization */
```

B. Parallelism

There are basically two type of parallelism: *data level parallelism* and *task level parallelism* [6]. Data level parallelism is when data can be divided into a certain number of partitions. There are two different methods to distributing data to threads: *static and dynamic*. In static distribution, the iterations are divided into chunk and chunks are statistically assigned to threads in round robin fashion.

In dynamic distribution the data is divided into chunks and each chunk is assigned when thread send request for it. The thread executes the chunk then requests another chunk until no chunk remains assigned. In general, static distribution has very low CPU overhead while dynamics distribution offers potentially better load balancing. In task level parallelism is when the tasks themselves are divided into a number of threads and task should be relatively independent to avoid synchronization overhead. *OpenMP* is designed to express data parallelism in which the threads perform the same task on different data. We have used the data level parallelism in the programs of parallel algorithms.

4 Evaluation of Parallel Algorithm

For designing the parallel algorithm, first step is to partition the problem into many parallel tasks, and second step is the performance of parallel algorithm. The criteria that have been used to evaluate the performance of parallel algorithm is speedup [7]. It is defined as

$$Sp = T_1 / T_P$$

Where, T_1 denotes the execution time of the best known sequential algorithm on single processor machine, and Tp is the execution time of parallel algorithm on P processor machine. In other word, *speedup* refers to how much the parallel algorithm is faster than the corresponding sequential algorithm. The linear or ideal speedup is obtained when $Sp=P$. When running the algorithm with linear speedup, doubling the number of processors, doubles the speedup. As this is ideal, it is a very good *scalability*.

5 Classical Parallel Algorithms

A. Calculation of π

The function $f(x) = \frac{1}{1+x^2}$ can be used to approximate the value of π using numerical integration. We get the value of π using the function $f(x)$. Consider the evaluation of definite integral

$$I = \int_0^1 \frac{1}{1+x^2}dx = \arctan(1) - \arctan(0) = \frac{\pi}{4}$$

where, $f(x)$ is called the integrand, a lower, and b is upper limits of integration.

Integration of function $f(x)$ can be evaluated numerically by splitting interval [a, b] into n equally spaced subintervals. We assume that the intervals are constant and its interval width is $h= (b-a)/n$. Method of integration (Simpson 1/3 rule) have been used to approximately compute the values of π [8]. The formula and the parallel algorithm are given below.

$$\int_a^b f(x)dx \approx \frac{h}{3}\left[f(x_0) + 4\sum_{\substack{i=1\\ i=odd}}^{n-1} f(x_i) + 2\sum_{\substack{i=2\\ i=even}}^{n-2} f(x_i) + f(x_n)\right]$$

Parallel Algorithms- Simpson 1/3 Rule:

```
Input a, b and N   // a and b are limits, N no. of intervals
h = (b-a) / N
sum= f(a) + f(b) + 4*f(a+h)
  parallel for i = 3 to N-1 step +2 do
        x = 2*f (a+ (i-1) *h + 4 * f (a+ i*h)
        sum = sum+ x
  End loop [i]
Int = (h/3) * sum
Print 'Integral is = ', Int
```

We implemented above parallel algorithm using *OpenMP*. In the program *#pragma* directives have been used for parallel computation and *reduction* clause for thread synchronization. Each thread calculates the value of *f(x)* in parallel to other threads.

B. Solution of System of Linear equations

Solution of system of linear equations is assignment of value to variables that satisfy the equations. To solve the system of linear equations, we considered the direct method: *Gaussian elimination*. It is a numerical method for solving the system of linear equations $AX = B$, where A is a known matrix of size $n\times n$, X is the required solution vector, and B is a known vector of size n. In the process, the system of equations $AX=B$ is reduced to an upper triangular system which is solved using back substitution. Consider the n linear equation in n unknowns as:

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = a_{1,n+1} \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = a_{2,n+1} \\ a_{31}x_1 + a_{32}x_2 + \dots + a_{3n}x_n = a_{3,n+1} \\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = a_{n,n+1}, \end{array} \tag{1}$$

where, $a_{i,j}$ and $a_{i,j+1}$ are known constants and x_i's are unknowns.

The equation (1) is equivalent to

$$AX = B \tag{2}$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{pmatrix} \begin{pmatrix} a_{1,n+1} \\ a_{2,n+1} \\ a_{3,n+1} \\ \dots \\ a_{n,n+1} \end{pmatrix}$$

The sequential algorithm consists of two phases. In the first phase, the original equations are reduced to an upper triangular form $AX = B$ where A is a matrix of size $n\times n$ in which all elements below the main diagonal are zeros.

In second phase the upper triangular matrix is used to obtain the value of unknown's x_i. In the algorithm, vector B is not taken separately. The vector B is stored in (n+1)th column of matrix A.

Sequential algorithm:

```
Input: Given Matrix a [1: n, 1: n+1]
Output: x [1: n]
// Traingularization process
for k = 1 to n-1
   for i = k+1 to n
```

```
                $m_{i,k} = a_{i,k} / a_{k,k}$
          for j = k to n+1
                $a_{i,j} = a_{i,j} - m_{i,k} * a_{k,j}$
          End loop [j]
      End loop [i]
  End loop [k]
 // Back substitution process
          $x_n = a_{n,n+1} / a_{n,n}$
    for  i = n-1 to 1 step -1 do
            sum = 0
        for j = i+1 to n do
              sum = sum + $a_{i,j} * x_j$
          End loop [j]
            $x_i = ( a_{i,n+1} - sum )/a_{i,i}$
   End loop[i]
```

In the sequential algorithm, the innermost loops indexed by *i* and *j* can be executed in parallel without affecting the result. In the parallel algorithm, we used *#pragma* directive to parallelize the loops.

Parallel algorithm:

```
 Input: Given Matrix a [1: n, 1: n+1]
 Output: x [1: n]
 // Traingularization process
 omp_set_num_threads(omp_get_num_procs());   //set the numbr of threads
 DWORD startTime=timeGetTime();              //get  start time
 for k = 1 to n-1
   #pragma omp parallel for private(i, t, k) shared(n, a) schedule(static,1)
     for i = k+1 to n
                $m_{i,k} = a_{i,k} / a_{k,k}$
          for j = k to n+1
                $a_{i,j} = a_{i,j} - m_{i,k} * a_{k,j}$
          End loop [j]
      End loop [i]
  End loop [k]

 // Back substitution process
          $x_n = a_{n,n+1} / a_{n,n}$
    for  i = n-1 to 1 step -1 do
            sum = 0
        for j = i+1 to n do
              sum = sum + $a_{i,j} * x_j$
          End loop [j]
            $x_i = ( a_{i,n+1} - sum )/a_{i,i}$
    End loop[i]
    DWORD endTime=timeGetTime();       //get  end time
    Print  “Number of equations= “, n ;
    Print “ Parallel execution took " , endTime-  startTime
  // print solution vector
     for(int i=0;i<n;i++)
          print x[i]
     End loop [i]
```

6 Performance Evaluation

We implemented the sequential and parallel algorithms for computing the value of π and the solution of system of linear equations and also computed their performances on multi-core system. The Intel C++ compiler 10.0 under Microsoft Visual Studio 8.0 used for compilations and executions. The Intel C++ compiler supports multithreaded parallelism with */Qopenmp* flag. All the experimental data presented in the Tables 1 & 2 have been collected on 2.4 GHz Intel@Core2-Duo processor machine. We have used the Origin6.1 software to plot the graph using the data obtained by the experiments.

7 Experimental Results

We implemented all the algorithms on a PC with Intel®Core-2 Duo machine, 2GB RAM, 2.66 GHz processor speed. Execution times of both the sequential and parallel algorithms have been recorded to measure the performance of parallel algorithm against sequential. In the first experiment we computed the value of π. We used π value from *Mathematica* to compare the accuracy of computed value. *Mathematica* is known for its capability to do computations with arbitrary precision (π=3.1415926495897932384626431).

Table 1. Performance comparison of sequential and parallel algorithm to compute the value of π

Sr. No.	No. of Intervals	Sequential Execution Time & Difference between results		Parallel Execution Time & Difference between results	
		Time (m. s.)	Difference	Time (m. s.)	Difference
1	1000	0	$-1.67*10^{-7}$	0	$-1.67*10^{-7}$
2	10000	0	$-1.67*10^{-9}$	0	$-1.67* 10^{-9}$
3	100000	15	$-1.67*10^{-11}$	0	$-1.67*10^{-11}$
4	1000000	78	$-4.44*10^{-16}$	32	$-2.04*10^{-16}$

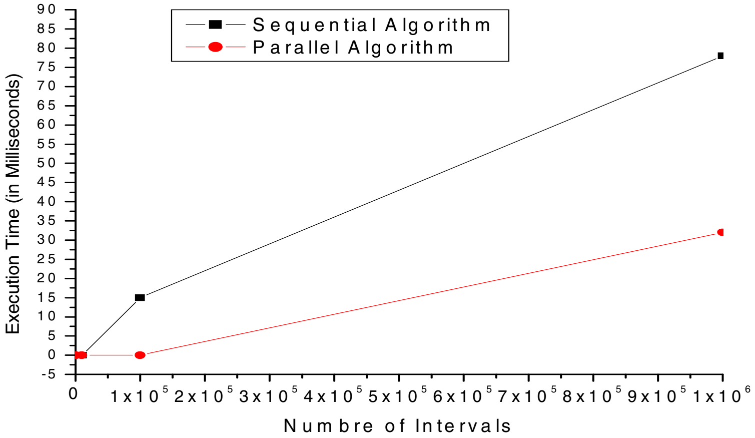


Fig. 1. Execution times of sequential and parallel computation of π

The data presented in Table 1 represents the execution time (in milliseconds) taken by the sequential and parallel algorithms and the difference between *Mathematica's* value of π. The results show that as the number of intervals is increased, the accuracy of result also increased. We plot the graph using the data in Table 1 to analyze the performance of parallel algorithm which is shown in fig 1. It shows that the parallel algorithm saves significant amounts of execution time and gives more efficient result. The parallel algorithm is faster than their corresponding sequential algorithm.

In second experiment, we implemented the sequential and parallel algorithms for finding the solution of system of linear equations. We tested both the algorithms on the equations of different sizes and recorded their execution times.

Table 2. Performance comparison of sequential and parallel algorithms

No. of Equations	Sequential Execution Time (m. s.)	Parallel Execution Time (m. s.)
100	15.5	0
125	31.5	15.5
150	46.5	31.25
175	70.5	36.5
200	124.5	46.5
225	183.25	70.25
250	249.5	93.25
275	327.5	124.5
300	379.25	172

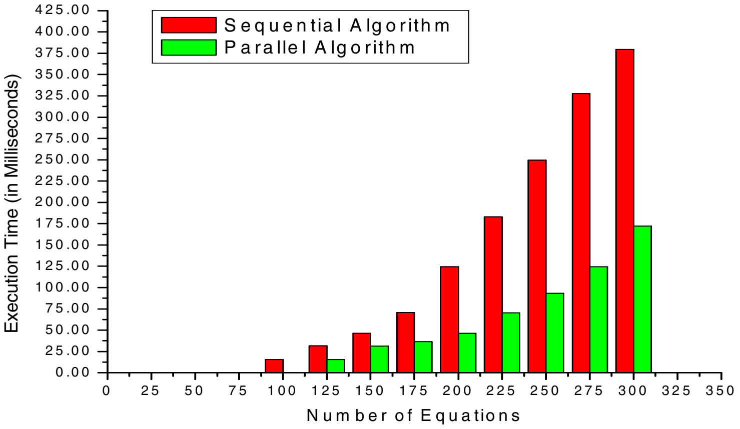


Fig. 2. Execution time of sequential and parallel algorithm of Gauss elimination algorithms

The data in Table 2 represent the execution times taken by sequential and algorithms for the solution of system of linear equations of different sizes. The result shows that the parallel algorithm is efficient than their corresponding sequential algorithm. We plot the graph using data in Table 2 to analyze the performance of parallel algorithm is presented in Fig.2. It shows that the parallel algorithm save significant amounts of execution time and gives more efficient results. The speedup of parallel algorithm on average is approximately twice than their corresponding sequential algorithm.

8 Conclusion

We designed and implemented the parallel algorithms using OpenMP for computing the value of π and for the solution of system of linear equations. Our experimental results achieve a noticeable performance in each case. Results show that the parallel algorithms are faster and efficient than their corresponding sequential algorithms.

The programmer has to put additional efforts to carefully divide the program of the algorithm into smaller section which can then be carefully assigned to threads that will execute in parallel on different processors. This technique provides an idea on how to use the parallel directive to measure the performance of parallel algorithm using *OpenMP*. The parallel algorithm contains the *#pragma* directives which exploit the processors available into the multi-core system. The parallel performance also depends on number of processors available in the system as well as on parallel algorithm. It concludes that parallel algorithm accelerates solution of numerical problems as compared to sequential algorithms.

References

[1] Pas, R.V.: Concept in Parallelism. In: IWOMP 2009. Purdue University Wes Lafayette, USA (2008)

[2] Wilkinson, B., Allen, M.: Parallel Programming. Pearson Education, Singapore (2002)

[3] Kulkarni, S.G.: Analysis of multi-Core system performance through OpenMP. In: National Conference on Advanced Computing and Communication Technology, IJTES, vol. 1(2), pp. 189–192 (2010)

[4] Smith, L., Bull, M.: Development of mixed model MPI/ OpenMP applications. Scientific Programming 9(2-3), 83–98 (2001)

[5] Barbara, C., Jost, G., Pas, R.V.: Using OpenMP: portable shared memory parallel programming. The MIT Press, Cambridge (2008)

[6] Chandra, R.: Parallel Programming in OpenMP. Morgan Kaufmann (2001)

[7] Quinn, M.J.: Parallel Programming in C with MPI and OpenMP. McGraw-Hill Higher Education (2004)

[8] Davis, P.J., Rabinowitz, P.: Methods of Numerical Integration. Academic Press (1975)

[9] Eason, G., Noble, B., Sneddon: On certain integrals of Lipschitz- Hankel type involving products of Bessel functions. Phil. Trans. Roy. Soc. London A247, 529–551 (1955)

Efficiency Improvement of a-Si:H/μc-Si:H Tandem Solar Cells by Adding a-SiOx:H Layer between Ag Nano Particles and the Active Layer

Ozcan Abdulkadir[1] and Dhinaharan Nagamalai[2]

[1] Assistant Professor, Dept. of Electrical & Electronic Engineering,
K.T.O. Karatay University, Konya, Turkey
akadirzcn@gmail.com
[2] Network Architect and Security Expert, Wireilla Net Solutions PTY Ltd, Australia

Abstract. In this article, we investigated the positive effect of a-SiOx:H layer between Ag surface plasmon polaritons and the active layer of a thin film a-Si:H/μc-Si:H tandem solar cell by isolating the metal nano particles that are responsible of creating surface recombination centres on the top cell. We fabricated four identical a-Si:H/μc-Si:H tandem cells having different thickness of a-SiOx:H layer like 0, 10, 20 and 30 nm just before the Ag nano particle development, and measured J-V characteristics of each to find out an optimum a-SiOx:H insulating layer thickness. We showed that the overall efficiency of a tandem cell with Ag nano particles could be improved up to 8.65% compared with the one having no a-SiOx:H layer. The most promising layer thickness for a small area tandem cell was obtained around 20 nm with an overall efficiency of 16.19%. An improvement of 14.6% in short circuit current density (JSC) and 2.64% in open circuit voltage (Voc) was achieved.

Keywords: tandem cell, micromorph, solar cell, a-Si:H, μc-Si:H, surface plasmon polariton, nano particle, cell efficiency, insulating layer, efficiency enhancement, solar cell.

1 Introduction

Thin film hydrogenated amorphous silicon (a-Si) and microcrystalline silicon (μc-Si); nowadays the latter also named as nanocrystalline silicon; tandem cells have being improved recently and preferred by many manufacturers thanks to their low cost and simpler production capabilities. Besides, the redshifted light absorbing ability of μc-Si:H solar cells makes them an attractive choice to fabricate tandem cells in order to compensate the poorer absorbance of a-Si:H cells.

Although their efficiencies are still much far away from crystalline silicon (c-Si) cells, this could be overcome by applying some light management techniques like growing textured front and/or back surface, back reflector and benefiting from surface plasmon polaritons (SPP) through nano particles. Among them the latter has a great role to improve efficiency. Enhancing efficiency due to SPP is caused by trapping plane waves of incident light and directed into the active layer. The nano particles

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 289–297.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

fabricated on top of the cell behave as light scattering elements thereby creating higher density of states in the cell that increases the light trapping probability of incoming photons[1]. Conversely they create surface recombination centres for carriers which results in a decrease of efficiency. We have overcome this worse effect by investigating a large bandgap energy (E_g) material such as an a-SiO_x:H insulating layer between the nano particles and the top cell. The E_g of a-SiO_x:H depends on the amount of oxygen-rich phase and starts from 2.32 eV and goes up to 2.70 eV[2].

A second efficiency enhancement technique is to apply an Ag back reflecting layer on the stainless steel (SS) substrate that would reflect the untrapped photons back into the intrinsic (i) layer of the bottom cell, thereby an additional increase in efficiency would be expected. Besides the thickness of i layers in tandem cells plays a great role on collection efficiency. Rigorous diffusion length measurements in polycrystalline and amorphous silicon were reported that they are lying in between 300-400 µm[3] and 2100 Å[4] respectively. Another discrepancy is that the top cell would be as thin as possible while the bottom as thick as possible. The bottom cell requirement can be satisfied by adding a back reflector thereby effectively doubling the photon path; and/or by passivating the top intermediate and bottom surfaces by mostly using thin ZnO and a-SiO_x:H layers as we had applied in this work.

The objective of this study was to find the most suitable a-SiO_x:H layer thickness by constructing four identical tandem cells with different thicknesses of a-SiO_x:H layer on top like 0, 10, 20 and 30 nm. We focused on investigating the effect of the insulating layer thickness on the cell performance.

2 Theory

2.1 Effect of Ag Nano Particles on Cell Efficiency

Light trapping is a critical parameter to enhance the cell performance of thin film solar cells. Naturally, as the thickness of i active layer is reduced, several light trapping methods should be necessary to compensate the decrease in photon absorption, especially in red and near IR portion of the spectrum. One of the most effective way to improve light trapping is to use nano particles developed at the top, middle or bottom of the cell. According to Bohren, C.F. and Huffman, D.R. the effective absorption and scattering cross section of nano particles located in a dielectric medium can be formulated as follows:

$$C_{abs} = \frac{2\pi}{\lambda}.\mathrm{Im}\,[\alpha] \tag{1}$$

$$C_{sca} = \frac{1}{6\pi}.\,(2\pi/\lambda)^4\,.\,\left|\alpha\right|^2 \text{ where} \tag{2}$$

$$\alpha = 3V.(\varepsilon/\varepsilon m - 1) / (\varepsilon/\varepsilon m + 2) \tag{3}$$

Here α is the polarizability of the particle which would be very high at specified frequency, V is the volume, ε is the dielectric permittivity of the metal nano particles and ε_m is the dielectric permittivity of the surrounding medium. Again, the scattering efficiency can be written as:

$$Q_{sca} = \frac{C_{sca}}{C_{sca}+C_{abs}} \quad \text{and} \tag{4}$$

$$Q_{sca} = \frac{C_{sca}}{\pi.R^2} \tag{5}$$

where R is the average radius of the nano particles. The scattering efficiency can be tuned to provide the maximum enhancement at a specified wavelength by varying grain size, shape and dielectric environment of nano particles. The grain size is a crucial factor that directly affects solar cell efficiency. As the size decreases beyond a critical value, they tend to convert the absorbed light into heat, thereby decreasing the efficiency. Conversely, if the size is increased more and more, higher order excitations of multipoles occur, resulting in another efficiency degradation. It is well known that larger sized particles deposited look more likely ellipsoid rather than spherical, causing the resonance frequency to be redshifted[6]. Several investigations have experimentally showed that the optimum grain size of Ag nano particles at a resonance wavelength of 750 - 800 nm in Si is around 20 nm[7]; with an efficiency improvement of ~9%[8,9].

2.2 Effect of a-SiO$_x$:H and Front Contact Layers

When nano particles are directly fabricated on Indium-Tin-Oxide (ITO) layer, the polarizing capability of SPPs decreases drastically since the conductivity of ITO layer is fairly higher than the top a-Si:H cell. Therefore it would be quite beneficial to construct a thin a-SiO$_x$:H layer of the order of tens of nanometers just before Ag deposition to passivate the top surface. The front contact has to combine both low resistance and high transparency to absorb as many photons as possible. In this work we applied a grid-like ITO layer for the front contact since its transparency (84%) and resistivity ($\sim 5\text{x}10^{-4}\Omega.\text{cm}^{-1}$) in the visible range would provide better results as reported in Ref. 10[10].

2.3 The Structure of a-Si:H/µc-Si:H Tandem Cell

Recently a-Si:H cell fabrication has been growing with great success since the materials used are less, very cheap and processing techniques such as plasma-enhanced chemical vapour deposition (PECVD) is being improved to allow large scale manufacturing. The main role of H in Si:H alloy is passivating the naturally broken Si bonds which are introduced by the absence of long range order present in crystalline Si (c-Si) structures. It was reported that the broken band passivation by H addition could reduce electrically active bond density from 10^{19} cm^{-3} to 10^{15} cm^{-3}[11]. Although the E_g of a-Si:H is very large compared to µc-Si:H or other well known photovoltaics (PV), it luckily offers a very high absorption in blue & green range, typically associated with direct bandgap semiconductors such as GaAs, GaSb, InAs, InP, etc. This could be another reason to prefer a-Si:H.

The disadvantage of poor absorption at higher wavelengths could be compensated by means of µc-Si:H bottom cell which is able to absorb less energetic photons having wavelengths more than 700 nm; thanks to its lower E_g of ~1.12eV.

To enhance the light trapping capability of the bottom cell, i layer thickness should be as large as possible. Additionally increased thickness also carries some cost disadvantages. Conversely considerably low drift length of <100 nm and diffusion length of <<100 nm[12] for holes will make it difficult to collect photo generated carriers in a thick device. This discrepancy forces us selecting about 250 nm thick bottom cell with a back reflector that effectively doubles the photon path.

2.4 Back Contact, Back Reflector and Buffer

To improve light trapping capability of the μc-Si:H bottom cell; especially for the weekly absorbed long wavelength photons; the light should be repeatedly scattered and/or reflected by the help of a back reflector. Although the substrate of stainless steel (SS) in our work would reflect some untrapped photons back into the bottom cell, the reflectance is poor compared to Al or Ag. Here we deposited a Ag layer of 50 nm thick to get the back contact with a fairly better reflectance. This layer also prevents the impurity diffusion towards the active bottom layer while deposition.

The ZnO layer between the top and bottom cell provides both buffering and surface passivation. It offers a textured-like layer since ZnO itself is a porous material, thereby an increase in scattering towards the μc-Si:H portion.

3 Experimental Details

We used a 127 μm-thick SS foil (ST5430BA) substrate having a surface dimension of 15 cm x 15 cm to configure the four amorphous/microcrystalline (micromorph) tandem cells simultaneously in order to provide identical cell structures. Each cell has

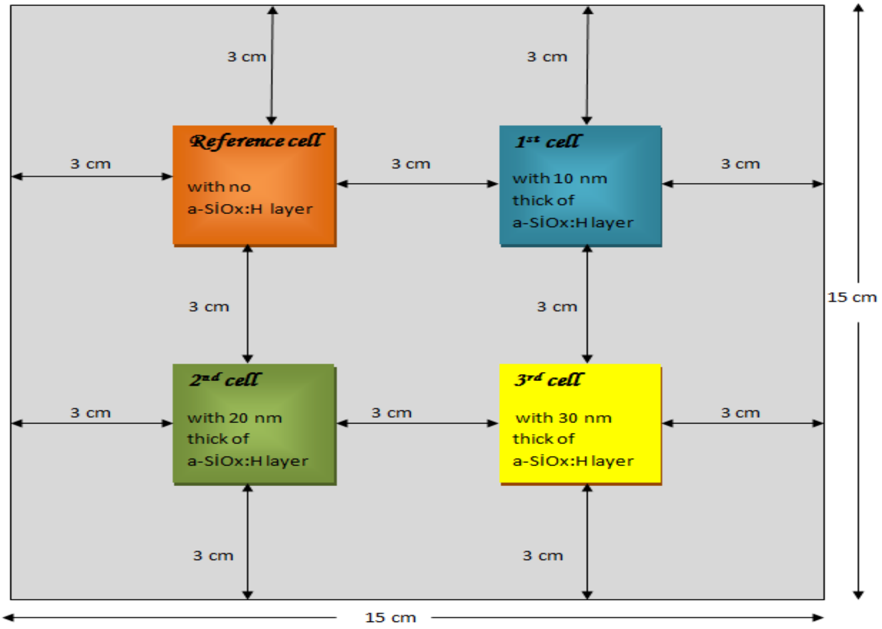


Fig. 1. Cell orientations on the SS substrate

a surface area of 3 cm x 3 cm and separated by 3 cm from each other. The orientations of the cells on the foil are shown in Fig.1.

The optimized a-Si:H and μc-Si:H cell i layer thicknesses were selected both as 250 nm. The layer dimensions are sketched in Fig. 2. The only difference among cells comes from a-SiOx:H layer thicknesses like 0 (for reference cell or cell 1), 10 (for cell 2), 20 (for cell 3) and 30 nm (for cell 4) respectively.

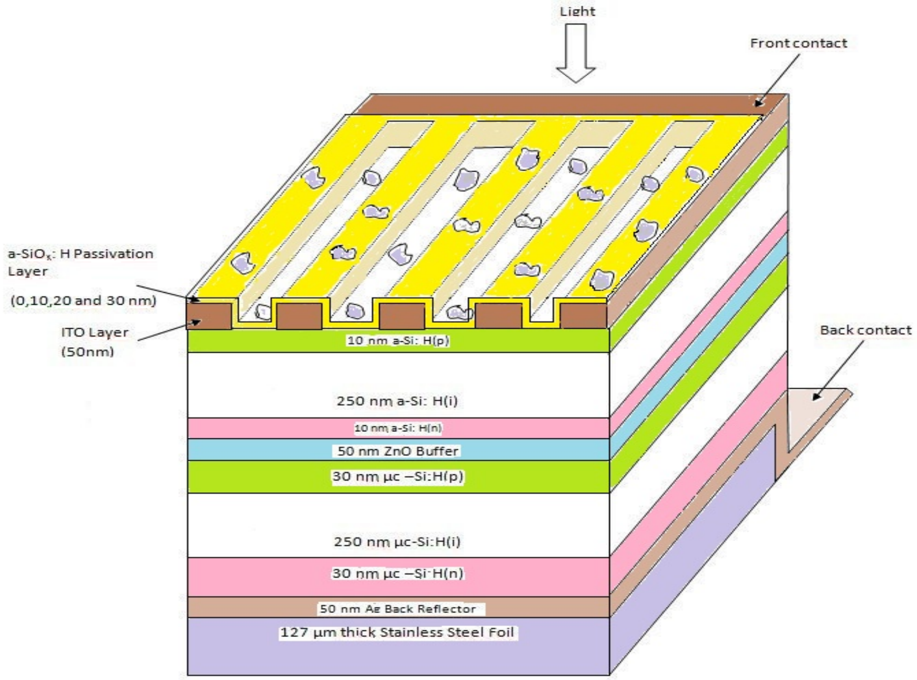


Fig. 2. The layer dimensions of the cells. Note that each cell has different a-SiO$_x$:H layer thicknesses.

We prepared the SS substrate by following the method cited in Ref. 13[13]; by firstly cleaning it in an ultrasonic bath with a soap solution in pure water, de ionized water bath and drying with a nitrogen flow.

In order to improve reflectivity together with the impurity shielding, we coated the SS with a 50 nm Ag layer by using thermal evaporation followed with a ZnO sputtering at 150^{0}C to prevent Ag agglomeration during μc-Si:H deposition.

Since crystallization quality highly depends on deposition pressure as well as the concentration of Silane (SiH_4), the deposition of the bottom cell was done in a well controlled chamber under the conditions of 200^{0}C and 200 Pa by means of radio frequency plasma enhanced chemical deposition (RF-PECVD) with 13.56 MHz and 72 mW/cm^2. The electrode gap was adjusted to 25 mm in order to achieve a slower microcrystal growth rate of approximately 7 Å/s thereby an expected average grain size of 100 Å. We used SiH_4 (75%) and H_2 (25%) mixture as reactant gas to grow i layer, hydrogen-diluted Diborane (B_2H_6) with a rate of $B_2H_6/(SiH_4 + H_2) = 0.7\%$ and

Phosphine (PH_3) with a rate of $PH_3/(SiH_4 + H_2) = 0.1\%$ to grow the p and n layers respectively.

The intermediate coupling layer of ZnO between the bottom and top cell was realized by RF sputtering at 150^0C.

Growing of the a-Si:H n, i and p layers was achieved again by using SiH_4 (75%) + H_2 (25%) as a main mixture, PH_3 (0.2%) and B_2H_6 (0.1%) with a rate of approximately 10 Å/s to deposit n and p layers. Deposition of the top cell was done under the conditions of 150^0C and 60 Pa.

The ITO grid layer having a pitch of 3 mm and a grid width of 0.5 mm was deposited by means of RF sputtering at room temperature.

The cells were sequentially masked to develop three different thickness of a-SiO_x:H intermediate passivation layer just before Ag deposition. We used again RF-PECVD at 150^0C through a decomposition of CO_2:SiH_4 (1:5) mixture under the pressure of 30 Pa with a growth rate of 2.5 Å/s. The electrode spacing was still kept at 25 mm.

The final step was to deposit the Ag layer of 5 nm. We constructed Ag nano particles by first thermal evaporation at a very low pressure of $1x10^{-5}$ Torr with a speed of 2 Å/s; and then annealing the samples two times each for 30 min in Nitrogen at 200^0C to allow nano particle growing. The statistical analysis through the secondary electron microscopy (SEM) showed that the average sizes of nano particles were on the order of 50 nm as expected. The shapes were like flattened hemisphere and the average spacing was ~150 nm. We will leave the effect of the grain size and shape on efficiency as future work since it is out of the topic.

4 Results and Discussions

The normalized scattering efficiency of Q_{sca} was calculated first to determine the effect of the medium in which Ag nano particles have been deposited. The calculation was done by means of the equations (1)-(5) and by using the complex refractive indexes derived in Ref 14[14]. Since the main goal is to compare the effect of a-SiO_x:H thickness on efficiency enhancement, we focused on the wavelength shift caused by the a-SiO_x:H layer. We found that there was a redshifting when 20 nm Ag nano particles were in SiO_x compared to the same size Ag nano particles in Si medium as seen in Fig.3.

It was shown that the improved scattering capability of the latter has a remarkable effect in shorter wavelength range; that means the a-SiO_x:H layer would help increasing absorbance performance of the top cell having a greater E_g compared with the bottom. We may conclude that the a-Si:H cell can absorb blue and green wavelength of the solar spectrum very efficiently as a-SiO_x:H layer is deposited on top, but poorer in red (λ>600 nm) and near IR, since higher absorption in longer wavelengths would require longer absorption lengths. Plasmonic enhancement of optical absorption of the top cell thanks to the Ag nano particles allows to construct thinner top cell.

The J-V parameters of the four cells were measured under the conditions of air-mass 1.5 (AM 1.5) illumination (100 mW/cm^2) and illustrated in Fig.4.

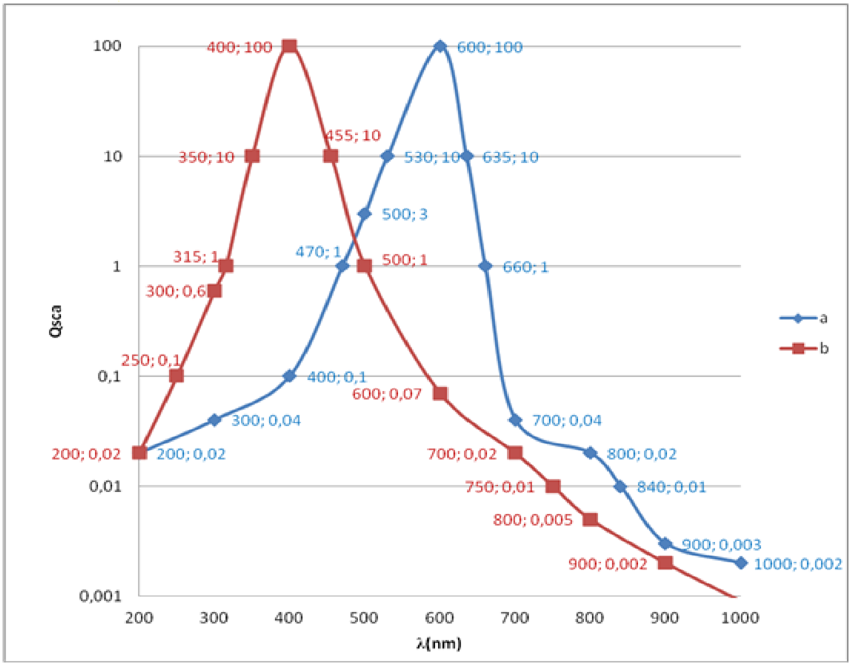


Fig. 3. The normalized scattering efficiency of the reference cell with no SiO_x layer (a) and the cells with SiO_x layer (b) between 20 nm Ag nano particles and the top cell

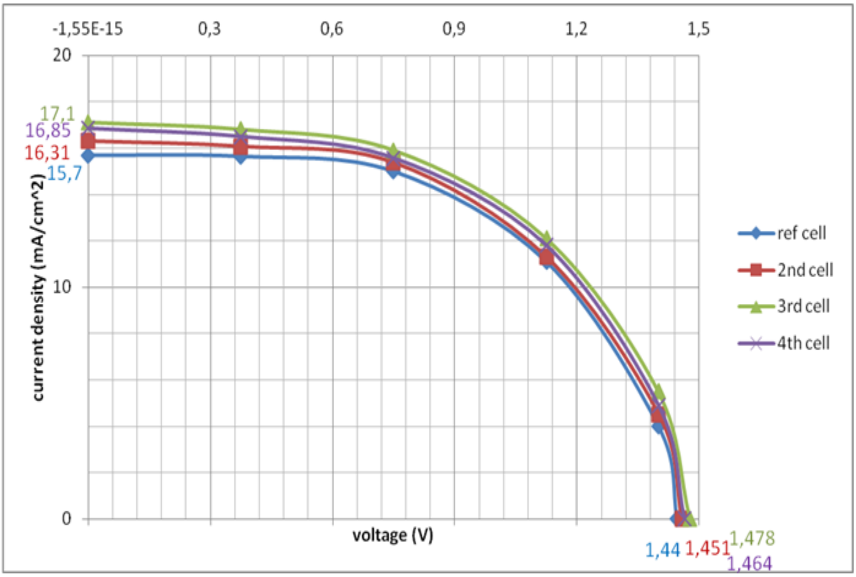


Fig. 4. J-V characteristics of four cells under AM 1.5 illumination conditions at room temperature

The reference cell (cell 1) with no a-SiO_x:H layer had a short circuit current density (J_{sc}) of 15.7 mA/cm^2, an open circuit voltage (V_{oc}) of 1.44 V and a fill factor (FF) Of 69.7%. The overall efficiency was observed as 14.9%. Cell 2 with a 10 nm thick a-SiO_x:H layer gives an improved J_{sc} of 16.31 mA/cm^2, a V_{oc} of 1.451 V and a (FF) of 72.2%. This improvement comes from the surface passivation property of a-SiO_x:H layer, but it seems the thickness couldn't satisfy sufficient elimination of surface recombination centres thereby expectedly causing some loss of EHPs which could penetrate the thin a-SiO_x:H layer.

It was observed that cell 3 with a-SiO_x:H layer thickness of 20 nm resulted in the highest J_{sc} of 17.1 mA/cm^2 and a V_{oc} of 1.478 V with a FF of 74.3%. The overall efficiency of this cell was as large as 16.19%. An improvement of 8.91% of J_{sc} and 2.64% of V_{oc} was observed. V_{oc} could be further increased by adjusting the crystallinity (the ratio of crystallized phase to amorphous phase), since V_{oc} of μc-Si:H i layer is inversely proportional to the crystallinity, as reported in Ref. 15[15].

The efficiency of cell 4 with insulating layer thickness of 30 nm which was measured as 16.14% tended to drop again towards the level of cell 2, like a gradual increase in V_{oc} but a decrease in J_{sc} as seen in Fig.4. The measured values of cell 4 were 16.85 mA/cm^2, 1.464 V and 72.9% for J_{sc}, V_{oc} and FF respectively. This gradual decrease was interpreted due to the degradation of the effect of SPP enhancement.

Consequently the Ag nano particles gave the best enhancement in cell 3 with 20 nm thick insulating layer of a-SiO_x:H compared to the other three. This thickness seems an optimal choice for maximizing the a-Si:H/μc-Si:H tandem solar cell efficiency.

5 Conclusion

The thickness of a-SiOx:H layer between the nano particles and the top cell is a crucial factor by affecting the overall efficiency. To select the optimum thickness we fabricated four cells simultaneously having different thickness of a-SiOx:H layer and we have demonstrated that the cell with a 20 nm thick of it exhibited the best performance with 16.19% efficiency which was as high as 8.65% compared to the reference cell.

The reason for such an improvement was attributed to the insulating and surface passivation capability of a-SiOx:H layer between the SPPs and the top cell. Also an improvement of 8.91% in J_{sc} and 2.64% in V_{oc} was achieved.

6 Future Work

SPP assisted tandem solar cells have been widely investigated since they provide promisingly higher efficiency and lower cost production possibilities. More methods such as developing an intermediate reflector among cells and/or applying textured bottom surface could be considered as future work. Besides the effect of the shapes of SPPs like elliptic or hemispheric grains could be examined.

Acknowledgements. We thank Mr. Musa DEMIRCI and Ms. Merve GUNES; who are the assistants in Electrical&Electronic Eng.g. Dept. at Karatay University; for their kind support while drawing the figures and graphs.

References

[1] Mie, G.: Beitrug Zur Optik Truber Medien, Speziell Kolloidaler Metallosungen. Annalen der Physik, series IV, pp. 35–377 (1908)

[2] Karajangsang, T., Kasashima, S.: Hetero-Junction Microcrystalline Silicon Solar Cell with Wide-Gap p μc-SiOx:H Layer. In: 35th IEEE PV Specialists' Conf., PVSC, pp. 1531–1534 (2010)

[3] Fuyuki, T., Kondo, H., et al.: Poly-Si Solar Cells Using Electroluminescence. In: PV Specialists' Conf., PVSC Record, pp. 1341–1345 (2005)

[4] Hock, M., McGill, J., Czubatyj, W., Singh, R.: Minority Carrier Diffusion Length in a-Si:H Based Alloys. J. of Applied Physics 53(9), 6270–6275 (1982)

[5] Bohren, C.F., Huffman, D.R.: Absorption and Scattering of Light by Small Particles. John Wiley & Sons, N.Y (1983)

[6] Royer, P., Goudonnet, J.P., Warmack, R.J.: Physics Review B: 35, 3753 (1987)

[7] Hao, H., Li, W., Xing, J., et al.: Enhanced Absorption in Nanocrystalline Silicon Thin Film Solar Cells Using SPPs. In: International Conf. on Materials for Renewable Energy & Environment, ICMREE, vol. 1, pp. 242–246 (2011)

[8] Biswas, R., Zhou, D., Curtin, B.: Surface Plasmon Enhancement of Optical Absorption of Thin Film a-Si:H Solar Cells. In: 34th IEEE PV Specialists' Conf., PVSC, pp. 557–560 (2009)

[9] Warrick, W., Kappera, R., Tayahi, M.: Enhanced Optical Absorption in Thin Film Solar Cells by Surface Plasmons. In: International Conf. on Photonics, ICP, pp. 1–5 (2010)

[10] Kim, D.I., Chun, H.G., et al.: Improved Optical and Electrical Property of ITO Film Prepared Using a Magnetron Sputter Type Negative Metal Ion Beam Deposition Method. In: Proceedings the 9th Russian-Korean International Symposium on Science and Technology, KORU (2005)

[11] Wronski, C.: Amorphous Si Photovoltaics: Order from Disorder. In: 28th IEEE PV Specialists' Conf., PVSC, pp. 1–6 (2000)

[12] Deng, S.X., Schiff, E.: a-Si Related Solar Cells. In: Handbook of PV Science & Engineering. John Wiley & Sons Ltd. (2003)

[13] Blösh, P., Chirila, A., et al.: Comparative Study of Different Back-Contact Design of High Efficiency CIGS Solar Cells on SS Foils. IEEE J. of Photovoltaics, 194–199 (October 1999)

[14] Weber, M.J.: Handbook of Optical Materials. The CRC Press (2003)

[15] Goya, S., Nakano, Y., et al: Development of a-Si/μc-Si Tandem Solar Cells. In: 3rd World Conf. on PV Energy Conversion, Osaka, Japan, May 11-18, pp. 1570–1573 (2003)

Unsupervised Hidden Topic Framework for Extracting Keywords (Synonym, Homonym, Hyponymy and Polysemy) and Topics in Meeting Transcripts

J.I. Sheeba[1,4], K. Vivekanandan[2,4], G. Sabitha[3,4], and P. Padmavathi[3,4]

[1] Research Scholar
sheeba@pec.edu
[2] Professor
k.vivekanandan@pec.edu
[3] Final Year MCA
[4] Department of Computer Science & Engineering, Pondicherry Engineering College, Puducherry – 605 014, India

Abstract. Keyword is the important item in the document that provides efficient access to the content of a document. It can be used to search for information or to decide whether to read a document. This paper mainly focuses on extracting hidden topics from meeting transcripts. Existing system is handled with web documents, but this proposed framework focuses on solving Synonym, Homonym, Hyponymy and Polysemy problems in meeting transcripts. Synonym problem means different words having similar meaning are grouped and single keyword is extracted. Hyponymy problem means one word denoting subclass is considered and super class keyword is extracted. Homonym means a word can have two or more different meanings. For example, Left might appear in two different contexts: Car left (past tense of leave) and Left side (Opposite of right). A polysemy means word with different, but related senses. For example, count has different related meanings: to say number in right order, to calculate. Hidden topics from meeting transcripts can be found using LDA model. Finally MaxEnt classifier is used for extracting keywords and topics which will be used for information retrieval.

Keywords: Keyword, Meeting transcripts, LDA, MaxEnt, Synonym, Homonym, Polysemy, Hyponymy.

1 Introduction

Keyword is a word occurs in text more often with some useful meaning. Keywords provide efficient information and sharp access to documents concerning their main topics. It can be used for various natural language processes like text categorization and information retrieval. However, most documents will not provide keywords. In particular, spoken documents mostly may not have keywords. On comparing written text and other speech data with meeting speech, meeting speech is much different [1]. There is a sudden increase in Communication, e-marketing, online services and other

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 299–307.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

entertainments, due to this Web data is available in many different forms, genres, and formats than before. This difference in data formats gives new challenges in mining and IR search.

The main challenges in this study are synonym, Homonym, Hyponymy and Polysemy problems. Synonyms are natural linguistic phenomena which NLP and IR researchers commonly find difficult to cope with. Synonym, that is, two or more different words have similar meanings, causes difficulty in connecting two semantically related documents. For example, the similarity between two (short) documents containing Tale and Story can be zero despite the fact that they can be very relevant. Hyponymy is a relation between two words in which the meaning of one of the words includes the meaning of the other word. For example Blue, Green is kinds of color. They are specific colors and color is a general term for them. Homonym means a word can have two or more different meanings. For example, Bank of India (Institution), River of bank (River). A Polysemy means word with different, but related senses. For example bank has different related meanings: blood bank, financial institution.

Query expansion in IR [3] helps to solve the synonym problem so that retrieval precision and recall will be improved. It retrieves more relevant and better documents by representing user queries with additional terms using a concept-based thesaurus, word co occurrence statistics, query logs, and relevance feedback. Dimension reduction and synonym problem can be solved by mapping vector space model to compact space using mathematical tool by Latent semantic analysis(LSA)[4,5]. Some studies use clustering as a means to cluster related words before classification and matching[6,7,8].The semantic correlation between words can be represented using taxonomy, ontology, and knowledge base for better classification or clustering.

In the existing system, we come up with a general framework to overcome the above challenge by utilizing hidden topics discovered from data sets. The main idea behind the framework is that, we collect a bulk of data set, and then build a model on both a small set of data and a rich set of hidden topics discovered from the universal data set. A better similarity measure between the documents for more accurate classification, clustering, and matching/ ranking can be given by these hidden topics. Topics inferred from a global data collection help to emphasize and guide semantic topics hidden in the documents in order to handle synonym problem[2].

In our proposed framework is going to solve the above four problems in the extracted keywords from the meeting transcripts.

2 Related Works

The number of related studies focused on solving Synonym problem. In this section, we give a short introduction of several studies that found most related to the work. The first group of studies focused on the similarity between very short texts. Sahami and Heilman [11] also calculated the relatedness between text snippets with the help of search engines and a similarity kernel function. Metzeler et al. [9] evaluated a large variety of similarity measures for short queries from Web search logs. Yih and Meek [10] considered this problem by improving Web-relevance similarity and the method in [11].

Gabrilovich and Markovitch [12] computed semantic relatedness using Wikipedia concepts. Before topic analysis models, word clustering algorithms were introduced to get better text categorization in different ways. Baker and McCallum [6] tried to condense dimensionality by class distribution-based clustering. Bekkerman et al. [7] combined distributional clustering of words and SVMs. Dhillon and Modha [8] introduced spherical k-means for clustering sparse text data. "Text categorization by boosting automatically extracted concepts" by Cai and Hoffman [13] is almost certainly the study most related to this framework. Their method attempts to evaluate topics from data using probabilistic LSA (pLSA) and uses both the original data and resulting topics to train two different weak classifiers for boosting. The difference is that they extracted topics only from the training and test data while we discover hidden topics from external large-scale data collections.

Another related work used topic-based features to improve the word sense disambiguation by Cai et al[14]. The Existing system uses TF-IDF for finding semantic features. The proposed framework tries to discover the semantic relations, but instead of using a classifier with a large taxonomy, we use hidden topics discovered automatically from meeting transcripts.

3 Proposed Framework

This Fig 1 represents Unsupervised Hidden topic framework it consists the following steps:

1. Meeting transcripts
2. Data Pre-processing
3. LDA Model
4. Synonym problem
5. Hyponymy problem
6. Polysemy problem
7. Homonym problem
8. MaxEnt classifier
9. Topic Extraction

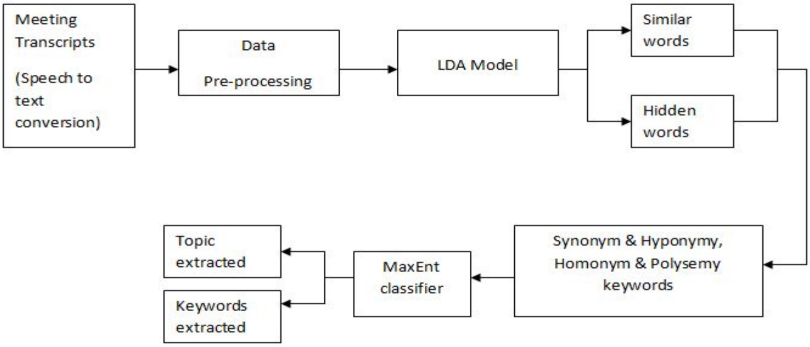


Fig. 1. Unsupervised Hidden Topic Framework

3.1 Meeting Transcripts

Meeting transcripts are the text file containing meeting speech in readable format. The audio dialogue from meeting is taken as an input to Nuance Dragon Naturally Speaking conversion software tool and speech is converted readable - written text file. Before conversion to text, the software is trained with some data.

3.2 Data Pre-processing

Here text file is taken as input, in the file stem and stop words are removed to give only the meaningful words. Then using tf-idf frequency of each word is calculated.

3.2.1 Stem Word

The form of a word after all affixes are removed; "thematic vowels are part of the stem". Thus, in this usage, the English word friendship contains the stem friend, to which the derivational suffix -ship is attached to form a new stem friendship, to which the inflectional suffix -s is attached. Porter Stemming algorithm is used to remove all affixes.

3.2.2 Stop Word

Stop words are words which are filtered out prior to, or after, processing of natural language data (text).Some search engines don't record extremely common words in order to save space or to speed up searches. These are known as "stop words."

3.2.3 TF-IDF

The **tf–idf** weight (term frequency–inverse document frequency) is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus.

The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.The term count in the given transcript is simply the number of times a given term appears in that transcript[15].

3.3 LDA Model

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each document **w** in a corpus D:

α is the parameter of the Dirichlet prior on the per-document topic distributions.
β is the parameter of the Dirichlet prior on the per-topic word distribution.
θ_i is the topic distribution for document i,
φ_k is the word distribution for topic k,
z_{ij} is the topic for the jth word in document i, and
w_{ij} is the specific word.

LDA Algorithm

```
For each topic k $\epsilon$ [1, K] do
    Generate $\vec{\varphi}_k \sim Dir\ (\vec{\beta})$
End for
For each document m $\epsilon$ [1, M] do
    Generate $\vec{\vartheta}_k \sim Dir\ (\vec{\alpha})$
    Generate $\vec{N}_m \sim poiss\ ()$
    For each word n $\epsilon$ [1,$N_m$] do
        Generate $z_{m,n} \sim Multi(\vec{\vartheta}_m)$
        Generate $w_{m,n} \sim Multi(\vec{\varphi}_z)$
    End for
End for [18].
```

Gibb sampling Algorithm

Gibbs sampler is an algorithm to generate a sequence of samples from the joint probability distribution of two or more random variables. The purpose of such a sequence is to approximate the joint distribution to approximate the marginal distribution of one of the variables, or some subset of the variables.

The Gibbs sampling algorithm generates an instance from the distribution of each variable in turn, conditional on the current values of the other variables[2].

Algorithm (Gibbs Sampling)

```
Specify an initial value $\boldsymbol{\psi}^{(D)} = (\boldsymbol{\psi}_1^{(D)},\ldots, \boldsymbol{\psi}_p^{(D)})$

Repeat for    j =1,2,…,$M$

Generate $\psi_1^{(j+1)}$ from $\pi\,(\boldsymbol{\psi}_1 | \boldsymbol{\psi}_2^{(j)}, \boldsymbol{\psi}_3^{(j)},\ldots, \boldsymbol{\psi}_p^{(j)})$

Generate $\boldsymbol{\psi}_2^{(j+1)}$ from $\pi\,(\boldsymbol{\psi}_2 | \boldsymbol{\psi}_1^{(j+1)}, \boldsymbol{\psi}_3^{(j)},\ldots, \boldsymbol{\psi}_p^{(j)})$

Generate $\boldsymbol{\psi}_p^{(j+1)}$ from $\pi\,(\boldsymbol{\psi}_p | \boldsymbol{\psi}_1^{(j+1)}, \boldsymbol{\psi}_3^{(j)},\ldots, \boldsymbol{\psi}_{p-1}^{(j+1)})$

Return the values. $\{ \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)},\ldots, \boldsymbol{\psi}^{(M)}\}$
```

Using this Gibb sampling algorithm one can yield relatively simple algorithms for approximate inference in high dimensional models like LDA. LDA method uses Gibb Sampling Algorithm and this algorithm takes much iteration to find more relevant words and hidden words as output[19].

3.4 Synonym Problem

A Synonym means different words having same meaning. From LDA module similar words are grouped and their topic inference is made. Here the topics of the similar words are extracted. Each similar group of words contain single topic and that topic is extracted as the output of this problem. By extracting topics we can solve synonym problem.

3.5 Hyponymy Problem

A Hyponym is a lower class, specific term whose referent is included in the referent of higher class term. Hyponymy is not restricted to objects, abstract, concepts, or nouns [16]. Here word's subclass is considered and its super class is extracted as the output. By extracting super class of each word Hyponymy problem is solved.

3.6 Homonym Problem

Homonym means same word having different meanings. In LDA model, keywords are grouped under hidden topics. These topics are labeled with generalized context. Homonym keywords are identified by comparing with hidden topic keywords. The corresponding topics name gives context of keywords and then calculated the frequency used for extraction. The outputs of this problem are keywords and different meaning words.

3.7 Polysemy Problem

Polysemy refers to a word that has two or more similar meanings. Different keywords are presented in the meeting transcripts. Related meaning keywords are identified by comparing with hidden keywords. These identified keywords are used for MaxEnt classifier.

3.8 MaxEnt Classifier

Maximum entropy is a general technique for estimating probability distributions from data. The principle in maximum entropy is that, the distribution is uniform as possible when nothing is known, will have maximal entropy. Constraints can be set using labeled training data. Constraints are represented as expected values of “features," any real-valued function of an example. It is a, machine learning framework used in classification. MaxEnt takes single observation; it extracts features and groups to one set. MaxEnt is robust and has been applied successfully to a wide range of NLP tasks, such as part-of speech (POS) tagging, Named Entity Recognition (NER), parsing etc. It even performs better than SVM. It is very fast in both training and inference [17].

MaxEnt classifier is trained and on the basis of probability estimation, high probability keywords are extracted from the meeting transcripts.

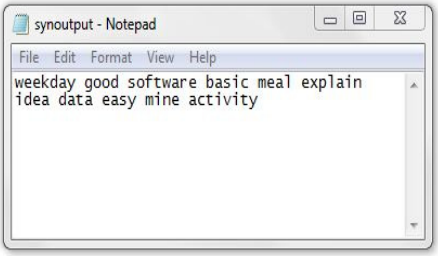
3.9 Topic Extraction

Topic Extraction means extracting overall topic of the transcript. First, we have prepared labeled test data that will contain topic name and keywords. This labeled data is used for topic extraction that is, labeled data compared with transcript keywords. The topic extraction can be done using LDA model. The most of the keywords in transcript are presented in particular topic. That topic can be extracted as overall topic of the transcripts.

4 Experiments and Results

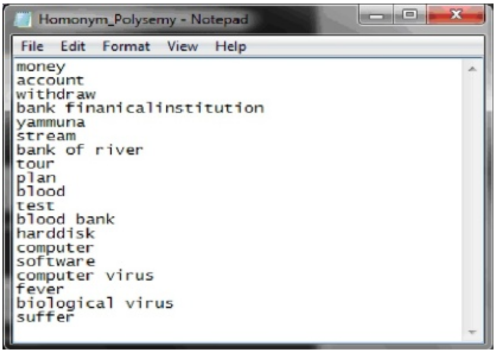
Using this approach keywords have been extracted from the meeting transcripts which describes about some topic. Nuance Dragon Naturally Speaking conversion software tool converts the audio dialogue to text format Data preprocessing is done and unwanted words are removed. LDA model provides more similar words under each topic as the result for finding hidden topic.

Synonym and Hyponym problem was solved by using Word net as training set. Homonym and Polysemy problem was solved by training dataset. Finally MaxEnt Classifier was trained using constraints and most probable keywords were extracted. Here Fig 2 represents after solving Synonym and Hyponymy problem results and Fig 3 shows after solving Polysemy & Homonym problem results.



```
weekday good software basic meal explain
idea data easy mine activity
```

Fig. 2. After solving Synonym & Hyponymy problem



```
money
account
withdraw
bank finanicalinstitution
yammuna
stream
bank of river
tour
plan
blood
test
blood bank
harddisk
computer
software
computer virus
fever
biological virus
suffer
```

Fig. 3. After solving Polysemy & Homonym problem

5 Conclusions

The unsupervised Hidden topic framework provides a solution to solve the Synonym, Hyponymy, Homonym and Polysemy problems in the meeting transcripts. Discovering the hidden topics makes the framework more efficient and reduces the complexity. Topic and keywords can be extracted more accurately by solving those problems. Because, executing the iteration in LDA model takes less time compared to other hidden topic model.

This framework focused on solving Synonym and Hyponym problems by reducing the similar keywords and provides better results. Homonym and Polysemy problem gives more accurate meaning to the keyword. Thus this framework extracts keywords in an effective and efficient way.

References

1. Liu, F., Pennell, D., Liu, F.: Unsupervised Approaches for Automatic keyword extraction, Boulder, Colorado. ACM (June 2009)
2. Phan, X.-H., Nguyen, C.-T., Le, D.-T., Nguyen, L.-M.: A Hidden Topic-Based Framework toward Building Applications with Short Web Documents. IEEE Transactions on Knowledge and Data Engineering 23 (2011)
3. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge Univ. Press, Springer (2008)
4. Deerwester, S., Furnas, G., Landauer, T.: Indexing by Latent Semantic Analysis. J. Am. Soc. for Information Science 41(6), 391–407 (1990)
5. Letsche, T.A., Berry, M.W.: Large-Scale Information Retrieval with Latent Semantic Indexing. Information Science 100(1-4), 105–137 (1997)
6. Baker, L., McCallum, A.: Distributional Clustering of Words for Text Classification. In: Proc. ACM SIGIR (1998)
7. Bekkerman, R., El-Yaniv, R., Tishby, N., Winter, Y.: Distributional Word Clusters vs. Words for Text Categorization. Machine Learning Research 3, 1183–1208 (2003)
8. Dhillon, I., Modha, D.: Concept Decompositions for Large Sparse Text Data Using Clustering. Machine Learning 42(1/2), 143–175 (2001)
9. Metzler, D., Dumais, S., Meek, C.: Similarity Measures for Short Segments of Text. In: Proc. 29th European Conference IR Research, ECIR 2007. ACM (2007)
10. Yih, W., Meek, C.: Improving Similarity Measures for Short Segments of Text. In: Proc. 22nd National Conference on Artificial Intelligence, AAAI (2007)
11. Sahami, M., Heilman, T.: A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets. In: Proc. 15th International Conference on World Wide Web. ACM (2006)
12. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In: Proc. 20th Int'l Joint Conference, Artificial Intelligence (2007)
13. Cai, L., Hofmann, T.: Text Categorization by Boosting Automatically Extracted Concepts. In: Proc. ACM SIGIR (2003)
14. Cai, J., Lee, W., The, Y.: Improving WSD Using Topic Features. In: Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLPCoNLL, Prague, pp. 1015–1023 (June 2007)

15. Term frequency-inverse document frequency, `http://www.wikipedia.com/`
16. `http://www.buzzle.com/articles/lexical-relations-hyponymy-and-homonymy.html`
17. `http://umass.academia.edu/AndrewMcCallum/Papers/49541/Using_Maximum_Entropy_for_Text_Classification`
18. `http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation`
19. Gibb Sampling Algorithm, `http://www.wikipedia.com/`

Comprehensive Study of Estimation of Path Duration in Vehicular Ad Hoc Network

R.S. Raw, Vikas Toor, and N. Singh

Ambedkar Institute of Advanced Communication Technologies & Research,
Delhi, India
rsrao08@yahoo.in, {vikastoor,nsingh1973}@gmail.com

Abstract. In this paper, we study the significance of path duration and link duration in Vehicular Ad hoc Networks (VANETs). In VANETs, the high mobility of nodes (vehicles) is the main issue of concern. Because of this mobility and connectivity graphs changes very frequently and it affects the performance of VANETs. Therefore, path duration can be used to predict the behavior of the mobile nodes in the network. Estimation of the path duration in VANETs can be a key factor to improve the performance of the routing protocol. Estimation of path duration is a challenging task to perform as it depends on many parameters including node density, transmission range, numbers of hops, and velocity of nodes. This paper will provide a comprehensive study for estimating the path duration in VANETs.

Keywords: VANET, MANET, Path Duration, Link Duration, Least Remaining Distance, Routing Protocols.

1 Introduction

VANET is a subclass of Mobile Ad hoc Networks (MANETs). VANETs are special in the sense of their mobile nodes, these nodes are vehicles on the roads and the mobility of these vehicles is very high. VANETs are real time network, used in the daily life or practically on the roads. The increasing traffic on the roads causes accidents and a matter of concern worldwide. Increasing vehicular crime is also an important issue for the nations worldwide. These accidents and crimes cause loss to life, economy, and security of human beings. VANET is one of the best solutions for all these problems. VANETs will help in making the road safer and well organized in future. The Intelligent Transportation System (ITS) has been working since long time to make the road safe, secure, and has improved efficiently in way to travel and transportation [1]. 802.11 WLAN technologies developed a Dedicated Short Range Communication (DSRC) in 2003 under the Federal Communication Commission (FCC) for VANET. DSRC service is using the 5.850-5.925 GHz for the communication between vehicles-to-vehicles (V2V) and vehicles to road-side units (V2R). VANETs make road secure by setting a communication between vehicles or their drivers and concerned authorities in periodic manner. VANETs will provide us following applications [2]:

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 309–317.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

1.1 Infrastructure to Vehicle Applications

- Violation warning
- Vehicle location information
- Back up route information
- Road blockade alarm

1.2 Vehicle-to-Vehicle Applications

- Electronic brake warning
- Oncoming traffic warning
- Vehicle stability warning
- Lane change warning
- Collision warning

VANET is a Vehicle-to-Vehicle (V2V) and Vehicle to Road-side units (V2R) communication system [1]. In V2V communication, vehicle exchange the information about each other's status. In V2R communication, road-side units exchange the information with vehicles about the traffic flow and route information. In VANETs, due to lack of any predetermined infrastructure, nodes are acting as a router in the network. These nodes can communicate with each other through multiple paths using direct links within the transmission range or using intermediate nodes to transfer data packet from source to destination. VANETs has all the characteristics of the MANETs but the key difference is that it has come from the high speed and uncertain mobility of the nodes along the network. VANET nodes have battery backup from the vehicles battery power, so there is no power management issue in VANETs. The mobility and density of nodes changes with time and location, like the speed of nodes is higher on highways in comparison to city roads. Density of nodes is higher in daytime in comparison to night; these factors always affect the topology and connectivity graph of the network.

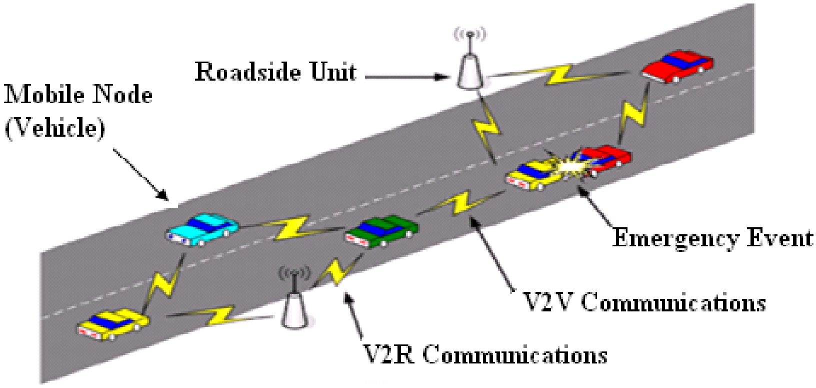


Fig. 1. VANETs Communication Scenario [3]

Protocols used in the ad hoc network can be classified into table driven routing protocol and on demand routing protocol. Table driven routing protocol are proactive in nature, which tries to maintain a path to every present node in the transmission range of the source node. Whereas on-demand routing protocol are reactive, whose path is set up only when it is required or requested [4].

The rapid change in the topology is a big problem for routing protocol in VANETs. As we know, routing is the process of finding the optimal path between source and destination node and then sending message in a timed manner [5]. As said earlier in this paper the route between source and destination can be of single hop (direct) or multi hops (intermediate) in ad hoc multi-hop network. The knowledge of relative position of nodes is very useful in delivering the message from one node to other nodes. The mobility of nodes and rapidly changing topology of the network creates problem in maintenance of routes in VANETs. Due to mobility and frequent change in network topology, one or more links along a path goes down, the path becomes invalid. This affects the on-going communication and increase the overhead. Frequent path breaks degrade the performance and efficiency of the network [5].

Path duration is the amount of time, the path stays available till one of the link along the path goes down [4]. If we are able to calculate or estimate the path duration of the routes in VANETs, this can help to deal with many challenges of the VANETs. As path duration is an important performance measure parameter, that can improve the performance of the network.

The rest of the paper is organized as follows. Section 2, present the literature survey of the path duration in ad hoc networks. Section 3, present the path duration analysis. Finally, section 4 conclude this paper.

2 Literature Survey

Significant work has been carried out in the estimation of the path duration in ad hoc networks. Several theoretical and mathematical works related to the estimation of path duration and different parameters which depends on it directly or indirectly has been done.

Estimation of path duration in the MANETs is problematic. In VANETs estimation of path duration becomes more difficult as the speed of the nodes is very high. As we have seen in MANET, intermediate node is critical to send the data from source to destination. The number of hops present between source and destination is the key element to calculate the path duration in the MANET like VANETs. In [6], an analytical approach to calculate the numbers of hops between source and destination and Euclidean distance in the uniformly distributed nodes through greedy forwarding are proposed. Greedy forwarding is also known as least remaining distance (LRD) that attempts to minimize the remaining distance to the destination with each hop. In [6], the average distance and progress per hop gradually varies with respect to current distance to the destination node and is a function of node density. The idea of hop count between two nodes with the Euclidean distance helps in the estimation of end to end delay with the help of per hop trans-receive latency.

In [4], in order to maximize the path duration in MANET, a scheme is proposed. It tries to maximize the expected path duration of the routes and also perform the path recovery in case of primary path failure by computing probability of a cached alternative path that may be available. Author shows that when the hop count along the path is large, the distribution of path duration can be calculated by an exponential distribution and parameter of the exponential distribution is given by the sum of the inverse of the expected duration of the links along the path. Secondly, avoiding nodes with shorter average link duration can help to choose the long lasting paths.

In [7], analysis of the path duration statistics and their impact on reactive MANETs protocol has been done for better understanding of the mobility in the network. It has been observed that the path duration, depends on Path Density Functions (PDFs) for the path of two or more hops can be approximated by an exponential distribution, which is parameterized by the relative speed of the mobility model, transmission range of the node, and number of hops in the path. In [8], the behavior of the communication links of a node in a multi-hop environment has been presented. In which, an approach is provided for better understanding the behavior of communication links in the presence of mobility or mobile nodes. In [9], the performance comparison of different position-based routing protocols in VANETs has been done and it shows that the features and concepts of routing protocol used. This provides us with a great help in choice of the right routing protocol and made calculation for the estimation of the path duration. This comparison also helps to choose the best routing protocol for the calculation of the path duration in VANETs.

In [10], the distribution of path durations in MANETs is studied with the help of Palm’s Theorem. Under a set of mild conditions, the distribution of path duration converges to an exponential distribution with appropriate scaling, as the number of hops increases. In [11], authors formulate the problem of optimal next-hop selection in a route between two vehicles on highway. In this, author also try to find the optimal number of hops in one link with the help of the optimal selection of next-hop in the maximum route lifetime based on vehicle speed and inter-node distance. To get the optimal path and expected lifetime, author propose a solution where two vehicles are moving on the highway. However, this research only focuses on one direction only. They ignore the scenario of opposite direction of vehicle movement.

3 Analysis of Path Duration in VANETs

Path duration is primarily based on the path selection as only after selecting a path, other parameters comes into the scenario of estimation. Therefore, the path selection is also an important aspect in VANETs. Path selection should be done, so that the distance between the source and destination should be minimum. However, in [5], a scheme to explore long lifetime of a link is proposed, which shows that the shortest path is not the best path when path duration is taken into account. The path duration of the route is also critically essential as the path breakage affects the communication in the mid of the transmission. A new route or path has to be set up for the further communication once a path failure happens. It degrades the performance of the ad hoc networks, as new route requires time and overhead both. To increase the efficiency and performance of the VANETs, knowledge of the path duration can help greatly. The path duration is also not an independent factor as it also depends on various other factors [7]. The parameters, which related to the path duration are examined under different models and protocols.

3.1 Path Duration under Different Ambience

In this paper, we focus on the study of the path duration in the VANETs. The path duration in MANETs is analyzed under different mobility model and routing protocol. Every mobility model and protocol shows some common parameters, which should be followed to calculate the path duration in ad hoc networks. The path duration is an effective and important factor to design a new routing protocol, which has high efficiency and throughput with less number of link breakage and high transmission rate.

Random waypoint mobility model is a free flow model where nodes are moving in random directions. Distribution of path duration in MANET using Random Waypoint Mobility Model has been carried out in [10]. It shows that under a set of mild conditions, the path duration distribution can be estimated by an exponential distribution as the number of hops along the path increases. Mathematically, it computes the link duration distribution with given speed of node and provides the correlation of the residual life of links. Finally, the result shows that the inverse of path duration can be estimated accurately by the sum of the inverse of the link duration of the links along the path, when number of hops is large.

Greedy Routing [6] approach in ad hoc networks is based on the Least Remaining Distance (LRD) forwarding method. This is one of the most used methods in the MANETs for the transmission of data in multi-hop routes. In LRD forwarding method, a forwarding node finds the position information of direct neighbors within the transmission range and selects one of these nodes, which is closest to the destination node as the next-hop node [12]. In others words, LRD selects the next hop node that attempts to minimize the remaining distance to the destination within transmission range of the source node with each hop. Greedy routing tries to cover the maximum distance of the route per hop in the multi-hop transmission and provide estimation of average distance progress per hop. In [6], authors shows that the node density and current distance to the destination is the function of average progress per hop and with the given hop count; the bounds on Euclidean distance can be computed numerically.

Ad hoc On-demand Distance Vector Routing Protocol (AODV) is the reactive routing protocol and selects the first available path. Distribution of path duration in MANETs using AODV has been carried out in [13]. In this schema, each node maintains a sequence number and broadcast ID. Here, sequence number shows the freshness of the path request. To avoid path failure, routes are ranked according to destination sequence number and inverse path duration values. All these above factors are related to link duration and hop counts. The path with the largest expected path duration can be calculated by the exponential distribution, when links are dependent and heterogeneous.

To calculate the path duration in VANETs, Border node based most Forward within Radius (B-MFR) is one of the best option. B-MFR [11] avoids using the interior nodes within the transmission range for packet forwarding. In B-MFR forwarding, a packet is sent to the next-hop node, which is present on the border of the transmission range towards the destination. The border node with the greatest progress on the straight line is chosen as next-hop node for transmitting packets further. Therefore, B- MFR forwards the packet to the border node that is closest to the destination node and attempts to minimize the number of hops. In B-MFR, the expected distance, expected number of hops between source and destination, and maximum progress towards destination is also mathematically estimated. These mathematical expressions are very useful and helpful in estimation of path duration in VANETs.

Edge node based Directional Routing Protocol (E-DIR) [14] is position based protocol, which is more suitable protocol for dense networks, where number of nodes per unit area are enough to provide connectivity. The edge node with smallest angle towards the destination is chosen as the next-hop node. The nodes should be present at the border area for the source node to send the data packet from source to destination as E-DIR uses the edge nodes for the next-hop transmission towards the destination. In this protocol, expected number of hops between source and destination are calculated mathematically. In E-DIR, numbers of hops are effectively minimized as it uses the edge nodes. Therefore, E-DIR shows that the density of nodes in the network is also an important factor for designing of new routing protocols so that these can increase and enhance the performance and efficiency of the networks.

In [15], we study the real world, practically collected data gathered from 20 user in real mobile ad hoc network for the analysis of the residual lifetime of the links and paths in the network. The experiment compares the result of the data with the two very commonly used mobility models for MANETs including Random Waypoint and Random Reference Group Mobility Model. The result shows that mobility of nodes and number of hops effects the lifetime of the links and routes. It also calculates the conditional route lifetime CDF (Cumulative Distribution Function) for routes, mathematically.

In [12], the path duration is estimated for the MANETs. It shows that path duration is not easy to estimate as there are many other parameters which are related to it. In [12], mathematical model is proposed and validated with the help of the simulation process. In MANETs, the mobility and speed is comparatively low than the VANETs. However, the mathematical model can help in many ways. It can provide the base and motivation to estimate the path duration in MANETs as well as in VANETs.

3.2 Parameters for Calculations

As per the analysis, the parameters on which path duration of the multi-hop routes in VANETs depends are as follows:

3.2.1 Least Remaining Distance and Shortest Path

In shortest path, we choose the next-hop node, which is at shortest distance to the destination and provide least number of hops. Similarly, in least remaining distance forwarding, choose the next-hop node, which is closest to the destination and has minimum remaining distance to the destination [6].

3.2.2 Link Residual Life

Link Residual Life is the time for which the direct link between two nodes is active and is a part of the route. Links are as far as both of nodes (source and next-hop node) in the transmission range of each other [16]. It can be defined as:

$$t = \frac{d}{v_r} \tag{1}$$

d is the distance between next-hop node and maximum transmission range of source node, v_r is the relative velocity of source node and next-hop node [12].

3.2.3 Link Distance

Distance between two nodes, which provide a link to a route can be defined as the link distance. Link distance depends on the protocol which is used in the VANETs [12]. Link distance will increase if we choose the border node of the transmission range as a next-hop node.

3.2.4 Node Density

Node density means number of vehicles per unit area of the transmission range. It affects the path duration as if we have sparse number of nodes. In this case, link formation towards the destination is a difficult job. If we have high node density network and we choose the border node as a next-hop node then it will suffer with the Edge Effect [12]. When any border node is moves out from the transmission range of the source node and path failure occurs, then it is known as Edge Effect in the network.

3.2.5 Velocity of Nodes

Direction of motion of a node and its speed both are crucial for the calculation of the path duration between the nodes. Velocity of nodes should not cause the link breakage or path breakage in VANETs as it can take the next-hop node out of transmission range. Link duration is also depends on the relative velocity of the nodes as it can increases the link distance. Equation (1) shows that relative velocity between nodes, which is inversely proportional to the link duration.

3.2.6 Number of Hops

Number of hops can be defined as the number of intermediate nodes in the route (source to destination). Number of hops depends on all the above discussed parameters. As every parameter, contributes to decide the number of hops between source and destination. Number hops should be as low as possible which decrease the chances of link breakage [10].

3.2.7 Average Progress per Hop

Average progress per hop is the average distance covered by the each hop present in the route. More the average progress per hops means less the number of hops. If there is less number of hops, it means there is less number of links and less chances of link breakage.

As we have seen, all the parameters are inter-related with each other as well as to the path duration of the multi-hop routes in VANETs. This shows that, in order to calculate the path duration in VANETs, these above parameters are crucial and must be considered in the calculations.

4 Conclusion

The path duration in VANETs is a key design parameter, which can be useful to improve the performance and throughput of the network. The path duration will be helpful in the process of path selection and for the transmission of packet from source to destination. But the mobility and uncertainty of the nodes in the VANETs is the most challenging issue for the calculation. Mobility changes the network topology and connectivity graph. Routing protocol has to find a route once again as soon as the route becomes a failure due to mobility. Once a route breaks, the communication is failed between source and destination, which affects the performance of the VANETs. Therefore, the estimation of the path duration for a particular route will provide the information and help to choose a suitable path for the transmission.

The path duration of the route can be calculated by considering the design parameters like node density, transmission range, number of hops and velocities of nodes etc. which affect the path duration of the VANETs. In future, path duration can be estimated in VANETs analytically by considering the above discussed parameters and routing protocols.

References

1. Lakshmi, K., Thilagan, K., Rama, K., Jeevarathinam, A., Priya, S.M.: Compariaon of Three Greedy Routing Algorithm for Efficient Packet Forwarding in VANET. IJCTA 3(1), 146–151 (2012)
2. Hartenstein, H., Laberteaux, K.P.: VANET: Vehicular Applications and Inter-Networking Technologies. Book of A John Wiley and sons ltd. publication (2010)
3. Rao, J.: Security in Vehiculat Ad hoc Networks (VANETs). CSE 825: Course Presentation (March 10, 2008)
4. Han, Y., La, R.J.: Maximizing Path Durations in Mobile Ad-Hoc Networks. In: 2006 40th Annual Conference on Information Science and System (March 2006)
5. Cheng, Z., Heinzelman, W.B.: Exploring long lifetime routing in ad hoc networks. In: 7th ACM international Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (2004)
6. De, S., Caruso, A., Chaira, T., Chessa, S.: Bounds on hop distance in greedy routing approach in wireless ad hoc networks. Int. J. Wireless and Mobile Computing 1(2), 131–140 (2006)
7. Sadagopan, N., Bai, F., Krishnamachari, B., Helmy, A.: PATHS: analysis of PATH duration Statistics and their impact on reactive MANET routing protocols. In: Proc. Mobihoc, 245–256 (2003)
8. Samar, P., Wicker, S.B.: On the Behavior of Communication Links of a Node in a Multi-Hop Mobile Environment. In: MobiHoc, pp. 145–156 (May 2004)
9. Raw, R.S., Das, S.: Performance comparison of position based routing protocols in vehicle-to-vehicle communication. International Journal of Engineering Science and Technology (IJEST) 3(1), 435–444 (2011)
10. Han, Y., La, R.J., Makowski, A.M.: Distribution of path durations in mobile ad hoc networks – Palm's Theorem at work. In: ITC Specialist Seminar on Performance Evaluation of Wireless and Mobile Systems, ITCSS (2004)

11. Raw, R.S., Lobiyal, D.K.: B-MFR Routing Protocol Vehicular ad hoc networks. In: IEEE ICNIT 2010, Philippines, Manila (2010)
12. Srinivasan, M.: Analytical estimation of path duration in mobile ad hoc networks. Department of Electrical and computer Engineering. Wichita State University (May 2008)
13. La, R.J., Han, Y.: Distribution of Path Durations in Mobile Ad-Hoc Networks and Path Selection. IEEE/ACM Transactions on Networking 15(5) (2007)
14. Raw, R.S., Lobiyal, D.K.: E-DIR: a directional routing protocol for VANETs in a city traffic environment. Int. J. Information and Communication Technology (IJCT) 3(2), 242–257 ISSN: 1466-6642
15. Lenders, V., Wagner, J., May, M.: Analyzing the Impact of Mobility in Ad Hoc Networks. In: ACM Conference, Florence, Italy (2006)
16. Kamaruzaman, N.N.S., Hasbullah, H.: Silent Alarm: Path Optimization Route Lifetime for VANET Multi-Hop Routing Protocol. In: International Conference on Network Applications, Protocols and Servicesm, Netapps 2008 (2008)

Securing DICOM Format Image Archives Using Improved Chaotic Cat Map Method

M. Arun Fera[1] and Suresh Jaganathan[2]

[1] PG Scholar, Department of Computer Science & Engineering,
Sri Sivasubramania Nadar College of Engineering, Chennai, Tamil Nadu, India
fera26@gmail.com
[2] Assistant Professor, Department of Computer Science & Engineering,
Sri Sivasubramania Nadar College of Engineering, Chennai, Tamil Nadu, India
whosuresh@gmail.com

Abstract. In healthcare industry, the patient's medical data plays a vital role because diagnosis of any ailments is done by using those data. The high volume of medical data leads to scalability and maintenance issues when using healthcare provider's onsite picture archiving and communication system (PACS) and network oriented storage system. Therefore a standard is needed for maintaining the medical data and for better diagnosis. Since the medical data reflects in a similar way to individuals personal information, secrecy should be maintained. Maintaining secrecy can be done by encrypting the data, but as medical data involves images and videos, traditional text based encryption/decryption schemes are not adequate for providing confidentiality. In this paper, we propose a method for securing the DICOM format medical archives by providing a better confidentiality. Our contribution in this method is of twofold: (1) Development of Chaotic based Arnold Cat Map for encryption/decryption of DICOM files and (2) Applying diffusion for those encrypted files. By applying this method, the secrecy of medical data is maintained and is tested with various DICOM format image archives by studying the following parameters i) PSNR-for quality of images and ii) Key-for security.

Keywords: Medical images, DICOM, Encryption, Healthcare, Confusion, Diffusion.

1 Introduction

Patient's data plays a major role in the healthcare industry. Nowadays they are stored in digital form. Storing them onsite (within the hospital network) is not an efficient solution for current and future trend because of issues such as scalability, and interoperability. Therefore there must be an off-site management of the patient's data. Some standard format must be followed for maintaining and transferring which led to the development of DICOM (Digital Imaging and Communications in Medicine) standard [1].

The main concern regarding the medical data is the confidentiality. Health Insurance Portability and Accountability Act (HIPAA) [1] is an act in US that tells about how secure the patient's medical data should be. It is mandatory to protect the medical data.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 319–328.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Many security issues come into picture like confidentiality, authentication, authorization and integrity [17]. The confidentiality is provided by encrypting the medical data before they are stored in offsite. Unlike text messages, image data have special features such as high redundancy, bulk capacity which generally make encrypted image data vulnerable to attacks via cryptanalysis and high correlation among pixels. As in this case, content encryption, where only the image data are encrypted, leaving file header and control information unencrypted is preferable, they usually are huge in size, which together makes traditional encryption methods difficult to apply and slow to process.

Two general principles that guide the design of cryptographic ciphers are diffusion and confusion [10]. Diffusion means spreading out the influence of a single plain-text digit over many cipher text digits, so that the statistical structure of the plain-text becomes unclear. Confusion means using transformations that complicate the dependence of the statistics of the cipher text on the statistics of the plain-text. They are closely related to the mixing and ergodicity properties of chaotic maps [16].

Rest of the paper is organized as follows: Section 2 provides details about the chaotic map employed for providing confusion. Section 3 provides details of improved Chaotic Maps which provide diffusion along with confusion. Section 4 explains the architecture of the proposed method for securing the medical images. Experimental results for the proposed method are presented in Section 5. At last Section 6 concludes our work with references.

2 Chaotic Cat Map

2.1 Usage of Confusion

Confusion is one of the important aspects of cryptography. Confusion is meant for confusing the statistical attackers to derive the original data from the statistics of the cipher data. The high initial value sensitivity and ergodicity properties of chaotic map are very essential in providing confusion for medical image data. Advantages of using confusion are i) sensitivity to initial conditions and control parameters and ii) pseudo-randomness and ergodicity. Also it has some disadvantages they are i) after the period of the cat map is reached, the original image appears and ii) once the key parameters are leaked, the adversary can easily decrypt the cipher data.

In the confusion process, many different 2D chaotic maps are used, such as the Baker map, the Cat map and the Standard, which must be used to realize the confusion of all pixels [15]. Some of the 3D maps currently in practice are just extensions of 2D chaotic maps [4]. Even chaotic maps can provide integrity. Chaotic maps properties are in close relation with the cryptosystem security. First, its parameter is used as confusion key. The higher the parameter sensitivity is, then higher the key sensitivity and the stronger the cryptosystem. Secondly, the initial-value sensitivity and state ergodicity of the chaotic map determine the confusion strength. In chaotic confusion process, initial value refers to the initial position of a pixel. Thus, the higher the initial value sensitivity is, then smaller the correlation between adjacent pixels and the more random, the confused image. Similarly, state ergodicity means that a pixel in certain position can be permuted to any position with the same probability. Thus, the higher the state ergodicity is, then more random the confusion process and the more difficult the statistic attack [8]. Therefore the chaotic map with high initial-value sensitivity and state ergodicity is preferred [9]. Other than the chaotic theory, encryption schemes for images are T-matrix and watermarking [11]. In [12], it is proved that chaotic based image encryption works efficiently than traditional AES based encryption. In [18], chaos based encryption is done with the help of traditional wavelet transform.

2.2 Mathematical Details [Confusion]

Let (X, d) be a metric space. Then a map f is said to be (Devaney) chaotic on X if it satisfies the following conditions:

- f exhibits sensitive dependence upon its initial conditions
- f is topologically transitive

The dependence on initial conditions is very important in chaos as it makes hard to determine long term behaviour of dynamical systems which show signs of chaos. If a chaotic output is generated by one set of initial conditions and then if it is changed with a little number of bits, then the output will change drastically. As mentioned previously, chaos is sometimes seen as meaning of random or unstable, but it is important to make sure that the randomness also exhibits the conditions from the definition of chaos [2].

The Arnold Cat Map is a discrete system [3] that stretches and folds its trajectories in phase space [5] as shown in the below equation.

$$\begin{pmatrix} x_{i+1} \\ y_{i+1} \end{pmatrix} = \begin{pmatrix} 1 & p \\ q & pq+1 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} \bmod N$$

where x_{i+1} and y_{i+1} are the pixel positions of the cipher image,x_i and y_i are the pixel positions of original image,N is the number of columns considered while applying the cat map.

The values p and q indicate the parameters which must be kept secret and act like the key values.One more important condition of cat map is that it must be area preserving. To achieve this, the determinant value should be 1 so that the reverse operation can be applied [6].

3 Improved Chaotic Cat Map

3.1 Diffusion

Due to some of the demerits of confusion, we go for diffusion. Diffusion [13] is another important aspect of cryptography which aims at providing additional security. In our proposed system, diffusion is done for the data that is got from the process of confusion. For diffusion function, a change of a pixel can spread to other pixels, which keeps the cryptosystem of high plain-text sensitivity. (0, 0) is the first pixel position in normal scan mode, which cannot be permuted by chaotic maps. So by applying diffusion process, the first pixel is always changed by addition operation with diffusion key Q_{i-1}. If a change happens in a pixels gray-level, then the change can cause great ones in other pixels through diffusion process [7]. Thus, the greater the changes caused by diffusion process are, then higher the cryptosystem plain-text sensitivity and the more difficult the systems security against differential attack [14].

3.2 Mathematical Details [Diffusion]

The relationship between the first pixels plaintext P_0, diffusion key Q_1 and ciphertext Q_0 is

$$Q_i^n = [D(P_0, Q_{-1})]^n \tag{1}$$

where D is the diffusion function. A powerful diffusion function is given as,

$$Q_i = P_i \oplus (4 * Q_{i-1} * (1 - Q_{i-1})) \tag{2}$$

where P_i is the current plain text pixel, Q_i is the current cipher text pixel and Q_{i-1} is the initial value of diffusion process which is used as a key for diffusion process. Here the formula is a kind of logistic cat map which provides pseudo randomness because of the XOR operation. Since we use a XOR operation which considers each and every bit of the input pixel value, it brings a stronger security by making the statistical relation among the plain images and the cipher images.Advantages of using Diffusion are i) since we use a kind of logistic map for diffusion, it provides a random behavior so that a tiny change in the plain image is reflected in more than one pixel in the cipher image [7] and ii) Pseudo-randomness and ergodicity. Disadvantages are i) usually the diffusion function takes some time to complete its operation because the real valued arithmetic operation consume much computation time and ii) once the key parameters are leaked, the adversary can easily decrypt the cipher data.

4 Architecture of the Proposed Method

Figure 1 shows the architecture of our proposed method. The proposed method uses both confusion and diffusion properties of cryptography. DICOM standard is used for both storing and exchanging the medical files such as scan images. By applying confusion and diffusion to the DICOM format medical images, their confidentiality is maintained. The keys used for confusion and diffusion must be known only to authorized persons. Some methods are proposed based on confusion for providing confidentiality but they are vulnerable to known plain image attack since confusion only rearranges the pixels. This method is against the known plain image attacks when diffusion is applied. The proposed approach takes two parts 1) encrypting the medical files and 2) decrypting the medical files. Both the parts are done with some mathematical equations which represent the chaotic based theory.

Part 1: Encryption

- Convert the DICOM file into a video sequence using third party software. Eg: Rubo DICOM viewer
- The video file is converted to individual frames

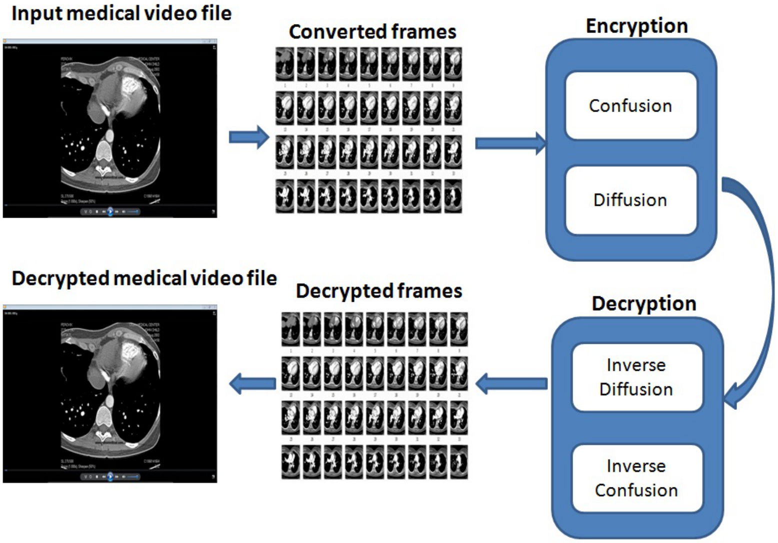


Fig. 1. Architecture of proposed method

- Extract the pixel coordinates starting from the left top for all the frames
- Apply the formula to perform confusion

$$\begin{pmatrix} x_{i+1} \\ y_{i+1} \end{pmatrix} = \begin{pmatrix} 1 & p \\ q & pq+1 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} \bmod \mathrm{N}$$

where x_{i+1} and y_{i+1} are the pixel positions of the cipher image,x_i and y_i are the pixel positions of original image, N is the number of columns considered while applying the cat map

- Apply the formula to perform diffusion

$$Q_i = P_i \oplus (4 * Q_{i-1} * (1 - Q_{i-1})) \tag{3}$$

where P_i is the current plain text pixel, Q_i is the current cipher text pixel and Q_{i-1} is the initial value of diffusion process which is used as a key for diffusion process

Part 2: Decryption

- Apply the formula to perform inverse diffusion

$$P_i = Q_i \oplus (4 * Q_{i-1} * (1 - Q_{i-1})) \tag{4}$$

where P_i is the current plain text pixel, Q_i is the current cipher text pixel and Q_{i-1} is the initial value of diffusion process which is used as a key for diffusion process

- Apply the formula to perform inverse confusion

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{pmatrix} pq+1 & -p \\ -q & 1 \end{pmatrix} \begin{pmatrix} x_{i+1} \\ y_{i+1} \end{pmatrix} \bmod \mathrm{N}$$

where x_{i+1} and y_{i+1} are the pixel positions of the cipher image,x_i and y_i are the pixel positions of original image,N is the number of columns considered while applying the cat map

- From the pixel values, construct the frames
- Convert the individual frames to a video file

5 Experimental Results

The proposed improved chaotic cat map method is applied to various DICOM format image archives and tests are conducted. The results are verified with PSNR values for QoS and with key for security.

Peak Signal to Noise Ratio(PSNR) is used as a quality parameter for reconstruction of compression images or videos. Here signal is in the original data and the noise is in the compressed data. Calculating PSNR values is used as an estimation to human awareness for reconstructing quality of compressed data or encrypted data. Two steps are involved in calculating PSNR values.

PSNR Calculation

Step 1: Calculate Mean Square Error [MSE]

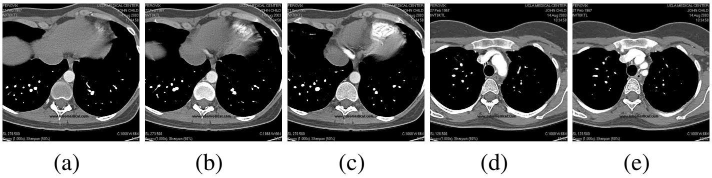
$$MSE = 1/mn \sum_{i=1}^{x} \sum_{j=1}^{y} \frac{A_{i,j} - B_{i,j}}{x * y} \tag{5}$$

where x,y are the width and height for images respectively. A and B are the input and the reproduced images respectively.

Step 2:

$$PSNR = 10 * log \frac{255^2}{MSE} \tag{6}$$

Figure 2 shows the test DICOM file (FEROVIX) used in this paper for evaluating the proposed algorithm.



(a) (b) (c) (d) (e)

Fig. 2. Sample Test Sequence

Sample test images are taken from a DICOM file which contain details of Ferovix CT scan images e.g. Lungs. These test images (shown in Column 1 of Figure 3) are encrypted using Arnold Chaotic Cat Map and its results are shown in Figure 3. When the images are encrypted using Arnold Chaotic Cat Map, the pixels are rearranged.

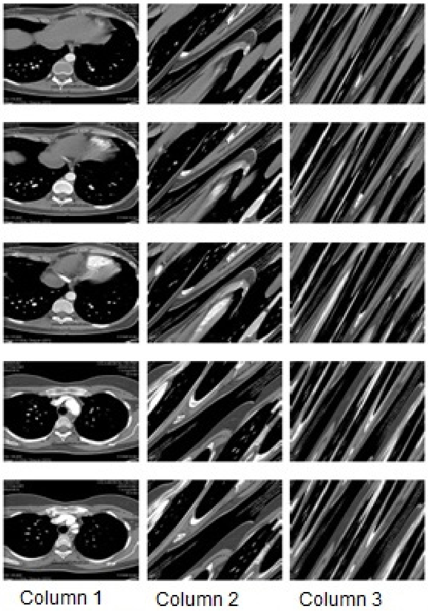


Fig. 3. Applying Chaotic Cat Map

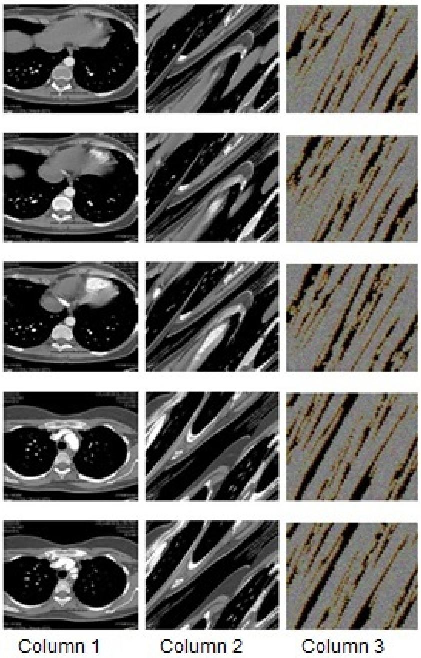


Fig. 4. After applying Proposed Method [Confusion and Diffusion]

Table 1. PSNR comparison for various methods

Bit Rate	Confusion	Difussion	Proposed
(KB)	PSNR		
350	33.83	42.99	38.41
360	33.26	42.6	37.93
380	33.35	42.6	37.97
390	33.14	42.35	37.74
400	32.94	39.06	36

Hence a shuffled image is obtained as input (shown in Column 2 of Figure 3). Images are encrypted using the key (1,1). If the key value is changed, the original is not obtained which is shown in Column 3 of Figure 3.

As stated earlier, applying Chaotic Cat Map only is not enough for security and is vulnerable to crypt-analysis. Hence the diffusion is applied with confusion, which makes the crypt-analysis harder, hence improved security. Figure 4 shows the image sequences after applying confusion (Column 2 of Figure 4) and diffusion (Column 3 of Figure 4).

Tested DICOM image sequences are encrypted with the available chaotic cat map method and also with the proposed method i.e. applying both confusion and diffusion. Table 1 shows the PSNR values obtained to check the viewable quality of DICOM images after decryption for confusion and after applying the proposed method. i.e. both confusion and diffusion. When confusion is applied to the image, the pixels values are rearranged when applying inverse operation, some pixels are not rearranged properly hence, there is a decrease in PSNR value (shown in blue color in Figure 5), when applying only diffusion the pixels value are changed and there is no rearrange of pixels and when doing inverse operation the original image is obtained, hence increase in PSNR value (shown in red color in Figure 5). When applying the proposed method i.e. combination of confusion and diffusion, the viewable quality of image after decryption

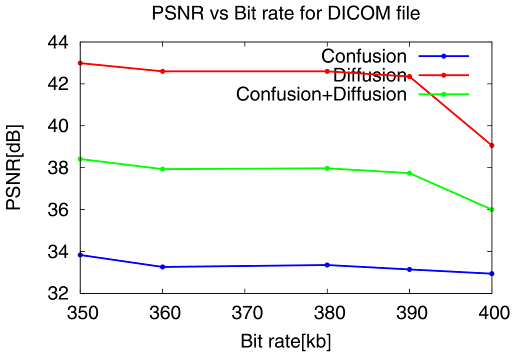


Fig. 5. Graph showing PSNR values for various methods

comes in between two extremes (shown in green color in Figure 5), having the quality viewable image reproduction.

6 Conclusion

In this paper we proposed a new method for providing two fold securities for maintaining the confidentiality of patient's medical data. Confusion is provided by means of Arnold Cat Map and diffusion is provided by means of a strong diffusion function. By employing this method, we found that this method suits well for medical images and is tested for QoS with PNSR and security level with keys. Future work entails providing a hash function based on chaos theory for maintaining the integrity of the medical data and storing the DICOM medical archives in the cloud environment due to their large size.

References

1. Teng, C.-C., Mitchell, J., Walker, C., Swan, A., Davila, C., Howard, D., Needham, T.: A Medical Image Archive Solution in the Cloud, pp. 431–434. IEEE (2010), doi:10.1109/ICSESS.2010.5552343
2. Wang, D., Zhang, Y.-B.: Image encryption algorithm based on s-boxes substitution and chaos random sequence. In: International Conference on Computer Modeling and Simulation (2009), doi:10.1109/ICCMS.2009.26
3. Peterson, G.: Arnolds Cat Map. Math. 45 Linear Algebra (Fall 1997)
4. Chen, G., Maob, Y., Chui, C.K.: A symmetric image encryption scheme based on 3D chaotic cat maps. Science Direct (2003), doi:10.1016/j.chaos.2003.12.022
5. Kashyap, S., Karthik, K.: Authenticating encrypted data, pp. 1–5. IEEE (2010), doi:10.1109/NCC.2011.5734696
6. Struss, K.: A Chaotic Image Encryption. Mathematics Senior Seminar 4901 (Spring 2009)
7. Wong, K.-W., Kwok, B.S.-H.: An Efficient Diffusion Approach for Chaos-based Image Encryption. In: 3rd International Conference Physics and Control, Physcon (2007)
8. Wong, K.-W., Kwok, B.S.-H., Law, W.-S.: A Fast Image Encryption Scheme based on Chaotic Standard Map. In: Cryptography and Security Conference, vol. 372(15), pp. 2645–2652 (2006), Cited as arXiv:cs/0609158v1
9. Ling, B., Liu, L., Zhang, J.: Image encryption algorithm based on chaotic map and S-DES, pp. 41–44. IEEE (2010), doi:10.1109/ICACC.2010. 5486998
10. Kocarev, L., Jakimoski, G., Stojanovski, T., Parlit, U.: From chaotic maps to encryption schemes. In: Proceedings of the 1998 IEEE International Symposium on Circuits and Systems, ISCAS 1998, vol. 4, pp. 514–517 (1998), doi:10.1109/ISCAS.1998.698968
11. Sharma, M.: Image encryption techniques using chaotic schemes: A review. International Journal of Engineering Science and Technology 2(6), 2359–2363, 1–10 (2010) ISSN: 0975-5462
12. Asim, M., Jeoti, V.: Image Encryption: Comparison between AES and a Novel Chaotic Encryption Scheme, pp. 65–69. IEEE (2007), doi:10.1109/ICSCN.2007.350697
13. Lian, S., Sun, J., Wang, Z.: Security analysis of a chaos-based image encryption algorithm. Science Direct, Physica A: Statistical Mechanics and its Applications 351(2-4), 645–661 (2005)

14. Ren, S., Gao, C., Dai, Q., Fei, X.: Attack to an Image Encryption Algorithm based on Improved Chaotic Cat Maps. In: 3rd International Congress on Image and Signal Processing, CISP 2010, pp. 533–536 (2010), doi:10.1109/CISP.2010.5647659
15. Mao, Y., Chen, G.: Chaos-Based image encryption (2005), `http://www.open-image.org/725publication/journal/CBIE.pdf`
16. Zhang, Y.-B.: Chaos based cryptography. An alternative to algebraic cryptography (2008)
17. Zhu, Z., Zhai, K., Wang, B., Liu, H., Jiang, H.: Research on Chaos-based Message Digest Method for Medical Images, pp. 1–510. IEEE (2009), doi:1109/CISP.2009.5301139
18. Zhu, Y., Zhou, Z., Yang, H., Pan, W., Zhang, Y.: A Chaos-Based Image Encryption Algorithm Using Wavelet Transform, pp. 1037–1040. IEEE (2010), doi:10.1109/IEMBS.2010.5628061

Brain Tumor Segmentation Using Genetic Algorithm and Artificial Neural Network Fuzzy Inference System (ANFIS)

Minakshi Sharma[1] and Sourabh Mukharjee[2]

[1] Assistant Professor in the Department of IT in GIMT Kanipla, Kurukshetra, India
[2] Associate Professor in the Department of Computer Science in Banasthali University, Rajasthan

Abstract. Medical image segmentation plays an important role in treatment planning, identifying tumors, tumor volume, patient follow up and computer guided surgery. There are various techniques for medical image segmentation. This paper presents a image segmentation technique for locating brain tumor (Astrocytoma-A type of brain tumor). Proposed work has been divided in two phases-In the first phase MRI image database (Astrocytoma grade I to IV) is collected and then preprocessing is done to improve quality of image. Second-phase includes three steps-Feature extraction, Feature selection and Image segmentation. For feature extraction proposed work uses GLCM (Grey Level co-occurrence matrix). To improve accuracy only a subset of feature is selected using Genetic algorithm and based on these features fuzzy rules and membership functions are defined for segmenting brain tumor from MRI images of .ANFIS is a adaptive network which combines benefits of both fuzzy and neural network. Finally, a comparative analysis is performed between ANFIS, neural network, Fuzzy, FCM, K-NN, DWT+SOM, DWT+PCA+KN, Texture combined +ANN, Texture Combined+ SVM in terms of sensitivity, specificity, accuracy.

Keywords: ANFIS, Brain tumor(Astrocytoma), sensitivity, specificity, accuracy, MR images, Neural network, Fuzzy, ANFIS, FCM, K-NN, GLCM, Genetic algorithm.

1 Introduction

Image segmentation plays an important role in medical field because it is important for treatment planning and identification of Brain Tumor, measures tissue volume to see tumor growth, patient follow up and computer guided surgery. Manual segmentation of magnetic resonance (MR) brain tumor images is a very challenging and time-consuming task [1,2,3,4]. Manual classification can cause human error, also result depends on human to human, time consuming process and results cannot be reproducible. So, an automatic or semi-automatic classification method is required because it reduces the load on the human observer, accuracy is not affected due to fatigue and large no. of images.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 329–339.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

For segmenting different body parts, different types of segmentation algorithm are present. But, proposed work focus literature related only to brain tumor segmentation. Monireh Sheikh Hosseini1 proposed a technique which presents a review of medical image segmentation using ANFIS[5]. He integrate the best features of fuzzy systems and neural network. A brief comparison with other classifiers, main advantages and drawbacks of this classifier are investigated. NOOR ELAIZA ABDUL KHALID [6] proposed a comparative study of Adaptive Network-Based Fuzzy Inference System (ANFIS), k-Nearest Neighbors (k-NN) and Fuzzy c-Means (FCM) in brain tumor segmentation. T. Logeswari [7] presents a brief comparison with other classifiers, main advantages and drawbacks of proposed classifier are analyzed. Rami J.Oweis[12] present the pixel classification of medical image using neuro fuzzy approach, which is based on spatial properties of the image features. N.Benamrane [13] has proposed an approach which combines Neural Networks, Fuzzy Logic and Genetic Algorithms as a hybrid system. For extracting image it uses region growing method.Ian Middleton[14] uses a neural network(a multi layer perceptron, MLP) and active contour model (‘snake’) to segment tumor in magnetic resonance (MR) images. Ramiro Castellanos [15] presents a image segmentation technique which uses adaptive fuzzy leader clustering (AFLC) algorithm.

Chin-Ming Hong [16] propose a novel neuro fuzzy network which use refined K-means clustering algorithm and a gradient-based learning rule to logically determine and adaptively tune the fuzzy membership functions for the employed neuro fuzzy network. S. Shen [17] presents a approach which is based on fuzzy c-means (FCM) clustering algorithm. In this algorithm, two factors of neighborhood attraction are the feature difference between neighboring pixels in the image, the other is the specification technique is applied on brain MR images before segmentation. The method enhances the contrast between different brain tissues.

1.1 Artificial Neural Networks

1.1.1 Learning

The proposed method shows high quality classification accuracy for images with simple components.

ANFIS is one of the widely used neuro-fuzzy systems. In this work, the neuro-fuzzy based approach namely adaptive neuro fuzzy inference system (ANFIS) is used for MR brain tumor classification.

2 Proposed Methodology

The methodology used for MR brain tumor images is Divided in to four steps and third step is further divided in to four parts as shown in fig. 1 and 2.

2.1 MR Image Database

MR image database consists astrocytoma type of brain tumor images of GRADE I to IV. These images are collected from web resource- **http://mouldy.bic.mni.mcgill.ca/brainweb/**

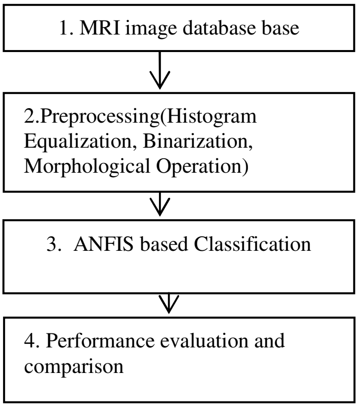

1. MRI image database base
2.Preprocessing(Histogram Equalization, Binarization, Morphological Operation)
3. ANFIS based Classification
4. Performance evaluation and comparison

Fig. 1. Proposed Methodology for Classification of Brain Tumor

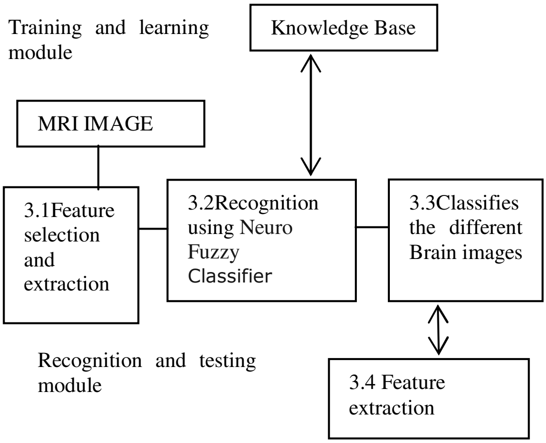

Training and learning module
Knowledge Base
MRI IMAGE
3.1Feature selection and extraction
3.2Recognition using Neuro Fuzzy Classifier
3.3Classifies the different Brain images
Recognition and testing module
3.4 Feature extraction

Fig. 2. Proposed Methodology for ANFIS based brain tumor classification

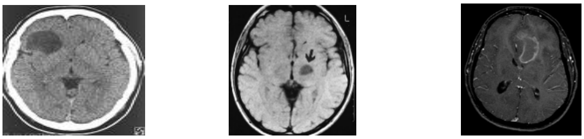


Fig. 3. Sample Data Set

2.2 Image Preprocessing

Image preprocessing involves different techniques to improve image quality before actual segmentation process. It removes irrelevant information like noise and enhances contrast to improve image quality. In the proposed work, three preprocessing techniques are used. They are-

a) Histogram Equalization

Image histogram is a graph which represents grey level frequencies of image. The histogram equalization is a technique that spreads out intensity values over the entire scale to obtain uniform histogram which in turn enhances the contrast of an image [11]. Histogram equalization used in this proposed work taken from MATLAB built-in function(histeq)[10].

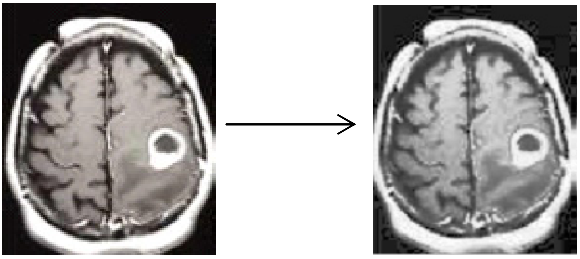


Fig. 4. Histogram Equalized Image

b) Binarization

Image binarization is used as preprocessor which converts grey scale image in to a binary image (either black or white) based on some threshold value. The pixel values above threshold value are classified as black and other are white[10].

$$G(x,y)=\begin{cases}1 & f(x,y) \geq T \\ 0 & f(x,y) < T\end{cases} \quad (1)$$

In the proposed work only one threshold value is chosen for the entire image which is based on intensity histogram (mean of intensity values are taken)

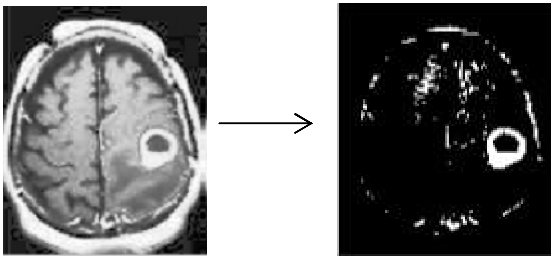


Fig. 5. Binarized image for the given grey scale image

c) Morphological Operations

This is used as a image preprocessing tools to sharpen regions and to fill gaps of binarized image. There are four basic morphological operations are defined like dilation, erosion, opening and closing. Here, proposed work uses only dilation and erosion. In erosion every pixel which touches background pixel is converted in to background pixel. Erosion turns object smaller. Mathematically erosion can be represented as,

$$(A\Theta B)(x)=\{x\text{€}\ X,\ x=a+b:\ a\ \text{€}A\ b\text{€}B\} \quad (2)$$

Where A represents matrix of binary image and B represents mask. Whereas, dilation change background pixel which touches object pixel is converted in to object pixel. Dilation combines multiple objects in one. Mathematically dilation can be represented as,

$$(A\ominus B)(x)=\{x€\ X,\ x=a+b:\ a\ €A\ b€B\} \quad (3)$$

The morphological algorithm used in this work is extracted from [11].

2.3 Feature Extraction

Features are the characteristics of the objects present in an image. Feature extraction is the procedure of extracting certain features from the pre-processed image. There are various techniques for measuring texture such as co-occurrence matrix, Fractals, Gabor filters, wavelet transform [9]. In this proposed work Gray Level Co-occurrence Matrix (GLCM) features are used to separate out normal and abnormal brain tumors. GLCM is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix)[8]. GLCM has following 20 features which are calculated using function available in MATLAB 7.0.4 for a given image:

```
GLCM2 = graycomatrix(image,'Offset',[2 0;0 2])
```

Where, image represents grey scale image. graycomatrix is the function available in MATLAB. It is used for calculating image feature values.

Table 1. Features Values of an given image

Feature No	Feature Name	Feature Values
1	**autocd**	**43.1530**
2	**contrd**	**1.8692**
3	**corrpd**	**0.1392**
4	**cpromd**	**34.6933**
5	**cshad1**	**5.2662**
6	**energd**	**0.1233**
7	**Dissid**	**0.6877**
8	**entrod**	**2.6980**
9	**homopd**	**0.65645**
10	**maxprd**	**0.6411**
11	**sosvhd**	**0.1973**
12	**savghd**	**44.9329**
13	**svarhd**	**13.2626**
14	**senthd**	**133.5676**
15	**dvarhd**	**1.8188**
16	**denthd**	**1.8927**
17	**inf1hd**	**1.2145**
18	**inf2hd**	**-0.0322**
19	**indncd**	**0.2863**
20	**idmncd**	**0.9107**

Table 1 shows feature values of an image which is calculated using above function.

3 Feature Selection

Feature selection helps to reduce the features extracted from GLCM which in turn improves the prediction accuracy, as well as computation time is also reduced. The main goal of feature selection is to select only relevant and informative features. Features are generally selected by search procedures. Popularly used feature selection algorithms are Sequential forward Selection, Sequential Backward selection, Genetic Algorithm and Particle Swarm Optimization. Here proposed work uses Genetic algorithm. Genetic algorithm is a heuristic search or optimization technique for obtaining the best possible solution in a vast solution space [21].

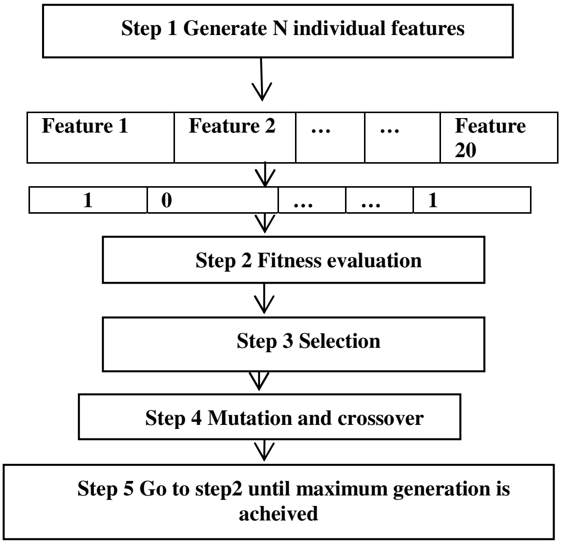

Step 1 Generate N individual features
Feature 1
Feature 2
…
…
Feature 20
1
0
…
…
1
Step 2 Fitness evaluation
Step 3 Selection
Step 4 Mutation and crossover
Step 5 Go to step2 until maximum generation is acheived

Fig. 6. GA Feature Selection Procedure

Following features are selected by Genetic algorithm:

1. Contrast: It calculates intensity contrast between a pixel and its neighbor pixel for the whole image. Contrast is 0 for a constant image.[8]

$$\text{Contrast}= \sum_{i,j}|i-j|^2\, p(i,j) \quad (4)$$

Where, P(I,j) pixel at location (i,j)

2. Angular Second Moment (ASM): It is a measure of homogeneity.

$$\text{ASM}=\sum_{i,j} p^2\,(i,j) \quad (5)$$

3. Homogeneity (HOM): It measures the variation between elements in the neighbourhood [8].

$$\text{HOM}=\sum_{\text{i,j}} \frac{\text{p(i,j)}}{1+|i-j|} \quad (6)$$

4. Inverse Difference Moment (IDM): It is the measure of local homogeneity.[8]

$$\text{IDM}= \sum_i \sum_j \frac{1}{1+(i-j)^2}\, p(i,j) \tag{7}$$

5. Energy (E): Returns the sum of squared elements in the GLCM. Energy is 1 for a constant image [8].

$$\text{E}=\sum_{i,j} p(i,j)^2 \tag{8}$$

6. Entropy (EN): It is a measure of randomness [8].

$$\text{EN}=\sum_{b=0}^{L-1} p(i,j) \log_2\{p(i,j)\} \tag{9}$$

Where, L is no. of different values which pixels can adopt[8].

7. Variance (VAR): It calculates deviation of the gray level values from the mean [8].

$$\text{VAR}=\sum_i \sum_j p(i,j)p(i,j) - \mu^2 \tag{10}$$

In the proposed work, seven GLCM features are calculated per image in four directions 0,45,95 135 and hence the number of input linguistic variables are seven. The number of output linguistic value is 2. Table 2 show a sample of features value for image 1 and image 2.Based upon this value normal and abnormal brain can be differentiated.

Table 2. Seven features with range (low and High) of image1 and image2

	Features	IMAGE1 Range(High-Low)	IMAGE2 Range(High-Low)
1.	Contrast	7.08e+00-6.98e+00	3.60e+00-4.53e-001
2	ASM	8.76e-001-8.72e-001	6.05e-001-6.72e-001
3	HOM	8.87e-001-8.62e-001	8.72e-001-8.62e-001
4	E	2.93e-001-2.85e-001	2.28e-001-2.26e-001
5	EN	2.68e-001-3.44e-001	2.72e-001-3.01e-001
6	VAR	8.96e-001-8.54e-001	9.06e-001-8.81e-001
7	IDM	9.92e-001-9.90e-001	9.94e-001-9.93e-001

A sample of fuzzy if-then rules framed for the MR brain tumor classification is shown below:ï

Rule 1: If x is CON1 and y is HOM1 and z is E1and w is EN1 and a is IDM1 and b is VAR1, then o/p = 1

Rule2: If x is CON2 and y is HOM2 and z is E21and w is EN2 and a is IDM2 and b is VAR3, then output = 2

Rule3: If x is CON3 and y is HOM3 and z is E3and w is EN3 and a is IDM3 and b is VAR31, then output = 3

The number of membership functions used in this work is 2 (low and high) and hence there are 49 rules framed for this image classification system. These fuzzy if-then rules form the input for the ANFIS architecture.

3.1 ANFIS Architecture

Adaptive Neuro-Fuzzy Inference System (ANFIS) is a very popular technique which includes benefits of both fuzzy and neural network (Jang,1993). According to [21], some advantages of ANFIS are:

- It refine fuzzy if-then rules for segmenting image
- It does not require human expertise all time.
- Provides more choices of membership function to use
- It provides fast convergence time

An ANFIS tune parameters and structure of FIS(fuzzy inference system) by applying neural learning rules The structure of ANFIS consists of 7 inputs and single output. The 7 inputs represent the different textural features calculated from each image. Each of the training sets forms a fuzzy inference system with 49 fuzzy rules. Each input was given two bell curve membership functions and the output was represented by two linear membership functions. The outputs of the 49 fuzzy rules comprised one single output, which represent output for that particular input image. The ANFIS architecture used in this work is extracted from [21].

The data set is divided into two categories: training data and testing data. The training data set consists of MRI brain images (Astrocytoma) from GRADE I to IV. These training samples are clustered in to four groups- white matter (WM), grey matter(GM), cerebrospinal fluid(CSF) and the abnormal tumor region using the fuzzy C-means (FCM) algorithm(Built-in function MATLAB).In the testing process, features are extracted and try to find best match. The algorithm used in this work is extracted from [21].

3.2 Performance Measures

Performance of different image segmentation algorithm can be analyzed in following terms:

True Positive (TP): Both Proposed Segmentation algorithm and radiologist results are positive

True Negative (TN): Both Proposed Segmentation algorithm and radiologist results are negative

False Positive (FP): Proposed Segmentation algorithm result is positive and radiologist results are negative.

False Negative (FN): Proposed Segmentation algorithm result is negative and radiologist results are positive.

Sensitivity = TP/ (TP+FN) *100%
Specificity = TN/ (TN+FP) *100%
Accuracy = (TP+TN)/ (TP+TN+FP+FN)*100 %

Table 3. Comparison of classification performance for the proposed technique and recently other work

Algorithms	**Sensitivity**	**Specificity**	**Accuracy**
DWT+SOM[7]	**95.13**	**92.2%**	**94.72**
DWT+PCA+KNN]	**96.2**	**95.3**	**97.2%**
Second order+ANN	**91.42**	**90.1**	**92.22**
Texture Combined+ANN	**95.4**	**96.1**	**97.22**
Texture Combined+SVM	**97.8**	**96.6**	**97.9**
FCM	**96%**	**93.3%**	**86.6**
K-Mean	**80%**	**93.12%**	**83.3**
Proposed (ANFIS+Genetic)	**96.6%**	**95.3%**	**98.67%**

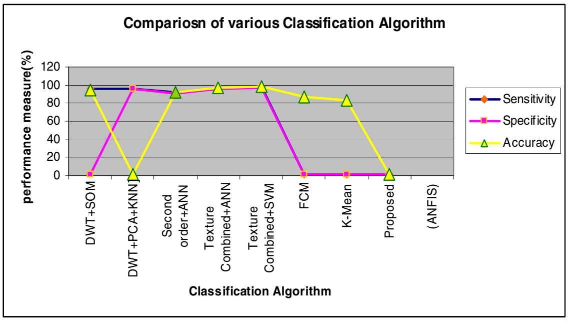


Fig. 7. Comparison of classification performance for the proposed technique and recently other work

4 Results

Proposed algorithm experimented on many images.Some of the results are shown in fig. 8.

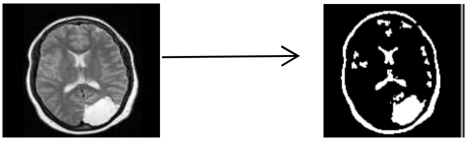


Fig. 8. Tumor Segmented from abnormal brain MRI image

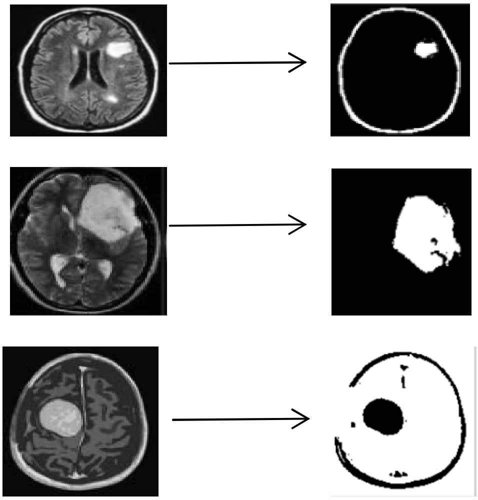


Fig. 8. (*continued*)

5 Conclusion

In this work, the application of ANFIS and genetic algorithm for MR brain tumor image classification is explored. Table 2 shows satisfactory results for proposed algorithm in terms of sensitivity, specificity, accuracy. The classification accuracy of proposed work as shown in fig.2.6. The future scope of this work is to enhance the ANFIS and genetic algorithm to achieve high classification accuracy, also measure thickness and volume of tumor.

References

[1] Jude Hemanth, D., Kezi Selva Vijila, C., Anitha, J.: Application of Neuro-Fuzzy Model for MR Brain Tumor Image Classification. Biomedical Soft Computing and Human Sciences 16(1), 95–102 (2010)

[2] Bose, N.K., Liang, P.: Neural Network Fundamentals with Graphs, Algorithms, and Applications. TMH, India (2004)

[3] Gonzalez, R.C., Richard, E.W.: Digital ImageProcessing, II Indian edn. Pearson Education, New Delhi (2004)

[4] Hosseini, M.S., Zekri, M.: A review of medical image classification using Adaptive Neuro-Fuzzy Inference System (ANFIS). Journal of Medical Signals and Sensors, 51–62 (2012)

[5] Khalid, N.E.A., Ibrahim, S., Manaf, M.: Brain Abnormalities Segmentation Performances contrasting: Adaptive Network-Based Fuzzy Inference System (ANFIS) vs K-Nearest Neighbors (k-NN) vs Fuzzy c-Means (FCM). Recent Researches in Computer Science, 285–290

[6] Logeswaria, T., Karnan, M.: An improved implementation of brain tumor detection using segmentation based on soft computing. Journal of Cancer Research and Experimental Oncology 2(1), 006–014 (2010)

[7] Haarlick, R.M.: Statistical and structural approaches to texture. Proceedings of the IEEE 67, 786–804 (1979)

[8] Saha, S.K., Das, A.K., Chanda, B.: CBIR using Perception based Texture and Color Measures. In: Proc. of 17th International Conference on Pattern Recognition, ICPR 2004, vol. 2 (2004)

[9] MATLAB, User’s Guide, The Math Works

[10] Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital image processing using MATLAB, pp. 82–83, 338–339, 336–351

[11] Oweis, R.J., Sunna, M.J.: A Combined Neuro Fuzzy Approach for Classifying Image Pixels in Medical Applications. Journal of Electrical Engineering 56, 146–150 (2005)

[12] Benamrane, N., Aribi, A., Kraoula, L.: Fuzzy Neural Networks and Genetic Algorithms for Medical Images Interpretation. In: IEEE Proceedings of the Geometric Modeling and Imaging-New Trends, pp. 259–264 (2006)

[13] Castellanos, R., Mitra, S.: Segmentation of magnetic resonance images using a neuro-fuzzy algorithm. In: IEEE Symposium on Computer-Based Medical Systems (2000)

[14] Hong, C.-M.: A Novel and Efficient Neuro-Fuzzy Classifier for Medical Diagnosis. In: IEEE International Joint Conference on Neural Networks, pp. 735–741 (2006)

[15] MATLAB, User’s Guide, The Math Works, Inc.

[16] Albayrak, S., Amasyal, F.: Fuzzy C-means clustering on medical diagnostic systems. In: International Turkish Symposium on Artificial Intelligence and Neural Networks (2003)

[17] Saha, S.K., Das, A.K., Chanda, B.: CBIR using Perception based Texture and Color Measures. In: Proc. of 17th International Conference on Pattern Recognition, ICPR 2004, vol. 2 (2004)

[18] Kim, H.-D., Park, C.-H., Yang, H.-C.: Genetic Algorithm Based Feature Selection Method Development for Pattern Recognition. In: SICE-ICASE, pp. 1020–1025 (2006)

[19] Jang, J.-S.R.: ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man and Cybernetics, 665–685 (1993)

Model Oriented Security Requirements Engineering (MOSRE) Framework for Web Applications

P. Salini[1,3] and S. Kanmani[2,3]

[1] Research Scholar Department of Computer Science and Engineering,
`salini@pec.edu`
[2] Professor, Department of Information Technology,
`kanmani@pec.edu`
[3] Pondicherry Engineering College, Pilaichavady, Puducherry-605014, India

Abstract. In the recent years, tasks such as the Security Requirements Elicitation, the Specification of Security Requirements or the Security requirements Validation are essential to assure the Quality of the resulting software. An increasing part of the communication and sharing of information in our society utilizes Web Applications. Last two years have seen a significant surge in the amount of Web Application specific vulnerabilities that are disclosed to the public because of the importance of Security Requirements Engineering for Web based systems and as it is still under estimated. Therefore a thorough Security Requirements analysis is even more relevant. In this paper, we propose a Model oriented framework to Security Requirement Engineering (MOSRE) for Web Applications and applied our framework for E-Voting system. By applying Modeling technologies to Requirement phases, the Security requirements and domain knowledge can be captured in a well-defined model and it is better than traditional process.

Keywords: Secure, Security Requirements, Security Requirements Engineering and Web Applications.

1 Introduction

The requirements must be clear, comprehensive, consistent and unambiguous. This statement has significance for security requirements and if you say application must be secure, it is not security requirements. It is hard to construct secure web applications or to make statements about security unless we know what to secure, against whom and at what extent. To this day, not one web application technology has shown itself invulnerable to the inevitable discovery of vulnerabilities that affect its owners' and users' security and privacy. Most security professionals have traditionally focused on network and operating system security. Assessment services have typically relied heavily on automated tools to help find holes in those layers. Security Requirements engineering (SRE), a phase that comes before design and programming, will play a more important role that determines the success of Web Applications Design.

In fact Security requirements engineering should be as complex and well thought out as the design and programming, yet its insufficiencies have led to many projects

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 341–353.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

with poor Security requirements and blamed as the major reason for many web applications' failures. Therefore, Security requirements engineering is now moving to the forefront of gaining increased significance in software engineering for services oriented web applications. Web applications requirements have new characteristics causing them to change more rapidly. This makes traditional Security requirements modeling and validation methods insufficient to provide adequate support for web applications. So it is essential to capture the corresponding security needs and requirements to fulfill business goals, build trustworthy systems, and protect assets. Most requirement documents were written in ambiguous natural languages which are less formal and imprecise and it is hard to analyze and integrate with artifacts in other phases of software life cycle.

The Security requirements of the web applications come from not only the general domain analysis and the personalized, diverse users' requirements, but also the availability of the related web services. Web applications Security requirements are also evolving while they are widely used. Most of the methodologies that have been proposed for the development of Web applications focus only on Non Security requirements and paying no attention to the Security requirements engineering. Therefore, SRE for Web applications is challenged to explore sound engineering approaches for eliciting, describing, validating and managing Security requirements of Web applications and its integration with the artifacts of other phases can be cost effectively improved and can effect a significant reduction of the problems currently encountered in the SDLC for Web Applications due to poor Security Requirements Engineering and Management.

1.1 The Importance of Web Applications Security

The importance of Securing Web Applications is that, often a web application is the only thing standing in the way of an attacker and sensitive business information, firewalls can only stop network service attacks and depending on the application an attacker may be able to View or manipulate sensitive information, Obtain unauthorized access to an application and also able to take control of the whole application.

Web Applications Security is about protecting assets and assets may be tangible or they may be less tangible. When to analyze your infrastructure and applications, you can identify potential threats and each threat presents a degree of risk. Web Applications Security is about risk management and implementing effective countermeasures. The Web Applications Security goals are Authentication, Authorization, Auditing, and (CIA) Confidentiality, Integrity and Availability. The security requirements identified in requirements engineering phase have to satisfy all these security goals of web applications. To build secure Web applications, threats, vulnerability and security requirements for network, host and applications should be identified. An ever-increasing number of attacks target your application. They pass straight through your environment's front door using HTTP. The reliance on firewall and host defenses are not sufficient when used in isolation. To secure web applications, means it involves security at three layers: the network layer, host layer, and the application layer.

The development of Secure Web applications has several characteristics that differ from the development of other kinds of applications [1]. On the one hand, many different kinds of stakeholders participate in the development process: analysts, customers, users, graphical designers, marketing, multimedia and security experts, etc. On the other hand, the main features of these systems are the navigational structure, the user interface and the personalization capability. So along with analysis of domain of the business, the infrastructure or the environment where you use web applications must be analyzed.

In this paper we present a framework for Model Oriented Security Requirements Engineering for Web Applications (MOSRE-WebApp). So, the remainder of this paper is structured as follows: First we establish a framework MOSRE-WebApp in Section 2 followed by Section 3 gives the application of MOSRE-WebApp framework to a E-Voting Web Application- a case study, while Section 4 presents the result analysis, discussion and comparison of SRE methods with proposed work and last Section 5 concludes with future works.

2 MOSRE-WebApp Framework

Web application has become more and more critical in every domain of the human society. Transportation, communications, entertainment, health care, military, e-commerce, and education; the list is almost endless. These systems are used not only by major corporations and governments but also across networks of organizations and by individual users. Such a wide use has resulted in these systems containing a large amount of critical information and processes which inevitably need to remain secure. Therefore, although it is important to ensure that Web Applications are developed according to the user needs, it is equally important to ensure that these applications are secure.

However, the common approach towards the inclusion of security within a Web Applications is to identify security requirements after analysis, means that security enforcement mechanisms have to be fitted into a pre-existing design, leading to serious design challenges that usually translate into the emergence of computer systems afflicted with security vulnerabilities. Recent research has argued that from the viewpoint of the traditional security paradigm, it should be possible to eliminate such problems through better integration of security and requirements engineering. Security should be considered from the early stages of the development process and security requirements should be defined alongside with the system's requirements specification.

The Security Requirements Engineering is the process of eliciting, specifying, and analyzing the security requirements for system fundamental ideas like "what" of security requirements is, it is concerned with the prevention of harm in the real world and considering them as functional requirements. Many methods have been developed that facilitate this kind of requirements analysis and the development of security requirements. The internet has already created social and economic opportunities for people around the world. But even there are many Challenges to Web Applications Security like threats, attacks, phishing spyware, worms, Trojans and virus which cause to denial of service hacking into and defacing web sites and destroying. Here we present the proposed work; an improved MOSRE-WebApp a model oriented Security Requirements Engineering Framework for Web Applications. So the completeness, consistency, traceability and reusability of Security Requirements can be cost effectively improved.

Our framework follows the spiral process model which is iterative and all phases of Requirements Engineering are covered in this framework.

2.1 Inception

Inception is to establish the ground work, before to start the elicitation and analysis of security requirements for web applications. Different steps are involved in the inception phase of MOSRE-WebApp.

Step 1 Identify the Objective of the Web Applications

The Web Applications objective must be identified from the customer requirements. This process will help to understand the domain of the application that customer needs.

Step 2 Identify the Stakeholders

The identification of stakeholders plays an important role in security requirements engineering. The stakeholders include the Architect, developer, customers/end users, security experts, requirements engineering team and other interested people. Each stakeholder is responsible to find the assets and security goals. The security experts help in finding the security requirements and security mechanisms to obtain high level of security to the Web Applications. The stakeholders will have multiple view points on the security requirements of the system. It may be conflicting security requirements and the stakeholders will help to prioritize the assets and security requirements of the system. So care to be taken to prepare the list of stakeholders, to improve the effectiveness of preliminary communication and collaboration between the stakeholders.

Step 3 Identify the Assets

The next step is to identify the assets of the targeted system. Assets may be business or system assets (e.g.: data, money, and password). From our survey it is found that assets identification is an important step in security requirements engineering. This could range from confidential data, such as customer or database, to Web pages or Web site availability.

The assets should be identified in the context of the software system, so the objective of software system is to be identified first. To identify the assets different techniques like interview, questionnaire, and brainstorming can be used. The stakeholders help in finding the assets. Assets should be viewed not only at developer or customer/end user perspective but also in attacker's point of view. Assets can be identified from existing documents. The identified assets have to be categorized and prioritized with regard to different stakeholders need. Assets can be categorized under Confidentiality, Integrity and Availability and prioritized as low medium and high level of preference. Example password can be categorized under confidentiality. After the list of assets is identified and categorized, the level of security to be implemented in the web application is fixed. Five levels of security can be implemented for web applications based upon the value of the assets.

Inception phase of security requirements engineering should be worked with high level of collaboration and care.

2.2 Elicitation

The next phase in security requirements engineering is elicitation, the stakeholders and requirements engineering team will work together to identify the problem, propose the solution and specify the set of security requirements. There are different steps involved in the elicitation phase of security requirements engineering.

Step 4 Select an Elicitation Technique

The elicitation phase starts some ground work to be done for selecting the elicitation technique. Requirements elicitation is called as capturing, requirements discovery or requirements acquisition. The process of requirements elicitation can be complex, mainly if the problem domain is unknown for the analysts. Some of the elicitation techniques are, misuse cases, Issue Based Information Systems (IBIS), Joint Application Development (JAD), Interviewing, Brainstorming, Sketching and Storyboarding, Use Case Modeling and Questionnaire and Checklist A suitable method can be chosen from these elicitation techniques based on the requirements engineering community or expert's choice, level of the security to achieve, cost –effort benefit and organizational policies.

Step 5 High level of Architecture Diagram of Web Applications

With the objective of web application we can identify the number of tiers in the web applications. So draw a rough architecture diagram with high level of abstraction of the web applications. Network or hierarchical style of Architecture can be chosen based on the application domain. This diagram can be extended in detail with low level of abstraction in the next phase of design.

Step 6 Elicit Non-Security goals and Requirements

Once the business goals are identified, and then the non-security goals and requirements of the web applications are to be elicited. The collaborative requirement gathering is adopted to gather non-security goals and requirements. A general classification of requirements for Web applications are Functional requirements and Non Functional requirements. Functional requirements are capabilities that a system must exhibit in order to solve a problem. We consider Security Requirements, as one of the functional requirements for a Web Application because Web Application has become a target of choice for hackers/hacking operations. The Gartner group estimates that 75% of attacks now target Web Applications. [1]

The non-security requirements are categorized as essential and non essential requirements and prioritized according to the Stakeholders preference.

Step 6 Generate Use Cases Diagram for the Web Applications

The non security requirements are gathered; for better understanding and then the use case modeling of the web applications should be developed. Use Case Modeling is a technique which was developed to define requirements [2]. A use case model consists of actors, use cases and relationships between them [3]. It is used to represent the environment by actors and the scope of the system by use cases (functional requirements). An actor is an external element to the system that interacts with the system as a black box. A use case describes the sequence of interactions between the system and its actors when a concrete function is executed. An actor can take part in several use cases and a use case can interact with several actors. The use case is the set of scenarios that encompass the non-security requirements of the system created by the developers and users of the system.

Step 7 Identify the Security Goals / Security Objectives

The security goals / security objectives can be identified with respect to assets, business goals and organizational principles that is the security policies of the organization. The list of security goals can be identified and the security goals can be of main goals and sub goals. The main goals are the top goals, e.g. Confidentiality, Integrity and Availability, that to be identified for the web applications based on the level of security we need.

There are many security sub goals/objectives for web applications and are based on the application domain and security policy of the organization, e.g. Prevent attackers from obtaining sensitive customer data, including passwords and profile information which comes under confidentiality. Prevent tampering, trail and access control which comes under the top security goal Integrity. The techniques like Facilitated Application Specification Technique (FAST), survey and interviews can be used to identify the security goals / security objectives.

Step 8 Identify threats and vulnerabilities

By identifying the assets, business goals and security goals the threats to the web applications can be identified. The overall system threats and vulnerabilities can be identified during this step. The list of threats and vulnerabilities can be developed for the web applications. The main threats to a Web application are: Profiling, Denial of service, Unauthorized access, Arbitrary code execution, Elevation of privileges, Information gathering, Sniffing, Spoofing, Session hijacking, SQL injection, Network eavesdropping, Password cracking, Viruses, Trojan horses, and worms. Some of the vulnerabilities to the web application are unnecessary protocols, Open ports, Web servers providing configuration information in banners, Weak IIS Web access controls including Web permissions, Weak NTFS permissions, Poor input validation in your Web applications, Unsafe, dynamically constructed SQL commands, Weak or blank passwords, and Passwords that contain everyday words.

Step 9 Risk Assessment

The next step is to assess and determine the risk when the threats and vulnerabilities occur. The impact of threats and vulnerabilities are analysed and risk determination process [18] is carried out. To do risk determination process any of risk assessment test models [5] like National Institute of Standards and Technology (NIST) model, NSA's INFOSEC Assessment Methodology, Butler's Security Attribute Evaluation method (SAEM) ,CMU's "V-RATE" method ,Yacov Haimes's RFRM model can be used or Microsoft risk based on DREAD method [6] can be used.

Step 10 Categorize and Prioritize the Threats and Vulnerabilities for mitigation

The threats and vulnerabilities can be Categorized with respect to the security goals and security policies of the organization and prioritized based on the level of security and assets to be secured, e.g. tamper threat Categorized under top security goals Integrity, unauthorized users under Confidentiality, and Integrity. This process can be done with the help of a survey or interview between the stakeholders.

Step 11 Generate Misuse Cases Diagram for the Web Applications

The detailed set of misuse case diagram [7] of the web applications should be developed that encompass the most significant threats to the system e.g. tamper misuse case, unauthorized users misuse case.

Step 12 Identify Security Requirements

The security requirements [19] are the counter measures that the Web Applications should have, as the functional requirements, e.g. Threat – password attack, wire tap, tamper, and Security goals – Availability - password attack, wire tap and Integrity - tamper. The Security requirements to prevent these threats are Prevent password attack, encrypt communication, authenticate, validate data and lock data.

Step 13 Generate Use Cases Diagram for the Web Applications considering Security Requirements

The security requirements are gathered; for better understanding, the use case diagram of the Web Applications should be generated, that encompass the security requirements of the system created by the developers and users of the system.

2.3 Elaboration

In this phase the detailed view of the web applications with security requirements can be understood with models and diagrams, which gives clear idea of the application in design and implementation phase.

Step 14 Generate Structural Analysis models

Next step of security requirements engineering is to develop different analysis models. These models form the solid foundation for the design of security requirements. The data models, flow models and behavioural models are the structural analysis models that can be used to show the functional requirements and data flow.

Step 15 Develop UML diagrams

Develop UML diagrams for detailed view of security requirements and for better understanding of the secure web applications. High level of class diagram and sequence diagrams can be developed. These diagrams can be used to generate code and test cases for testing the security requirements. The navigational model consists of a navigation class diagram and a navigation structure diagram. Security based modelling can be done using SecureUML and UMLsec .

2.4 Negotiation and Validation

In this phase the security requirements are categorized as essential and non essential requirements and prioritized according to the level of security and Stakeholders preference of security requirements. Then rough effort time and cost are estimated to implement security requirements.

The validation is done by the security experts and engineers with the requirements of the stakeholders. Review or Walk-through is a technique which consists in reading and correcting the requirements definition documentation and models. Such a technique only validates the good interpretation of the information. Traceability Matrix consists of a comparison of the application objectives with the requirements of the system [8]. A correspondence is established between objectives and how they are covered by each requirement. This way, inconsistencies and non-covered objectives will be detected.

2.5 Specification

Specification is the last phase in security requirements engineering framework. The security requirements specifications are modeled and they are validated with the stakeholders and this specification forms the source for the design of security requirements. This phase is executed in parallel with each other phases of requirements engineering. Scenario or use case modeling can be used to specify the functional requirements with security requirements and non functional requirements for web applications.

In this MOSRE-WebApp framework, object modeling is used to model the components of the web applications and the concept of encapsulation with the function and data in data modeling, reusability of some of the security requirements against different threats, and the functions can be extended to implement the security requirements ,the concept of inheritance is adopted here.

3 Application of MOSRE-WebApp Framework to a E-Voting System- A Case Study

Manual voting systems have been deployed for many years with enormous success. If those systems were to be replaced with Electronic Voting Systems, we have to be absolutely sure that they will perform at-least as efficient as the traditional voting systems without any security issues. Failures or flaws in Online Voting Systems will put at risk to Democracy in the country implementing them. The main focus of security requirements engineering is on defining and describing what a software system should do to satisfy the informal requirements provided by a statement of need. In this paper, we will define and describe what the secure Online Voting System should do to ensure a secure, robust, accurate, secure and quality-based design and implementation.

Security Requirements are defined during the early stages of system development as a specification of what level of security should be implemented. In other words, they represent what the system should do and have security from the stakeholders' point of view. Performing a good security analysis on E-Voting, web application is an essential step in order to guarantee a reasonable level of protection. However, different attacks and threats may be carried out depending on the operational environment in which the system is used, i.e. the procedures that define how to operate the systems.

An e-voting system should consider the following minimum requirements:

1. To ensure that only persons with the right to vote are able to cast a vote.
2. To ensure that every vote cast is counted and that each vote is counted only once.
3. To maintain the voter's right to form and to express his or her opinion in a free manner, without any coercion or undue influence.
4. To protect the secrecy of the vote at all stages of the voting process.
5. To guarantee accessibility to as many voters as possible, especially with regard to persons with disabilities.
6. To increase voter confidence by maximizing the transparency of information on the functioning of each system.

The MOSRE-WebApp Framework was applied to E-voting web application to gather functional requirements which includes security requirements. The architecture of the e-voting system is shown in the Fig. 1.

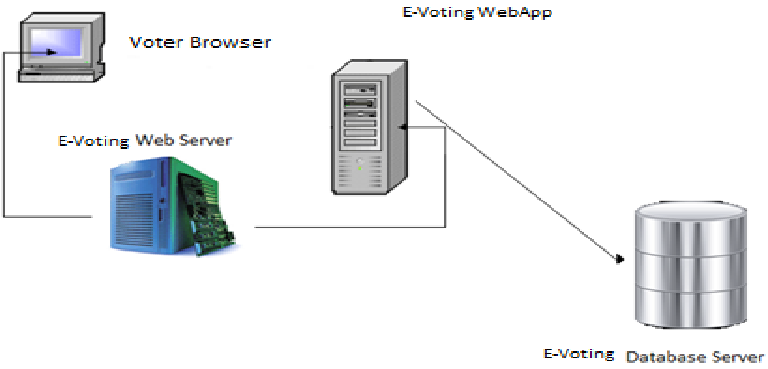


Fig. 1. Simple E-Voting Architecture

Each step of the MOSRE-WebApp Framework was applied to E-voting web application. The partial list of the E-Voting System security requirements are given below

Security requirements based on business assets:

- The voting system should include controls to prevent deliberate or accidental attempts to replace code such as unbounded arrays and strings
- Election process should not be subject to any manipulation including even a single vote manipulation
- The system should provide accurate time and date settings
- The system should not allow improper actions by voters and election officials

- The system should not allow Local Election Officials(LEOs) to down load votes to infer how voters in their precinct have voted
- The system should provide means for protecting and securing recounts of ballots cast in elections
- The system should not allow voter submissions to be observed or recorded in any way that is traceable to the individual voter
- The system should ensure that election results would be verifiable to independent observers.
- This implies that published election results correspond to the ballots cast by legitimate voters
- The system should not allow tampering with audit logs

Security requirements based on system assets:

- Use secure authentication, such as Windows authentication, that does not send passwords over the network.
- Use secure communication channels
- Use remote procedure call (RPC) encryption
- Use a segmented network, which can isolate eavesdropping to compromised segments
- Firewall policies that block all traffic except expected communication ports
- Disabling unused services
- Promptly applying the latest software patches
- Running processes with least privileged accounts to reduce the scope of damage in the event of a compromise.

We have given the list of some security requirements identified and they are based on the business and system assets.

4 Discussion

In the previous section we have identified the list of some security requirements and they are based on the business and system assets by applying MOSRE-WebApp Framework for Online Voting system. Based on the identified list of threats, vulnerabilities and security requirements we found that using our MOSRE-WebApp Framework for web applications we will be able to get better set of security requirements. There are many methods to elicit security requirements but concentrating less on the phases of requirements engineering [15, 16, 17, 20 and 22]. In this section we compare results obtained from MOSRE-WebApp Framework, Haley and colleagues security requirements engineering framework [11]. We consider the percentage of vulnerabilities, threats and security requirements found with each method as the parameters for comparison.

Table 1 Shows the comparison of MOSRE-WebApp Framework with Haley and colleagues security requirements engineering framework.

Table 1. MOSRE-WebApp Framework with Haley and colleagues SRE Framework

Parameters	Proposed MOSRE-WebApp Framework	Haley and His Colleagues SRE Framework
Completeness of SR	Yes	No
Interaction between other req.	Yes	No
Resolve Conflicts	Yes	No
Stakeholders and Vulnerability Identification	Yes	No
Approach	Model based	Problem based
Multilateral	Yes	No
Risk Assessment	Yes	No
Complexity	Simple	Complex
RE Phases	Includes all	Only Elicitation and Analysis
Traceability to Design	Easier	Hard
Categorize and prioritize	Yes	No
Elicitation techniques	Used	No
Modeling	Part of Framework	Not used
Misuse cases and use cases Diagrams	Used	No

From a technical point of view, the most difficult task of the methodology is where security objectives are identified from functional descriptions, such as functional requirements. This has been the observation from several projects using MOSRE-WebApp Framework to elicit security requirements. MOSRE-WebApp Framework requires expertise on at-least three dimensions: (i) information structuring and analysis, (ii) requirements engineering, and (iii) security. There as on is that it is rarely intuitive what the overall security goals and objectives are, and it is not easy to simply extract these from highly abstract system information, incomplete sets of functional Requirements and early draft system architecture. MOSRE-WebApp Framework provides some support, with use case, misuse case models.

5 Conclusion and Future Work

Security Requirements have to be considered in the early phase of Requirements Engineering [12, 13, and 14], so a Model oriented Security Requirements Engineering framework is developed for Web Application and evaluated for an E-Voting Web Application, The main aim of MOSRE-WebApp is to extend security requirements engineering by seamlessly integrating elicitation, traceability and analysis activities. The motivation for this is that requirements engineering activities are often executed by other people than those writing the code, and often without much contact between the two groups. This applies in particular to security requirements, which is a major quality, attribute of today’s system. It is therefore important to develop both the ability of the people involved in the development to identify potential security aspects, and the capabilities of the development team to solve these needs in practice through secure design.

As future work the Security Requirements identified from RE Phase should be carried to Design phase because good design will give Vulnerability free Web Applications and implement them. We also intent to do penetration testing and find the results based how far our application is vulnerable.

Acknowledgments. We thank our paper reviewers for their valuable comments and for selecting our paper for publication.

References

1. CLUSIF, Web Application Working Group, Web application security, managing web application security risks, Technical Studies (March 2010), `http://www.clusif.asso.fr/`
2. Jacobson, I.: Modeling with Use Cases: Formalizing Use Case Modelling. Journal of Object-Oriented Programming (1995)
3. UML. Unified Modeling Language. Version 1.5 (2003), `http://www.omg.org`
4. Meier, J.D., Mackman, A., Dunner, M., Vasireddy, S., Escamilla, R., Murukan, A.: Improving Web Application Security:Threats and Countermeasures. Microsoft Corporation (June 2003)
5. Mead, R., Houg, E.D., Stehney, T.R.: Security Quality Requirements Engineering (Square) Methodology, tech. report CMU/SEI-2005-TR-009, Software Eng. Inst., Carnegie Mellon Univ. (2005)
6. Swiderski, Frank, Syndex: Threat Modeling. Microsoft Press (2004)
7. Sindre, G., Opdah, A.L.: Eliciting security requirements with misuse cases. Requirements Eng. 10, 34–44 (2005)
8. José Escalona, M., Koch, N.: Requirements Engineering for Web Applications – A Comparative Study. Journal of Web Engineering 2(3), 193–212 (2004)
9. Lee, H., Lee, C., Yoo, C.: A Scenario-based Object-oriented Methodology for Developing Hypermedia Information Systems. In: Sprague, R. (ed.) Proceedings of 31st Annual Conference on Systems Science (1998)
10. Bieber, M., Galnares, R., Lu, Q.: Web Engineering and Flexible Hypermedia. In: The Second Workshop on Adaptive Hypertext and Hypermedia, Hypertext 1998, Pittsburg, USA (1998)

11. Haley, C.B., Laney, R., Moffett, J.D., Nuseibeh, B.: Security Requirements engineering: A Framework for Representation and Analysis. IEEE Transaction on Software Eng. 34(1), 133–152 (2008)
12. Dubois, E., Mouratidis, H.: Guest editorial: security requirements engineering: past, present and future. Requirements Eng. 15, 1–5 (2010)
13. Fabian, B., Gurses, S., Heisel, M., Santen, T., Schmidt, H.: A comparison of security requirements engineering methods. Requirements Eng., Special Issue Security Requirements Engineering 15, 7–40 (2010)
14. Houmb, S.H., Islam, S., Knauss, E., Jurjens, J., Schneider, K.: Eliciting security requirements and tracing them to design: An integration of Common Criteria, heuristics, and UMLsec. Requirements Eng., Special Issue Security Requirements Engineering 15, 63–93 (2010)
15. Hadavi, M.A., Hamishagi, V.S., Sangchi, H.M.: Security Requirements Engineering; State of the Art and Research Challenges. In: Proceedings of the International Multi Conference of Engineers and Computer Scientists, IMECS 2008, Hong Kong, vol. I, pp. 19–21 (March 2008)
16. Wang, H., Jia, Z., Shen, Z.: Research in security requirements engineering process, pp. 1285–1288. IEEE (2009)
17. Jain, S., Ingle, M.: Software Security Requirements Gathering Instrument. International Journal of Advanced Computer Science and Applications (IJACSA) 2(7), 116–129 (2011)
18. Chandrabose, A., Alagarsamy, K.: Security Requirements Engineering – A Strategic Approach. International Journal of Computer Applications (0975 – 8887) 13(3), 25–32 (2011)
19. Pandey, D., Suman, U., Ramani, A.K.: Security Requirement Engineering Issues in Risk Management. International Journal of Computer Applications (0975 – 8887) 17(5), 12–14 (2011)
20. Firesmith, D.: Engineering Security Requirements. Journal of Object Technology 2(1), 53–68 (2003), `http://www.jot.fm/issues/issue_2003_01/column6`
21. Apvrille, A., Pourzandi, M.: Secure Software Development by Example. IEEE Security & Privacy 3(4), 10–17 (2005)
22. Graham, D.: Introduction to the CLASP Process. Build Security (2006), `https://buildsecurityin.us-cert.gov/daisy/bsi/articles/best-practices/requirements/548.html`

Axial T2 Weighted MR Brain Image Retrieval Using Moment Features

Abraham Varghese[1], Reji Rajan Varghese[2], Kannan Balakrishnan[3], and J.S. Paul[4]

[1] Computer Science and Engg., Asiet, Kalady
[2] Co-operative Medical College, Cochin
[3] Cochin University of Science and Technology, Cochin
[4] Indian Institute of Information Technology and Management, TVM

Abstract. Magnetic resonance images play a vital role in identifying various brain related problems. Some of the diseases of the brain show abnormalities predominately at a particular anatomical location which on MR appears at a slice at defined level. This paper proposes a novel technique to locate desired slice using Rotational, Scaling and Translational (RST) invariant features derived from a ternary encoded local binary pattern (LBP)image. The LBP image is obtained by labeling each pixel with a code of the texture primitive based on the local neighborhood. The ternary encoding on LBP identifies the boundary of the uniform region and thus reduces the time for calculating moments of different order. The distance function based on the RST features extracted from LBP between query and database image is used to retrieve similar images corresponds to the query image.

Keywords: Feature Reduction, Rotational Scaling and Translational (RST) invariant features, Local Binary pattern, Eccentricity, Precision & Recall.

1 Introduction

Due to the developments in various imaging technologies, the number of images produced from different sources especially in medical field increases in a alarming rate. Nowadays Physician mostly relies on images for the diagnosis purpose. The accuracy in the medical diagnosis depends upon the information available on the image. Therefore it is very necessary to develop an appropriate information system to manage large collection of images [1-3].

One of the key issues in image management system is to locate a desired image in a large and varied collection of images. Patient-to-patient search, which can compare multiple patients and retrieve relevant cases among them can be used as a as a training tool for medical students for their follow-up studies and research purposes. The accuracy in diagnosis also can be improved by comparing similar images across the patients. Furthermore, in the medical domain, especially on brain images, experts focus on a region of-interest (single slice) or a volume-of-interest (several contiguous slices) in order to identify the cause of a pathology. In such cases, patient-to-patient search problem can further be simplified as retrieving the relevant slice given a query,

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 355–363.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

which can be used in diagnosis of structure specific diseases in the brain. Identification of similar slices from a volume of brain images may be taken as a first step in diagnosis of brain related problems. There are several methods depicted in the literature for the brain Image retrieval. The searches for medical tumors by their shape properties have been described in [4].Classifications of lesions in CT brain scans have been done in [5]. CBIR system for Traumatic brain injury (TBI) CT images has been proposed by Shimia et.al [6]. In this web-based system, user can query by uploading CT image slices and, retrieval result is a list of TBI cases ranked according to their 3D visual similarity to the query case. In [7] brain slice retrieval problem has been proposed using principal component analysis. But it requires image registration. A wavelet-based retrieval solution for brain images is introduced by Traina et al.[8]. A shape-based retrieval method that needed manual delineation of the anatomical structures is proposed in [9-10]. The geometric features and Fourier descriptors are used to represent brain images of pediatric patients, but it needs registered images [11]. The co-occurrence matrix representations, color quantization and wavelet responses are used to retrieve CT and MR images of different tissues [12-13]. LBP and KLT feature tracker have been used to extract region of interest form a brain MR images. LBP computes structure features in all local regions of the image while KLT identifies salient points in the image [15].

2 Methodology

First we give an overview of LBP and Moment Invariants which form a base for our work.

2.1 Local Binary Pattern

Local binary patter (LBP) has been used as a texture descriptor for various image retrieval problems. The LBP method can be regarded as a truly unifying approach. Instead of trying to explain texture formation on a pixel level, local patterns are formed. Original LBP is formed by taking difference between the gray value of a pixel (g_c) and the gray values of P pixels (g_k) in a local neighborhood [15].

$$LBP = \sum_{i=0}^{k-1} s(g_i - g_c)2^p, \begin{Bmatrix} s(x) = 1, x \geq 0 \\ 0, x < 0 \end{Bmatrix} \tag{1}$$

A number of variants of LBP have been evolved. The smallest value from a n-1 bitwise shift operation on the binary pattern of n bits gives a rotational invariant descriptor [16]. The standard LBP has been replaced by circular neighborhood as a rotation invariant descriptor, but in some problems the anisotropic structural information is an important information source. In order to attain this, another variant named elliptical binary pattern (EBP) has been proposed. Ternary encoding and Quinary encoding, are proposed for the evaluation of the local gray-scale difference [17].

Ternary encoding (T): Instead of binary encoding, here the difference is encoded as 3 values correspond to a threshold σ.

$$T = \begin{Bmatrix} 1, u - x > \sigma \\ 0, |u - x| \leq \sigma \\ -1, u - x < -\sigma \end{Bmatrix} \tag{2}$$

2.2 Moment Invariants

Hu [18] proposed Moment Invariants (MI) for two-dimensional pattern recognition applications. Two-dimensional moments of order (*p* + *q*) for digital image *f* (*x*, *y*) is defined as follows.

$$m_{pq} = \sum_p \sum_q x^p y^q f(x, y) \tag{3}$$

where, *p*, *q* = 0,1,2..... The summations are over the values of spatial co-ordinates *x* and *y* spanning the entire image. The moments in Eq(2.3) are not in general invariant under translation, rotation or scale changes in the image *f* (*x*,*y*). Translation invariance can be achieved by using central moment defined as follows.

$$\mu_{pq} = \sum \sum (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{4}$$

where $\bar{x} = \frac{m_{10}}{m_{00}}$, $\bar{y} = \frac{m_{01}}{m_{00}}$.

Information about image orientation can be derived by first using the second order central moments to construct a covariance matrix.

Covariance = $\begin{pmatrix} \mu'_{20} & \mu'_{11} \\ \mu'_{11} & \mu'_{02} \end{pmatrix}$ where $\mu'_{20} = \frac{\mu_{20}}{\mu_{00}}$,

$\mu'_{02} = \frac{\mu_{02}}{\mu_{00}}, \mu'_{11} = \frac{\mu_{11}}{\mu_{00}}.$

The Eigen of the covariance matrix can easily be shown to be

$$\lambda i = \frac{\mu'_{20} + \mu'_{02}}{2} \pm \frac{\sqrt{4\mu'^2_{11} + (\mu'_{20} - \mu'_{02})^2}}{2} \quad . \tag{5}$$

The relative difference in magnitude of the Eigen values are thus an indication of the eccentricity of the image, or how elongated it is.

The eccentricity is

$$\sqrt{1 - \frac{\lambda_2}{\lambda_1}} \tag{6}$$

The scaling invariant may be obtained by further normalizing μ_{pq} as

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\frac{p+q}{2}+1}} \tag{7}$$

Using this, a set of seven Rotational Scaling & Translational invariant features (RST) are derived by Schalkoff [19].

2.3 Filtering

The complexity of the retrieval problem increases with increase in number of features. Therefore it is very essential to retrieve relevant features from the images. Researchers proposed descriptors, such as the SIFT descriptor [20], Surf descriptor [21], KLT feature tracker to find the interesting regions of an image. We extract the relevant features in the following way.

Let I be an image of size M X N. Consider a window of size w x w where w = 3,5,7,9. The gray value of central pixel(g_c) of each window is compared with that of neighborhood pixels(g_n). If $g_n > g_c + \sigma$ or $g_n < g_c - \sigma$, where $\sigma > 0$, then g_n is assigned a value 1, otherwise 0. Thus a binary map image(BI) is formed. The filtered image(FI) is formed by assigning a value to the central pixel based on the gray value of the neighborhood pixels.

FI $= \frac{1}{Nb}\sum_{n=0}^{n-1} s(g_n - gc)gn,$ $\begin{Bmatrix} s(x) = 1, x \geq 0 \\ 0, x < 0 \end{Bmatrix}$ where Nb represents the number of 1's in the window. Thus each pixel is labeled with the code of the texture primitive that best matches the local neighborhood. Due to the presence of threshold, it identifies some edges in the image. In the case of a brain image with ventricular region, it identifies ventricular shapes.

2.4 Threshold Chosen

The threshold value σ is chosen locally based on the range of each window. i e $\sigma = (g_{max} - g_{min})/2$, where g_{max}- is the maximum gray value & g_{min}- is the minimum gray value of the of the window. It is observed that it is invariant to monotonic gray level change and rotation. The Fig 1. shows the original image and filtered image.

2.5 Feature Extraction

The filtered image *(FI)* has been divided into *l* disjoint blocks of size *w x w*. The seven RST moment Invariants [19] and eccentricity is calculated for each window. The eccentricity is calculated for each window using eqn (6) based on the Eigen value determined from the covariance matrix of the second order moments. There are $\left\lfloor \frac{N}{w} \right\rfloor \times \left\lfloor \frac{N}{w} \right\rfloor \times 8$ features for each image. Distance matrix is formed by the feature vectors of the query image with that of the images in the database using Euclidean distance function. The distance matrix is arranged in ascending order and the smallest will be assigned rank1, next smallest rank2 and so on The overall procedure is shown in the diagram, Fig 2.

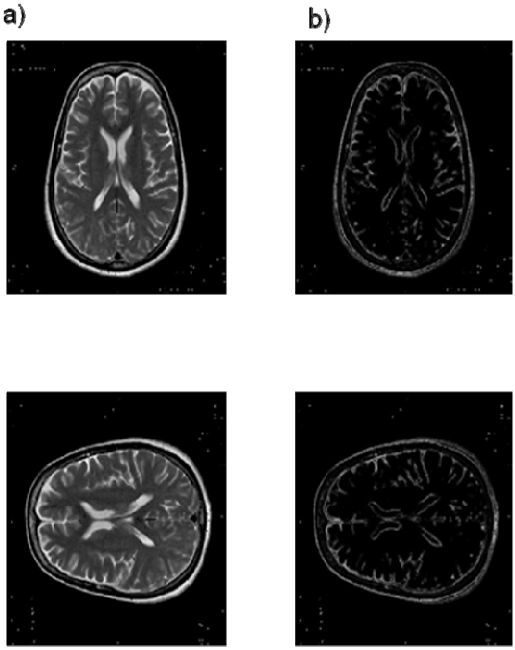


Fig. 1. a) original Image b) filtered Image

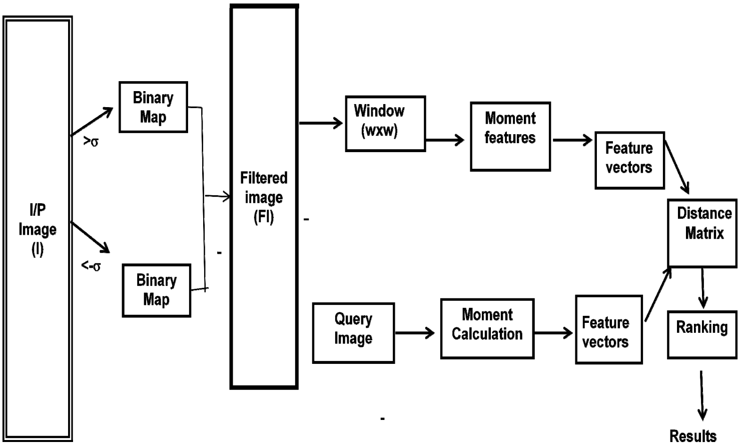


Fig. 2. Block Diagram of locating desired images

2.6 Performance Evaluation

Precision and Recall is calculated based on P- R graph.
Precision = No of relevant images retrieved / Total no of images retrieved
Recall = No of relevant images retrieved / Total no of relevant images retrieved

The precision and recall alone does not give full information. A precision score of 1.0 means all retrieved items are relevant, but it gives no information regarding whether all relevant items are retrieved. A recall score of 1.0 means all relevant items are retrieved, but it fails to give information regarding number of irrelevant images retrieved. Therefore a combined measure Precision-Recall is used. The weighted average of precision and recall is also used to evaluate the performance.

3 Results

We categorized 500 T2 weighted Brain unregistered MR images of different persons into 3 classes and evaluated the performance of the method. Class1 images represent the top of the ventricle. Class 2 and class3 is associated with ventricles. The slices used in this work, were acquired on a 1.5 Tesla, MR scanner from Pushpagiri Medical College Tiruvalla, INDIA. The T2 weighted images (TR/TE(eff.) of 3500-4500/ 85-105(eff.)ms) were collected using Fast Spin Echo (FSE) sequences with a matrix size of 320 X 224 (Frequency X Phase) and a NEX (Averages) of 2.

It is observed the time taken for extracting moments of different order of filtered image takes lesser time than that of original image. As the order increase time also increases. Fig 3 shows this

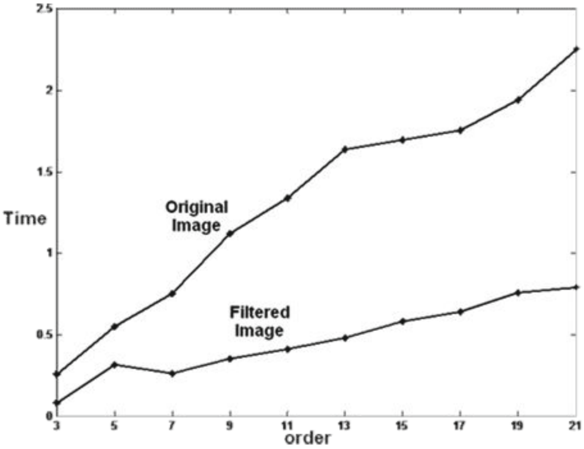


Fig. 3. Time for calculating moments of different order

We have taken 8 features for each window. It has been observed that the eccentricity value of similar images are closer compare with dissimilar images. One image from each class has been taken and its precision and recall is calculated. Fig 4 shows the P-R graph of the query images from different classes.

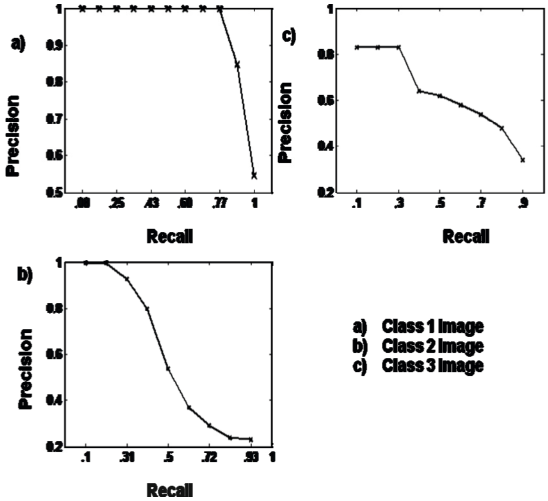


Fig. 4. Precision Vs Recall of images in different classes

It is also observed that accuracy of the retrieval increase as the window size (Number of directions) increases. after reaching an optimum value it starts decreasing. Fig 5 shows this

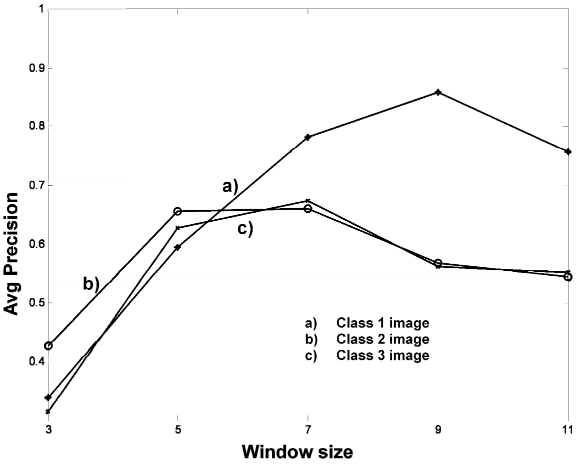


Fig. 5. The average precision vs window size of query images in different classes

The top 5 images retrieved in each class is shown in Fig 6.

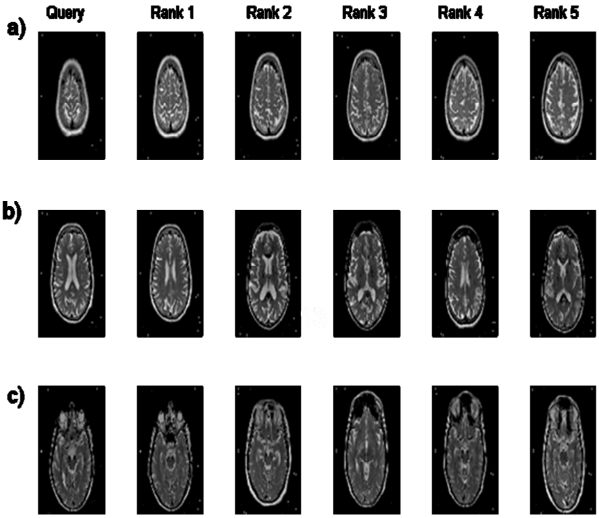


Fig. 6. The top5 images retrieved from each classs a) class 1 image b) class 2 image c) class 3 image

4 Conclusion

The paper illustrates a method to locate relevant slices from MR image database. The moment features extracted from a ternary encoded LBP image is used to locate relevant slices. The proper choice of features from LBP images and proper relevance feed mechanism may improve the results. Also this can be extended for retrieving T2 weighted coronal and sagittal slices from the MR image database.

References

[1] Müller, H., Michoux, N., Bandon, D., Geisbuhler, A.: A review content-based image retrieval systems in medical applications—Clinical benefits and future directions. Int. J. Med. Inf. 73(1), 1–23 (2004)

[2] Lehmann, T.M., Wein, B., Dahmen, J., Bredno, J., Vogelsang, F., Kohnen, M.: Content-based image retrieval in medical applications: A novel multi-step approach. In: Proceedings SPIE, vol. 3972, pp. 312–320 (2000)

[3] Smeulders, A.W.M., Worring, M., Santint, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)

[4] Korn, P., Sidiropoulos, N., Faloutsos, C., Siegel, E., Protopapas, Z.: Fast and effective retrieval of medical tumor shapes. IEEE Transactions on Knowledge an Data Engineering 10(6), 889–904 (1998)

[5] Liu, Y., Dellaert, F.: Classification-driven medical image retrieval. In: Proceedings of the ARPA Image Understanding Workshop (1997)

[6] Li, S., Gong, T., Wang, J., Liu, R., Tan, C., Leong, T.Y.: 3D Content-based CT Image Retrieval System for Traumatic Brain Injury. In: Proceedings SPIE (2010)

[7] Bucci, G., Cagnoni, S., Domenicis, R.D.: Integrating content-based retrieval in a medical image reference database. Comput. Med. Imag. Graph. 20(4), 231–241 (1996)

[8] Traina, A., Castañón, C., Traina Jr., C.: Multiwavemed: a system for medical image retrieval through wavelets transformations. In: Proc. 16th IEEE Symp. Comput., Based Med. Syst., NewYork, USA pp. 150–155 (2003)

[9] Robinson, G.P., Tagare, H.D., Duncan, J.S., Jaffe, C.C.: Medical image collection indexing: shape- based retrieval using kd-trees. Comput. Med. Imag. Graph. 20(4), 209–217 (1996)

[10] Petrakis, G.M., Faloutsos, C.: Similarity searching in medical image databases. IEEE Trans. Knowl. Data Eng. 9(3), 435–447 (1997)

[11] Huang, H.K., Nielsen, J.F., Nelson, M.D., Lui, L.: Image-matching as a medical diagnostic support (DST) for brain diseases in children. Comput. Med. Imag. Graph. 29(2-3), 195–202 (2005)

[12] Felipe, J.C., Traina, A.J.M., Traina Jr., C.: Retrieval by content of medical images using texture for tissue identification. In: 16th IEEE Symp. Comput.-Based Med. Syst., New York, USA, pp. 175–180 (2003)

[13] Müller, H., Rosset, A., Vallét, J.-P.: Comparing feature sets for content based image retrieval in a medical case database. In: Proc. SPIE Med. Imag., PACS Imag. Inf., San Diego, USA, pp. 99–109 (2004)

[14] Unay, D., Ekin, A., Jasinsch, R.S.: Local Structure-Based Region-of-Interest Retrieval in brain MR Images. IEEE Transactions on Information Technology in Biomedicine 14(4) (July 2010)

[15] Ojala, T., Pietikäinen, M., Harwood, D.: Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, ICPR 1994, vol. 1, pp. 582–585 (1994)

[16] Ojala, T., Pietikäinen, M., Harwood, D.: A Comparative study of texture measures with classification based on featured distribution. Pattern Recogn. 29(1), 51–59 (1996)

[17] Nanni, L., Lumini, A., Brahnam, S.: Local binary patterns variants as texture descriptors for Medical image analysis. Artificial Intelligence in Medicine (2010)

[18] Hu, M.K.: Visual Pattern Recognition by Moment invariants. IRE Trans. Info. Theory IT-8, 179–187 (1962)

[19] Schalkoff, R.: Digital Image Processing and Computer vision (1989)

[20] Lowe, D.: Distinctive image features from scale- invariant keypoints. Int. J. Comput. Vis. 60(2), 91–110 (2004)

[21] Herbert, A., Tuytelaars, T., Van Gool, L.: Speeded-Up Robust Features (SURF). Elseiver (2008)

Securing Data from Black Hole Attack Using AODV Routing for Mobile Ad Hoc Networks

V. Kamatchi, Rajeswari Mukesh, and Rajakumar

Department of Computer Science and Engineering,
Meenakshi College of Engineering, Chennai, India
kamatchiv@gmail.com, rajimukesh95@yahoo.co.in, rajamce@yahoo.com

Abstract. Mobile Adhoc Networks has become an important and exciting technology in recent years in which security has become an important issue. Black hole Attack is one of the possible and severe security attacks in mobile ad hoc networks which block the communication of secret data. Black hole attack directly attacks the node's data traffic on the path and intentionally drops, alters or delays the data traffic passing through that node. In another type of black hole attack which falsely replies for the route request which comes from the source that it has enough routes to the destination even it does not have path to the destination. This paper deals with prevention of both types of black hole attacks and secure data communication using secret sharing and Random Multipath Routing Techniques.

Keywords: Mobile Adhoc Networks, AODV, Blackhole attack, Destination Sequence Number, Random Routing.

1 Introduction

A Mobile Ad Hoc Network (MANET) is a self configuring Network and they are very attractive for military communications in hostile battlefield environments. In such situations, the ability to reliably communicate secret information in the presence of attacker is very difficult. Attackers may attempt both passive and active type of attacks to gain nonauthorized access to classify or modify or disrupt the information process. Nodes usually share the same physical media; they transmit signals and acquire them at the same frequency band. However, due to characteristics like dynamic topology and lack in centralized management security, MANET is exposed to various kinds of attacks. Black holes are the places or the areas within which the attacker can either passively intercept or actively block information delivery and black hole is a malicious node that falsely replies for any route requests and drops all the receiving packets which are forwarded towards the destination. Each and every mobile node in an ad-hoc network moves arbitrarily and acts as both a router and a Host.

The interconnections between nodes have the capacity of changing on a continuous and arbitrary basis. Nodes within the same radio range communicate directly via wireless links, but the nodes that are far away use other nodes as relays.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 365–373.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

1.1 BlackHole Attack

Black hole attack is one of many possible attacks in MANET. It is a kind of Denial of Service attack. This attack can be easily lessen by setting the promiscuous mode of each node and to see if the next node on the path forward the data traffic as expected is shown in Fig 1. In the other type, a malicious node sends a forged Route Reply (RREP) packet to a source node which initiates the route discovery to act as a destination node is given in Fig 2. When a source node receives multiple RREP from the same node it compares the destination sequence number contained in RREP packets and decides the greatest one as the most recent routing information and selects the route contained in that RREP packet. When sequence numbers become equal it will selects the route with the smallest hop count number. The data traffic starts flowing toward the attacker when the identity of destination node is spoofed by attacker since attacker sends Highest Destination Sequence number RREP to the source node.

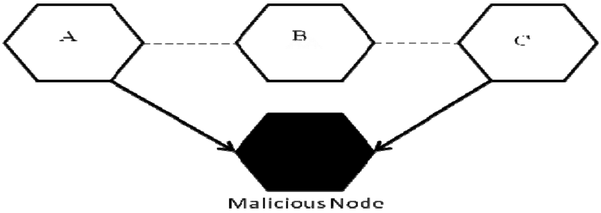


Fig. 1. BlackHole Packet Interception from Nodes

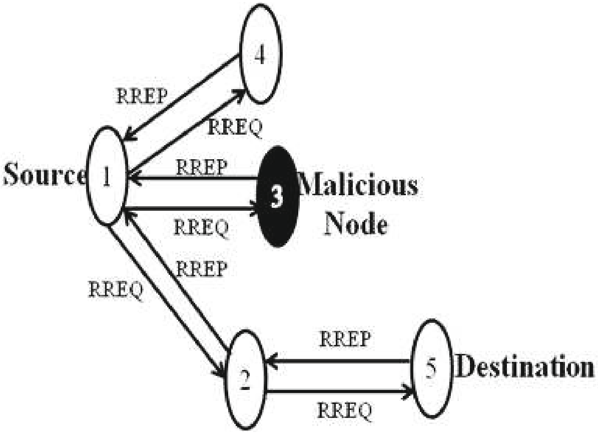


Fig. 2. BlackHole False Route Reply

1.2 AODV and Destination Sequence Number

AODV is a reactive Protocol used to discover the routes to start data communication. It builds the routes using route request and route reply cycles.DSN is a 32-bit integer [1] associated with every route and is used to decide the freshness of a particular route. The larger sequence numbers contain fresh route information. Node N3 will now send it to node. Both the nodes N1 and N2 do not have a route to Destination D. They should broadcast the RREQ message again to its neighbors.

Figure 3 shows that RREQ message broadcasted by the node N1 is also be received by node M which is considered as a blackhole. Thus, node M being malicious node, would generate a false RREP control message and send it to node N1 with a very high destination sequence number, that subsequently would be sent to the node S[2]. However, in simple AODV, as the destination sequence number is high, the route from node N1 will be considered to be fresher and hence node S would start sending data packets to node N1 then blackhole node stops the delivering the secret data which come from source node and starts dropping the packets intentionally.

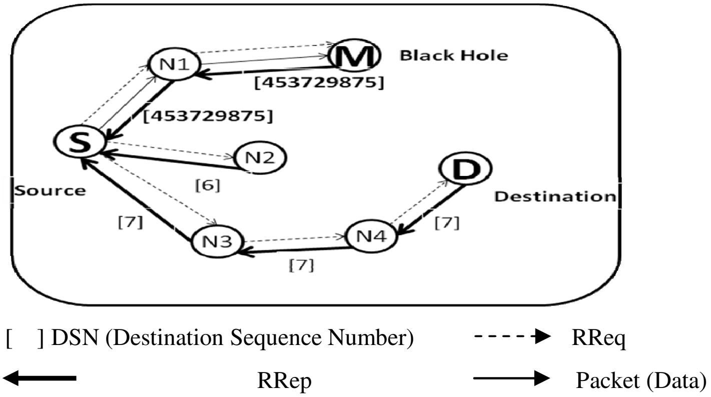


Fig. 3. BlackHole Forged RREP to the Source Node

2 Related Work

There has been numerous research results published in the literature survey that aims at finding the Black hole attacks is discussed.

Prashant B. Swadas, Payal N. Raj, [3] proposed"DPRAODV" that is "detection, prevention and reactive AODV". It is to inform the other nodes in the network about the security from blackhole. The ALARM packet is sent when the node is selected as anamoly. So that the malicious node's RRep is discarded. Then the Routing Table details will be discarded. This approach increases routing overhead and the average end to end delay.

Sanjay Ramaswamy, Huirong Fu, Manohar Sreekantaradhya, John Dixon and Kendall Nygard [4] proposed a method for identifying multiple black hole nodes.The authors have proposed a solution for cooperative blackhole attacks. They introduced a cross checking and data routing information table.The source node checks it own DRI whether intermediate node is reliable or not. If Source node uses Intermediate or neighbor node to send the packet then the packet is a reliable one. The cost of cross checking is more.

Mohammad Al-Shurman, Seong-Moo Yoo and Seungjin Park [5] proposed two different approaches to solve the black hole attack. In the first solution by utilizing the redundancy of the network the sender node will be verified for the authenticity. The idea behind following this solution is to find multiple routes for the destination. Once the valid route is identified the buffered packets will be transmitted through that route.The solution increases time delay. In the second solution, the last sent packet sequence number and the last received packet sequence number is stored in the routing table. When any packet is transmitted or arrived it is updated. If there is any mismatch or deviation in the sequence number then an ALARM will notifies the black hole node existence to neighbor.

Rei Heng, Cheng and Shun Chao Chang, Chang Wu Yu, Tung-Kuang, Wu, [6] proposed a procedure called" distributed and cooperative procedure "to detect black hole node.In this each node detects local anomalies. It collects information to construct an estimation table which is maintained by each node containing information regarding nodes within power range. The simulation result show the higher black hole detection rate and achieves better packet delivery.

3 The Proposed Solution

The solution proposed here is that the packets are delivered using multiple random routes after sharing of packets by secret sharing method .A four stage approach is considered. They are: 1.Preventing blackhole using Receive RouteReply method, 2.Secret sharing of information, 3.Randomized propagation of each information share using Random multipath routing, and 4.Normal routing toward the destination.

3.1 Preventing BlackHole Using Receive Route Reply (RRR) Method

SN-Source Node
DSN-Destination Sequence Number
SSN-Source Sequence Number
NID-Node ID
MN-ID-Malicious Node ID
RT-Routing Table

```
SN Broadcasts RREQ
 SN Receives RREP
  SN Stores DSN and NID in RT
   Retrieve First entry from RT
        IF (DSN>>>=SSN)
                {
                        MN-ID=NID
                        Black Hole Node
                }
        ELSE
                {
                         Normal Node
                }
```

When the Node with largest Sequence number is received by the source it is considered as a black hole and that route toward that black hole is discarded and the routing table is flushed or updated and sorted according to the destination sequence number.

3.2 Secret Sharing Method

When a node wants to send a packet to the destination, it first breaks or divides the packet into”N”shares, according to a (T; N)-threshold secret sharing algorithm or Shamir’s algorithm. In brief, this algorithm divides a message into N pieces, such that the original message can be reconstructed from any T pieces, where $T \leq N$, while any number of shares (pieces) less than T cannot yield any information about the original message.

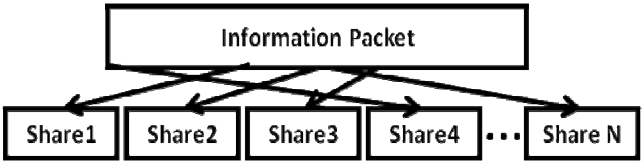


Fig. 4. Secret Sharing

Algorithm: Shamir’s Secret sharing :(T, N) Threshold Scheme

I. Setup Phase: Source

1. Chooses a large prime q
2. Selects a polynomial π Є $\prod$t-1 over Zq* such that $\pi (0) \equiv S \pmod q$
3. Computes si ≡ π (i) (mod q), i = 1,. . . ,n.
4. Distributes si to the shareholders Di, i = 1,…, n

II. Reconstruction Phase: Any group Γ of t shareholders

1. Compute π (0) ≡ $\sum$iЄr siLi (0) (mod q)

Note that Li (0) ≡ $\prod$ jЄr ,j≠i, j/j–i (mod q) are nonsecret constants and can be precomputed.

3.3 Randomized Propagation of Each Share

Randomized multipath routing algorithm that can overcome the fixed path problems. This Algorithm shows that multiple paths will be computed in a random way each time when an information packet needed to be sent to the destination, therefore the number of routes selected by different shares of different packets changing.. However, the algorithm ensures that the randomly generated routes are as dispersive as possible, i.e., the routes are geographically separated as far as possible such that they have high likelihood ofnot simultaneously passing through a black hole. Each share is then transferred to some randomly selected neighbors which is from updated routing table after deleting largest sequence number then that neighbor continually relay the share to other selected neighbors which are random. It chooses multiple possible numbers of paths to reach the destination.

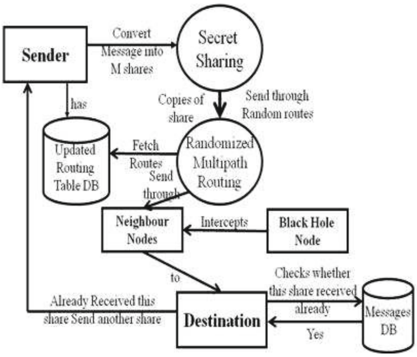

Sender
Convert Message into M shares
Secret Sharing
has
Copies of share
Send through Random routes
Updated Routing Table DB
Fetch Routes
Send through
Randomized Multipath Routing
Neighbour Nodes
Intercepts
Black Hole Node
to
Checks whether this share received already
Already Received this share Send another share
Destination
Messages DB
Yes

Fig. 5. Shamir Secret sharing and Random multipath routing Method

3.4 Normal Routing Phase

Normal Routing is done with AODV routing Protocol. The data are sent to these randomly selected multiple routes with “n shares’ of data so even if the attacker present in the route, he can hear only partial data among “n” number of shares. The paths are chosen randomly every time the sender node needs to send the data. If the sender receives ‘n’ shares of the same message, and then the energy consumption is more in receiver.

The sender needs to send all the ‘n’ shares of the messages to the receiver, even if the receiver receives the message, which causes unwanted energy consumption in sender node. So the unwanted energy consumption in both the sender and receiver node needs to be reduced, so, we propose a new method to achieve the minimum energy consumption. If the receiver receives some shares of the message from ‘n’ shares, then it should intimate to the sender that the receiver received the message. After receiving the confirmation message from the receiver, the sender node stop the current message transmission and starts to send another ‘n’ shares of the message. So the unwanted transmission in the both sender and receiver node is reduced.

4 Simulation and Results

The simulation is done with the help of NS-2 (v-2.34) network simulator. The implementation of the protocol has been done using C++ language in the backend and TCL language in the frontend on the Red hat Linux operating system.

Table 1. Simulation Parameters

Number of Mobile Nodes	20
Topology	1500m x 1500m
Number of Black Hole Node	2
Pause Time	5s
Traffic	Constant Bit Rate
Maximum Speed of Node	20 m/s
Packet Rate	4 Packets/s
Routing Protocol	AODV

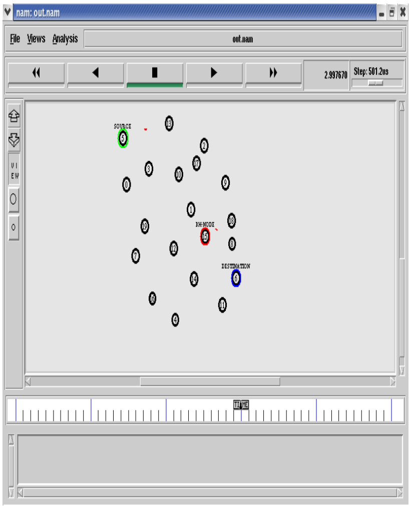


Fig. 6. Prevention of Blackhole using DSN

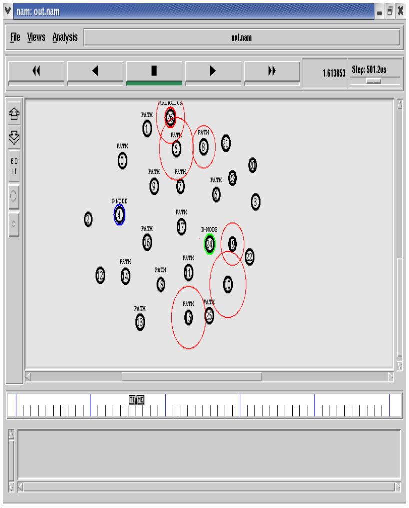


Fig. 7. Random Multipath routing

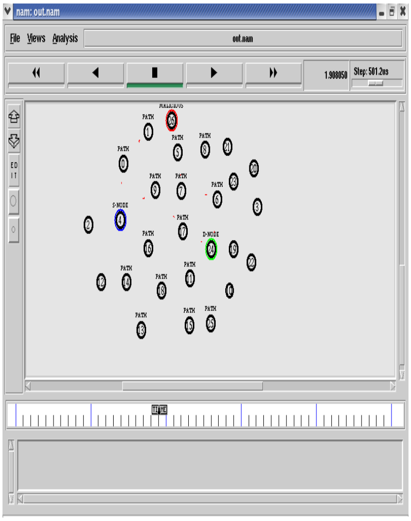


Fig. 8. With Blackhole Packet transmission using Random routes

The Figures 6, 7, and 8 shows how the blackhole route is discarded from routing table and random paths are chosen from refreshed routing table and secret shares are delivered to the destination using random routes.

4.1 Packet Delivery Ratio

PDR is the ratio of the number of CBR packets received by the destination to the number of CBR packets sent by the source.

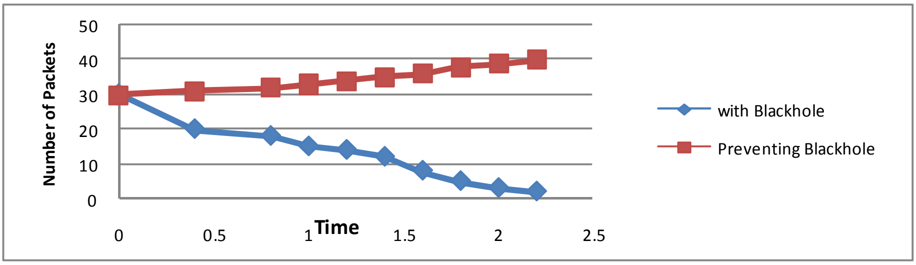


Fig. 9. Packet Delivery Ratio

Figure 9 shows the effect of blackhole on throughput of received packets after deleting the blackhole using the single path and the effect of using random routes to secure the packets with increased throughput. Clearly throughput has been increased to the maximum.

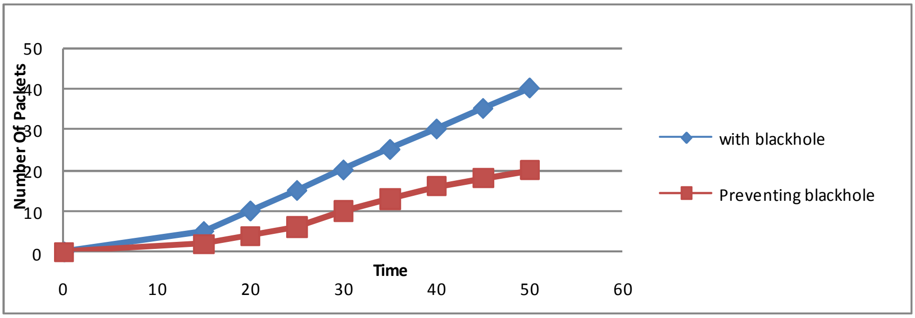


Fig. 10. Packet Loss Ratio

4.2 Packet Loss Ratio

As speed increases, the position of a node will clearly change more rapidly. This will cause more and more packets to time out before reaching their destinations. By sending the packets through random routes the blackhole interception rate is reduced to low thus the delay in sending of packets is reduced.

5 Conclusion and Future Work

By using Random dispersive routes maximum throughput is achieved with reduced delay even after blackhole presence. Energy consumption at both sender and the receiver is reduced and high security is achieved. Communication between sender and receiver is achieved with minimal energy factor. Future work can include the areas to develop simulations to analyze the performance of the proposed solution based on the various security parameters like mean delay time, packet overhead, memory usage, mobility, increasing number of malicious node, increasing number of nodes and scope of the black hole nodes.

References

[1] Himral, L., Vig, V.: Preventing AODV Routing Protocol from Black Hole Attack. International Journal of Engineering Science and Technology 3(5), 3927–3932 (2011)

[2] Claveirole, T., de Amorim, M.D., Abdalla, M., Viniotis, Y.: Securing wireless sensor networks against aggregator compromises. IEEE Communications Magazine, 134–141 (April 2008)

[3] Raj, P.N., Swadas, P.B.: DPRAODV: A dynamic learning system against black hole attack in AODV based MANET. International Journal of Computer Science Issues (IJCSI) 2(3), 54–59 (2009)

[4] Ramaswamy, S., Fu, H., Sreekantaradhya, M., Dixon, J., Nygard, K.: Prevention of cooperative black hole attack in wireless ad hoc networks. In: International Conference, ICWN 2003, Las Vegas, Nevada, USA, pp. 570–575 (2003)

[5] Yoon, S.-M., Park, S., Al-Shurman, M.: Black Hole Attack in Mobile Ad Hoc Networks. In: Proceedings of the 42nd Annual Southeast Regional Conference, ACM South East Regional Conference, pp. 96–97 (2004)

[6] Yu, C.W., Wu, T.-K., Cheng, R.-H., Chang, S.C.: A Distributed and Cooperative Black Hole Node Detection and Elimination Mechanism for Ad Hoc Networks. In: Washio, T., Zhou, Z.-H., Huang, J.Z., Hu, X., Li, J., Xie, C., He, J., Zou, D., Li, K.-C., Freire, M.M. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4819, pp. 538–549. Springer, Heidelberg (2007)

ISim: A Novel Power Aware Discrete Event Simulation Framework for Dynamic Workload Consolidation and Scheduling in Infrastructure Clouds

R. Jeyarani[1], N. Nagaveni[1], S. Srinivasan[2], and C. Ishwarya[3]

[1] Coimbatore Institute of Technology, India
symphonyjr@gmail.com, nagavenipalanisamy@yahoo.com
[2] Tata Consultancy Services, India
srinivasan.soundararajan@tcs.com
[3] Alcatel Lucent Limited, India
ishwarya.chandrasekar@alcatel-lucent.com

Abstract. Today's cloud environment is hosted in mega datacenters and many companies host their private cloud in enterprise datacenters. One of the key challenges for cloud computing datacenters is to maximize the utility of the Processing Elements (PEs) and minimize the power consumption of the applications hosted on them. In this paper we propose a framework called ISim, wherein a Datacenter manager playing the role of a Meta-scheduler minimizes power consumption by exploiting different power saving states of the processing elements. The considered power management techniques by the ISim framework are dynamic workload consolidation and usage of low power states on the processing elements. The meta-scheduler aims at maximizing the utility of the cores by performing dynamic workload consolidation using context switching between the cores inside the chip. The Datacenter manager makes use of a prediction algorithm to predict the number of cores that are required to be kept in active state to fulfil the input service request at a given moment, thus maximizing the CPU utilization. The simulation results show, how power can be conserved from the host level till the core level in a datacenter with the optimal usage of different power saving states without compromising the performance.

Keywords: Cloud computing, power conservation, VM provisioning, workload prediction, dynamic workload consolidation.

1 Introduction

Enterprise datacenters is one of the most rapidly growing emitters of greenhouse gas pollution. With the vast power consumption by the datacenters and the significant cooling effect required to maintain them, they become one of the major sources of global warming. Google has consumed over 2 billion kilowatt-hours (kWh) worth of energy in 2010 [1]. According to a report released by Pike research [2] with the shift of enterprise infrastructure to cloud, the datacenter power consumption will decrease by 31 percent between 2010 and 2020. A report issued by the Carbon Disclosure

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 375–384.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Project [3], supported by AT&T, finds that a company that adopts cloud computing can reduce its energy consumption, lower its carbon emissions and decrease its capital expenditure on IT resources while improving operational efficiency.

The dynamic power management techniques such as DVFS and DCD can be implemented in the hardware as part of an electronic circuit. But the implementation and reconfiguration of dynamic power management algorithms and policies in hardware are difficult in addition to poor visibility towards the environment. To address this issue, the organizations such as Intel, Toshiba and Microsoft have published Advanced Configuration and Power Interface (ACPI) specification to provide standardized nomenclature for various power saving states and also have defined software interfaces for managing them [4]. It is an open standard and it defines unified operating system centric device configuration and power management interface. The interface can be used by software developers to leverage flexibility in adjusting system power states.

ACPI defines five system states [5]. A PE conserves different amount of power at different sleep states. Many modern power-aware processors and microcontrollers have built-in support for active, idle or standby and sleep operating modes [6]. The active power is the power consumed while the chip is doing useful work. The standby power is the power consumed while the chip is idle. In sleep mode, the supplies to unused circuits are turned-off. The time taken to come to live state from any sleep state is called wake up latency, and the power drawn during wake up time is comparatively insignificant. The power conservation and wake up latency are inversely related in a particular sleep state [7]. To generalize the sleep states supported in various systems, the proposed work categorizes them into two classes of sleep states viz. the shallow sleep states and the deep sleep state. To maximize the power conservation it is required that the computing resources in the datacenter are well utilized and the idle resources are kept in appropriate power saving states.

As the incoming workload to the datacenter is highly dynamic in nature, the number of cores required for a VM request at a given instance is predicted with the help of historical data and provisioned for immediate allocation and the remaining processing cores are transitioned to the appropriate power saving states thus preventing the underutilisation of the cores and maximizing the power conservation. The proposed ISim framework allows seamless modelling, simulation and experimentation of emerging cloud computing infrastructure and power aware services.

ISim is a discrete event driven simulator framework for realizing power aware datacenters. ISim toolkit offers the following features which helps to save power and consolidate the workload in the cloud datacenter.

- Simulation of cloud environment that supports dynamic VM provisioning in datacenter
- Workload generation algorithm to generate three types of input workload pattern namely smooth, drastic and mixed variation in resource requirement, processing time and batch size
- Run Time Prediction of number of cores, chips and hosts that are required to be kept in active state and adaptive provisioning of these resources.
- Implementation of Chip level power-aware VM allocation policies using bin packing method
- Dynamic workload consolidation without performance degradation

The unique features of ISim toolkit facilitate, researchers and cloud infrastructure providers to simulate and test their power-aware scheduling and consolidation policies in a controlled and easy to set-up environment.

2 Literature Survey

Power conservation can be done from transistor level within a microprocessor to external level involving cooling infrastructure. Hence the power conservation methods are classified either as external power saving or internal power saving methods based on the level at which they are done. Dynamic Voltage and Frequency Scaling is a kind of internal power saving method which exploits the various operating states of a processing core based on the load. Apart from using operating states, IBM Power systems define three sleep states namely nap, sleep and winkle to conserve power.

Earlier works on power and energy aware management focused on increasing the battery life time for mobile devices [8, 9]. Current cloud infrastructure such as Amazon EC2 [10] does not consider power aware resource allocation of VM for its users. The proposed work concentrates on power saving at PE level, Chip level and host level using various sleep states thereby facilitating the provider to save significant amount of power without violating Service Level Agreements (SLA) [11, 12, 13].

Pinheiro et al. [14] have proposed a technique called “load concentration” to save power while providing the QoS requirements. The author uses a static estimation method to predict the performance by considering the demand for various resources. This method is inefficient if the total demand exceeds the available resource capacity leading to throughput degradation. Our work predicts the demand for resources of the incoming jobs dynamically based on the recent history hence assuring the demand fulfilment while conserving power.

Elnozahy et al. [15] studied the problem of power management in a web application environment that has constant SLA. The authors considered CPU as a major source of power consumption and focused on the power management considering the CPU alone. The authors proposed various policies to scale the voltage and turn on/off the idle nodes. This approach does not consider the wake up latency and its power consumption. The significant wake up latency is a major drawback causing SLA violations. The proposed method is also limited to the fact that it cannot handle dynamic workload leading to inefficient decision making and wastage of power and energy.

Srikantaiah et al [16] highlighted the importance of key observable characteristics such as resource utilization, performance and energy consumption in designing an effective consolidation strategy. Their work considered CPU and disk resource combination for consolidation and determined optimal utilization level at which maximum power can be saved. However, the proposed model does not consider the performance degradation due to consolidation. Our work introduces a reduced bin packing algorithm for efficient consolidation of PEs, with limited migration and negligible transition overhead which is more suitable for a generic cloud environment and it is also not application or workload dependent.

Nathuji and Schwan [17] proposed a new power management technique called soft resource scaling in virtualized systems where hardware scaling is emulated by providing a VM less time for the resource utilization. The authors proposed local and global policies for managing power. At the local level guest VM's power management policy is leveraged and QoS is maintained by changing the power state in accordance with guest OS policies. This becomes insignificant if the guest OS is non power aware. The global policies manage multiple hosts and consolidate them using migration in order to save significant power. However the author does not describe how the policy is applied for efficient migration of the VMs.

3 ISim Framework

3.1 System Architecture

Fig. 1. Depicts the proposed system architecture for ISim and the interaction between the different components of ISim. The heart of ISim is the datacenter manager which acts as a Meta-scheduler performing the runtime demand prediction, adaptive resource provisioning, VM scheduling and dynamic workload consolidation onto the hosts in a datacenter.

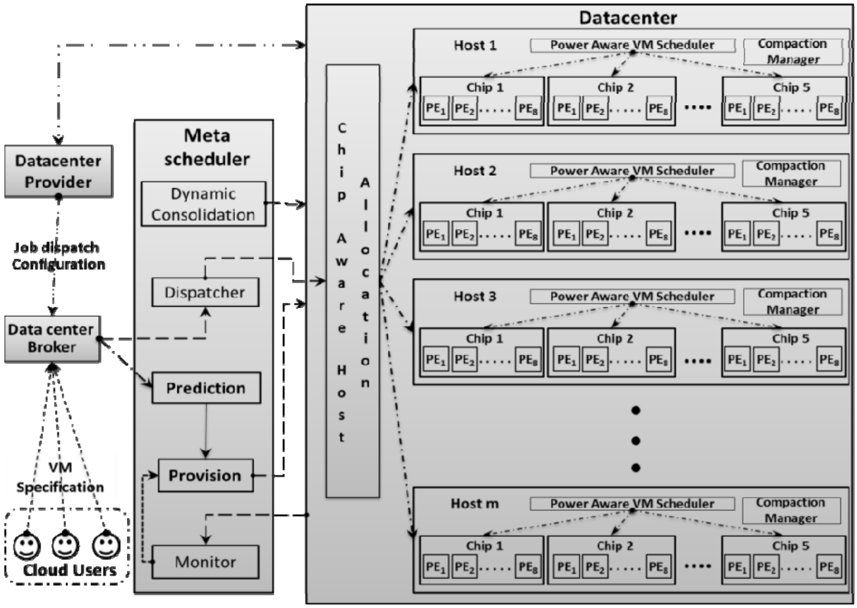


Fig. 1. Architecture of ISim Framework

3.2 ISim Class Hierarchy

A class diagram hierarchy of the ISim package represented using UML notation is shown in Fig. 2. The specification of each class contains up to three parts: attributes, methods and internal classes. In the class diagram, attributes and methods are prefixed with characters '+', '-', and '#', indicating access modifiers public, private, and protected, respectively. The ISim package implements the following classes.

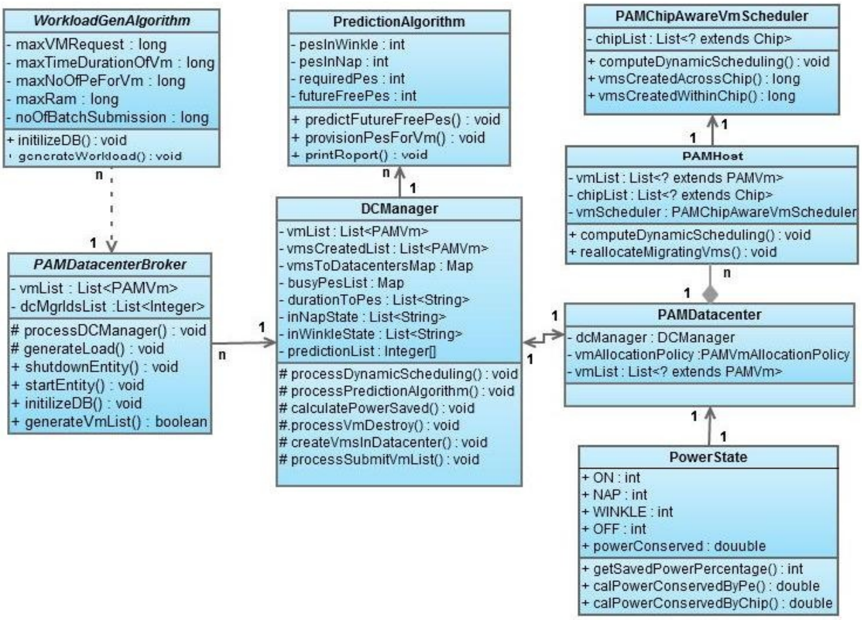

WorkloadGenAlgorithm
- maxVMRequest : long
- maxTimeDurationOfVm : long
- maxNoOfPeForVm : long
- maxRam : long
- noOfBatchSubmission : long
+ initilizeDB() : void
+ generateWorkload() : void
n
1
PAMDatacenterBroker
- vmList : List<PAMVm>
- dcMgrIdsList :List<Integer>
# processDCManager() : void
# generateLoad() : void
+ shutdownEntity() : void
+ startEntity() : void
+ initilizeDB() : void
+ generateVmList() : boolean
n
1
PredictionAlgorithm
- pesInWinkle : int
- pesInNap : int
- requiredPes : int
- futureFreePes : int
+ predictFutureFreePes() : void
+ provisionPesForVm() : void
+ printReport() : void
n
1
DCManager
- vmList : List<PAMVm>
- vmsCreatedList : List<PAMVm>
- vmsToDatacentersMap : Map
- busyPesList : Map
- durationToPes : List<String>
- inNapState : List<String>
- inWinkleState : List<String>
- predictionList : Integer[]
# processDynamicScheduling() : void
# processPredictionAlgorithm() : void
# calculatePowerSaved() : void
#.processVmDestroy() : void
# createVmsInDatacenter() : void
# processSubmitVmList() : void
1
1
PAMChipAwareVmScheduler
- chipList : List<? extends Chip>
+ computeDynamicScheduling() : void
+ vmsCreatedAcrossChip() : long
+ vmsCreatedWithinChip() : long
1
1
PAMHost
- vmList : List<? extends PAMVm>
- chipList : List<? extends Chip>
- vmScheduler : PAMChipAwareVmScheduler
+ computeDynamicScheduling() : void
+ reallocateMigratingVms() : void
n
1
PAMDatacenter
- dcManager : DCManager
- vmAllocationPolicy :PAMVmAllocationPolicy
- vmList : List<? extends PAMVm>
1
1
PowerState
+ ON : int
+ NAP : int
+ WINKLE : int
+ OFF : int
+ powerConserved : douuble
+ getSavedPowerPercentage() : int
+ calPowerConservedByPe() : double
+ calPowerConservedByChip() : double

Fig. 2. ISim Class Hierarchy

Class ISim.PE: This is used to represent CPU/Processing Element. The capability of PE is defined in terms of Million Instructions Per Second (MIPS) rating.

ClassISim.chip: An instance of this class simulates a collection of PEs. This class can be instantiated to model Muti Chip Module (MCM) and Multi Core System (MCS).

Class ISim.PAMHost: It represents uniprocessor or shared memory multiprocessor machines.

Class ISim.Datacenter: An instance of this class simulates a collection of hosts. This class can be instantiated to model large scale, heterogeneous datacenter. It is associated with PAMVmAllocation Policy class.

Class ISim.Datacenter Characteristics: This represents the static properties of a datacentre such as architecture, operating systems, cost per memory, cost per storage, cost per bandwidth, etc.

Class ISim.PAMVmAllocation Policy: This class is used to find the host for executing the submitted job. It can be specialised as heuristic or non- heuristic allocation Policy. Generally, Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Simulated annealing (SA), etc., can be implemented under heuristic allocation policy. Any one of the bin packing algorithms such as Best Fit, Worst Fit, First Fit, etc., can be implemented as a part of non-heuristic allocation algorithms.

Class DC Manager: This is a key entity performs two major functionalities namely workload prediction and resource provisioning.

Class PAMDataCenterBroker: This acts as an interface between the user and the Datacenter Manager. It is instantiated to supply batches of workload to Datacenter Manager.

Class PAMChipAwareVmScheduler: This class is an abstract class that can be implemented either as a space shared or time shared PAMVmScheduler. Its main functionality is to assign the incoming workload to the available PEs in the hosts in the datacenter.

Class DynamicPAMVmScheduler: This class is instantiated to perform dynamic compaction periodically by packing the workloads within few hosts.

Class PAMVmSchedulerBackFilling: This class implements conservative backfilling policy which reduces the makespan of a batch of workloads.

4 Proposed Policies for Efficient VM Provisioning Using ISIM

4.1 Dynamic Runtime Prediction

The dynamic runtime prediction module predicts dynamic arrival pattern of VM requests based on the recent history of arrivals, and uses it for provisioning resources. As the resource demand in a cloud datacenter is observed to be a time series, statistical prediction methods for time series is used. The moving average method is used to predict the total number of required PEs for the next batch of VM requests. In simple moving average method, a window size is fixed and the average demand in the window gives the forecasted demand.

4.2 Adaptive Resource Provisioning

Based on the runtime prediction, the Meta-Scheduler performs adaptive provisioning to fulfil forthcoming demands. Let the scheduling process start at time t and the request batching duration be t_o. The steps for adaptive provisioning are shown below.

Algorithm for Adaptive Resource Provisioning

```
Input: PEs in SS state, PEs in DS state and predicted RequiredPEs representing forecasted demand for the next batch
  Repeat
  a.Get thepredictedRequiredPEs for the next dispatch time t+t0
  // extra provisioning = X% of SS state
  b.SetPEsToBeProvisioned ← predictedRequiredPEs + extra provisioning
  c. Find the PEs that will become free at t+t0; Let it be denoted as futureFreePEs;
  d. If futureFreePEs are more than PEsToBeProvisioned then excessPEs are transitioned to DS state
```

```
          else// This number is either satisfied from PEs
         in SS state or from PEs in  DSS
      Find the number ofinsufficientPEsfor allocation
      e.if (insufficient PEs are less than PEs in SS
state)
      allocateinsufficientPEs from PEs in SS state
         else
      allocateinsufficientPEs from PEs in DS state
      while (there are VM requests )
```

4.3 Dynamic Workload Consolidation

As the dynamic consolidation of VMs with the help of VM migrations between hosts involves lots of overhead, dynamic consolidation through the context switching of VMs is performed within each host. The context switching avoids the following cases. Those VMs which have lesser remaining execution time than threshold are not context switched. The VMs that have already been relocated more than a fixed number of times are not context switched. A VM that is already running inside a single chip, but the consolidation results in placing it across more than one chip are also not context switched.

Algorithm Dynamic Workload Consolidation

```
Input: chipList of a host and set of VMs in each chip
Output: The result of the compaction process. Success is
returned, if it results in freeing at least a chip,
else failureis returned
// donorList represents source chips, from which VMs are
relocated
//receiverList represents a set of target chips, to which
VMs are relocated
// n denotes the number of chips in a host
Step 1: Sort the chips based on numberOfBusyPEs in the
ascending order
Step 2:for (I = 1 to n) do
add chip1 … chipi to ReceiverList
add chipi+1…chipn to DonorList
if the numberOfFreePEs in receiverList can accommodate
the numberOfBusyPEs in  donorList
begin
       i. VMs are context switched such that VMs in Do-
          norList uses PEs in ReceiverList
      ii. result  = success
      iii.return result
end
end for
```

5 Experimental Setup and Simulation Results

The simulated environment is set up using ISim toolkit. The simulator is extended to include the entities such as Chip, Datacenter Broker, Datacenter Manager (DC Manager and power aware VM allocation policies. For a resource pool containing 'm' hosts and each host having one chip with eight cores, the power consumption is modeled as 20 watts per processor core.

5.1 Power Conservation with Single Power Saving State

Three kinds of input workload patterns namely smooth variation, drastic variation and mixed variation with required number of PEs are generated for submitting to the datacenter. The power aware VM scheduler uses the Best Fit Decreasing allocation policy to allocate the input VM requests onto the hosts in the datacenter. This experiment is conducted to show the realization of significant power conservation with single power saving state. For a given random workload pattern at first, power conservation is computed using a provisioning policy that does not consider any power saving state. The same is repeated with a provisioning policy with shallow sleep state called Nap. Fig. 3 shows the comparison of power conservation with single power saving state and with host shutdown.

From the experiments it can be clearly seen that by using the power states of the cores good amount of power can be conserved with better resource prediction. Since the resource pool consists of multicore and multi chip modules, the idle cores, chips and modules can be transitioned to shallow sleep state contributing to significant power conservation.

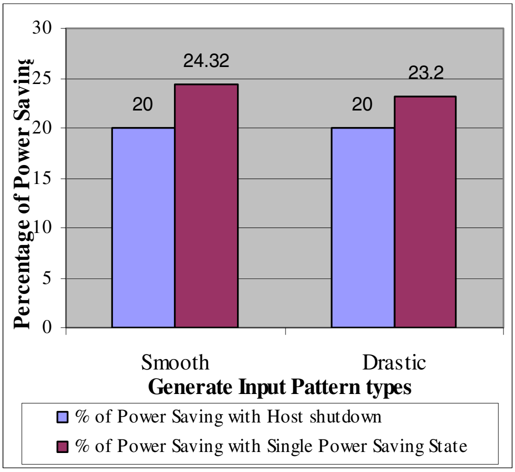


Fig. 3. Power Saving with host shut down and Single Power saving State

5.2 Power Saving with Multiple Power States

This experiment was conducted to exploit multiple power saving states such as shallow and deep sleep states. Based on the prediction, the required PEs are transitioned to shallow sleep state. To manage the prediction error, an additional percentage of PEs are kept in deep sleep state. The chips and hosts that are not used by the VMs in near future are shut down state. This results in improved power conservation compared to single power saving state and is depicted in Fig. 4.

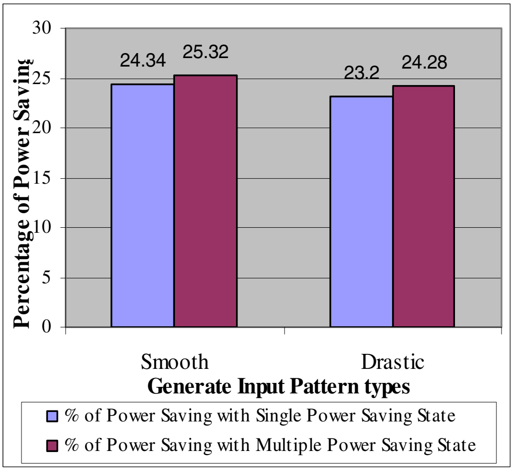


Fig. 4. Power Conservation under Multiple Power Saving States

6 Conclusions and Scope for Future Work

The key challenge in cloud computing is to manage unexpected demand and resource provisioning on the fly in an efficient manner. The proposed ISim toolkit facilitates the cloud provider to manage his hosts in the cloud datacenter efficiently with huge power conservation without compromising the performance. The meta-scheduler of ISim supports adaptive provisioning policy along with power aware allocation policy to provision the VMs on the host dynamically. Several experiments were conducted to show the effectiveness of the proposed framework in saving power and managing cloud datacenter. The future work aims to implement application profiling as well as resource profiling at cloud level by introducing a monitor tool to observe the pattern of application execution.

References

1. Boulton, C.: Google Reveals Energy Consumption to Tout Green Efforts (2011), `http://www.eweek.com/c/a/Green-IT/Google-Reveals-Energy-Consumption-to-Tout-Green-Efforts-854799/` (accessed September 20, 2011)

2. Cloud Computing Could Cut Datacenter Energy Consumption by Nearly One-Third by 2020, `http://www.pikeresearch.com/newsroom/cloud-computing-could-cut-data-center-energy-consumption-by-nearly-one-third-by-2020` (accessed September 20, 2011)
3. Carbon Disclosure Project. Building a 21st century communications economy (2011), `https://www.cdproject.net/en-US/WhatWeDo/Pages/21st-Century-Comms-Economy.aspx` accessed September 24, 2011
4. Venkatachalam, V., Franz, M.: Power Reduction Techniques for Micro Processor Systems. ACM Computing Surveys 37(3), 195–237 (2005)
5. Ware, M., Rajamani, K., Floyd, M., Brock, B., Rubio, J.C., Rawson, F., Carter, J.B.: Architecting for Power Management: The IBM POWER 7 Approach. In: IEEE 16th International Symposium High Performance Computer Architecture, HPCA (2010)
6. Cardosa, M., Korupolu, M.R., Singh, A.: Shares and Utilities based Power Consolidation in virtualized Server environments. In: IFIP/IEEE International Symposium on Integrated Network Management, New York, pp. 327–334 (2009)
7. Le, H.Q., Starke, W.J., Fields, J.S., O'Connell, F.P., Nguyen, D.Q., Ronchetti, B.J., Sauer, W.M., Schwarz, E.M., Vaden, M.T.: IBM POWER6 microarchitecture. IBM Journal of Research and Development 51 (2007)
8. Neugebauer, R., McAuley, D.: Energy is just another resource: Energy accounting and energy pricing in the nemesis OS. In: 8th IEEE Workshop on Hot Topics in Operating Systems, pp. 59–64 (2001)
9. Zeng, H., Ellis, C.S., Lebeck, A.R., Vahdat, A.: ECOSystem: managing energy as a first class operating system resource. ACM SIGPLAN Notices 37(10), 132 (2002)
10. Amazon Elastic Compute Cloud (Amazon EC2) (2011), `http://www.amazon.com/ec2/` (accessed September 23, 2011)
11. Kant, K.: Datacenter evolution A tutorial on state of the art issues and challenges. Computer Networks 53, 2939–2965 (2009)
12. Kim, K.H., Beloglazov, A., Buyya, R.: Power-Aware Provisioning of Virtual Machines for Real-Time Cloud Services. In: Concurrency and Computation: Practice and Experience. Wiley Press, New York (2011)
13. Nurmi, D., Wolski, R., Grzegorczyk, C., Obertelli, G., Youseff, S.S.L., Rodnov, D.Z.: The Eucalyptus Open Source Cloud Computing System. In: 9th IEEE/ACM International Symposium on Cluster Computing and the Grid (2009)
14. Pinheiro, E., Bianchini, R., Carrera, E.V., Heath, T.: Load balancing and unbalancing for power and performance in cluster-based systems. In: Workshop on Compilers and Operating Systems for Low Power, pp. 182–195 (2001)
15. Elnozahy, E., Kistler, M., Rajamony, R.: Energy-efficient server clusters. Power-Aware Computer Systems, 179–197 (2003)
16. Srikantaiah, S., Kansal, A., Zhao, F.: Energy aware consolidation for cloud computing. Cluster Computing 12, 1–15 (2009)
17. Nathuji, R., Schwan, K.: Virtualpower: Coordinated power management in virtualized enterprise systems. ACM SIGOPS Operating Systems Review 41(6), 265–278 (2007)

Processing RDF Using Hadoop

Mehreen Ali[1], K. Sriram Bharat[1], and C. Ranichandra[2]

[1] MS (Software Engineering), VIT University, Vellore, India
mehreen1591@yahoo.com, sriramkondeti@gmail.com
[2] Assistant Professor (Senior), VIT University, Vellore, India
cranichandra@vit.ac.in

Abstract. The basic inspiration of the Semantic Web is to broaden the existing human-readable web by encoding some of the semantics of resources in a machine-understandable form. There are various formats and technologies that help in making it possible. These technologies comprise of the Resource Description Framework (RDF), an assortment of data interchange formats like RDF/XML, N3, N-Triples, and representations such as RDF Schema (RDFS) and Web Ontology Language (OWL), all of which help in providing a proper description of concepts, terms and associations in a particular knowledge domain. Presently, there are some existing frameworks for semantic web technologies but they have limitations for large RDF graphs. Thus storing and efficiently querying a large number of RDF triples is a challenging and important problem. We propose a framework which is constructed using Hadoop to store and retrieve massive numbers of RDF triples by taking advantage of the cloud computing paradigm. Hadoop permits the development of reliable, scalable, proficient, cost-effective and distributed computing using very simple Java interfaces. Hadoop comprises of a distributed file system HDFS to stock up RDF data. Hadoop Map Reduce framework is used to answer the queries. MapReduce job divides the input data-set into independent units which are processed in parallel by the map tasks , which then serve as inputs to the reduce tasks. This framework takes care of task scheduling, supervising them and re-execution of the failed tasks. Uniqueness of our approach is its efficient, automatic allocation of data and work across machines and in turn exploiting the fundamental parallelism of the CPU cores. Results confirm that our proposed framework offers multi-fold efficiencies and benefits which include on-demand processing, operational scalability, competence, cost efficiency and local access to enormous data, contrasting the various traditional approaches.

Keywords: Semantic Web, Distributed Computing, Map-Reduce Programming, SPARQL, Graph Data, Performance Evaluation.

1 Introduction

Cloud computing is a budding concept in the IT and data dispensation communities. Various new ventures are utilizing cloud computing service to contract out data safeguarding, which results in noteworthy financial benefits. Enterprises usually stock up and access data at far-off sites in the “cloud”. As the reputation of cloud computing

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 385–394.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

nurtures, the service providers come across several challenges. They have to maintain gigantic magnitudes of heterogeneous data while providing well-organized information retrieval. Hence the prominence for usage of cloud computing is scalability and query efficiency.

Semantic Web technologies [2] are being urbanized to present data in standardized way such that the data can be retrieved and understood by both human and machine. Traditionally, web pages are published in simple html files and hence are not suitable for interpretation. Instead, the machine regards these html files as a container of keywords. Researchers are cultivating these Semantic Web technologies to address such shortcomings. The most well-known standards are Resource Description Framework1 (RDF) and the SPARQL Protocol and RDF Query Language (SPARQL). RDF is the standard for piling up and expressing data and SPARQL is a query language to retrieve data from an RDF stockpile. Via OWL (Web Ontology Language) ontologies, different schemas, classes, data types and relationships can be represented without forgoing the benchmark RDF/SPARQL interface. Figure 1 reflects the fundamental organisation of Semantic Web.

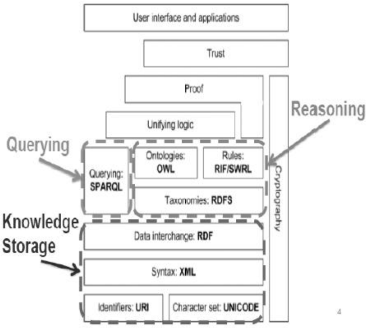


Fig. 1. Semantic Web Layer Cake

Semantic web datasets are mounting exponentially. More than any other domain, in the web arena, scalability is supreme. Nevertheless, high speed response time is also critical in the web society. Cloud computing concept offers a way out that can achieve both the goals. Present industrial tools and technologies do not dwell well in Cloud Computing scenery.

A distributed system can be fabricated to triumph over the scalability and performance issues of present Semantic Web frameworks. Datasets are being disseminated to help in providing such scalable solutions. Yet, till date, there isn't any distributed storehouse for storing and managing RDF data. We put forward a solution with a generic distributed storage system which makes use of a Cloud Computing platform. And then propose a way to tailor the system and schema explicitly to meet the demands of semantic web data. Finally, we recommend constructing a semantic web storehouse using such a storage provision.

In this paper we spotlight on the design characteristics inherent to using the MapReduce software framework to assemble highly-parallel, high-performance and scalable data management systems. Our views and discussions in these fields are stimulated by our understanding of designing, constructing and evaluating our preliminary implementation of a triple store which is vigorous, robust, scalable and distributed.

A triple-store is a data storage and recovery environment for graph data, conventionally represented in RDF formats [15]. Our triple store stores graph data [17] as RDF triples and retorts to queries over this data using SPARQL query language. We make use of the Hadoop implementation of MapReduce to build it.

2 Hadoop: The Best Implementation of MapReduce

MapReduce is a software framework used to process and generate large datasets. Users state a map function that divides data into key/value pairs and a reduce function that amalgamates all key/value pairs on the basis of the key.

MapReduce software framework is effortlessly parallelizable for implementation on bulky clusters of commodity equipments. This facilitates the production of high-performance, highly-scalable applications. The most accepted MapReduce implementation is Hadoop. Hadoop controls the management of data on compute nodes by making use of the Hadoop Distributed File System (HDFS), scheduling the program's execution over a set of machines, managing machine breakdowns, and handling the obligatory inter-machine communication. This paves way for the design and implementation of high-level functionality by means of the MapReduce framework to put up high-performance and highly scalable applications.

A main feature of the MapReduce software framework, as articulated in the Hadoop implementation, is the utilization of a unique, centralized node, referred to as the NameNode which guides the posting of data onto compute nodes using HDFS, allocates compute jobs to the diverse nodes, traces fiascos and supervises the shuffling of data after the Map step concludes.

The unit of computation in Hadoop is referred to as a job. Here the users submit jobs to Hadoop’s JobTracker module. Every job consists of two phases: Map and Reduce. The Map phase takes a key-value pair as input and possibly will output zero or more key-value pairs. In Reduce phase, the values of every key are assembled into compilations traversable by an iterator. The generated key-iterator pairs are in turn passed to the Reduce method, which as well outputs zero or more key-value pairs. When a job is presented to the JobTracker, Hadoop attempts to place the Map processes near to the entered data in the cluster. All the Map process and Reduce process work autonomously without communicating and hence beneficial for both swiftness and ease.

3 Design Goals

Our principal data system design driving force is the knack to persist and speedily query hefty data graphs. To align with Semantic Web data principles, we consider graphs characterized as subject-predicate-object triples [2][6]. A small illustration graph can be seen in Figure 1 that contains 7 triples –

John lives in Pune, John owns an object car0, car0 is a car, car0 was made by Ford, car0 was made in Delhi, Delhi is a city and Pune is a city.

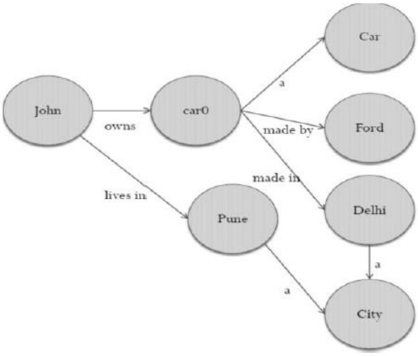


Fig. 2. A Small Graph of Triple Data

We make use of SPARQL [18] -it is the benchmark Semantic Web query language. SPARQL semantics are broad-spectrum and similar to the well-known SQL. An example SPARQL query for the above graph data is the following:

SELECT ?person WHERE
{
?person :owns ?car .
?car :a :car .
?car :madeIn :Delhi.
}

The above query written in SPARQL language has three clauses and looks for all equivalences to the variable *?person* such that `?person` owns a unit specified by the variable `?car` which is a car and was made in Delhi. The above query corresponds to the directed graph as seen in Figure 2.

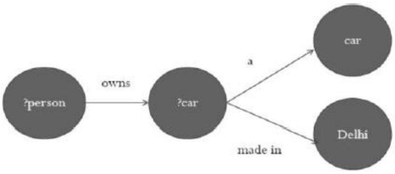


Fig. 3. A Directed Graph Representation of a Query

Processing of SPARQL queries in the perspective of a data graph such as the one above involves identifying which variables in the query clauses can be related to nodes in the data graph in order to let the query clauses align with the data triples [16]. This process of alignment for query processing is moderately universal across various data depictions and query languages. An example of this sort of alignment for our example query can be seen in Figure 4.

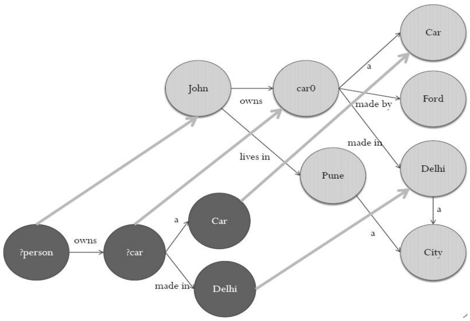


Fig. 4. Alignment of SPARQL Query Variables with Triple Data

The main functional design objectives for the triple-store are to:

1. Serve as an unrelenting store for triple data in RDF set-up.
2. Provide as a SPARQL endpoint to process SPARQL queries.

There have been several other design approaches for triple-stores with comparable design goals. Several of these triple-stores have accomplished superior performance on solo compute-node systems [14] by making use of designs based on memory mapping index information [9]. Nevertheless, disk and memory restrictions have motivated the necessity for distributed computing tactics to triple-stores [10][12].

4 Proposed Architecture of Our Framework

Our architecture comprises of two components. The upper part of Figure 5 portrays the data pre-processing component and the lower part portrays the query answering one. There are three sub-components for data generation and pre-processing. Conversion of RDF/XML to N-Triples serialization format is done using N-Triples Converter module.

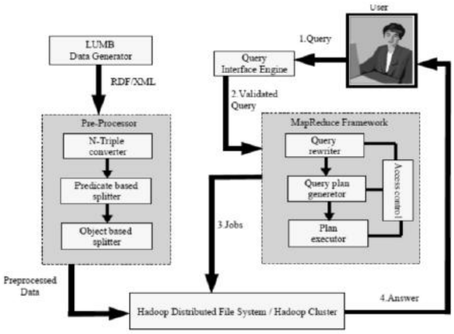


Fig. 5. The System Architecture

The PS module splits the N-Triples data into predicate files. These predicate files are in turn fed into the POS module which divides the predicate files into smaller files based on the type of objects. Our MapReduce framework has three subcomponents in it. It inputs the SPARQL query from the user and feeds it to the Query Rewriter and Query Plan Generator. This module picks the input files to decide how many MapReduce jobs are required and then pass the data to the Plan Executer module which runs the jobs using MapReduce framework. It then dispatches the query result from Hadoop to the user.

4.1 Data Partitioning and Triple Placement

In order to use a distributed computing approach to information management system design, it is generally infeasible to pass large volume input data directly to and from the user. This data passing would involve the coordinated movement of data onto and off of the compute nodes as and when the data needs to be processed. The large magnitude of data makes this approach unfeasible owing to data churn. As a result, hefty input (data and queries) and output (results of query) data sets are required to be stored directly onto the compute nodes. This direct storage of data on the compute nodes is done natively by means of the Hadoop implementation of MapReduce which involves placing data in the HDFS distributed file system.

Data is maintained in the form of flat files in the HDFS file system i.e. each line of the triple-store text file represents all triples related with a different subject. Though this approach of persisting triple data in the form of flat text files is rudimentary when evaluated against other data management approaches, it brings a level of automated robustness which is achieved by replicating the data and MapReduce operations across various nodes [7]. The data is stored in a simple, easy to read set-up that imparts itself to easier, user centric drill-down investigation of query results returned from the triple store. [13]

4.2 Query Execution

MapReduce offers only plain data manipulation techniques by splitting the data into key-value pairs, and combining all values with the same keys. For complex query processing, there is a need for data management systems to iterate over clauses of the queries to incrementally bind variables of the query to the literals in the triple store while fulfilling all the constraints of the query. All iterations consist of a MapReduce operation intended for a single clause in the query.

The initial map step maps the data in the triple store to a list of variable bindings such that the first clause of the query is satisfied. The importance of the Map step is the list of variable bindings. Once done, the Reduce step discards duplicate results and uses variable binding as the key to save them to disk.

The intermediary query binding steps iteratively bind variables to literals as and when new variables are introduced by processing succeeding query clauses and additionally sieving out the previous bindings which cannot satisfy the new clauses. These intermediate steps carry out MapReduce operations over the triple data as well as the earlier bound variables saved to disk.

The ith intermediate Map step's job is to identify all variables in the triple-data which satisfy the ith clause and thus save this result with the key as any of the variables in the ith clause which surfaced in previous clauses. The value of this Map step is the bindings of the variables that have not been seen in the previous clauses. The iteration of this Map step in addition rearranges the results of the prior variable bindings which have been saved to disk to the equivalent name of a variable key present in the ith clause that materialized in previous clauses.

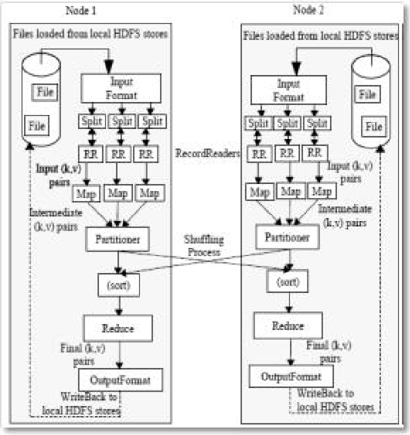


Fig. 6. Iteration of MapReduce to Process SPARQL Queries

Consequently, the ith Reduce step performs a join function on top of the intermediary results from the Map step by iterating over the entire results pairs from the prior clause and the recent clause by means of the assignment of the same key.

These map-reduce-join iterations continue until the entire clauses in the query are administered and variables are assigned such that they satisfy all the query clauses. Our triple-store design saves the intermediate results of the query processing onto local disk to increase the swiftness in processing of similar queries which might come into view later. The finishing step of MapReduce carries out filtering operations on the bound variable assignments in order to satisfy the main `SELECT` clause of the respective SPARQL query.

5 System Implementation and Initial Results

In this segment, we draft an initial design of our system and show preliminary results. In order to test the performance of our proposed design based on MapReduce framework for a scalable data management system, we built a premature version of triple store using the Cloudera Hadoop(CDH3) that we deployed on top of an Amazon EC2 [1] cloud environment comprised of 20 XL nodes running RedHat Linux and Cloudera Hadoop.

5.1 Lehigh University Benchmark(LUBM)

We made use of the LUBM benchmark [5] to assess the performance of our triple store. This benchmark generates synthetic data concerning the publishing, coursework and advising activities of the faculty and students belonging to different departments in various universities.

Once the data triples were stocked up onto the triple store, we assessed our framework's performance in responding to 1st, 9th and 14th query of LUBM as these were the queries used for studying the performance of previous triple-store. 1st query is basically very plain which asks for the all students that take a particular course. Its response is a very small set of triples. 9th query is fairly complicated query comprising of a triangular pattern of relationships and it asks for all the teachers, students and courses such that the teacher is the adviser of the student who takes a course taught by the teacher. 14th query is reasonably simple as it just asks for all the undergraduate students but the response generated includes very large set of triples.

5.2 Summary of Results

Our triple store evaluation attained the following query response time for the 3 queries for 6000 universities which corresponds to approximately 800 million triples using the LUBM benchmark when deployed on an Amazon EC2 cloud with 20 compute nodes:

Query 1: 323 sec. (approx 0.1 hr.); **Query 9**: 560 sec. (approx 0.2 hr.); **Query 14**: 100 sec. (approx 0.03 hr.)
For evaluation, in the triple store study the industrial single-machine DAMLDB triple-store achieved the following performance on the same queries in combination with the Sesame2 and Jena3 Semantic Web frameworks in order to assist in query processing.

Sesame+DAMLDB took:

Query 1: approx 0.1hr.; **Query 9**: approx 1 hr.; **Query 14**: approx 1 hr.
For Jena+DAMLDB, owing to the difficulty in loading triples into this dataset, we cannot have data on performance over 550 million triples. However based on observed trends this triple-store probably would require the following response times:

Query 1: approx 0.001 hr., **Query 9**: approx 1 hr.; **Query 14**: approx 5 hr.

It is to be noted that the only query where our framework performed evidently worse than DAMLDB was on query 1. Query 1 returns a very small subset of values bound to variables. Even though MapReduce is conventionally used to build indices, its implementation of Hadoop provides little support for accessing data stocked up in HDFS files. Conversely, DAMLDB makes use of a number of exceptional indexing optimizations for plain queries like that for Query 1. Our triple store does not implement such optimization techniques. Barring this one drawback, our approach performed better than other existing technologies owing to the vastly parallel and distributed implementations of the MapReduce framework.

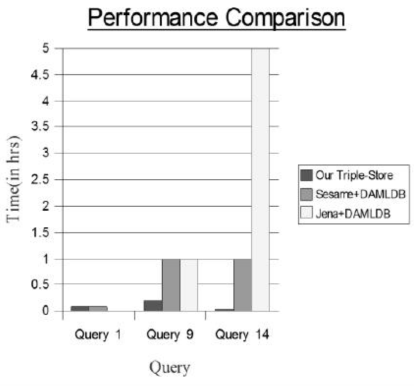


Fig. 7. Graph Comparing Performances of Various Triple-Stores

6 Future Directions

On basis of our understanding of the initial operation of our proposed framework, we have quite a few short and long-term actions needed to further improve the performance and acceptance equally from design as well as software framework perspective.

Foremost improvement needed is a more effectual method for indexing data. In addition to this, we need a supporting alternative to Map-Reduce that supports native indexing as a replacement for the basic Map operations over all the stored data elements.

In addition to this we can make improvements in performance of our design which can be made available by caching the partial results locally for high performance parallel operations as well as globally using a entity like NameNode which will track the local caching of the partial results. To achieve this we require additional potential in the software framework to keep track of the partial results which were cached beforehand and perhaps to decide which cached results can possibly be discarded to produce free disk space in the cloud. This should be done only if it is a deployment concern.

A major advancement that can be done to support the design of information systems would be a substituting software framework that aids in data linking. As an alternative of storing lists of data in the form of flat files in constructs similar to HDFS, a software framework can be built that can offer a inherent linked-data construct which couples the data elements and the pointers to related data. The resulting linked-data framework will provide more rapid localized processing of queries without a need for exhaustive searching of the data set for each and every query request.

References

[1] Amazon. Amazon EC2 Instance Types (2010), `http://aws.amazon.com/ec2/instance-types/`

[2] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American Magazine (May 17, 2001)

[3] Dean, J., Ghemawat, S.: MapReduce: Simplified data processing on large clusters. In: Proceedings of the USENIX Symposium on Operating Systems Design & Implementation, OSDI, pp. 137–147 (2004)

[4] DeWitt, D., Stonebraker, M.: MapReduce: A major step backwards, `http://database-column.com`, `http://databasecolumn.vertica.com/database-innovation/mapreduce-a-major-step-backwards/` (retrieved August 28, 2010)

[5] Guo, Y., Pan, Z., Heflin, J.: LUBM: A benchmark for OWL knowledge base systems. Journal of Web Semantics 3(2), 158–182 (2005)

[6] Grigoris, A., van Harmelen, F.: A Semantic Web Primer, 2nd edn. The MIT Press (2008)

[7] Urbani, J., Kotoulas, S., Oren, E., van Harmelen, F.: Scalable Distributed Reasoning Using MapReduce. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 634–649. Springer, Heidelberg (2009)

[8] Hendler, J.: Web 3.0: The Dawn of Semantic Search. IEEE Computer (January 2010)

[9] Kolas, D., Emmons, I., Dean, M.: Efficient Linked-List RDF Indexing in Parliament. In: The Proceedings of the Scalable Semantic Web (SSWS) Workshop of ISWC 2009 (2009)

[10] Li, P., Zeng, Y., Kotoulas, S., Urbani, J., Zhong, N.: The Quest for Parallel Reasoning on the Semantic Web. In: Liu, J., Wu, J., Yao, Y., Nishida, T. (eds.) AMT 2009. LNCS, vol. 5820, pp. 430–441. Springer, Heidelberg (2009)

[11] LinkingOpenData (2010), `http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

[12] Mika, P., Tummarello, G.: Web Semantics in the Clouds. IEEE Intelligent Systems 23(5), 82–87 (2008)

[13] Husain, M., McGlothlin, J., Masud, M.M., Khan, L., Thuraisingham, B.: Heuristics Based Query Processing for Large RDF Graphs Using Cloud Computing. Journal of Latex Class Files 6(1) (January 2007)

[14] Project Voldemort (2010), `http://project-voldemort.com/`

[15] RDF. Resource Description Framework (RDF) (2010), `http://www.w3.org/RDF/`

[16] Rohloff, K., Schantz, R.: High-Performance, Massively Scalable Distributed Systems using the MapReduce Software Framework: The SHARD Triple-Store. In: International Workshop on Programming Support Innovations for Emerging Distributed Applications, PSIEtA (2010)

[17] Rohloff, K., Dean, M., Emmons, I., Ryder, D., Sumner, J.: Evaluation of Triple-Store Technologies for Large Data Stores. In: 3rd International Workshop on Scalable Semantic Web Knowledge Base Systems, SSWS 2007, Vilamoura, Portugal (2007)

[18] SPARQL. SPARQL Query Language for RDF (2010), `http://www.w3.org/TR/rdf-sparql-query/`

An Ant Colony Optimization Based Load Sharing Technique for Meta Task Scheduling in Grid Computing

T. Kokilavani[1] and D.I. George Amalarethinam[2]

[1] Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India
vani78_ram@yahoo.com
[2] Director, Department of MCA, Jamal Mohamed College, Tamil Nadu, India
di_george@jmc.edu

Abstract. Grid Computing is the fast growing industry, which shares the resources in the organization in an effective manner. Resource sharing requires more optimized algorithmic structure, otherwise the waiting time and response time are increased, ansd the resource utilization is reduced. In order to avoid such reduction in the performance of the grid system, an optimal resource sharing algorithm is required. The traditional min–min algorithm is a simple algorithm that produces a schedule that minimizes the makespan than the other traditional algorithms in the literature. But it fails to produce a load balanced schedule. In recent days, ACO plays a vital role in the discrete optimization problems. The ACO solves many engineering problems and provides optimal result which includes Travelling Salesman Problem, Network Routing, and Scheduling. This paper proposes Load Shared Ant Colony Optimization (LSACO) which shares the load among the available resources. The proposed method considers memory requirement as a QoS parameter. Through load sharing LSACO reduces the overall response time and waiting time of the tasks.

Keywords: Grid Computing, Ant Colony Optimization, Argentine Ants, Resource Sharing.

1 Introduction

As the scientific problem grows very complex in the modern computing technology, it requires more computing power and more storage space. Based on these basic requirements, an organization requires higher computational resource when dealing with current technological methodology. The past technologies such as distributed computing, parallel computing are not suitable for recent advancement. The modern computer industry operates with very large amounts of data which utilise more processing power and high storage volumes of data. Therefore, the Grid computing is proposed as effective resource management to the organization. In grid computing, the network status and the resource status are to be managed effectively. If the network status or resource status are not in feasible level, then the total computation time will be increased dramatically. In grid computing, the user will encounter thousands of computers to utilize in effective and efficient manner. The Grid architectures are serving as a middleware technology for various purposes like resource allocation management, job scheduling, data

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 395–404.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

management, security and authorization. Grid resource brokers [1] are responsible for gathering resource information, discovery, resource management and job scheduling. Various scheduling methods exist which try to optimize time, cost, resource utilization and load balancing. Fig. 1 shows the research activities related to Grid scheduling methods. Programming in the grid computing involves more complexities which not only require a single-machine application but needs additional features. Some of the additional aspects in the grid computing are 1) Dividing and combining data and results, 2) Data security, 3) Application security, 4) Testing, and 5) Redundancy and capacity planning.

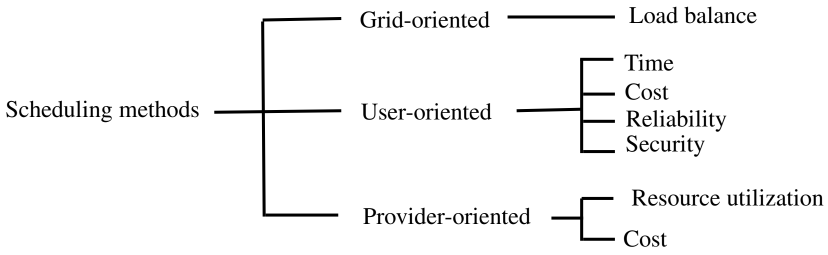


Fig. 1. Various Research Activities in the Grid Environment

The purpose of task scheduling in the grid computing is to balance the load of entire grid system in such a way that completing all the assigned workload as soon as possible and feasible than other system. It is impossible for anyone to manually assign these loads in the large computing resources of grid system. As the environment status of grid architecture is changing frequently, the traditional job scheduling algorithm such as ”First Come First Serve” (FCFS), ”Shortest Job First” (SJF), etc., may not be suitable for the dynamic environment in grids. Therefore, job scheduling in the grid environment is a very important issue. The Non-traditional algorithms differ from the conventional traditional algorithms in that it produces optimal results in a short period of time [2]. There is no best scheduling algorithm for all grid computing systems. An alternative is to select an appropriate scheduling algorithm to use in a given grid environment based on the characteristics of the tasks, machines and network heterogeneity. The Grid scheduler should consider the QoS requirement of a task to find a perfect match of resource [3]. The QoS constraint is the responsibility of scheduler and it does not focus on the objective function. This paper proposes an efficient job scheduling algorithm based on Ant Colony Optimization which considers memory as a QoS requirement and shares the load among the resources for optimizing the resource usage in the grid environment.

2 Related Work

Braun *et al.* [4] have studied the relative performance of eleven heuristic algorithms for task scheduling in grid computing. They have also provided a simulation basis for researchers to test the algorithms. Their results show that Genetic Algorithm (GA) performs well in most of the scenarios and the relatively simple Min-Min algorithm performs next to GA and the rate of improvement is also very small. The simple algorithms proposed by Braun are Opportunistic Load Balancing (OLB), Minimum Execution Time (MET), Minimum Completion Time(MCT), min–min, max–min.

Wei-Neng *et al.* [5] proposed Ant Colony System (ACS) for Grid computing, in which the author considers the scheduling of tasks in terms of more than one quality of service (QoS) parameters. The ACO proposed by them enables the users to specify their QoS preferences as well as define the minimum QoS thresholds for a certain application and the ACS is tested in ten task applications with at most 120 tasks. Based on the characteristics of workflow scheduling, they have designed seven new heuristics for the ACO approach and propose an adaptive scheme that allows artificial ants to select heuristics based on pheromone values. This ACO proposed by them decreases the cost by 10–20% compared with the existing Deadline based approach.

Ruay-Shiung Chang at al [6] proposed a Balanced Ant Colony Optimization (BACO) algorithm for job scheduling in the Grid environment. The main contributions of this paper is to balance the entire system load while trying to minimize the make span of a given set of jobs, the BACO focuses on the make-span and system load balance. According to the author the experimental results in this paper shows that BACO can outperform other job scheduling algorithms.

AliEn RB [7] is a Grid Broker which handles File transfer optimization, fault tolerance by multithreading, and Push and pull task assignment. In Apples [1], the Parameter study support, event-driven rescheduling, Centralized adaptive scheduling with heuristics, and self-scheduled work queues are handled. In EZ-GRID Broker [8], job handling, transparent file transfer, self-information service with dynamic and historical data, Policy Engine Framework for provider policies are proposed.

A load balancing algorithm aims to increase the utilization of resources with light load or idle resources thereby freeing the resources with heavy load [9]. The algorithm tries to distribute the load among all the available resources. At the same time, it aims to minimize the makespan with the effective utilization of resources. Only few of scheduling algorithms for grid task are focusing the problems with a variety of QoS parameter.

Stefka *et al.* [10] proposed a high throughput computing scheduling algorithm based on ACO. The authors says that ACO algorithm can be interpreted as parallel replicated Monte Carlo (MC) systems. The scheduling algorithm designed by them is for distributed systems shared asynchronously by both remote and local users. They have developed 3 simulated grid examples and one ant to evaluate the newly proposed ACO algorithm for grid scheduling.

Graham Ritchie *et al.* [11] proposed an ant colony optimisation (ACO) algorithm that, when combined with local and tabu search, can find shorter schedules on benchmark problems than other techniques found in the literature.

Zhihong Xu *et al.* [12] suggested that the structure of Grid makes the QoS of Grid more difficult than network and distribute computing environment. They proposed a category scheduling based on ant algorithm that makes use of user category, resource category and task category. The algorithm proposed by them improved the successful scheduling rate, system performance and Quality of Service (QoS) to a certain extent.

3 Ant Colony System

The ant algorithm is inspired from the behaviour of real ants. A colony of cooperative ants moves from their nest to find food. The ants deposit a chemical pheromone on its path while moving from nest to food. When more ants move in the same direction the strength of pheromone increases on the path. Other ants use this to choose the shortest path. The idea of ants was used by Marco Dorigo to develop an Ant Colony System which aimed to search for an optimal path in a graph [13]. He formed artificial ants to find a solution for the Travelling salesman problem. The algorithm used the past history to update the pheromone value. Based on the pheromone value the shortest path was identified. The state transition rule used by ant system, called a random-proportional rule, is given by Equation (1) which gives the probability with which ant k choose the resource r for job s.

$$p_k(r,s) = \begin{cases} \frac{\tau(r,s)\times[\eta(r,s)]^\beta}{\sum\tau(r,s)\times[\eta(r,s)]^\beta}, & \text{if } s \in J_k(r) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where τ is the pheromone, η is the heuristic value, and J_k(r) is the set of jobs that remain to be scheduled by ant k.

The pheromone is updated according to Eq. (2):

$$\begin{aligned} \tau(r,s) &\leftarrow (1-\alpha)\times\tau(r,s) + \sum(1-\alpha)\times\tau(r,s) \\ \Delta\tau_K(r,s) &= \begin{cases} \frac{1}{CT_K}, & \text{if resource found} \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{2}$$

Deneubourg *et al.*proposed an ant colony system based on Argentine ants. According to the author most of the ants lay a pheromone trail only when they return from food source to nest. But argentine ants lay the pheromone trail both when leaving from and returning to the nest from food source. They proposed a model based on the behavior of these ants and formulated a equation by assuming that two paths are available for a ant. Fig. 2 shows the two bridge and the formula is given in Eq. (3).

$$\text{Prob}_A = \frac{(k+A_i)^n}{(k+A_i)^n + (k+B_i)^n} (\text{Prob}_A + \text{Prob}_B = 1) \tag{3}$$

$$\begin{aligned} A_{i+1} &= A_i + \delta, \\ B_{i+1} &= B_i + (1-\delta), \quad (A_i + B_i = i) \end{aligned} \tag{4}$$

where δ is a stochastic variable that takes a value of 1 or 0 based on the probability of Prob_A and Prob_B respectively. A is increased by 1 unit if ant chooses path A, otherwise B is increased by 1 unit if ant chooses path B. Based on the values on n and k, successive ants choosing the path may vary.

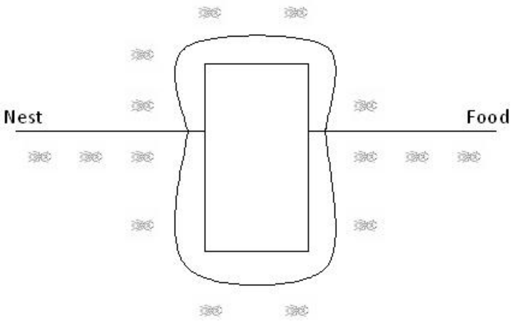


Fig. 2. Ant moving from Nest to Food

4 Proposed Work

This paper proposes a new load sharing approach based on the behavior of argentine ants. In deneubourg work they propose that the ants choose either path A or path B based on the probability value. This work considers path A as resource1 and path B as resource2. It also assumes that both resources are capable of executing the given tasks. The resources should be chosen based on the RAM requirement of task and the availability of RAM in the given resource. The traditional method allots any one resource based on the task requirements. In Ant Colony Optimization also, based on the probability and the requirements the task choose either resource1 or resource2. But this paper aims to share the load on both resources which will minimize both the makespan and wait time of the tasks. So it subdivides the given task and allots a percentage of tasks to both resources based on the probability value obtained by applying the random proportional rule. The random proportional rule used to identify the percentage of task allotted in a resource is given in Eqs. (5) and (6).

$$P_1 = \frac{(R_1 + k)^h}{(R_1 + k)^h + (R_2 + k)^h} \tag{5}$$

$$A_1 = P_1 \times \mathrm{TR}_i \tag{6}$$

where A_1 is the amount of task allotted in resource1 and TR_i is the memory requirement of task$_i$. Based on the value of the coefficients k and h the probability of choosing a resource for scheduling a task will change. The value of k and h should be carefully chosen so that it shares the load among the resources. A deep study was done by varying the values of k and h and it was identified that by choosing the values of $k = 1$ and $h = 0$ load balancing is achieved.

The formula can be generalized if more number of resources is available. The pheromone value of each resource is updated based on the percentage of task allotted. The higher percentage the resource allotted, the value of pheromone incremented more. The probability rule used for more number of resources is given in Eq. (7).

$$P_j = \frac{(R_i + k)^h}{\sum_{i=1}^{n} (R_i + k)^h} \tag{7}$$

The tasks scheduled here are considered to be Meta tasks. Meta tasks don't have dependency relationship among them. So they can be easily subdivided and shared on many resources for parallel execution. The algorithm for the proposed work is as follows:

Procedure LSACO

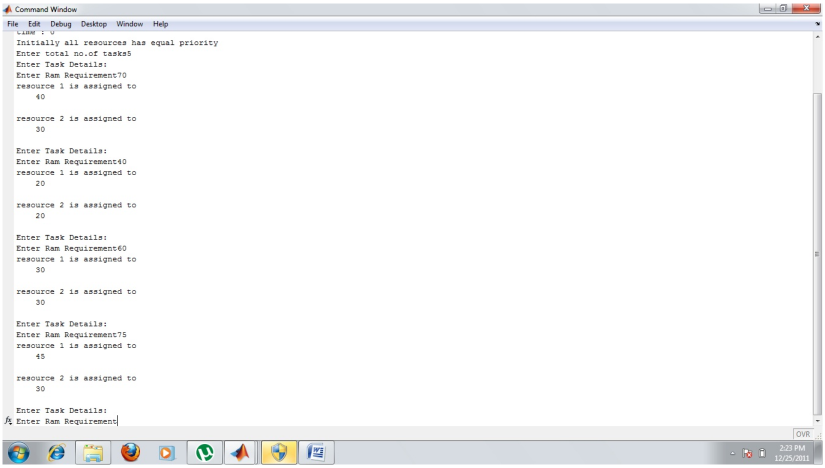
```
1. Initialize ACO Parameters
2. Read the Resource characteristics for all the
    available resources
3. For each task
4. Read task requirement $TR_i$
5. For each resource calculate the probability value
       $P_j = \frac{(R_i + k)^h}{\sum_{i=1}^{n} (R_i + k)^h}$
6. Calculate allotment percentage based on
   task requirement
   using $A_j$,  $A_j = P_j * TR_i$
7. Update pheromone value for each
    resource if it is allotted $R_i = R_i + 1$
8. End for
9. End for
```

5 Experiments and Results

The proposed work is implemented in MATLAB 7.0. The simulation was done with 2 resources and 5 tasks. The tasks were tested under two situations. First test case assumes that the tasks arrive at same time whereas in the second test case the tasks are assumed to arrive at different times. In both test cases it was determined that the proposed method minimizes wait time. The tests also show that the response time for LSACO is lesser than the traditional method in both test cases and it does not vary based on the job arrival time. The traditional methods will not schedule the task if the task requirement is more than the available memory size of the large resource because they choose any one resource and not all resources for executing a task. But LSACO will look for the sum of available memory size in all available resources and it will reject the task for scheduling only if the task requirement is more than the available memory size of the

Table 1. Resource Characteristics

Resource	Available RAM
R1	70
R2	30



```
time : 0
Initially all resources has equal priority
Enter total no.of tasks5
Enter Task Details:
Enter Ram Requirement70
resource 1 is assigned to
    40

resource 2 is assigned to
    30

Enter Task Details:
Enter Ram Requirement40
resource 1 is assigned to
    20

resource 2 is assigned to
    20

Enter Task Details:
Enter Ram Requirement60
resource 1 is assigned to
    30

resource 2 is assigned to
    30

Enter Task Details:
Enter Ram Requirement75
resource 1 is assigned to
    45

resource 2 is assigned to
    30

Enter Task Details:
Enter Ram Requirement
```

Fig. 3. Resource Allotment in LSACO

Table 2. Arrival of job at same time

Task	Resource Require-ment	Percentage of R1 Allotted in LSACO	Percentage of R2 Allotted in LSACO	Wait Time in LSACO	Wait Time in Traditional Method	Response Time in LSACO	Response Time in Traditional Method
1	70	40	30	0	0	40	70
2	40	20	20	40	70	60	110
3	60	30	30	60	110	90	170
4	75	45	30	90	170	135	245
5	80	50	30	135	245	185	325

sum of all resources. The resource characteristics are shown in Table 1. Fig. 3 shows the implementation window. The task requirements and the test results are shown in Table 2 for jobs arriving at same time. Table 3 shows the test results and arrival time of jobs for jobs arriving in different time. The graph showing the comparison results for wait time and response time produced by LSACO and traditional method are given in Figs. 4, 5, 6 and 7. Fig. 8 shows that LSACO reduces the overall wait time and response of jobs in both test cases compared to the scheduling in traditional method.

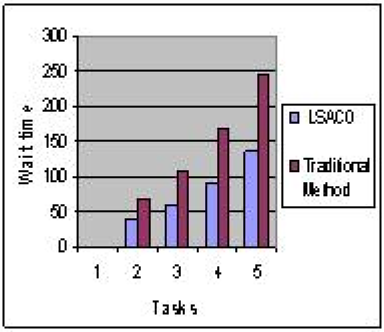


Fig. 4. Comparison of wait time when jobs arrive at same time

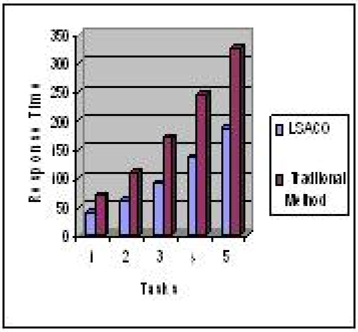


Fig. 5. Comparison of response time when jobs arrive at same time

Table 3. Arrival of job at different times

Task	Arrival Time	Percentage of R1 allotted in LSACO	Percentage of R2 allotted in LSACO	Wait Time in LSACO	Wait Time in Traditional Method	Response Time in LSACO	Response Time in Traditional Method
1	0	40	30	0	0	40	70
2	30	20	20	10	40	60	110
3	45	30	30	15	65	90	170
4	45	45	30	45	125	135	245
5	60	50	30	75	185	185	325

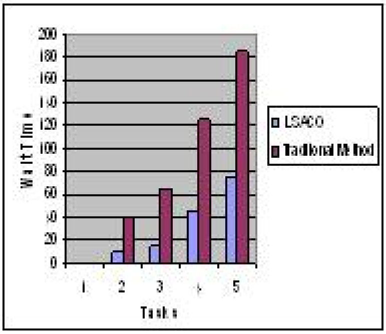


Fig. 6. Comparison of wait time when jobs arrive at different times

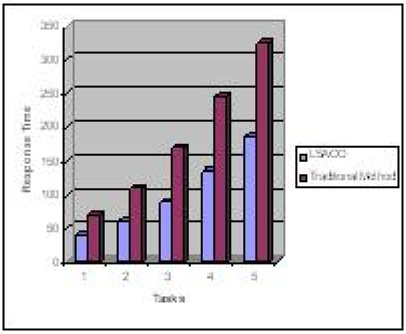


Fig. 7. Comparison of response time when jobs arrive at different times

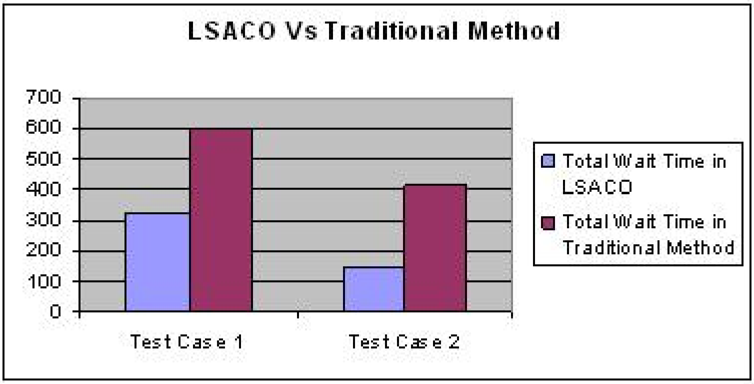


Fig. 8. Comparison of total wait time

6 Conclusion

The Ant Colony Optimization has proved to produce optimized results for many problems in science and engineering area. In this paper a new approach based on the argentine ant's behavior is proposed. The traditional methods like min–min, max–min tries to reduce the overall response time by giving an optimized schedule. They fail to produce a load balanced schedule. Moreover they do not share the load among the available resources. Thus the proposed method uses Ant Colony System approach to share the load among the resources for Meta Tasks. The proposed work uses one of the QoS parameter memory requirements to subdivide the task and schedule them in parallel among the available Grid resources. The experiments and results show that LSACO reduces both the wait time and response time of tasks. This method also shows that the resources are used effectively. That is the idle time of resources is minimized. In this paper only two test cases are considered. The work can be further extended by considering the heterogeneity of resources and tasks and the simulations can also consider the ETC matrix to get the perfect wait time and response time of the tasks. It can also be extended for dependent tasks using DAG approach.

References

1. Casanova, H., Obertelli, G., Berman, F., Wolski, R.: The AppLeS parameter sweep template: user-level middleware for the grid. In: Proceedings of the ACM/IEEE Conference on Supercomputing (2003)
2. Kokilavani, T., George Amalarethinam, D.I.: Applying Non-Traditional Optimization Techniques to Task Scheduling in Grid Computing. International Journal of Research and Reviews in Computer Science 1(4), 34–38 (2010)
3. Agarwal, A., Kumar, P.: Multidimensional Qos Oriented Task Scheduling In Grid Environments. International Journal of Grid Computing & Applications (IJGCA) 2(1), 28–37 (2011)

4. Braun, T.D., Siegel, H.J., Beck, N., Boloni, L.L., Maheswaran, M., Reuther, A.I., Robertson, J.P., et al.: A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems. Journal of Parallel and Distributed Computing 61(6), 810–837 (2001)
5. Chen, W.-N., Student Member, IEEE, Zhang, J., Senior Member, IEEE: An Ant Colony Optimization Approach to a Grid Workflow Scheduling Problem With Various QoS Requirements. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews 39(1), 29–43 (2009)
6. Chang, R.-S., Chang, J.-S., Lin, P.-S.: An ant algorithm for balanced job scheduling in grids. Future Generation Computer Systems 25, 20–27 (2009)
7. Saiz, P., Buncic, A., Peters, J.: AliEn Resource Brokers. In: Proceedings of the Third International Workshop on in High-Energy and Nuclear Physics, CHEP 2003 (2003)
8. Kertész, A., Kacsuk, P.: A Taxonomy of Grid Resource Brokers, pp. 201–210.
9. Kokilavani, T., George Amalarethinam, D.I.: Load Balanced min–min Algorithm for Static Meta-Task Scheduling in Grid Computing. International Journal of Computer Applications (0975–8887) 20(2) (April 2011)
10. Fidanova, S., Durchova, M.K.: Ant Algorithm for Grid Scheduling Problem. In: Lirkov, I., Margenov, S., Waśniewski, J. (eds.) LSSC 2005. LNCS, vol. 3743, pp. 405–412. Springer, Heidelberg (2006)
11. Ritchie, G., Levine, J.: A hybrid ant algorithm for scheduling independent jobs in heterogeneous computing environments. American Association for Artificial Intelligence (2004)
12. Xu, Z., Gu, J.: Research on Ant Algorithm Based Task Category Scheduling in Grid Computing. In: Second International Conference on Intelligent Networks and Intelligent Systems, pp. 498–501 (2009)
13. Dorigo, M., Gambardella, L.M.: Ant Colony system: A Cooperative Learning Approach to the Travelling Salesman Problem. IEEE Transactions on Evolutionary Computation 1(1), 1–24 (1997)

Optimum Sub-station Positioning Using Hierarchial Clustering

Shabbiruddin[1], Sandeep Chakravorty[2], and Amitava Ray[3]

[1] Department of Electrical & Electronics Engg. Sikkim Manipal Institute of Technology, Sikkim, India
shabbiruddin85@yahoo.com
[2] School of Electronics & Electrical Engg. Lovely Professional University, Punjab, India
sandeep_chakravorty@yahoo.com
[3] Department of Mechanical Engg. Sikkim Manipal Institute of Technology, Sikkim, India
amitavaray.siliguri@gmail.com

Abstract. Selection of optimum location of a sub-station and distribution of load points to each available sub-station has been a major concern among researchers but all have made either the use of man machine interface or have made some approximations. Here in this paper, a soft computing approach Hierarchial Clustering method is used for grouping the various load points as per the number of distribution sub-stations available. The method further gives an optimum location of the distribution sub-station taking into aspects the distances of the various load points that it is feeding. The results of the discussed technique will lead to a configuration of distribution substations depending on the no. of load points and sub-stations required. It will have an effect of lowering the long range distribution expenses as it will lead to optimum feeder path. The application of the proposed methodology to a case study is presented.

Keywords: Distribution Planning, Hierarchial Clustering Method (HCM), Dendrogram.

1 Introduction

The development, planning and control of the electrical networks have a significant importance in the electric power system, presently being of great interest. Choosing an optimum location of a distribution sub-station and grouping the various load points to be fed from a particular distribution sub-station has always been a concern to the distribution planners. In general, the decisions in the planning of power distribution system include:

- Optimal location of sub-stations
- Optimal allocation of load
- Optimal allocation of sub-station capacity

Normally, any engineering problem will have large number of solutions out of which some of are feasible and some are infeasible. The designer s task is to get the best

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 405–412.
springerlink.com

© Springer-Verlag Berlin Heidelberg 2013

solution out of the feasible solutions.The available literature consists of work of only few researchers on the field of distribution planning. Most of them are based on mathematical programming such as transportation, transshipment algorithms [4,5], mixed integer programming [6], dynamic programming [7] etc. Literature reveals that only near optimal solutions have been obtained by these mathematical programming methods because almost all methods have made some approximations on the model of distribution planning, moreover all of these methods are often complicated and time consuming. In the work done by K.K.Li and T.S. Chung [3] genetic algorithm have been used to find the optimum location of sub-station to meet the load demands of 13 load points whose coordinates and MVA demands are given. The work done by Belgin Turkay and Taylan Artac [1], and by J.F.Gomez *et.al.*,[2] has also been in the same area. In all the above cases planning of laying the feeders or distribution planning has been done either by man machine interface or heuristic algorithm.

The work and techniques developed for the solution of the planning problem for primary distribution circuits can be found in [8] and [9]. Initially all the proposed methods were mainly based upon the generation and evaluation of possible solutions, oriented to small size problems, and requiring important efforts for the production of the alternatives to be evaluated. Among these the heuristic zone valuation and the generation of service areas methods may be mentioned. They rely completely upon the experience of the planning engineer and have the disadvantage that the best alternative may not be considered.

Other techniques such as Heuristic search methods have been developed [10], [11], showing faster performance than the conventional optimization techniques but with some limitations in the goodness of the solutions to the problem that are obtained.

In [9] and [12] the potential of the GA's Is shown in comparison with classical optimization techniques to solve the planning problem in a very complete and detailed formulation considering the nonlinearity of the cost function, the limits of the voltage magnitudes and a term in the objective function to take into account the reliability of the system, reporting significant improvements in the solution times. An integer variable coding scheme was used to facilitate the consideration of different conductor sizes and sub-station sizes also new genetic operators were proposed to improve the performance of the algorithm. In [13] the approach is expanded to consider the multiple development stages as well as multiple objectives. In [14] an evolutionary approach is applied to the design of a medium voltage network using a detailed model of the network.

Clustering is the technique of grouping together rows that share similar values across a number of variables. It is a wonderful exploratory technique that helps in understanding the clumping structure of the data. Hierarchical clustering may be represented by a two dimensional diagram known as Dendrogram which illustrates the fusion or divisions made at each successive stage of analysis. Hierachical clustering is suitable for small tables, up to several hundred rows. One can choose the number of clusters after the tree is built. Several agglomerative techniques are single linkage clustering, complete linkage clustering, average linkage clustering, centroid method and Ward's hierarchical clustering method. Differences between methods arise because of the different ways of defining distance (or similarity) between clusters.

In the present work the first section discusses a 13 load point problem having different MVA values. The location of the load points are shown with a graphical representation. The next section presents the Hierarchial clustering algorithm. The next section shows the application of the algorithm with the load points. The load points are arranged into different number of clusters depending on the availability of the sub-station. Further the cluster centers obtained are used for the location of the sub-station. The limitation of the work is with the distribution of load points and the location of sub-station it could not be used for the feeder layout.

2 Proposed Methodology

Architecture of Hierarchial Clustering Method

The agglomerative Hierarchial clustering method is demonstrated below:

Step-1: Initialize i to the total number of objects, $i = m$
Step-2: Imagine the points in n dimensional plane, where n is the number of variables (Plotting will be easy only with two dimensional plane)
Step-3: Find the distance between each pair of points.
Step-4: Identify the two points (p,q) which are having the least distance between them
Step-5: Find the centroid of the two points (p,q). let them be c_i and the distance between them be d_i
Step-6: Identify the next two nearest clusters of points and group the together. Then, find the centroid of the newly formed cluster.
Step-7: set $i = i$-1
Step-8: If $i > 1$, then go to step–6 otherwise go to step-9
Step–9: Draw Dendrogram of the sequence of cluster formation.
Step–10: As per the clustering criterion, determine the number of clusters and the object of the clusters. An example of the clustering criterion may be the sudden jump in the distance between clusters while adding a cluster to another cluster. Draw a line in the Dendrogram where there is sudden jump in the distance between any two clusters and identify the final set of clusters accordingly.

3 Case Study

The work done by S. Chakravorty *et.al.*,[15] where a group of thirteen load points are to be fed from two sub-stations depending on the capacity and the load demands. The table below shows the data of the thirteen load points considered.

Table 1. The coordinates of the various load points with their respective load demands in MVA

Load points	X coordinates	Y coordinates	Load demands in MVA
1	8	7	5
2	10	7	12
3	11	8	7
4	6	9	5
5	1	1	7
6	3	1	11
7	5	2	8
8	7	2	3
9	1	3	4
10	5	4	12
11	2	5	6
12	3	7	3
13	9	5	4

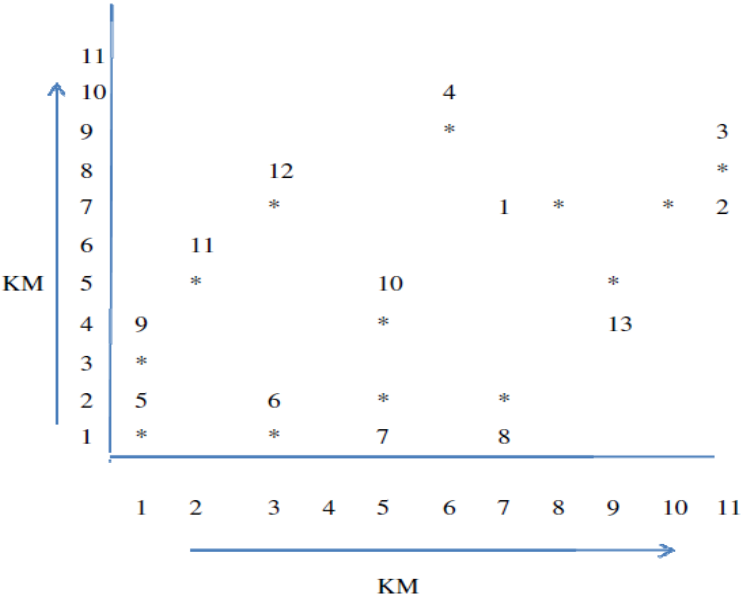


Fig. 1. Pictorial representation of the problem

4 Algorithm Application and Result Analysis

Initially total number of clusters equals to total number of points. C1,C2,C3, C4,C5,C6,C7,C8,C9,C10,C11,C12,C13. From fig 1 the distance is found out between each pair of points in order to identify the points having very less distance between them. It is to decide before how many clusters are to be formed after clustering, in this case it is decided to divide the load points into two clusters so that each cluster of load points are feeded by one sub-station. Considering the points having minimum distance and working according to the algorithm the following results are obtained.

Iteration no.	Nearest clusters	Centroid of nearest clusters	Distance between nearest clusters
1	C2 C3	(10.5,7.5)	1.414
2	C5 C9	(1,2)	2
3	C7C10	(5,3)	2
4	C11 C12	(2.5,6)	2.236
5	C1 C13	(8.5,6)	2.236
6	C5-9 C6	(2,1.5)	2.236
7	C7-10 C8	(6,2.5)	2.5
8	C1-13 C4	(7.25,7.5)	3.605
9	C5-9-6 C7-10-8	(4,2)	4.123
10	C1-13-4 C2-3	(8.875,7.5)	3.25
11	C5-9-6-7-10-8 C11-12	(3.25,4)	4.272

The 10th and 11th iteration gives the two clustered load points with center of the clusters which can be used for location of the sub-station. Thus from the results it is clear that load points (1,2,3,4,13) are in cluster 1 while load points(5,6,7,8,9,10,11,12) are in class Cluster 2.

The cluster center obtained for cluster 1 is (8.875, 7.5) and cluster center for cluster 2 is (3.25,4) these cluster center will be used for placing the sub-station.

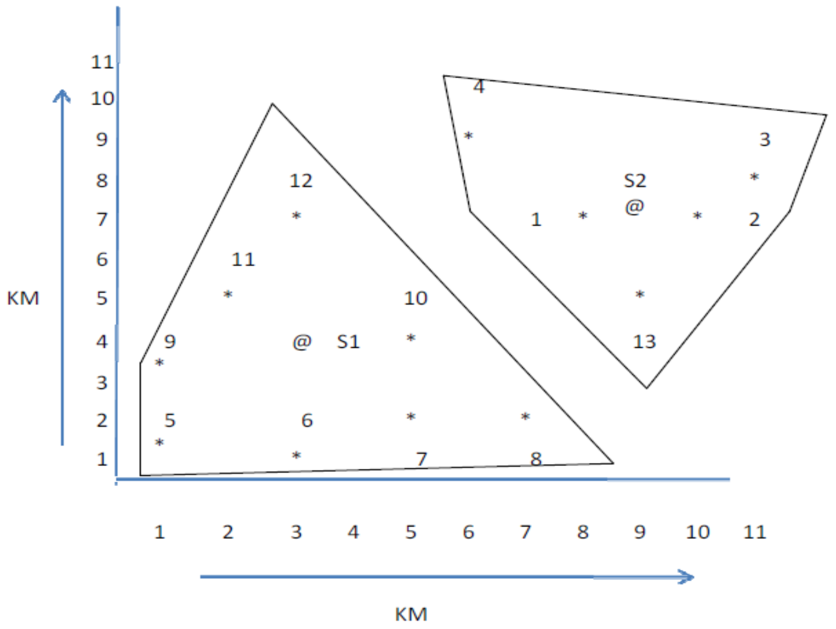


Fig. 2. Clustered load points with location of sub-station

5 Discussion and Conclusion

A new methodology, based upon the Hierarchial Clustering method (HCM) algorithm, is proposed for the planning of electrical power distribution system. Thus by applying Hierarchial Clustering method, various load points which are at different location can be grouped into number of clusters depending on the number of distribution sub-stations available. Also the location of the sub-station can be determined. The technique suggested is simpler than all the existing methods.

The work done by S. Chakravorty *et.al*, [15] in which the load distribution was proposed using the concept of genetic algorithm in which the capacity of the sub-station was pre assumed on the basis of which the distribution of the load points was carried out. The above mentioned drawback is removed in the present work, clustering of the load points are done irrespective of the capacity of the sub-station. One may decide on the capacity of the sub-station depending on the load points required to be fed from the sub-station. The technique is shown as a flexible and powerful tool for the distribution system planning engineers. The result encourages the use and further development of the methodology.

References

1. Turkay, B., Artac, T.: Optimal Distribution Network Design Using Genetic Algorithm. Electric Power Components and Systems 33, 513–524 (2005)
2. Gomez, J.F., et al.: Ant Colony System Algorithm for the Planning of Primary Distribution Circuits. IEEE Transactions on Power Systems 19(2) (May 2004)

3. Li, K.K., Chung, T.S.: Distribution Planning Using Rule Based Expert System Approach. In: IEEE International Conference on Electric Utility Deregulation and Power Technologies, DRPT 2004 (April 2004)
4. Crawford, D.M., Holt, S.B.: A Mathematical Optimization Technique For Locating Sizing Distribution Sub-stations, and Driving Their Optimal Service Areas. IEEE. Trans. on Power Apparatus and Systems PAS 94(2), 230–235 (1975)
5. El-Kady, M.A.: Computer Aided planning of Distribution Sub-station and Primary Feeders. IEEE. Trans. on Power Apparatus and Systems PAS 103(6), 1183–1189 (1984)
6. Gonen, T., Ramirez-Rosado, I.J.: Optimal Multi Stage Planning of Power Distribution Systems. IEEE Trans. on Power Delivery PWRD-2(2), 512–519 (1987)
7. Partanen, J.: A Modified Dynamic Programming Algorithm for Sizing, Locating and Timing of Feeder Reinforcements. IEEE Trans. on Power Delivery 5(1), 227–283 (1990)
8. Khator, S.K., Leung, L.C.: Power Distribution Planning: A review of models and issues. IEEE Trans. Power Syst. 12, 1151–1159 (1997)
9. Bernal-Agustin, J.L.: Aplicacion de Algoritmos Geneticos al Diseno Optimo de Sistemas de Distribucion de Energia Electrical. Ph.D. dieesrtation, University de Zaragoza, Espana (1998)
10. Boardman, J.T., Meekiff, C.C.: A branch and bound formulation of an electricity distribution planning problem. IEEE Trans. Power App. Syst. 104, 2112–2118 (1985)
11. Nara, K., et al.: Distribution system expansion planning b multi-stage branch exchange. IEEE Trans. Power Syst. 7, 208–214 (1992)
12. Carvalho, P.M.S., Ferreira, L.A.F.M.: Optimal distribution network expansion planning under uncertainty by evolutionary decision convergence. Int. J. Elect. Power Energy Syst. 20(2), 125–129 (1998)
13. Dorigo, M., Gambardella, L.M.: Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Trans. Evol. Comput. 1, 29–41 (1997)
14. Diaz Dorado, E., Cidras, J., Miguez, E.: Application of evolutionary algorithms for the planning of urban distribution networks of medium voltage. IEEE Trans. Power Syst. 17, 879–884 (2002)
15. Chakravorty, S., Ghosh, S.: An Improvised Method for Distribution of Loads and Configuration of Distribution Sub Station. International Journal of Engineering Research and Industrial Applications 2(II), 269–280 (2009)
16. Chakravorty, S., Ghosh, S.: Fuzzy Based Distribution Planning Technique. Journal of Electrical Engineering 9, 38–43 (2009)
17. Chakravorty, S., Ghosh, S.: Distribution Planning Based on Reliability and Contingency Criteria. International Journal of Computer and Electrical Engineering 1(2), 156–161 (2009)
18. Chakravorty, S., Ghosh, S.: A Novel Approach to Distribution Planning in an Unstructured Environment. International Journal of Computer and Electrical Engineering 1(3), 362–367 (2009)
19. Chakravorty, S., Ghosh, S.: A Hybrid Model of Distribution Planning. International Journal of Computer and Electrical Engineering 1(3), 368–374 (2009)
20. Chakravorty, S., Ghosh, S.: Power Distribution Planning Using Multi-Criteria Decision Making Method. International Journal of Computer and Electrical Engineering 1(5), 622–627 (2009)
21. Chakravorty, S., Thukral, M.: Optimal Allocation of Load Using Optimization Technique. In: Proceedings of International Conference CISSE, Bridgeport, USA, pp. 435–437 (2007)
22. Chakravorty, S., Thukral, M.: Choosing Distribution Sub Station Location Using Soft Computing Technique. In: Proceedings of International Conference on Advances in Computing, Communication and Control – 2009, Mumbai, India, pp. 53–55 (2009)

23. Shabbiruddin, Chakravorty, S.: Distribution of Loads and Setting of Distribution Sub Station Using Clustering Technique. In: Proceedings of International Conference on Advances in Computing, Communication and Control–2011, pp. 88–94. Springer, Heidelberg (2011)
24. Shabbiruddin, Chakravorty, S.: Load Distribution Among Distribution Substation and Feeder Routing Using Fuzzy Clustering and Context Aware Decision Algorithm. Journal of Electrical Engineering 11, 57–67 (2011)
25. Shabbiruddin, Kibria, G.: An Efficient Method for Speed Control of DC Shunt Motor using Response Surface Methodology (RSM) Approach. Journal of Control Engineering and Technology 1, 11–16 (2011)

Comparison of PCA, LDA and Gabor Features for Face Recognition Using Fuzzy Neural Network

Dhanya S. Pankaj[1] and M. Wilscy[2]

[1] Rajagiri School of Engineering and Technology
[2] Department of Computer Science, University of Kerala, Kariyavattom
`{dhanyaspankaj,wilsyphilipose}@gmail.com`

Abstract. A face recognition system identifies or verifies face images from a stored database of faces when a still image or a video is given as input. The recognition accuracy depends on the features used to represent the face images. In this paper a comparison of three popular features – PCA, LDA and Gabor features - used in literature to represent face images is given. The classifier used is a Fuzzy Neural Network classifier. The comparison was performed using AT&T, Yale and Indian databases. From the experimental results, the LDA features provide better Recognition Rates in the case of face images with less pose variations. Where more pose variations are involved, the Gabor features performed better than LDA features. For recognition tasks where recognition of trained individuals and rejection of untrained individuals are considered, the LDA features provide better results in terms of very low False Acceptance Rates and False Rejection Rates.

Keywords: Face Recognition, PCA, LDA, Gabor, Fuzzy Neural, IAFC.

1 Introduction

Face Recognition has been an active research area due to both its scientific challenges and wide range of potential applications such as biometric identity authentication, human-computer interaction, and video surveillance [1]. Face recognition is one of the most widely used biometric identification systems. Biometric systems verify or recognize the identity of a living person on the basis of some physiological characteristics, such as fingerprints or facial features, or some aspects of the person's behavior, like his/her handwriting style or keystroke patterns [2]. Face recognition is one of the few biometric methods that possess the merits of both high accuracy and low intrusiveness. It has the accuracy of a physiological approach without being intrusive [2]. Face Recognition can be typically used for three different tasks - Identification, Verification and Watch-list. In identification tasks, the system identifies a person from a database of known faces whereas in verification tasks, the system verifies the claimed identity of a person. In watch-list tasks, the system finds whether the individual is present in the watch-list or not.

Various appearance based and feature based approaches have been proposed in literature for Face Recognition [3]. Principal Component Analysis (PCA) [4] and

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 413–422.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Linear Discriminant Analysis (LDA) [5] are the most commonly used appearance based methods for face recognition. Various Gabor wavelet based features are being successfully used recently for face recognition [6]. A class of face recognition algorithms employs feature extraction methods like PCA, LDA, Gabor wavelets [6] etc and various classifiers like probabilistic [7], hidden Markov models [8], Neural Networks [9], Support Vector Machines [10] etc. The performances of the various algorithms vary depending on the features used for representing the face images and the classifiers used for classifying them.

Most of the studies in the literature compare the appearance based features like PCA, LDA, Independent Component Analysis (ICA) etc for face recognition [12-14]. Also the features are compared using metrics like Euclidean distance, City block distance, Cosine angle, Mahalanobis distance etc. This paper presents a comparison between the most commonly used features – PCA, LDA and Gabor features– for face recognition. Gabor wavelet features and Neural Networks are being widely employed recently for face recognition tasks. This study is motivated by the lack of comparison in literature between the three given features and using Neural Networks. This comparison is performed using an efficient Neural Network based classifier for face recognition. A Fuzzy Neural Network based on Integrated Adaptive Fuzzy Clustering (IAFC) [11] is employed as the classifier. The network uses a combination of similarity measures to form the class boundaries and hence can distinguish the classes effectively. The comparative study was performed as part of a study of the IAFC algorithm for face recognition [15]. From the results obtained, LDA features give better performance in identifying trained persons and rejecting untrained persons which implies that LDA features are suitable for the identification tasks. The Gabor features perform better than LDA features in recognizing the trained individuals in the case of Indian face database which has more pose variations compared to the AT & T (ORL) and the Yale face databases. When less pose variations are involved, LDA features provided high recognition rates. This indicates that the Gabor features are more suitable for verification tasks with much pose variations. The LDA features perform better in the case of verification tasks with less pose variations.

The rest of this paper is organized as follows: A brief description of the features used in the face recognition system is explained in section 2. Section 3 explains the Fuzzy Neural Network based face recognition system. Section 4 reports the experimental results and comparison of the features and Section 5 concludes the paper.

2 Feature Extraction Algorithms

This section briefly explains the three different features used for representing the face images.

2.1 Principal Component Analysis(PCA)

From the m-dimensional vector representation of the training images, PCA finds a t-dimensional (t<<s) sub space whose basis vectors correspond to the maximum variance direction in the original image space. All training face images are projected

onto the new subspace called face space to find a set of weights that describes the contribution of each vector. The weights form the feature vector for the face image. The process is explained below.

Let the database of n training images be represented by n vectors Z = (Z1, Z2, …, Zn) of size m each. The mean vector $\overline{Z}$ is calculated as in equation 1.

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \tag{1}$$

The covariance matrix is defined as in equation 2.

$$\Gamma = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)\left(Z_i - \overline{Z}\right)^T = \phi\phi^T \tag{2}$$

The eigen values and eigen vectors of the covariance matrix Γ are calculated. Let E = (E1, E2, .., Et) be the t eigen vectors corresponding to the t largest eigen values. For the n patterns Z, their corresponding their corresponding PCA features X can be obtained by projecting Z onto the eigen space as in equation 3.

$$X = E^T Z \tag{3}$$

Thus the patterns of dimension m (m=p x q) is reduced to the dimension t. (t << m, t is taken to be (No of classes * No of training patterns per class) - No of classes.). PCA maximizes the intra class as well as the inter class scatter while performing the dimensionality reduction.

2.2 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis finds the vectors in the underlying space that best discriminate among classes. For all samples of all classes the between-class scatter matrix S_b and the within-class scatter matrix S_w are defined by equation 4-5.

$$S_w = \sum_{j=1}^{R}\sum_{i=1}^{M_j}(x_i^j - \mu_j)(x_i^j - \mu_j)^T \tag{4}$$

$$S_b = \sum_{j=1}^{R}(\mu_j - \mu)(\mu_j - \mu)^T \tag{5}$$

where j denotes the class while i denotes the image number. μ_j is the mean of class j while μ is the mean of all classes. M_j is the number of images in class j and R is the number of classes. LDA computes a transformation that maximizes the between-class scatter while minimizing the within-class scatter, in other words maximizing the ratio Sb/Sw . This ratio is maximized when the column vectors of the projection matrix (W_{LDA}) are the eigenvectors of S_w^{-1} S_b. However, in practice, S_w is often singular since the data are image vectors with large dimensionality while the size of the data set is much smaller. To alleviate this problem, PCA is first applied to the data set to reduce its dimensionality. LDA is then applied to form the feature vectors. The final transformation is $W^T = W_{LDA}^T W_{PCA}^T$.

2.3 Gabor Features

The Gabor wavelet representation of an image involves convolution of the image with a family of Gabor kernels at different spatial frequencies and different orientations. To encompass the different spatial frequencies, spatial localities, and orientation selectivities, the resulting convolution representations are normally concatenated as an augmented feature vector. A face image is typically represented as the convolution result of the face image with 40 Gabor wavelets (five scales, each with eight orientations). Keeping only the magnitude values in the representation, this gives a "h × w × 40" vector, where h × w is the length of the face vector. The "h × w × 40" vector usually has a very large dimensionality. To reduce the dimensionality of this vector, uniform sampling of the original Gabor features is used.

The 2-D Gabor wavelets (kernels and filters) can be defined as in equation 6:

$$\psi_j\left(\vec{x}\right) = \frac{\left\|\vec{k_j}\right\|^2}{\sigma^2} \exp\left(-\frac{\left\|\vec{k_j}\right\|^2 \left\|\vec{x}\right\|^2}{2\sigma^2}\right) \times \left[\exp\left(i\vec{k_j}\cdot\vec{x}\right) - \exp\left(-\frac{\sigma^2}{2}\right)\right], \quad j = 1 \to L \tag{6}$$

$$\vec{k_j} = k_m e^{i\phi_n}, \quad k_m = \frac{0.5\pi}{\left(\sqrt{2}\right)^m}, \quad \varphi_n = n\frac{\pi}{8}, \quad L = m \times n \tag{7}$$

where m and n define the scale and orientation of the Gabor wavelets, ||.|| denotes the norm operation, and x represents the pixel position.

The Gabor wavelet representation of an image is the convolution of the image with a family of Gabor wavelets. Let $I(x)$ be the $h \times w$ gray-level image, $\psi_j(x)$ be the jth Gabor wavelet, and $O_j(x)$ be the jth Gabor feature image of image $I(x)$ corresponding to the jth Gabor wavelet $\psi_j(x)$ as in equation 8.

$$O_j(x) = I(x) * \psi_j(x) \quad j = 1 \rightarrow L \tag{8}$$

where $(x) = (x, y)$, $*$ denotes the convolution operator, and $L = m \times n$ is the number of Gabor wavelets. Here, $L = 40$.

3 Face Recognition System

The Face Recognition System used for comparison involves a feature extractor and the Fuzzy Neural Network classifier. The various features used for representing the face images are discussed in section II.

The Fuzzy Neural Network based on Integrated Adaptive Fuzzy Clustering (IAFC) [11] is similar to Adaptive Resonance Theory-1 architecture and finds the cluster structure embedded in the data sets. The IAFC forms well defined class boundaries in the case of a face recognition system where the similar individuals may form closely located classes. The system consists of two phases – training and testing. A subset of images of each person is used to train the system. The training patterns of each person in the database are given as input to the classifier and the Fuzzy Neural Network classifies them into classes where each class corresponds to an individual. The weights and other parameters of the network (centroids) are updated during this training process. The clustering is an iterative process and the number of iterations is set as 10 empirically. Once the training is over, the test pattern is given as input to the Fuzzy Neural Network. The weight and the centroid updation process are not carried out during the testing process and the number of iteration is set as 1. The neural network classifies the input vector into one of the existing classes if the vigilance criteria are met or rejects it, otherwise.

The training algorithm consists of three major procedures: deciding a winning cluster, performing the vigilance test, and updating the centroids of a winning cluster. In the training phase, the network parameters are initialized. An initial direction for the convergence of the weights and cluster centroids is provided by initializing these values using the training face patterns. The training input pattern and the normalized input pattern are presented to the fuzzy neural network in parallel. The dot product between the normalized input pattern and the bottom up weights, which is same as the normalized cluster centroids, is performed and the output neuron that receives the maximum value for the dot product is chosen as the winning cluster. Since only the angle between the input pattern and the cluster centroids is considered here regardless of its location, IAFC performs a test to find the relative proximity of the pattern to the winning cluster by calculating the fuzzy membership value μ_i of the pattern in the clusters. If the value of μ_i in the winning cluster is less than a threshold σ, the algorithm finds the winning cluster by finding the Euclidean Distance between the input pattern and the top-down weights, which is same as the cluster centroids. Once the winning cluster is selected, the IAFC algorithm performs a vigilance test as in the equation 9.

$$e^{-\gamma\mu_i} \left\| x - v_i \right\| <= \tau \tag{9}$$

where γ is a factor that controls the shape of clusters, x is the input pattern and v_i is the centroid of the ith winning cluster and τ is the vigilance parameter. If the vigilance test is satisfied, the winning cluster centroid is updated [11] else the test is performed for other clusters until any one of it satisfies the test. Otherwise the input pattern is rejected as not belonging to any of the existing classes by forming a new class.

During the testing phase, the test pattern is given as input to the network. The cluster centroids or weights are not updated during testing. The pattern is classified to any of the existing classes or rejected based on the vigilance criterion. If the vigilance test is satisfied for any of the existing classes, the input pattern is classified into one of the classes in the database. Otherwise, the pattern is rejected which indicates that the input pattern is not present among the training classes.

The input and the test patterns presented to the network are the face feature vectors extracted using any of the 3 feature extractors – PCA, LDA or Gabor wavelets.

4 Experimental Results and Discussion

Experiments are carried out using the publicly available AT&T (ORL) [16], Yale [17] and Indian Face [18] databases. The AT&T database contains ten different images each of 40 distinct individuals, varying in lighting, facial expressions and facial details (glasses/no glasses). The Yale face database contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration. The Indian Face Database contains 10 different images of each of 60 distinct subjects. The images vary in orientation (pose) and facial expression.

The Face Recognition System using Fuzzy Neural Network based on IAFC is trained using the training patterns of persons from the databases. From the databases, a set of individuals are selected as registrants and others are selected as non-registrants. A subset of the images of the registrants is used for training the network. The cluster centroids and the bottom-up and top-down weights are initialized using the average of the training patterns of each individual in the training set and then the training is performed. Once the system is trained, i.e. after the defined number of iterations, the weights and the cluster centroids are saved. During testing, a test pattern is presented to the network. The system performs the vigilance test using the combined similarity measure and fuzzified learning rule [11] and groups the test pattern to any of the existing classes if the vigilance test is satisfied. Otherwise the pattern is rejected as non-registrant.

Two sets of evaluations are performed with and without the rejection criterion. In the first case, the vigilance threshold is set to high such that the system obtains maximum recognition and no rejection is considered. The first set of evaluation is performed on the set of known individuals (registrants) from the database. i.e. on the persons whose images are used for training the network. The performance of the system was measured by calculating the Recognition Rate. The Recognition Rate is the ratio of the number of test images correctly matched to the total number of test images, multiplied by 100.

The results obtained for the first evaluation for the three databases is shown in Fig.1. The results obtained using the three different features are compared.

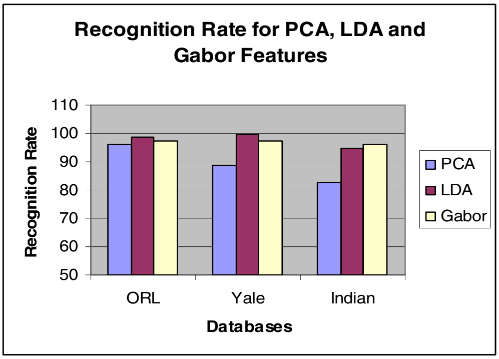


Fig. 1. Comparison of recognition rates using PCA, LDA and Gabor for three databases

From the results obtained, the three features provided comparable recognition rates in the case of AT&T (ORL) database where image variations are less. In the case of Yale face database which has more background illumination variations, the LDA features provided better recognition rates. The Gabor features provided the maximum recognition rate for the Indian face database which involves more pose variations compared to the other two databases. This indicates that the LDA features are suitable for face recognition tasks where images do not have much pose variation. In the case of face recognition tasks where more pose variations are involved, Gabor features perform better compared to the LDA features in identifying an individual. The face recognition tasks where recognition of individuals is of primary concern and rejection of unknown or untrained individuals is of less concern may employ the LDA or Gabor features depending upon the amount of pose variations. The Face Recognition System using the Gabor wavelet features are computationally expensive compared to the systems using the other two features and the size of the feature vector is large in the case of Gabor wavelets.

In the second set of evaluation, the vigilance threshold is adjusted to include the rejection criteria. The registrants are used for training the network. The entire set of known persons (registrants) and unknown persons (non-registrants) are considered for testing. The registrants should be identified by the network and the non-registrants should be rejected by the network. The performance of the system was measured by calculating the False Acceptance Rate (FAR) and False Rejection Rate (FRR). False Acceptance Rate is the probability that the system incorrectly matches the input image to a non-matching individual in the database. False Rejection Rate is the probability that the system fails to detect an actual match between the input image and a matching individual in the database. A practical face recognition system performs well when it has low values for FAR and FRR. When the vigilance threshold set for matching is high, the system achieves low FAR but high FRR and vice versa. A point where the FAR and the FRR curves meet is defined as the Equal Error Rate (EER) and low values for EER indicates a good face recognition system. The FAR and FRR curves are plotted and the Equal Error Rate is calculated. The vigilance threshold is adjusted so as to obtain the minimum value for the ERR.

The FAR and FRR curves and the ERR value obtained for the second evaluation for the Indian Face database is shown in Fig.2 and 3. The FAR and FRR curves are plotted similarly for the AT&T and Yale face databases also and the ERR values obtained are summarized in Table 1.

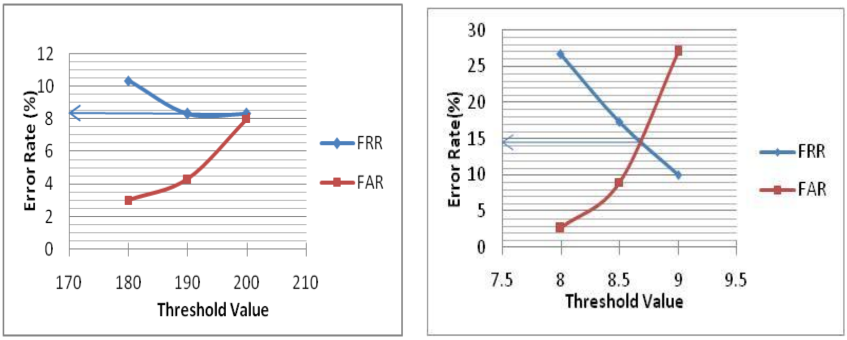


Fig. 2. Calculation of ERR from the FAR and FRR curves for the Indian Face Database for the LDA features and Gabor features

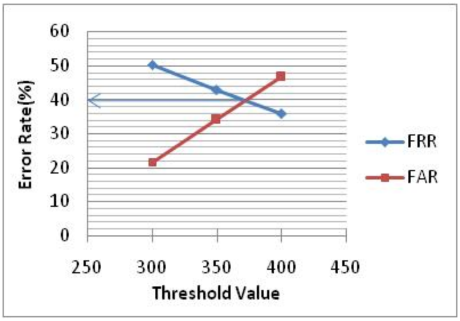


Fig. 3. Calculation of ERR from the FAR and FRR curves for the Indian Face Database for the PCA features

From the ERR values obtained, the LDA features provided the minimum ERR values for the all the three databases. This indicates that the LDA features provided better False Acceptance Rates and False Rejection Rates compared to the other features. The PCA features are considered optimum in representing a face image and hence provided high Recognition Rates as observed in Fig.1. However, the PCA features reported high ERR as indicated in Table 1, which shows that PCA is not as efficient as LDA or Gabor features in discriminating the individuals. The False Acceptance and False Rejection Rates increases in the case of PCA features where more pose variations and background variations are present in the images as indicated by the results obtained for the Yale and Indian database images. The Gabor features which provided high recognition rates in Fig.1 are not as good as LDA features in discriminating individuals as indicated by the high Equal Error Rates in Table 1. This indicates that the LDA features are more suitable for face recognition applications where very low values of both False Acceptance Rates and False Rejection Rates are required. i.e. the LDA features are more suitable for face recognition tasks where rejection and recognition are both equally important.

Table 1. Comparison of Equal Error Rates using PCA, LDA and Gabor features

Database	Feature	ERR(%)
AT&T	PCA	25
	LDA	**3.7**
	Gabor	6.4
Yale	PCA	25
	LDA	**1.3**
	Gabor	17
Indian	PCA	40
	LDA	**8.4**
	Gabor	14.5

From results obtained for the face recognition system using IAFC, the PCA, LDA and Gabor features obtained comparable recognition rates for the trained individuals. The Gabor features exhibited high recognition rates for the Indian face database and LDA features exhibited high recognition rates for the Yale and AT&T face databases. Hence the LDA and Gabor features perform better than PCA for face recognition applications where identification of trained individuals is of primary concern. However in face recognition applications where rejecting the unknown individuals reliably while correctly identifying the known individuals is required, LDA features are more suitable. The results indicate that the LDA features are better than PCA and the Gabor features in discriminating the registrants and the non-registrants. This is also supported by the fact that the LDA extracts the most discriminative features from the training classes compared to the optimal feature representations by PCA or Gabor wavelets. The discriminating features are more important in the case of a face recognition application where low ERR rate is of primary concern, compared to the representative features.

5 Conclusion

A comparative study of three most popular face features – PCA, LDA and Gabor - for face recognition is presented in this paper. The comparison is performed based on the performance evaluation of a face recognition system using these features. The classifier used for recognition is a Fuzzy Neural Network classifier based on Integrated Adaptive Fuzzy Clustering Model. The experiments are conducted using three different face databases – AT&T, Yale and Indian face databases. The LDA features performed better in experiments where both recognition of trained individuals and rejection of untrained individuals are considered. LDA features provided the minimum error rates compared to Gabor and PCA features. However in experiments where rejection is not of primary concern, Gabor features performed better when more pose variations are involved. If rejection is not considered and not much pose variations are involved, LDA features provided good recognition rates. However, LDA feature extraction process requires a number of training patterns for forming the within class and between class scatter matrices and hence is not suitable for the cases where the number of training samples is less. The LDA process often requires PCA as a preprocessing step to deal with singularity problem. The Gabor features faces the problem of high dimensionality of the feature vector and hence requires more computational time compared to PCA and LDA features.

References

1. Su, Y., Shan, S., Chen, X., Ga, W.: Hierarchical Ensemble of Global and Local Classifiers for Face Recognition. IEEE Transactions on Image Processing 18 (2009)
2. Lin, S.-H.: An Introduction to Face Recognition Technology. Informing Science Special Issue on Multimedia Informing Technologies – Part 2 3(1) (2000)
3. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.J.: Face Recognition: A Literature Survey. ACM Computing Surveys, 399-458 (2003)
4. Lu, J., Yuan, X., Yahagi, T.: A Method of Face Recognition Based on Fuzzy c-Means Clustering and Associated Sub-NNs. IEEE Transactions on Neural Networks 18(1) (2007)
5. Zhao, W., Chellappa, R., Krishnaswamy, A.: Discriminant analysis of Principal Component For Face Recognition. IEEE Trans. Pattern Anal. Machine Intel. 8 (1997)
6. Du., S., Ward, R.K.: Improved Face Representation by Nonuniform Multilevel Selection of Gabor Convolution Features. IEEE Transactions on Systems, Man, and Cybernetics—Part b: Cybernetics 39(6) (2009)
7. Moghaddam, B.: Principal manifolds And Probabilistic Subspaces For Visual Recognition. IEEE Trans. Pattern Anal. Machine Intel. 24(6), 780–788 (2002)
8. Othman, H., Aboulnasr, T.: A Separable Low Complexity 2D HMM with Application to Face Recognition. IEEE Trans. Pattern. Anal. Machine Intel. 25(10), 1229–1238 (2003)
9. Er, M., Wu, S., Lu, J., Toh, L.H.: Face recognition with Radial Basis Function (RBF) Neural Networks. IEEE Trans. Neural Networks 13(3), 697–710 (1999)
10. Lee, K., Chung, Y., Byun, H.: SVM Based Face Verification With Feature Set of Small Size. Electronic Letters 38(15), 787–789 (2002)
11. Kim, Y.S., Mitra, S.: An Adaptive Integrated Fuzzy Clustering Model for Pattern Recognition. Journal Fuzzy Sets and Systems (65), 297–310 (1994)
12. Delac, K., Grgic, M., Grgic, S.: Independent Comparative Study of PCA, ICA, And LDA on the Feret Data Set. International Journal of Imaging Systems and Technology 15(5), 252–260 (2005)
13. Cho, H., Moon, S.: Comparison of PCA and LDA Based Face Recognition Algorithms Under Illumination Variations. In: ICCAS-SICE, pp. 4025–4030 (2009)
14. Luo, B., Hao, Y.-J., Zhang, W.-H., Liu, Z.-S.: Comparison of PCA and ICA in Face Recognition. In: International Conference on Apperceiving Computing and Intelligence Analysis, pp. 241–243 (2008)
15. Pankaj, D.S., Wilscy, M.: Face Recognition Using Fuzzy Neural Network Classifier. Advances in Parallel Distributed Computing, 53–62 (2011)
16. AT&T. The Database of Faces, `http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html`
17. Yale Face Database, `http://cvc.yale.edu/projects/yalefaces/yalefaces.html`
18. Jain, V., Mukherjee, A.: The Indian Face Database, `http://vis-www.cs.umass.edu/~vidit/IndianFaceDatabase`

Investigations on Object Constraints in Unified Software Development Process

Meena Sharma[1] and Rajeev G. Vishwakarma[2]

[1] Institute of Engineering & Technology, Devi Ahilya University, Indore, India
meena@myself.com
[2] SVITS, Rajiv Gandhi Technical University of MP, Indore, India
rajeev@mail.com

Abstract. The Object constraints can be described as the expressions that are used to insert important data in object oriented models. The Object Management Group founded a worldwide standard for object-oriented analysis and design artifacts specifically the diagrams. The specification and standard, known as the Unified Modeling Language, comprises model diagrams and the allied and associated semantics. Unified Modeling Language is meant for modeling and, the Object Constraint Language is the specification and standard for specifying expressions. These expressions add essential, crucial and critical information to object-oriented models and other object modeling workproducts. In Unified Software Development Process we have analysis & Design discipline to have complete architecture and design of the system. Analysis & Design discipline is followed by implementation discipline. The activities of the implementation phase are mainly captured in construction phase. In unified software development life cycle the expressions in the design model are forward engineered to produce or generate the source code. The source code that is generated depends a lot on the platform and the technology that are selected for the software development. We have investigated how the expressions are developed and incorporated in the models in an electronic business solution. Further object constraint language is used to form the expressions that are attached to the object oriented model and artifacts. Then models are forward engineered to generate the code. In a nutshell we produce code from abstract, models or diagrams making use of object constraints language that is to achieve round trip engineering format between model and code. In this paper we have generated expressions from object constraint perspective for a reward point system as applied to a customer in e-business. We have also evolved the Context Definition, Initial Values & Derivation Rules, Query Operations, Attributes and Operations in this regard.

Keywords: UML, OCL, Model, Object Oriented Analysis & Design, Unified Software development Process, Context Definition, Initial Values & Derivation Rules, Query Operations, Attributes and Operations.

1 Introduction

This part deals with the essential elements that we need to write object constraints. This should be noted down that we can mark the object constraints with the help of

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 423–432.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

the unified modeling language. We consequently herby explain Expressions, Types, and Values. In Object Constraint Language (OCL), every value, either it is an object, or a component instance, or a datavalue, has a particular type, which defines the process or function that can be applied to the object. Types in OCL are divided into two groups- Predefined types, as defined in the standard library, (including the: Basic types and Collection types) and User-defined types.

The predefined basic types are Integer, Real, String, and Boolean. Their descriptions are similar to those in many acknowledged languages. The preclassified collection types are Collection, Set, Bag, OrderedSet, and Sequence. They are utilized to state the accurate results of a navigation through associations in a class diagram. We need to be familiar with these types to write more complex expressions. User-defined types, such as Customer or RewardScheme, are defined by the user in the UML diagrams. Every model element that can be instantiated (for example each class, interface, component, or datatype) - in a UML diagram becomes a type in OCL also automatically. Each OCL expression either a user-defined type or a predefined OCL type symbolizes a value; consequently, it has a type as well. Each OCL expression has a result: the output that results from evaluating the expression. The type of the result value is equivalent to the type of the expression. OCL marks the difference between value types and object types. Both are types, and both denote instances, but there is one significant differentiation: value types define instances that never modify. For example, integer with the value 10 will never change its value that is it will not become an integer with a value of 100. Object types, or Classifiers, stand for types that define instances. The value of the instances can be changed over time and states. An instance of the class Customer can amend the value of its attribute (for example name) and even than it may remain the same instance. For instance, Martin Fowler [Fowler, 97] calls object types as reference objects and value types as value objects.

One more main quality of value types involves identity. For value types, the value identifies the instance, consequently the name. Two incidences of a value type that have the same value are by definition and other by the same instance. Two incidents of an object type are the same instance if they have the same object identity. In other words, value types take account of value-based identity; object types encompass reference-based identity. Value types include the predefined basic types and the predefined collection types of OCL as well. The user defined types may be classified by either value types or may be object types. UML datatypes are value types that include enumeration types. Object types include UML classes, components, and interfaces are object types.

2 OCL

OCL is a kind of modeling language that facilitates construct software models. It is identified as a customary add-on to the UML. UML is the Object Management Group (OMG) standard for object-oriented analysis and design. The expressions developed in OCL rely on the types like the classes, interfaces, and so on. These types are defined in the UML models and diagrams. The use of OCL therefore requires that we should have input of some features of UML.

Expressions written in OCL attach essential information to object-oriented models and other artifacts pertaining to object modeling. This information often cannot be put across in a diagram. Earlier in UML, this expression writing was limited to constraints, where a constraint is defined as a control or restriction on one or more values or their part of an object-oriented model or system. Later on in UML 2, the perception and view is that more and more additional information should be included in a model along with the constraints alone. With the help of UML and OCL we can associate constraints and information in a model for defining queries, referencing values, or stating conditions and business rules in a model. In order to write these expressions in a clear and unambiguous manner we make use of OCL.

The OCL version 2.0, has been formally classified in the Object Constraint Language Specification [OCL03] and it has been very well accepted by the OMG. The, models should be good quality, firm, dependable, and logical. Using the mixture of UML and OCL, we can build such models.

3 Unified Software Development Process

We have researched and produced an object constraint code framework for an e-business application in unified paradigm. Object constraints are employed as the formalization of the specifications in the model. The Unified Process (UP) is not just a process, but rather an expandable framework which may be tailored and customized for specific organizations or projects. Now we describe the important characteristics of UP. It is iterative and incremental. The Unified Process is an iterative and incremental development process. The Elaboration, Construction and Transition phases are divided into several time boxed iterations. The Inception phase may also be partitioned in to more than one, iterations for a large project. Each iteration results in an increment or executable, which is a release of the system that has additional or improved and enhanced functionality compared to the earlier release. Even though most iterations will comprise work in most of the development process disciplines (Business Modeling, Requirements, Analysis & Design, Implementation, Testing and others) the relative work, tasks and importance will change over the life of the project. UP is use case driven. In the Unified Process, use cases are the key elements to capture the functional requirements and to characterize the contents of the iterations. Each of the iterations needs a set of use cases or use case scenarios from requirements discipline to the implementation discipline, test and deployment.

UP is architecture centric. The Unified Process persists that architecture resides at the heart of the software development life cycle and project team's efforts to accomplish the system. Since no single model is adequate to wrap up all aspects of a system, the Unified Process allows multiple architectural models and views. One of the most significant deliverables of the unified process is the executable architecture baseline. This architectural baseline is developed during the Elaboration phase. This partial implementation of the system is necessary to validate the architecture and serve as a basis for rest of the development. UP is risk focused. The Unified Process necessitates the project development team to focus on dealing with the majority of critical risks at an early stage in the project life cycle. The deliverables of each iteration, particularly in the Elaboration phase, have to be chosen in order to make certain that the maximum risks are addressed and attended first. Now we will focus on the fact why we should necessitate the object constraint.

4 Context Definition

The context definition identifies and states the model entity for which the OCL expression is defined. Typically this is a class, interface, datatype, or component. Sometimes it may be an operation, and rarely is it an instance. It is always a particular element defined in a UML model or diagram. This element is termed the context of the expression. OCL expressions can be integrated and incorporated simply in the model straight forward in the diagrams. OCL expressions can also be made available in a separate text file. In both the cases it contains a context definition. In the UML diagram, a dotted line is used to symbolize the context definition that connects the model element and the OCL expression. In Figure 1, we show the expressions and their contexts.

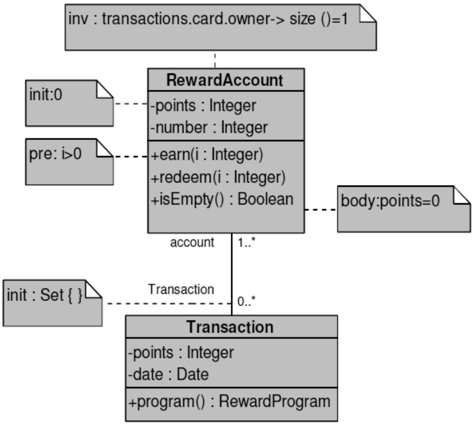


Fig. 1. Expressions & their contexts

5 E-Business – Case Study Investigations

As a real world case study of an ebusiness, we have modeled a computer software system for a imaginary company called Award and Reward (A&R). A&R manages rewardreward point programs for business organizations that offer their customers various kinds of shopping and service rewardes. Often, the extras take the form of reward points or air miles, petrol, gas and many others. But other rewards on the basis of rewards points collected are possible as well. For example: discounted rates, a luxury sedan rental for the equal price as a standard rental car, additional or ad-on service on an airline, and so on. Anything a company is willing to offer can be a service rendered in a reward program. Figure 1 shows the UML class model that A&R employs for most of its customers The model shown here is a classic platform independent model from the perspective and context of software architecture. It does not demonstrate any dependency on the programming language to be applied to make the system.

The key and fundamental class in this model is RewardProgram. A system that administers one reward program will hold only one instance of this class. In the case of A&R, many instances of this class will be present in the system. A company that offers its customers membership in a reward program is called a ProgramaAssociate. Several companies can join the same program. In this case, customers who join the reward program can make good profit from services offered by any of the member companies.

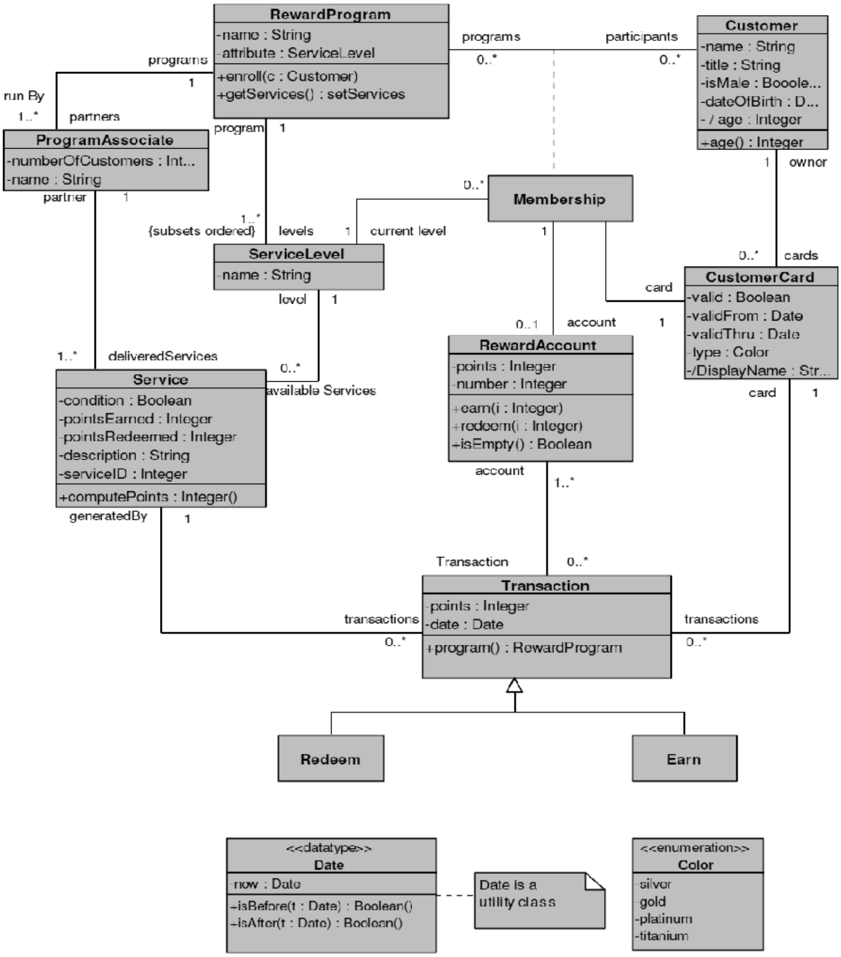


Fig. 2. Class Model for an eBusiness

Every customer of a program associate can join the reward program by submitting a form and receiving a membership card. The objects of class *Customer* represent the persons who have joined the program. *CustomerCard* class represents membership card. Where, the card is issued to one person. Card utilization is not checked, so one card could be used for all the members of the family or business. Most reward programs allow customers to save reward points. Each individual program associate chooses when and how many reward points are allocated for a particular purchase. Accumulated reward points can be used to "buy" specific services from one of the program associate. To account for the reward points that are gathered by a customer, every membership can be associated with a *RewardAccount*. A variety of transactions on this account are potential. For example, in the RewardProgram, Silver and Gold CustomerCard has four program partners: a food chain, a line of petrol pumps, and an airline service. At the food chain outlet, the customer can use reward points to pay the food bill. The customer earns five reward points for any regular purchase over the amount of Rs. 100. The petrol pump stations offer a discount of 5 percent on each petrol filling. Customers can save reward points for free flights with the airline company. For each flight that is usually paid by the customer, the airline offers two reward point for each 20 miles of flight.

In these circumstances, there are two kinds of transactions. First, there are transactions in which the customer accumulates reward points. In the model (Figure 2, this type of Transaction is modeled by a subclass of the class Transaction called Earning). Second, there are transactions in which the customer uses reward points for purchasing. In the diagram of class model, they are represented by instances of the subclass Redeem of Transaction class. The earnings generated by the customers need to be recorded. This is recorded as two simultaneous transactions on the RewardAccount, one Earning and one Redeeming for the same number of points.

6 Initial Values and Derivation Rules

A very fundamental addition to the diagram shown in Figure 2 is to take rules in account that state initial values for attributes and association ends. Initial value rules can be stated very comfortably and easily. First of all, we have to point out the class that has got the attribute or association end. This class is termed as the context. Then we describe the expression that states our initial value rule. For instance, initial value of reward points in a *RewardAccount* will always be zero points. Similarly the initial value of the validity of the card at the time of issue will be – valid card only.

```
context RewardAccount::points
init: 0
context CustomerCard::valid
init: true
```

Another important and crucial part of the model is the specification for establishing the value of derived elements. A model possibly will hold derived attributes and derived associations. For both derived attributes and derived associations, an invented *derivation rule* can be specified. A derivation rule is a rule that specifies the value of a derived element. Once more, the context of the expression is the class that has got the attribute or association end. The expression comes after the context that states the derivation rule. For example, the derived attribute printedName of *CustomerCard* is decided on the basis of the name and title of the card owner.

```
context CustomerCard::printedName
derive:owner.title.concat('').concat(owner.name)
```

7 Query Operations

Operations that do not change the state of the system are termed as query operations. They just return a value or return set of values. It is very difficult to to give definition of the result of query operations in UML diagram. Therefore we write or define body expressions in OCL. In the context we provide the operation name, parameters, and return type (its signature). For instance, assuming that the class RewardProgram has a query operation getServices. getServices returns all services proposed and offered by all program associate in this ebusiness:

```
context RewardProgram::getServices(): Set(Service)
body: partners.deliveredServices->asSet()
```

In the above example, the association-end deliveredServices of ProgramAssociate that has got a set of Services is used. For all instances of ProgramAssociate that are associated with the RewardProgram instance of which the operation getServices is called. The result of query operation wills the sets of Services that are collected and combined into one set. In the body expression, the parameters of the operation are to be used. For instance, assume tahe we are looking for advanced version of the getServices operation. This operation takes as a parameter a program associate object and returns the services delivered by the parameter object, if it is an associate in this program. In this case, the refined and advanced operation can be specified as follows:

```
context RewardProgram::getServices(pa: ProgramAssociate):
Set(Service)
body: if partners->includes(pa)
      then pa.deliveredServices
      else Set
      endif
```

The result of this query operation is the set of Services held by the parameter pa, or an empty set if the parameter instance of ProgramAssociate is not an associate of the RewardProgram for which the query operation is called. This should be noted that derived attribute and a query operation that has no parameters are same only difference is of notation. The query operation requires that it is written using braces.

8 Defining New Attributes and Operations

This way of defing the attribute results in a derived attribute. The name and type of the attribute, and the derivation rule are included by the expression that defines the attribute. For example, we might want to introduce an attribute called earnings in class *RewardAccount*, which would be the result of the sum up of all the amount attributes of all the transactions on the account as defined in the class *Transaction*. This attribute can be specified and defined by the below given expression. The code fragment after the equal sign states the derivation rule for the new attribute:

```
context RewardAccount
def: turnover : Real = transactions.amount->sum()
```

When the derivation rule is applied, its output will be a single real number that is computed by summing the value of the amount attribute in all transactions associated with the *RewardAccount* instance that holds the newly defined attribute. We can define the operation in OCL; we should note that the operation that is defined in OCL is always will be a query operation. The code fragment that represents the expression just after the equality sign defines the body expression. Let us consider for an instance, we might wish to introduce the operation getServicesAsPerLevel in the class RewardProgram. This query operation returns the set of all delivered services for a particular level in a reward point program:

```
context RewardProgram
def: getServicesAsPerLevel(levelName: String):
Set(Service)
    =levels->select(name=levelName).availableServices-
>asSet()
```

The effect and output of the body expression is computed and taken from a collection and selectionof the levels associated with the instance of RewardProgram for which the operation getServicesAsPerLevel is called. It returns only the services available from the *ServiceLevel* whose name is equal to the parameter levelName.

9 Summary

Unified modeling language is critical to model driven architecture. We have investigated that OCL is critical to and why UML alone is not enough. We investigated how OCL can be incorporated into model in order to forward engineer correct code in ebusiness solution. To demonstrate the use of object constraints along with the modeling we took a case study for an ebusiness enterprise that is based on a reward program. There are many advantages to make UML model crisp and more specified by applying OCL to real-world modeling challenges. Using a combination of UML and object constraints allows developers to realize the effective, consistent, and coherent models. Finally we have all the complete and correct code in hand to satisfy our customer. We have investigated that looking in to constraints applied to the modeling we can improve the quality of the product. It is oblivious that improving the quality of modeling definitely improves the quality of analysis and design. It is found during our research that when we apply tracing back, we found lesser bugs in the model (with constraints) as compared to the more number of bugs when we traced back to model from code without using object constraints.

References

1. Akehurst, D.H., Bordbar, B.: The Unified Modeling Language, Modeling Languages, Concepts and Tools. In: 4th International Conference, Toronto, Canada (2001)
2. Balsters, H.: Modelling Database Views with Derived Classes in the UML/OCL-framework. In: Stevens, P., Whittle, J., Booch, G. (eds.) UML 2003. LNCS, vol. 2863, pp. 295–309. Springer, Heidelberg (2003)
3. Blaha, M., Premerlani, W.: Object-Oriented Modeling and Design for Database Applications. Prentice-Hall (1998)
4. Booch, G.: Object-Oriented Analysis and Design with Applications, 2nd edn., Benjamin/Cummings (1994)
5. Booch, G., Rumbaugh, J., Jacobson, I.: The Unified Modeling Language User Guide. Addison-Wesley (1999)
6. Carnegie Mellon University/Software Engineering Institute, The Capability Maturity Model: Guidelines for Improving the Software Process. Addison-Wesley (1995)
7. Clark, A., Warmer, J. (eds.):Object Modeling with the OCL: The Rationale behind the Object Constraint Language. LNCS, vol. 2263. Springer, Heidelberg (2002)
8. Coleman, D., Arnold, P., Bodoff, S., Dollin, C., Chilchrist, H., Hayes, F., Jeremaes, P.: Object-Oriented Development: The Fusion Method. Prentice-Hall (1994)
9. Cook, S., Daniels, J.: Designing Object Systems—Object Oriented Modeling with Syntropy. Prentice-Hall (1994)
10. D'Souza, D.F., Wills, A.C.: Objects, Components, and Frameworks with UML: The Catalysis Approach. Addison-Wesley (1999)
11. UML/EJB Mapping specification, Java Community Process Document JSR 26 (2001)
12. Eriksson, H., Penker, M.: Business Modeling with UML. Business Patterns at Work. John Wiley & Sons (2000)
13. Fowler, M.: UML Distilled: Applying the Standard Object Modeling Language. Addison-Wesley (1997)
14. Graham, I.: Migrating to Object Technology. Addison-Wesley (1995)
15. Jacobson, I., Booch, G., Rumbaugh, J.: The Unified Software Development Process. Addison-Wesley (1999)
16. Kleppe, A., Warmer, J., Bast, W.: MDA Explained; The Model Driven Architecture: Practice and Promise. Addison-Wesley (2003)
17. Liskov, B., Wing, J.: A Behavioral Notion of Subtyping. ACM Transactions on Programming Languages and Systems 16(6), 1811–1841 (1994)
18. Meyer, B.: On Formalism in Specifications. IEEE Software (January 1985)
19. Meyer, B.: Object-Oriented Software Construction. Prentice-Hall (1988)
20. Meyer, B.: Design by Contract. In: Advances in Object-Oriented Software Engineering, Prentice-Hall, pp. 1–50. Prentice-Hall (1991)
21. Meyer, B.: Applying Design by Contract. IEEE Computer (October 1992)
22. Object Constraint Language Specification, version 1.1, OMG document (1997)

23. Object Constraint Language Specification, version 2.0, OMG document (2006)
24. Object Constraint Language Specification, version 2.3 Beta, OMG document (2011)
25. Pinet, F., Duboisset, M., Soulignac, V.: Using UML and OCL to maintain the consistency of spatial data in environmental information systems. Elsevier (2007)
26. Richters, M.: A Precise Approach to Validating UML Models and OCL Constraints, Logos Verlag Berlin (2001)
27. Rumbaugh, J., Blaha, M., Premelani, W., Eddy, F., Lorensen, W.: Object-Oriented Modeling and Design. Prentice-Hall (1991)
28. Booch, G., Rumbaugh, J., Jacobson, I.: Unified Modeling Language Reference Manual. Addison-Wesley (1999)
29. Selic, B., Gullekson, G., Ward, P.T.: Real-Time Object-Oriented Modeling. John Wiley & Sons (1994)
30. UML 1.1 Specification, OMG documents (1997)
31. UML 2.0 Specification, OMG documents (2005)
32. UML 2.2 Specification, OMG documents (2007)
33. UML -ISO Released Versions of UML, Ver 1.4.2 (2005)
34. Warmer, J., Kleppe, A.: The Object Constraint Language: Getting Your Models Ready for MDA, 2nd edn. Addison-Wesley Professional (2003)

Vehicle Safety Device (Airbag) Specific Classification of Road Traffic Accident Patterns through Data Mining Techniques

S. Shanthi[1] and R. Geetha Ramani[2]

[1] Senior Lecturer, Department of Computer Science and Engineering
Rajalakshmi Institute of Technology, Kuthambakkam, Chennai, India
`psshanthiselvaraj@gmail.com`
[2] Professor and Head, Department of Computer Science and Engineering
Rajalakshmi Engineering College, Thandalam, Chennai, India
`rgeetha@gmail.com`

Abstract. Rich developing countries suffer from the consequences of increase in both human and vehicle population. Road accident fatality rates depend upon many factors which could vary for different countries. It is a very challenging task and investigating the dependencies between the attributes become complex because of many environmental and road related factors. In this research work we applied data mining classification technique RndTree and RndTree using ensemble methods viz. Bagging, AdaBoost and Multi Cost Sensitive Bagging (MCSB) to carry out vehicle safety device based classification of which RndTree using Adaboost gives high accurate results. The training dataset used for the research work is obtained from Fatality Analysis Reporting System (FARS) which is provided by the University of Alabama's Critical Analysis Reporting Environment (CARE) system. The results reveal that RndTree using Adaboost improvised the classifier's accuracy.

Keywords: Data Mining, Classification Algorithms, Bagging, Boosting, MCSB, Road Accident Data, RndTree.

1 Introduction

Data Mining [4] has attracted a great deal of attention in the information industry and in society due to the wide availability of huge amounts of data and there is a need for converting such data into useful information and knowledge. The information and knowledge [4] gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration.

Application of data mining techniques on social issues has been a popular technique of late. Road traffic injuries also place a huge strain on national health systems, many of which suffer from insufficient levels of resources [16]. Historically, many of the measures in place to reduce road traffic deaths and injuries are aimed at protecting car occupants [16]. Many literature analyses the road related factors which increase the death ratio.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 433–443.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Ensemble methods such as bagging and boosting are used to improve the accuracy of the weak classifiers [4]. In this paper, we focus on vehicle safety device based classification by applying RndTree classification algorithm and ensemble methods viz. Bagging, AdaBoost and Multi Cost Sensitive Bagging through which to identify best ensemble method. Among these algorithms RndTree using AdaBoost algorithm gives better results. The rest of this paper is organized as follows. Section I lists the summary of the related works in classification and ensemble algorithms. Section II illustrates the methodology used which includes the training dataset description, system design, classification algorithms, and ensemble algorithms and classifier accuracy measures. In section IV we present and discuss the experiment results. Finally the section V concludes the paper.

2 Related Work

The main reason to employ ensemble methods is to improve the accuracy of the weak classifiers. Various studies have been conducted to emphasize the use of ensemble methods such as bagging and boosting.

Diverse ensembles have a better potential for improvement on the accuracy than non-diverse ensembles. Also it has been shown that based on the diversity measures boosting gives better results than bagging [7]. The voting classification algorithms, such as Bagging and AdaBoost are successful in improving the accuracy of certain classifiers for artificial and real-world datasets [3].

Ensemble methods are very popular method for improving the performance of any weak learning algorithm. The most popular weak learners are decision trees, for example C4.5 or CART, and decision stumps [9]. In [9] AdaBoost integrating with C4.5 is presented to classify missing data. In 1997 the multiplicative weight-update technique [15] was proposed to derive a new boosting algorithm which is used to solve the problem of learning functions whose range, rather than being binary, is an arbitrary finite set or a bounded segment of the real line. The combination of the AdaBoost and random forests algorithms were used for constructing a breast cancer survivability prediction model [5]. It was proposed to use random forests as a weak learner of AdaBoost for selecting the high weight instances during the boosting process to improve accuracy, stability and to reduce over fitting problems [5]. Performance measurements (e.g., accuracy, sensitivity, and specificity), Receiver Operating Characteristic (ROC) curve and Area Under the receiver operating characteristic Curve (AUC) were used to measure the efficiency of the proposed classifier [5].

In 2000, [14] compared the effectiveness of randomization, bagging, and boosting for improving the performance of the decision-tree algorithm C4.5. It was proved that the randomized boosting with C4.5 produces best result [14]. Random Projection Technique is used on various applications to speed up the training process of AdaBoost [2] especially when the input dimension of data is high.

Various application domains are used [10] to study the different relationships and groupings among the performance metrics, thus facilitating the selection of performance metrics that capture relatively independent aspects of a classifier's performance. Factor analysis is applied [10] to the classifier performance space.

In our research work we focused on vehicle safety device (Airbag) specific classification to find accident patterns in road accident data using various ensemble algorithms. Next section illustrates the methodology used in our research work which includes RndTree, Bagging, AdaBoost and MCSB algorithms.

3 Methodolgy

This research work focuses on vehicle safety device (Airbag) specific classification of road accident patterns. The existing classification algorithm RndTree is adopted for classification. The RndTree algorithm produced classification results with 8.2% misclassification rate. Since RndTree had misclassification rate, ensemble methods viz. bagging, AdaBoost and MCSB have been incorporated with RndTree to improve the accuracy. The details of the work are given in the following sub sections.

3.1 Training Dataset Description

We carry out the experiment with road accident training dataset obtained from Fatality Analysis Reporting System (FARS) [17] which is provided by Critical Analysis Reporting Environment (CARE) system. This safety data consists of U.S. road accident information from 2005 to 2009. It consists of 272831 records and 23 attributes.

To train the classifiers we have selected accident details for two states California and New York totally 63327 records with 17 attributes. The selected dataset with 63327 records is divided into training dataset which consists of 47761 samples and test dataset which consists of 15566 samples. The list of attributes and their description is given in the Table 1.

Table 1. Training Dataset Attributes Description

Attributes	Description
Year	Year of accident
Month	Month of accident
Day	Day of accident
Manner_of_Collision	Manner of collision
Person_Type	Driver/Passenger
Seating_Position	Seating Position
Age_Range	Age range of the person involved
Gender	Male/Female
Injury_Severity	Injury Severity
Transported_By	Transported by emergency vehicle or not
AirBag	Location of the airbag
Protection_System	Type of the protection system used
Dead_on_Arrival	Status of the person at the arrival to hospital
Year_of_Death	Year of death
Month_of_Death	Month of death
Drug_Test	Type of drug test
Related_Factors	Road related factors

We have applied the classification algorithms using AirBag attribute as the class attribute. Next sub section deals about the system model used in this study.

3.2 System Design

This section describes the steps used in this research work. The steps used in this work are depicted in Fig.1.

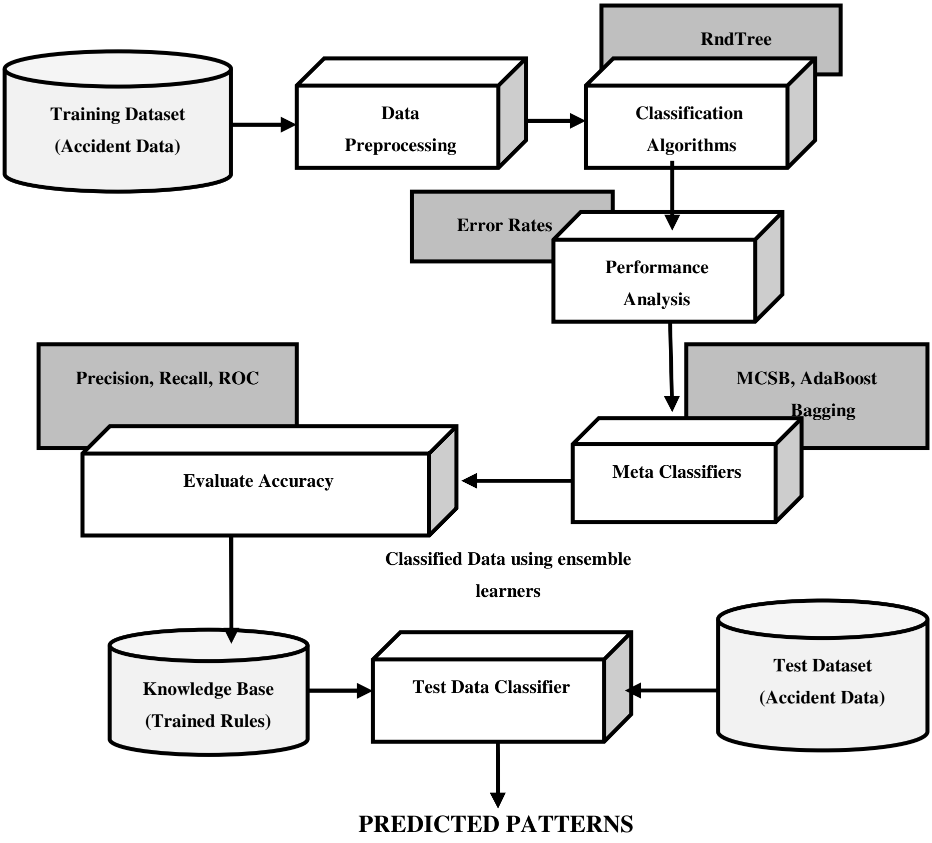


Fig. 1. Steps Involved in this study

After preprocessing the training set is given as input to the weak learner i.e. RndTree. The results shown that RndTree gives 8.2% misclassification rate. Thus to improve the accuracy of RndTree, Meta learners or ensemblers viz. bagging, AdaBoost and MCSB have been incorporated with RndTree and analysed for each ensembler's accuracy. The results are evaluated using accuracy measures such as precision, recall and ROC. It is found that Adaboost using RndTree significantly improves the accuracy. Test dataset is applied to evaluate the results.

3.3 RndTree Classification Algorithm

This section illustrates the RndTree algorithm. The accuracy of RndTree decision tree algorithm is better than that of other classification algorithms [13]. The advantage of decision tree algorithms is it is easy to derive the rules.

Random tree [11] can be applied to both regression and classification problems. The method combines bagging idea and the random selection of features in order to construct a collection of decision trees with controlled variation.

3.4 Bagging Algorithm

Bagging is an acronym for Bootstrap Aggregating. It is a method for generating multiple versions of a predictor [8] and using these to get an aggregated predictor. Given a certain model form like a classification tree, take repeated bootstrap samples of the original data set, refitting the model each time [1]. Each of these models will be slightly different, and the predictions at each point will vary from model to model. The bagged prediction for each point is an average over the predictions of all the models for that point. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets [8].

3.5 AdaBoost Algorithm

It is a boosting algorithm which is used to improve the accuracy of some learning method [4]. Accuracy is achieved by iteratively building weak models. Weak models are combined to deliver better model.

3.6 Multi Cost Sensitive Bagging (MCSB)

It is a sensitive supervised learning algorithm [14]. It uses the bagging for the estimation of the posterior probability. The misclassification cost is applied on each individual classifier and a predicted attribute is created with the bagging phase. Then it is used as a class attribute in the final construction of a classifier [14]. The main parameters of this algorithm are the cost matrix and the number of replication.

3.7 Accuracy Measures

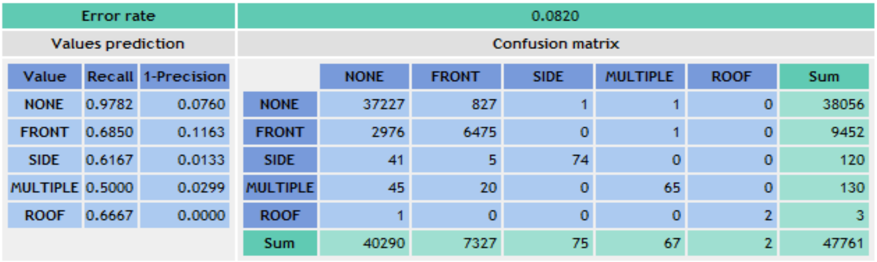
The accuracy of a classifier on a given set is the percentage of test set tuples that are correctly classified by the classifier. The confusion matrix is a useful tool for analyzing the efficiency of the classifiers [4]. ROC curve is a plot of TPR against FPR (False Positive Rate) which depicts relative trade-offs between true positives and false positives [4]. In next section we discuss the results we obtained in our research work.

4 Experimental Results

We have used Tanagra for our experimental study. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area [14]. We have divided the accident dataset into two parts: Training dataset which consists of 47761 records and test dataset which consists of 15566 records, totally 63327 records.

4.1 Experimental Results of Base Classifier, RndTree

In this phase we have applied basic decision tree algorithm RndTree to classify the training dataset. The results of these models are evaluated based on their error rates, precision and recall. The error rate of the RndTree is 8.2% which is given in the Fig.2.



Error rate	0.0820

Values prediction

Value	Recall	1-Precision
NONE	0.9782	0.0760
FRONT	0.6850	0.1163
SIDE	0.6167	0.0133
MULTIPLE	0.5000	0.0299
ROOF	0.6667	0.0000

Confusion matrix

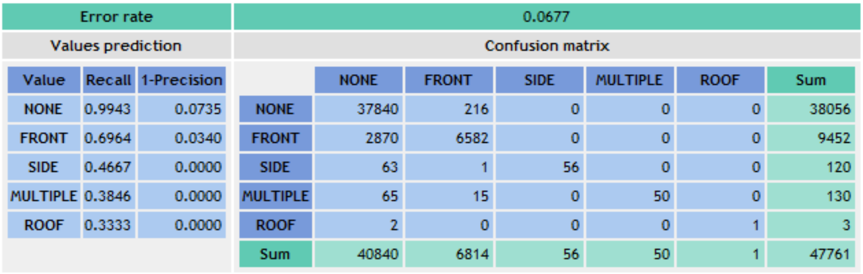
	NONE	FRONT	SIDE	MULTIPLE	ROOF	Sum
NONE	37227	827	1	1	0	38056
FRONT	2976	6475	0	1	0	9452
SIDE	41	5	74	0	0	120
MULTIPLE	45	20	0	65	0	130
ROOF	1	0	0	0	2	3
Sum	40290	7327	75	67	2	47761

Fig. 2. Error Rate of RndTree

In next section we discuss the results of ensemble classifiers using RndTree.

4.2 Experimental Results of Ensemble Classifiers

Since RndTree produced 8.2% misclassification rate we applied three ensemble methods viz. Bagging, AdaBoost and MCSB to improve the accuracy of RndTree. The error rate of the Bagging is given in Fig.3.



Error rate	0.0677

Values prediction

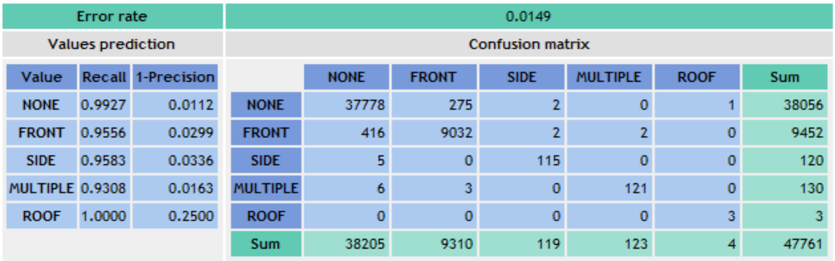
Value	Recall	1-Precision
NONE	0.9943	0.0735
FRONT	0.6964	0.0340
SIDE	0.4667	0.0000
MULTIPLE	0.3846	0.0000
ROOF	0.3333	0.0000

Confusion matrix

	NONE	FRONT	SIDE	MULTIPLE	ROOF	Sum
NONE	37840	216	0	0	0	38056
FRONT	2870	6582	0	0	0	9452
SIDE	63	1	56	0	0	120
MULTIPLE	65	15	0	50	0	130
ROOF	2	0	0	0	1	3
Sum	40840	6814	56	50	1	47761

Fig. 3. Error Rate of Bagging using RndTree

The error rate of bagging using RndTree is 6.77% which is better than that of the base clasifier, RndTree. It improvised the accuracy of RndTree from 91.8% to 93.23%. The error rate of AdaBoost using RndTree is given in Fig.4.



Error rate	0.0149

Values prediction

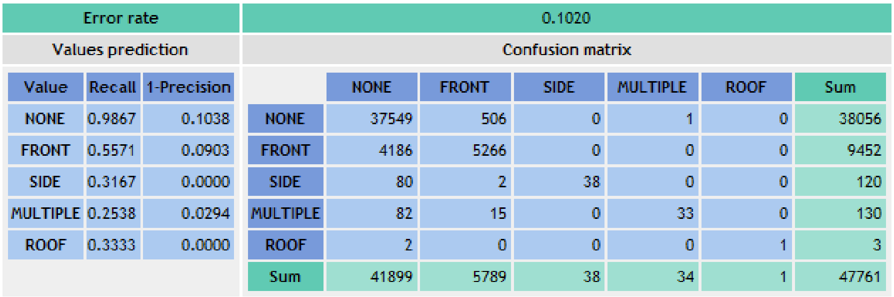
Value	Recall	1-Precision
NONE	0.9927	0.0112
FRONT	0.9556	0.0299
SIDE	0.9583	0.0336
MULTIPLE	0.9308	0.0163
ROOF	1.0000	0.2500

Confusion matrix

	NONE	FRONT	SIDE	MULTIPLE	ROOF	Sum
NONE	37778	275	2	0	1	38056
FRONT	416	9032	2	2	0	9452
SIDE	5	0	115	0	0	120
MULTIPLE	6	3	0	121	0	130
ROOF	0	0	0	0	3	3
Sum	38205	9310	119	123	4	47761

Fig. 4. Error Rate of AdaBoost using RndTree

The error rate of AdaBoost using RndTree is 1.49% which is better than that of the base clasifier, RndTree and bagging. It improvised the accuracy of RndTree from 91.8% to 98.51%. The error rate of MCSB using RndTree is given in Fig.5.



Error rate	0.1020

Values prediction

Value	Recall	1-Precision
NONE	0.9867	0.1038
FRONT	0.5571	0.0903
SIDE	0.3167	0.0000
MULTIPLE	0.2538	0.0294
ROOF	0.3333	0.0000

Confusion matrix

	NONE	FRONT	SIDE	MULTIPLE	ROOF	Sum
NONE	37549	506	0	1	0	38056
FRONT	4186	5266	0	0	0	9452
SIDE	80	2	38	0	0	120
MULTIPLE	82	15	0	33	0	130
ROOF	2	0	0	0	1	3
Sum	41899	5789	38	34	1	47761

Fig. 5. Error Rate of Multi Cost Sensitive Bagging using RndTree

The error rate of MCSB using RndTree is 10.2% which is greater than that of all the other classifiers. It decreased the accuracy of the base classifier from 91.8% to 89.8%. Fig.6 gives the comparison of accuracies of all the four classifiers.

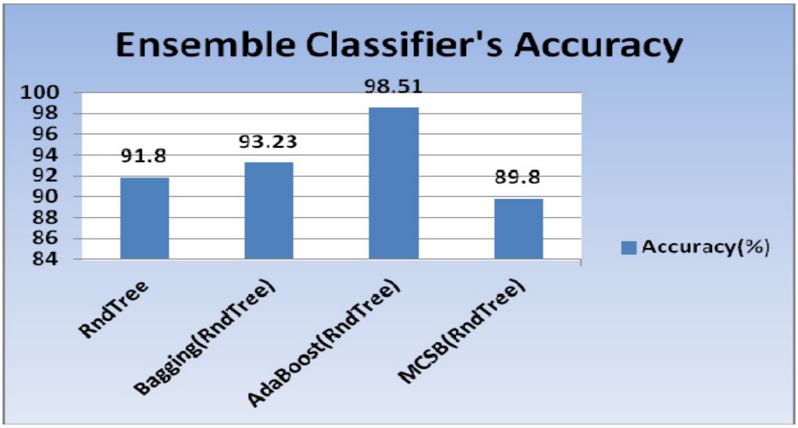


Fig. 6. Ensemble Classifier's Accuracy- Training Dataset

Fig.6 reveals that the AdaBoost using RndTree Classfier outperforms all other elsemble algorithms to perform vehicle safety measures (AirBag) based classification in road accident data set. The number of iterations at which the bagging, AdaBoost and MCSB accuracies are converged is given in the Table 2.

Table 2. Number of Iterations of Ensemble Algorithms

Classifiers	**Iterations**	**Accuracy**
Bagging	25	93.23
AdaBoost (RndTree)	5	98.51
MCSB	45	89.8

Table 3. ROC Curve Results

Sample size : 47761 Positive examples : 9452 Negative examples : 38309												
Score Attribute	**RndTree (Score 1)**			**Bagging (Score 2)**			**AdaBoost (Score 3)**			**MCSB (Score 4)**		
AUC	**0.9719**			**0.9824**			**0.9977**			**0.857**		
Target size (%)	Score	FPR	TPR	Score	FPR	TPR	Score	FPR	TPR	Score	FPR	TPR
0	1	0	0	1	0	0	0.9107	0	0	1	0	0
5	1	0	0.2526	0.72	.0004	0.251	0.6443	0	0.2526	1	0.0018	0.2452
10	1	0	0.5053	0.6	0.002	0.4978	0.6169	0	0.5053	0.8	0.0044	0.4873
15	0.5195	0.0202	0.676	0.48	0.007	0.7292	0.5911	0	0.7579	0.4706	0.0304	0.6348
20	0.4286	0.0517	0.8011	0.32	0.029	0.8928	0.494	0.0105	0.9678	0.1132	0.0684	0.7333
25	0.3333	0.0914	0.8927	0.2	0.076	0.9537	0.3852	0.065	0.9999	0	0.1235	0.7627
30	0.1818	0.1381	0.9561	0.12	0.134	0.9742	0.3507	0.1273	0.9999	0	0.1819	0.7788
35	0.043	0.1927	0.9876	0.08	0.194	0.9843	0.3234	0.1896	0.9999	0	0.2394	0.798
40	0.043	0.2523	0.9984	0.04	0.255	0.9889	0.2974	0.252	1	0	0.2975	0.8153
45	0	0.3143	1	0.04	0.316	0.9922	0.2728	0.3143	1	0	0.356	0.831
50	0	0.3766	1	0.04	0.378	0.9952	0.2477	0.3766	1	0	0.4142	0.8479
55	0	0.439	1	0	0.44	0.9962	0.2223	0.439	1	0	0.4723	0.8647
60	0	0.5013	1	0	0.502	0.9965	0.1956	0.5013	1	0	0.531	0.8795
65	0	0.5636	1	0	0.564	0.9968	0.1701	0.5636	1	0	0.5901	0.8925
70	0	0.626	1	0	0.627	0.9972	0.1448	0.626	1	0	0.6497	0.9039
75	0	0.6883	1	0	0.689	0.9978	0.1201	0.6883	1	0	0.7082	0.9193
80	0	0.7506	1	0	0.751	0.9982	0.0962	0.7506	1	0	0.7671	0.9331
85	0	0.813	1	0	0.813	0.9984	0.0727	0.813	1	0	0.8251	0.951
90	0	0.8753	1	0	0.876	0.9987	0.049	0.8753	1	0	0.8838	0.9657
95	0	0.9376	1	0	0.938	0.9995	0.0228	0.9376	1	0	0.942	0.9822
100	0	1	1	0	1	1	0	1	1	0	1	1

4.3 Experimental Results of Accuracy Measures

In this research work we used sensitivity, specificity, precision which should be high and FPR should be low to have high accuracy. Different classification algorithms may have their own characteristics on the same dataset [6]. In this work, the performance of AdaBoost using RndTree shows better results than that of bagging and MCSB. AdaBoost using RndTree is more specific and more sensitive. The results of training and test data are same.

Table 3 lists the AUC values of all classifiers depends on different sizes of the training data.

Though AUC of all the classifiers are greater than 0.7, the AUC of AdaBoost using RndTree (0.9934) is higher than other classifiers which conforms that AdaBoost using RndTree is comparatively good Meta classifier than other Meta classifiers. We got same results for test data.

Fig.7 explains the performance measures using ROC curves. The Score 3 (AdaBoost using RndTree) gives the curve which is nearer to the perfection point, 1.

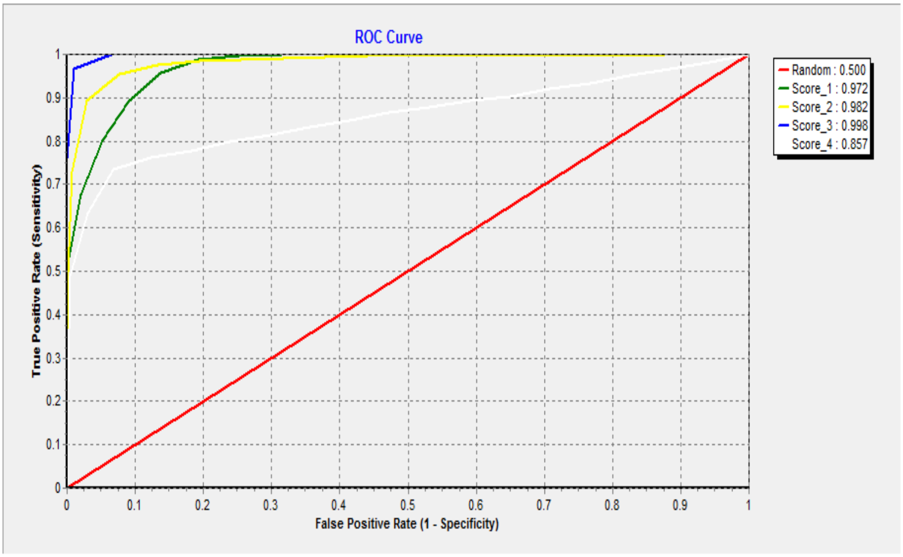


Fig. 7. Classifier's Accuracy- Test Dataset

4.4 Evaluation of Experimental Results Using Test Data

The correctness of the results of the foresaid classifiers have been evaluated using test dataset which consists of 15566 records. The results are depicted in Fig. 8.

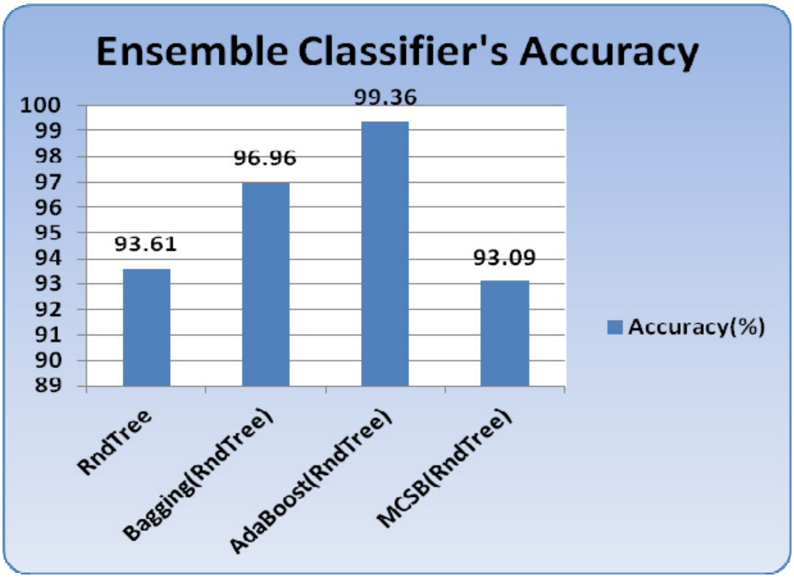


Fig. 8. Classifier's Accuracy- Test Dataset

5 Conclusion

In this paper we analyzed road accident training dataset using RndTree and ensemble methods viz. bagging, AdaBoost and MCSB to find patterns using vehicle safety measures (AirBag). Among the algorithms AdaBoost using RndTree gives high accuracy. The accuracy is evaluated based on precision, recall and ROC curves. The results showed that among ensemble algorithms (Bagging, Boosting and MCSB) the AdaBoost using RndTree improved accuracy from 91.8% to 98.51%.

References

1. Bagging and Boosting, `http://onlinelibrary.wiley.com`
2. Paul, B., Athithan, G., Narasimha Murty, M.: Speeding up AdaBoost classifier with random projection. In: Seventh International Conference on Advances in Pattern Recognition, pp. 251–254 (2009)
3. Bauer, E., Kohavi, R.: An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 36, 105–139 (1999)
4. Han, J., Kamber, M.: Data mining: concepts and techniques. Academic Press, ISBN 1-55860-489-8
5. Thongkam, J., Xu, G., Zhang, Y.: AdaBoost algorithm with random forests for predicting breast cancer survivability. In: International Joint Conference on Neural Networks (2008)
6. Wen, J., Zhang, X., Xu, Y., Li, Z., Liu, L.: Comparison of AdaBoost and logistic regression for detecting colorectal cancer patients with synchronous liver metastasis. In: International Conference on Biomedical and Pharmaceutical Engineering, December 2-4 (2009)

7. Kuncheva, L.I., Skurichina, M., Duin, R.P.W.: An experimental study on diversity for bagging and boosting with linear classifiers. Information fusion, Science Direct 3(4), 245–258
8. Breiman, L.: Arcing Classifiers. The Annals of Statistics 26, 801–849 (1998)
9. Miao, Z., Pan, Z., Hu, G., Zhao, L.: Treating missing data processing based on neural network and AdaBoost. In: IEEE International Conference on Grey Systems and Intelligent Services, Nanjing, China, November 18-20 (2007)
10. Seliya, N., Khoshgoftaar, T.M., Van Hulse, J.: A study on the relationships of classifier performance metrics. In: IEEE International Conference on Tools with Artificial Intelligence, pp. 59–66 (2009)
11. Random Tree Algorithm, `http://www.answers.com`
12. ROC Space, http://en.wikipedia.org/wiki/File:ROC_space-2.png
13. Shanthi, S., Geetha Ramani, R.: Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms. Int. Journal of Computer Applications 35(12), 30–37 (2011)
14. Dietterich, T.G.: An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning 40, 139–157 (2000)
15. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139 (1997)
16. World Health Organization: Global status report on road safety: time for action, Geneva (2009)
17. FARS analytic reference guide, `http://www.nhtsa.gov`

Design of Video Conferencing Solution in Grid

Soumilla Sen and Ajanta De Sarkar

Department of Computer Science and Engineering,
Birla Institute of Technology, Mesra, Kolkata 700 107
soumillasen@gmail.com, adsarkar@bitmesra.ac.in

Abstract. In grid environment, resources and users from different administrative domains are integrated and coordinated each other. Grid uses open standard, general-purpose protocols and interfaces to provide nontrivial quality of services. Videoconferencing is communication between two or more people from geographically different locations by simultaneous two-way video and audio transmissions. This is a type of 'visual collaboration' with audio-video communication. It differs from telephone calls through video display. Audio and Video communications is achieved in form of conferencing through Internet. This paper proposes design of a Video Conferencing Solution (VCS), which enables the users to explore the power of collaborative conferencing in grid environment. This solution is capable of setting up live meetings between a host and a number of participants.

Keywords: Gaia methodology, grid, video conferencing.

1 Introduction

A computational grid is designed to deliver larger computational resources to the users within affordable cost. Grids are used to solve large-scale computational and data-intensive problems by creating virtual organizations, sharing, accessing and combining different sets of resources and specialized devices in secured way. Occasionally, audio-video based collaborative conferencing might necessary in dynamic and diverse resource pool of grid environment. Thus the execution of video conferencing in such a dynamic and heterogeneous environment is a challenging problem. The proposed video conferencing solution (VCS) is based on a Multi-Agent System (MAS).

VCS is a collection of autonomous interacting software agents can offer services among distributed computational resources through deployment of these interconnected agents onto the resources. These resources are available and allocated in grid environment. Major task of *VCS* is to create super administrator, group administrator and user. *VCS* users complete registration for joining a meeting as host or participant as well as leaving meeting. Super administrator is solely responsible for creating groups, license for the groups and administrator for group of users. Group administrator manages meeting, users of the group. Details of these functionalities of all these roles are discussed in Section 4.

In this research work, Gaia methodology has been used to carry out the detailed design of the individual agent structure and the multi-agent development process in

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 445–457.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

the *VCS*. Therefore different types of analysis and design phases of *VCS* have been presented in this paper with *environment model*, *role model*, *interaction model* and *service model* respectively. Organization of this paper is as follows. Related work is presented in Section 2. Section 3 presented an overview of Gaia methodology. Detailed design of the video conferencing solution using Gaia methodology is presented in Section 4. Section 5 concludes the paper with direction of future work.

2 Related Work

In recent times there has been a surge of interest in network based video conferencing system. Research and development in this area has expanded significantly in the past few years. The various popular available solutions may be enumerated as follows: Skype [4], PlaceCam [3], Vennfer [5], Woovoo [6], GoToMeeting [2] and Dimdim [1]. These products are Internet based and supports two way live audio video communications. The number of participants supported, codec, protocol, bandwidth requirement varies from product to product. Some products also provides extra feature like white board, test massaging, audio-video archiving and playback.

In [12], heterogeneous multimedia clients can join in the same real-time sessions. This Global Multimedia Collaboration System provides support for a variety of protocols and applications, including H.323 clients, SIP clients and Access Grid rooms. It facilitates audio and video communications among participating clients in a real-time conference. A roadmap for the future of videoconferencing within e-science is presented in [8]. In this research, the concept of Access Grid Studio-based Videoconferencing is studied. The proposed *VCS* supports Unicast and Multicast both environments with audio / video standard (AAC / H.264) respectively. *VCS* uses HTTP (Hyper Text Transfer Protocol) for communication with meeting server and RTP (Real Time Transfer Protocol) for audio-video streaming. The total minimum bandwidth requirement for *VCS* is around 150Kb/s while an optimum bandwidth requirement is around 300Kb/s. Novelty of this research is that it is based on multi-agent system (MAS) technology and proposed to be implemented in grid environment.

Comparing all the agent-oriented methodologies is often difficult, because they might address different aspects in different scenarios. In this research, Gaia methodology is specially chosen to design the system in details. Main focus of Gaia is on individual and autonomous agent structure and interaction of these agents in organized society to achieve a common specified goal. Gaia [15] does not explicitly deal with the activities of requirement gathering phase and implementation issues. As requirement gathering is already done in this research, analysis and design is discussed elaborately in this paper. Implementation of this *VCS* would be done with the FIPA-compliant agent platform, Jade framework [11] in Grid environment.

3 Overview of Gaia Methodology

Gaia is a methodology [14] for agent-oriented software analysis and design. It deals with both the macro (social) aspect and the micro (agent internals) aspect of a multi-agent system and devotes a specific effort to model the organizational structure and the organizational rules that govern the global behavior of the agents in the organization. Gaia is intended to allow an analyst to go systematically from a statement of requirements to a design that is sufficiently detailed that it can be implemented directly. Analysis and design can be thought of as a process of developing increasingly detailed models of the system to be constructed.

The objective of this paper is to describe the proposed *VCS* for execution in Grid environment. The main concepts in Gaia can be divided into two categories: *abstract* and *concrete*. These concepts are developed in three different phases – the analysis phase, the architectural design phase and the detailed design phase. Abstract entities are developed at the time of analyzing the requirements of the system and conceptualizing it, these entities may not have any direct realization within the system. The abstract concepts in Gaia include *roles, permissions, responsibilities, protocols, activities, liveness properties* and *safety properties*. Concrete entities, in contrast, are used in the detailed design phase, and will typically have direct counterparts in the runtime system.

According to Gaia methodology, designing and developing of the *role* and *interaction models* require a few number of steps. In the detailed design phase *service model* is being developed. It depicts main services of each agent. Design of video conferencing solution is discussed elaborately phase wise in the next section.

4 Design of Video Conferencing Solution

Video Conferencing Solution is designed with a group of inter acting multiple autonomous agents, which performs several tasks to provide the live conference among group of people from different locations. A software engineering approach, namely Gaia methodology has been considered for detailing the design of the *VCS*. The steps for modeling the system using this methodology are described in this section. It composed of three different phases, namely: *analysis phase*, *architectural design phase* and *detailed design phase*. Analysis phase briefs about corresponding roles of each sub organization. In architectural design phase, rules of liveness and safety described. It also finalizes role model and interaction model. At last, agent model and service model are developed in detailed design phase.

Analysis Phase

The objective of the analysis phase is to develop perception of the system and its structure. In this section, first the organizations within the system are identified, and then an environment model is developed as proposed by Gaia.

Organizations

According to Gaia, the first step in the analysis phase of VCS is to determine whether multiple organizations have to coexist in the system and perform autonomously. Sub-organizations can be found, which fulfill any of these conditions:

i. Exhibit a behavior specifically oriented towards the achievement of a given sub-goal.
ii. Interact loosely with other segments of the system.
iii. Require competences that are not needed in other parts of the system.

The two major sub organizations of the system can be defined in analysis phase and the goal of each sub-organization is shown in Table 1.

Table 1. Sub-organizations in the Video Conferencing Solution

Name	Description
Meeting Provider	This sub-organization provides Group, License, User, Multicast IP, Meeting and Client Installer software
Meeting Executor	This sub-organization executes the meeting by enabling audio video interchange and desktop sharing

The Environment Model

In the next step of the analysis phase, the corresponding roles in specific sub organization are modeled as follows:

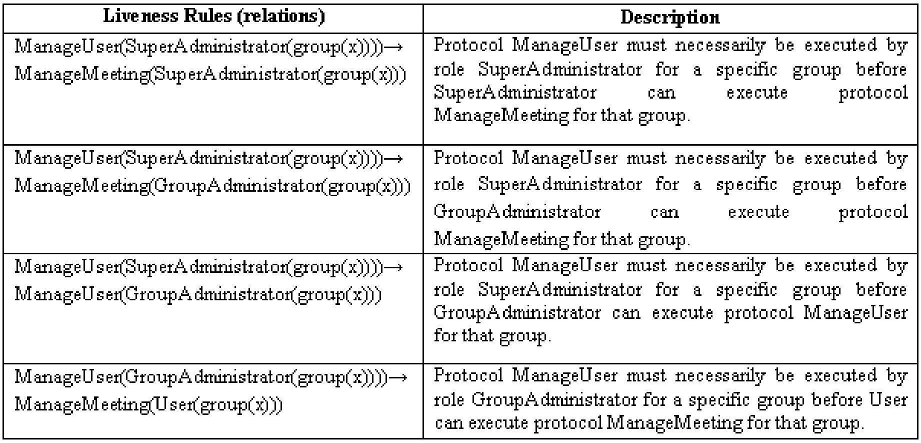
1. SuperAdministrator (SA), GroupAdministrator (GA) and User in *Meeting Provider* sub organization
2. Host, ActiveParticipant (AP), PassiveParticipant (PP) and Transponder (TP) in *Meeting Executor* sub organization

Next the organizational rules of the above mentioned sub-organizations are defined. Organizational rules are seen as responsibilities of the organization as a whole. As in the role model, organizational rules also have safety and liveness rules. The organizational rules for liveness and safety are shown in Table 2 and Table 3 respectively. They have been used and considered in the architectural design phase, which will be described in the next section.

Architectural Design Phase

On the basis of the analysis, this section describes the architectural design of the *VCS* that completes and refines the preliminary models and makes actual decisions about the organizational structure and models the *VCS* based on the specifications produced.

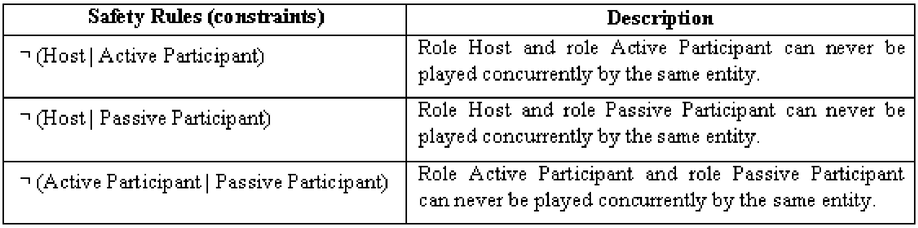
Table 2. Liveness (Relations) Organizational Rules



Liveness Rules (relations)	Description
ManageUser(SuperAdministrator(group(x))))→ ManageMeeting(SuperAdministrator(group(x)))	Protocol ManageUser must necessarily be executed by role SuperAdministrator for a specific group before SuperAdministrator can execute protocol ManageMeeting for that group.
ManageUser(SuperAdministrator(group(x))))→ ManageMeeting(GroupAdministrator(group(x)))	Protocol ManageUser must necessarily be executed by role SuperAdministrator for a specific group before GroupAdministrator can execute protocol ManageMeeting for that group.
ManageUser(SuperAdministrator(group(x))))→ ManageUser(GroupAdministrator(group(x)))	Protocol ManageUser must necessarily be executed by role SuperAdministrator for a specific group before GroupAdministrator can execute protocol ManageUser for that group.
ManageUser(GroupAdministrator(group(x))))→ ManageMeeting(User(group(x)))	Protocol ManageUser must necessarily be executed by role GroupAdministrator for a specific group before User can execute protocol ManageMeeting for that group.

As it is already mentioned that *Meeting Provider* sub organization is responsible for the overall provisioning which includes management of Group, License, User, Multicast IP, Meeting and Client Installer software. There are three types of roles in the system SuperAdministrator, GroupAdministrator and User. There is only one SuperAdministrator in the whole system. SuperAdministrator creates a new group for the customer or user and licenses are created whenever a new customer purchases license. SuperAdministrator will also create one User with GroupAdministrator role for the group. SuperAdministrator may create number of Users less than equal to the licenses being purchased for the group. Alternatively GroupAdministrator also may create the other Users.

Table 3. Safety (Constraints) Organizational Rules

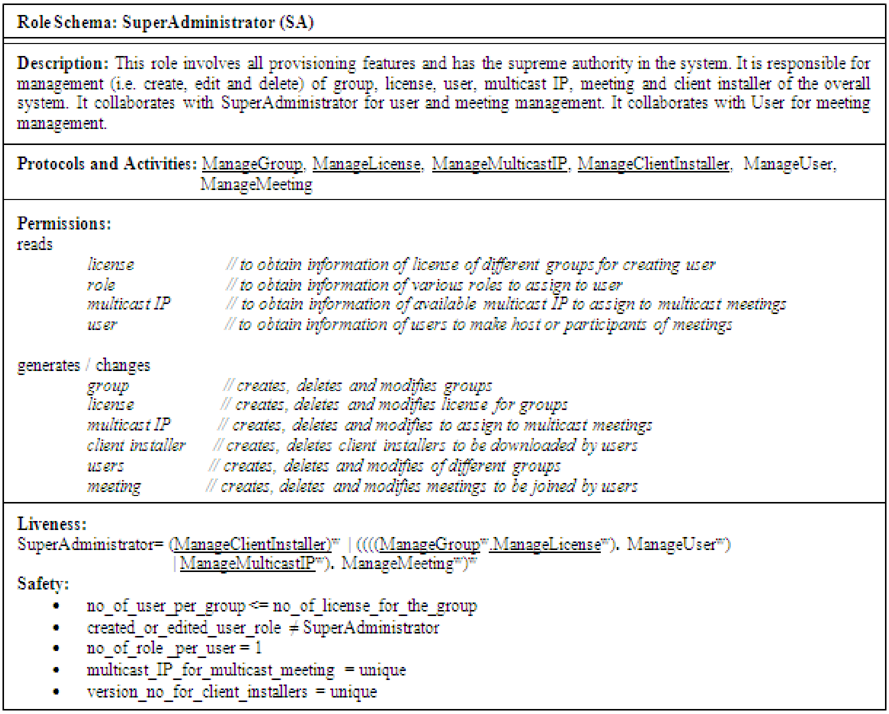


Safety Rules (constraints)	**Description**
¬ (Host | Active Participant)	Role Host and role Active Participant can never be played concurrently by the same entity.
¬ (Host | Passive Participant)	Role Host and role Passive Participant can never be played concurrently by the same entity.
¬ (Active Participant | Passive Participant)	Role Active Participant and role Passive Participant can never be played concurrently by the same entity.

Meetings may be of three types Multicast, Unicast, Multicast-Unicast. SuperAdministrator also creates a number of unique Multicast IP for the group which will be required to create multicast and unicast-multicast meetings. SuperAdministrator, GroupAdministrator and User any one can create meeting by selecting the host, participant and Multicast IP (for multicast and unicast-multicast meeting types). User may edit and delete meetings created by him/her. GroupAdministrator may edit and delete users and meetings of his/her group. SuperAdministrator also manages the Client Installer software as well as edit and delete Group, License, User, Multicast IP, and Meeting.

Meeting Executor sub organization is responsible for running the meeting participated by host and other participants. Running meeting includes organizing and hosting meeting by enabling audio video interchange and desktop sharing. Host transmits his/her audio video to all other participant of the meeting. It also can share his/her desktop to other participants to give a presentation. Participant can raise their hand to become “active”. Host may select any of the participants as “active”. ActiveParticipant (AP) also transmits his/ her audio video to host as well as other participants. Transmitted audio video may reach directly to the recipient or via n number of hops in-between depending upon the distance between the sender and receiver.

Next on the basis of the analysis and organizational structure, the agent role model of the *VCS* is described. The role schemas are presented using the templates from [15, 14]. At first, roles of the *Meeting Provider* sub organization are discussed. SuperAdministrator role has the ultimate authority of the system. SuperAdministrator performs these activities Group Management, License management, User management Multicast management, Meeting management, Client Installer management. Whenever a new customer purchases license SuperAdministrator will create one User with GroupAdministrator role for the group. GroupAdministrator may create the other users with User role. GroupAdministrator also may create meeting by selecting the host, participant and Multicast IP (for multicast and unicast-multicast meeting types) for the group. GroupAdministrator may edit and delete users and meetings of his/her group. Role user belongs to a particular group. User may create meeting by selecting the host, participant and Multicast IP (for both unicast / multicast meeting types) from his/her group. User may also edit and delete meetings created by him/her. Corresponding role schemas for SuperAdministrator (SA), GroupAdministrator (GA) and User are shown in Figure 1, Figure 2 and Figure 3.



Role Schema: SuperAdministrator (SA)

Description: This role involves all provisioning features and has the supreme authority in the system. It is responsible for management (i.e. create, edit and delete) of group, license, user, multicast IP, meeting and client installer of the overall system. It collaborates with SuperAdministrator for user and meeting management. It collaborates with User for meeting management.

Protocols and Activities: ManageGroup, ManageLicense, ManageMulticastIP, ManageClientInstaller, ManageUser, ManageMeeting

Permissions:
reads
license *// to obtain information of license of different groups for creating user*
role *// to obtain information of various roles to assign to user*
multicast IP *// to obtain information of available multicast IP to assign to multicast meetings*
user *// to obtain information of users to make host or participants of meetings*

generates / changes
group *// creates, deletes and modifies groups*
license *// creates, deletes and modifies license for groups*
multicast IP *// creates, deletes and modifies to assign to multicast meetings*
client installer *// creates, deletes client installers to be downloaded by users*
users *// creates, deletes and modifies of different groups*
meeting *// creates, deletes and modifies meetings to be joined by users*

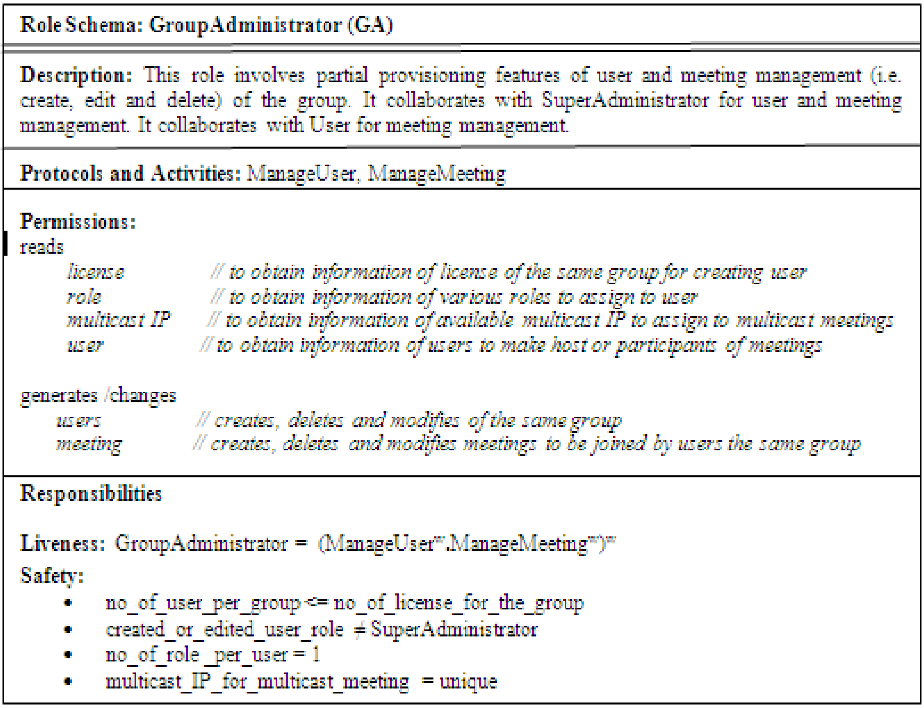
Liveness:
SuperAdministrator= $(\underline{\text{ManageClientInstaller}})^{\omega}$ | $((((\underline{\text{ManageGroup}}^{\omega}.\underline{\text{ManageLicense}}^{\omega}).\ \text{ManageUser}^{\omega})$
$|\ \underline{\text{ManageMulticastIP}}^{\omega}).\ \text{ManageMeeting}^{\omega})^{\omega}$

Safety:
- no_of_user_per_group <= no_of_license_for_the_group
- created_or_edited_user_role ≠ SuperAdministrator
- no_of_role _per_user = 1
- multicast_IP_for_multicast_meeting = unique
- version_no_for_client_installers = unique

Fig. 1. Role schema of SuperAdministrator (SA)

Roles of *Meeting Executor* consist of Host, Active Participant, Passive Participant and Transponder. Host is selected at the time of meeting creation, though it can be changed by editing the meeting at any point of time. When a meeting starts there is Host and other passive participants present in the meeting. All passive participants receive host's audio video. In addition host can also share desktop. Passive participants may raise hand to become active. One of the passive participants is selected as active by the host. ActiveParticipant does interactive communication with the host. If any participant previously selected as active, it will be turned passive again automatically, if host selects another participant as Activate participant. One of the passive participants is selected as active by the host. Passive participant receives audio video from host as well as active participant, receives desktop sharing from host. It cannot transmit audio video. But it can raises hand to become active. Transmitted audio video, desktop sharing and hand raise may reach directly to the recipient or via n number of hops in-between depending upon the distance between the sender and receiver. There will be only one Transponder per network. Transponder is essentially a client agent playing role of Host, AP or PP in the same meeting. Transponder may work independently or collaborates with other Transponders. It receives audio video from Host and ActiveParticipant, as well as desk top sharing (DS) and hand raise of PassiveParticipant. It also capable of retransmit all these at the same time. The role schemas for Host, ActiveParticipant (AP), Passive participant (PP) and Transponder (TP) are depicted in Figure 4, Figure 5, Figure 6 and Figure 7 respectively.

Interaction model follows the role model. The functionalities, activities and responsibilities of the roles are described in role model. Interaction model depicts the interaction between the roles on the basis of protocol. It also describes the characteristics and dynamics of each protocol. So it completely defines the protocol through which the roles will interact in the organization. Interaction model of *VCS* is shown in Figure 8.



Role Schema: GroupAdministrator (GA)

Description: This role involves partial provisioning features of user and meeting management (i.e. create, edit and delete) of the group. It collaborates with SuperAdministrator for user and meeting management. It collaborates with User for meeting management.

Protocols and Activities: ManageUser, ManageMeeting

Permissions:

reads

license	*// to obtain information of license of the same group for creating user*
role	*// to obtain information of various roles to assign to user*
multicast IP	*// to obtain information of available multicast IP to assign to multicast meetings*
user	*// to obtain information of users to make host or participants of meetings*

generates /changes

users	*// creates, deletes and modifies of the same group*
meeting	*// creates, deletes and modifies meetings to be joined by users the same group*

Responsibilities

Liveness: GroupAdministrator = $(ManageUser^{\omega}.ManageMeeting^{\omega})^{\omega}$

Safety:

- no_of_user_per_group <= no_of_license_for_the_group
- created_or_edited_user_role ≠ SuperAdministrator
- no_of_role _per_user = 1
- multicast_IP_for_multicast_meeting = unique

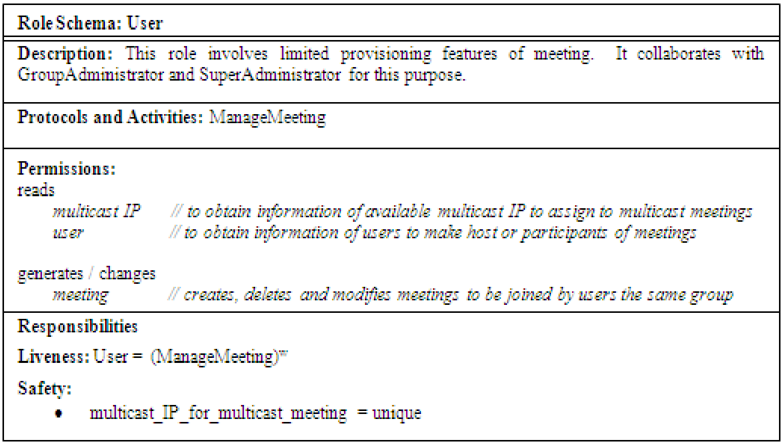
Fig. 2. Role Schema of GroupAdministrator (GA)

Detailed Design Phase

According to Gaia, the detailed design commences just after development of the role model and interaction model. In the detailed design phase, actually one-to-one correspondence between roles and agent classes is made to create agent model. It is also possible to combine closely roles into one agent. This is advantageous because of less number of classes and instances eventually reduce conceptual complexity.

However, this has to be done without:

- ❑ affecting the organizational efficiency,
- ❑ violating organizational rules and
- ❑ Creating problems (that is, without exceeding the amount of information it is able to store and process in a given amount of time).



Role Schema: User

Description: This role involves limited provisioning features of meeting. It collaborates with GroupAdministrator and SuperAdministrator for this purpose.

Protocols and Activities: ManageMeeting

Permissions:
reads
multicast IP // *to obtain information of available multicast IP to assign to multicast meetings*
user // *to obtain information of users to make host or participants of meetings*

generates / changes
meeting // *creates, deletes and modifies meetings to be joined by users the same group*

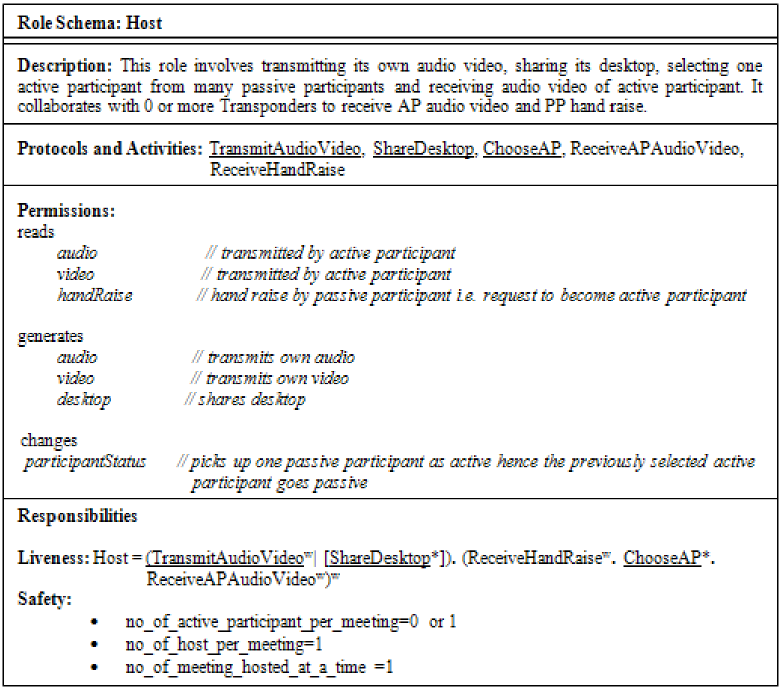
Responsibilities

Liveness: User = (ManageMeeting)$^{\omega}$

Safety:

- multicast_IP_for_multicast_meeting = unique

Fig. 3. Role Schema of User



Role Schema: Host

Description: This role involves transmitting its own audio video, sharing its desktop, selecting one active participant from many passive participants and receiving audio video of active participant. It collaborates with 0 or more Transponders to receive AP audio video and PP hand raise.

Protocols and Activities: TransmitAudioVideo, ShareDesktop, ChooseAP, ReceiveAPAudioVideo, ReceiveHandRaise

Permissions:
reads
audio // *transmitted by active participant*
video // *transmitted by active participant*
handRaise // *hand raise by passive participant i.e. request to become active participant*

generates
audio // *transmits own audio*
video // *transmits own video*
desktop // *shares desktop*

changes
participantStatus // *picks up one passive participant as active hence the previously selected active participant goes passive*

Responsibilities

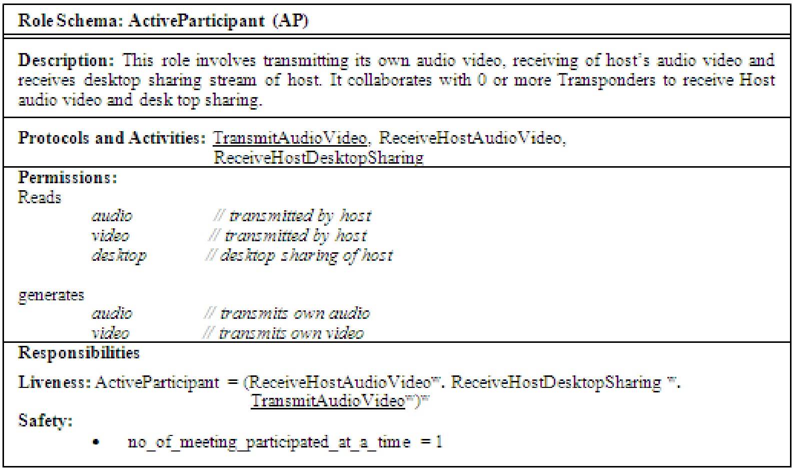
Liveness: Host = (TransmitAudioVideo$^{\omega}$| [ShareDesktop*]). (ReceiveHandRaise$^{\omega}$. ChooseAP*. ReceiveAPAudioVideo$^{\omega}$)$^{\omega}$

Safety:

- no_of_active_participant_per_meeting=0 or 1
- no_of_host_per_meeting=1
- no_of_meeting_hosted_at_a_time =1

Fig. 4. Role Schema of Host

Detailed design phase actually develops agent model and service model. In case of agent model, agent is a software program or class which would be implemented during implementation stage. In *Meeting Provider* sub organization, each role is mapped to its corresponding agent, while, group of roles are mapped to a single agent in *Meeting Executor* sub organization. Agent model of *Meeting Provider* and *Meeting Executor* are depicted in Figure 9 and Figure 10 respectively. In case of service model, the specific services of agents are identified. There are four properties related to each service, namely: input, output, pre-condition and post-condition. Input-output are actually derived from the interaction model or protocols, while pre and post conditions are restrictions on the execution and completion of the services respectively.



Role Schema: ActiveParticipant (AP)

Description: This role involves transmitting its own audio video, receiving of host's audio video and receives desktop sharing stream of host. It collaborates with 0 or more Transponders to receive Host audio video and desk top sharing.

Protocols and Activities: TransmitAudioVideo, ReceiveHostAudioVideo, ReceiveHostDesktopSharing

Permissions:

Reads

audio *// transmitted by host*

video *// transmitted by host*

desktop *// desktop sharing of host*

generates

audio *// transmits own audio*

video *// transmits own video*

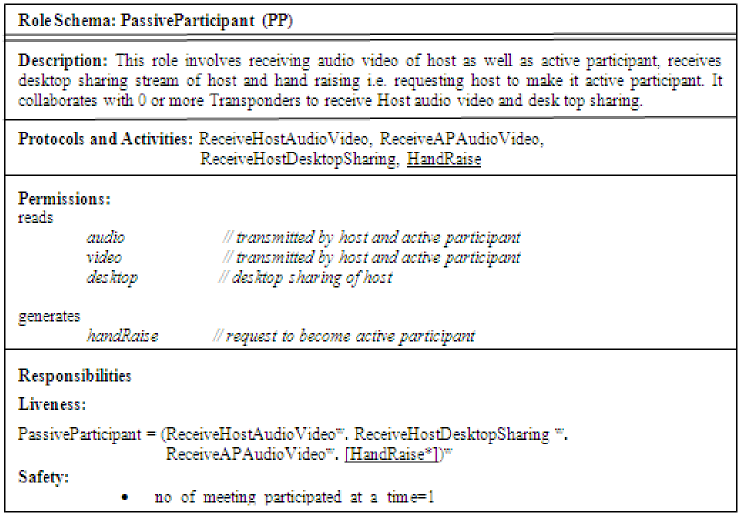
Responsibilities

Liveness: ActiveParticipant = (ReceiveHostAudioVideo$^{\omega}$. ReceiveHostDesktopSharing $^{\omega}$. TransmitAudioVideo$^{\omega}$)$^{\omega}$

Safety:

- no_of_meeting_participated_at_a_time = 1

Fig. 5. Role Schema of ActiveParticipant (AP)



Role Schema: PassiveParticipant (PP)

Description: This role involves receiving audio video of host as well as active participant, receives desktop sharing stream of host and hand raising i.e. requesting host to make it active participant. It collaborates with 0 or more Transponders to receive Host audio video and desk top sharing.

Protocols and Activities: ReceiveHostAudioVideo, ReceiveAPAudioVideo, ReceiveHostDesktopSharing, HandRaise

Permissions:

reads

audio *// transmitted by host and active participant*

video *// transmitted by host and active participant*

desktop *// desktop sharing of host*

generates

handRaise *// request to become active participant*

Responsibilities

Liveness:

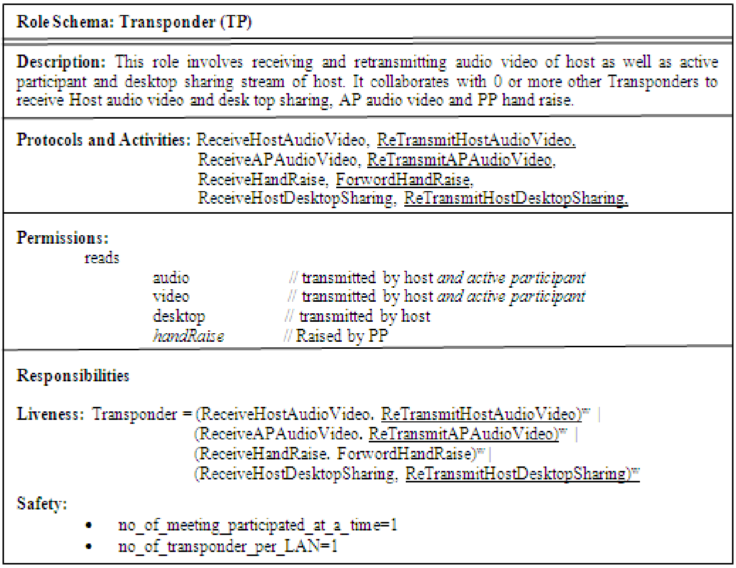
PassiveParticipant = (ReceiveHostAudioVideo$^{\omega}$. ReceiveHostDesktopSharing $^{\omega}$. ReceiveAPAudioVideo$^{\omega}$. [HandRaise*])$^{\omega}$

Safety:

- no of meeting participated at a time=1

Fig. 6. Role Schema of Passive Participant (PP)

These services are derived from the protocols, activities and liveness properties of the roles that the agent implements. As a rule of tomb, there will be one service for each parallel activity of execution that the agent has to execute. The service model of VCS is represented in Table 4.



Role Schema: Transponder (TP)

Description: This role involves receiving and retransmitting audio video of host as well as active participant and desktop sharing stream of host. It collaborates with 0 or more other Transponders to receive Host audio video and desk top sharing, AP audio video and PP hand raise.

Protocols and Activities: ReceiveHostAudioVideo, ReTransmitHostAudioVideo,
ReceiveAPAudioVideo, ReTransmitAPAudioVideo,
ReceiveHandRaise, ForwordHandRaise,
ReceiveHostDesktopSharing, ReTransmitHostDesktopSharing.

Permissions:
reads

audio	// transmitted by host *and active participant*
video	// transmitted by host *and active participant*
desktop	// transmitted by host
handRaise	// Raised by PP

Responsibilities

Liveness: Transponder = (ReceiveHostAudioVideo. ReTransmitHostAudioVideo)$^{\omega}$ |
(ReceiveAPAudioVideo. ReTransmitAPAudioVideo)$^{\omega}$ |
(ReceiveHandRaise. ForwordHandRaise)$^{\omega}$ |
(ReceiveHostDesktopSharing. ReTransmitHostDesktopSharing)$^{\omega}$

Safety:

- no_of_meeting_participated_at_a_time=1
- no_of_transponder_per_LAN=1

Fig. 7. Role Schema of Transponder (TP)

5 Conclusion and Future Work

This paper clearly portrays the design of a video conferencing solution based on multi-agent system. This solution is capable of setting up live meetings between a host and a number of participants. *VCS* is a collection of autonomous interacting software agents can offer services in grid environment through deployment of these interconnected agents onto the resources. In order to do design and analysis of *VCS,* a software engineering approach using Gaia Methodology is used. The *role model*, *interaction model*, *agent model* and *services model* are discussed and developed in this research work. The current work demonstrates individual and autonomous agent structure as well as illustrates the interaction of the agents in proposed *VCS*. The main intention of future work is to implement the *VCS* in grid environment with the FIPA-compliant agent platform, Jade framework through collaboration of a few interacting agents.

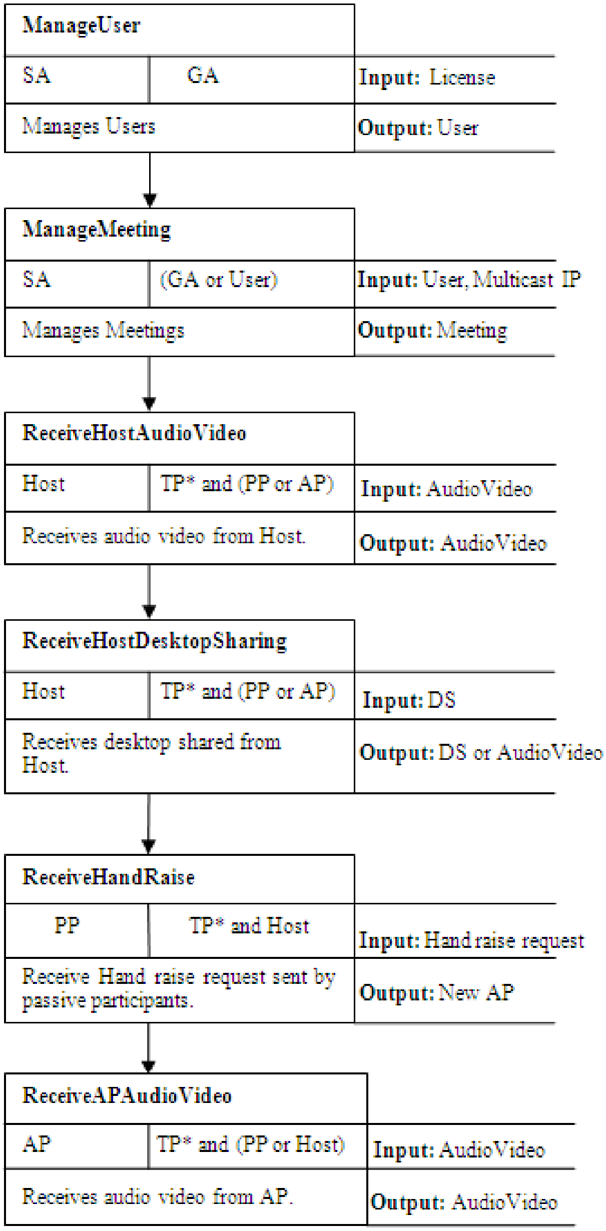


Fig. 8. Interaction model of protocol Manage User and Manage Meeting

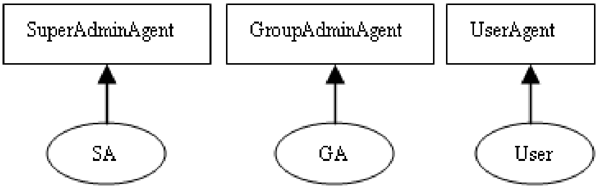


Fig. 9. Agent Model of Meeting Provider sub organization

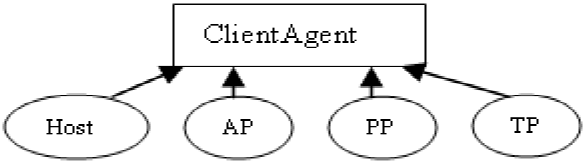
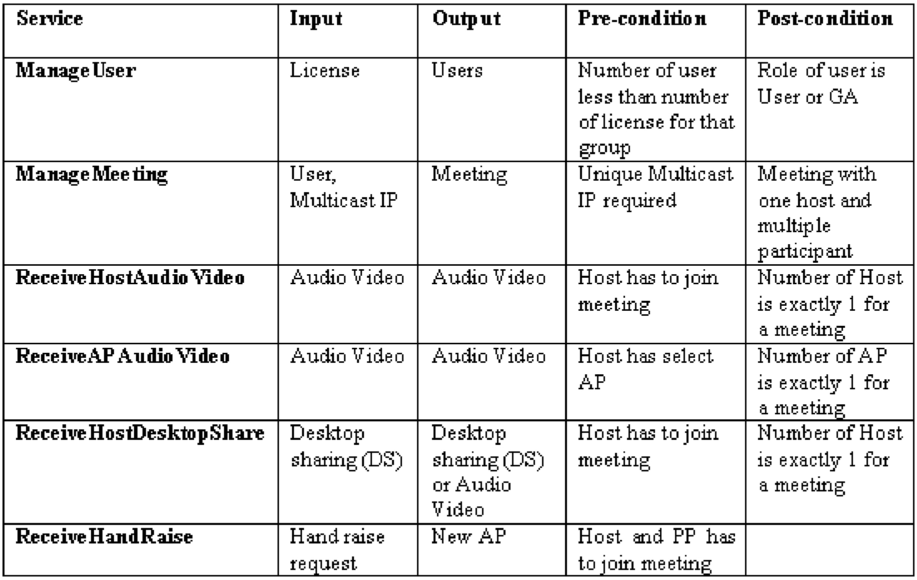


Fig. 10. Agent Model of Meeting Executor sub organization

Table 4. Service Model of Video Conferencing Solution



Service	**Input**	**Output**	**Pre-condition**	**Post-condition**
ManageUser	License	Users	Number of user less than number of license for that group	Role of user is User or GA
ManageMeeting	User, Multicast IP	Meeting	Unique Multicast IP required	Meeting with one host and multiple participant
ReceiveHostAudioVideo	Audio Video	Audio Video	Host has to join meeting	Number of Host is exactly 1 for a meeting
ReceiveAPAudioVideo	Audio Video	Audio Video	Host has select AP	Number of AP is exactly 1 for a meeting
ReceiveHostDesktopShare	Desktop sharing (DS)	Desktop sharing (DS) or Audio Video	Host has to join meeting	Number of Host is exactly 1 for a meeting
ReceiveHandRaise	Hand raise request	New AP	Host and PP has to join meeting	

References

1. Dimdim Video Conferencing Solutions, `http://www.dimdim.com`
2. GoTo Meeting Video Conferencing Solutions, `http://www.gotomeeting.in`
3. PlaceCam Video Conferencing Solutions, `http://www.placecam.com`
4. Skype Video Call Solutions, `http://www.skype.com`
5. Vennfer Video Conferencing Solutions, `http://www.intellisysin.com`
6. Woovoo Video Conferencing Solutions, `http://www.oovoo.com`

7. Cernuzzi, L., Cossentino, M., Zambonelli, F.: Process Models for Agent- based Development. Journal of Engineering Applications of Artificial Intelligence 18(2), 205–222 (2005)
8. Cooper, C.: Multi-Site Videoconferencing for the UK e-Science Programme, A Roadmap for the Future of Videoconferencing within e-Science. (UKERNA / Oxford Brookes University) and Michael Daw (University of Manchester) (2002)
9. Koutsabasis, P., Darzentas, J.S., Spyrou, T., Darzentas, J.: Facilitating User–System Interaction: the GAIA Interaction Agent Proceedings. In: 32nd Hawaii International Conference on System Sciences (1999)
10. Rodriguez, L., Hume, A., Cernuzzi, L., Insfrán, E.: Improving the Quality of Agent-Based Systems: Integration of Requirements Modeling into Gaia. In: 2009 Ninth I International Conference on Quality Software. IEEE (2009), doi:10.1109/QSIC.2009.43 278 2009, 1550-6002/09 $26.00
11. Moraitis, P., Spanoudakis, N.: The Gaia2Jade Process for Multi-Agent Systems Development. Applied Artificial Intelligence 20(2), 251–273 (2006)
12. Uyar, A., Wu, W., Bulut, H., Fox, G.: An Integrated Videoconferencing System for Heterogeneous Multimedia Collaboration. Department of Electrical Engineering and Computer Science. Syracuse University Community Grid Lab, Indiana University (2003)
13. Huang, W., El-Darzi, E., Jin, L.: Extending the Gaia Methodology for the Design and Development of Agent-based Software Systems. In: 31st Annual International Computer Software and Applications Conference, COMPSAC 2007. IEEE (2007) 0-7695-2870-8/07 $25.00
14. Wooldridge, M., Jennings, N.R., Kinny, D.: The Gaia Methodology for Agent-Oriented Analysis and Design. Journal of Autonomous Agents and Multi-Agent Systems 3(3), 285–312 (2000)
15. Zambonelli, F., Jennings, N.R., Wooldridge, M.: Developing Multiagent Systems: The Gaia Methodology. ACM Transactions on Software Engineering Methodology 12(3), 317–370 (2003)

Survey of Computer-Aided Diagnosis of Thyroid Nodules in Medical Ultrasound Images

Deepika Koundal[1,3], Savita Gupta[2,3], and Sukhwinder Singh[2,3]

[1] Research scholar
[2] Professor (CSE)
[3] UIET, Panjab University
Chandigarh, India

Abstract. In medical science, diagnostic imaging is an invaluable tool because of restricted observation of the specialist and uncertainties in medical knowledge. A thyroid ultrasound is a non-invasive imaging study used to understand the anatomy of thyroid gland which is not possible with other techniques. Various classifiers are used to characterize thyroid nodules into benign/malignant based on the extracted features to make correct diagnosis. Current classification approaches are reviewed with classification accuracy for thyroid ultrasound image applications. The aim of this paper is to review existing approaches for the diagnosis of Nodules in thyroid ultrasound images.

Keywords: Thyroid Nodule, TIRADS, Ultrasound Images, Computer-Aided Diagnosis, Feature extraction, Classification.

1 Introduction

Thyroid nodules are swells that appear in the thyroid gland and can be due to the growth of thyroid cells. The nationwide relative frequency of thyroid cancer among all the cancer cases is 0.1%-0.2%. As per this statistics, it is concluded that thyroid related cancer is a serious disease which can lead to death, with increasing incidence rates every year. Hence, early detection is important for effective diagnosis. For diagnosing thyroid diseases, Ultrasound (US) and Computer Tomography (CT) are two of the most popular imaging modalities. US imaging is inexpensive, non-invasive and easy to use. However, US image contains echo perturbations and speckle noise, which could make the diagnostic task harder. The boundary of malignant tumor often merged with the surrounding tissues. Therefore, some Computer-Aided Diagnosis (CAD) system is necessary to increase reliability and reduce invasive operations in order to delineate nodules, classifying benign/malignant and estimating the volumes of thyroid tissues.

Generally, CAD systems are consisting of various stages like pre-processing, segmentation, feature extraction and classification. The boundaries of the tumors in US images are unclear and hard to distinguish due to artifacts such as speckle, reverberation echo, acoustic shadowing and refraction. Thus, it is necessary to suppress speckle noise before segmentation. Image segmentation plays an important role for automatic

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 459–467.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

delineation of important regions used for analyzing anatomical structure, tissue types and pathological regions. Accuracy of segmentation is important because many crucial features for discriminating benign and malignant lesions are based on the contour, shape and texture of the lesion. These features can be effectively extracted after the lesion boundary is correctly detected. Thus, an accurate segmentation method is essential for a correct diagnosis. Segmentation of thyroid nodule in ultrasound images is given in [9-12].

The rest of this paper is organized in following sections. Section 2 describes the Feature Extraction and Selection for identification of thyroid nodules in ultrasound images. The classification techniques are explained in Section 3 and conclusions are given in Section 4.

2 Feature Extraction and Selection

Feature extraction is used to find a feature set of tissue that can accurately distinguish lesion/non-lesion or benign/malignant. Recently, various feature extraction methods were proposed from which lot of features from medical images can be obtained. However, it is difficult to select significant features from the extracted features. There is no single feature that can accurately determine whether a nodule is benign or malignant. In addition to features that can be derived from the inside of the nodule, the tissue texture around the margin of the nodule is also important. The growth of malignant tumors tends to distort the surrounding tissue texture, while benign nodules tend to have smooth surfaces with more uniform texture around them. Different shapes and margins have different likelihoods of malignancy. Thus, texture features have the potential to capture characteristics that are diagnostically important but are not easily visually extracted. Feature selection is a process of feature reduction by removing irrelevant, redundant or noisy data and has an immediate effect on application by accelerating the classification algorithm. A typical feature selection process consists of four basic steps: namely, subset generation, subset evaluation, stopping criterion and result validation. The feature space could be very large and complex, so extracting and selecting the most effective features are very important.

In literature, different authors have extracted different types of features from thyroid tissue. In [12], six textural features are extracted from the selected ROIs. These textural features including Haar wavelet features, homogeneity feature, histogram feature, Block Difference of Inverse Probabilities (BDIP) and Normalized Multi-Scale Intensify Difference (NMSID) will be used in the RBF neural network to classify the thyroid region. In [23], two features associated with the Rayleigh distribution parameter, four wavelet energy coefficients, four radon transform parameters are computed for each rectangular window. These features are also combined with the longitudinal middistance measure for each thyroid gland. This distance corresponds to the vertical distance measured between the borders of the thyroid at its middle section.

In [24], the features such as mean, variance, Coefficient of Local Variation Feature, Histogram Feature, NMSID Feature, and Homogeneity are extracted and are used to train the classifiers such as ELM and SVM.

Table 1. Feature Extraction Approaches For Thyroid Ultrasound Image Analysis

REF.	FEATURE EXTRACTION APPROACH
[42][43][44]	Grey Level Histogram
[45]	Muzzolini 'S Features
[42][45]	Co-Occurrence Matrix
[11]	Radon Transform
[46]	Local Binary Patterns
[47]	Fuzzy Local Binary Pattern
[7]	Mean, Variance, Coefficient Of Local Variation Feature, Histogram Feature, Normalized Multi Scale, Intensity Different NMSID Feature, Homogeneity
[48]	Statistical Pixel Level Features
[14]	Morphology And Tissue Reflectivity
[11]	Intensity And Statistical Textural Feature
[48]	Textural Features

In [46,47], texture patterns appearing in US images can be represented by a fuzzy distribution of Local Binary Patterns, referred to as Fuzzy Local Binary Patterns (FLBP) features [8]. The original approach of Local Binary Pattern (LBP) [12] has been proven to be sensitive to small variations of the pixel intensities usually caused by noise. The FLBP is an enhanced extension of the LBP approach, capable of better coping with speckle noise [8], a common characteristic of all US images [13]. In US images a substantial amount of information concerning the pathology of the examined tissue is contained in image echogenity [11]. Several studies on US medical images have been using echogenity features based on grey-level histograms (GLH) and Fuzzy grey-level histograms (FGLH). Morphological features describe the shape and the boundary regularity of each nodule and comprised several 1st order statistics of the boundary radius along with area, smoothness, concavity, and symmetry and fractal dimension [12].

In [21],a set of twenty morphological features (Mean radius, Radius entropy, Radius standard deviation, Perimeter, Area Circularity Smoothness Convex hull mean radius Concavity Number of concave points Symmetry Fractal dimension) and wavelet local maxima (first order histogram, Mean value, entropy, central moment 3rd degree, kurtosis, skewness, variance, standard deviation) are extracted from segmented nodule. The various feature extraction approaches for thyroid nodule in ultrasound images are summarized in Table 1.

3 Classification

The suspicious regions will be classified as lesion/non-lesion or benign/malignant based on the selected features by various classification methods. The Thyroid Imaging Reporting and Data System (TIRADS) is a standardized US characterization and reporting data system of thyroid lesions for clinical management. The TIRADS is based on the concepts of the Breast Imaging Reporting Data System (BIRADS) of the American College of Radiology [2]. The categories are as follow: TIRADS 1: normal thyroid gland, TIRADS 2: benign conditions (0% malignancy), TIRADS 3: probably benign nodules (< 5% malignancy), TIRADS 4: suspicious nodules (5–80% malignancy rate). A subdivision into 4a (malignancy between 5 and 10%) and 4b (malignancy between 10 and 80%) was optional, TIRADS 5: probably malignant nodules (malignancy>80%), TIRADS 6: included biopsy proven malignant nodules.

There are different neural networks used in image segmentation such as Back Propagation neural network, Hopfield neural networks and Self-Organizing Maps (SOM). Various previous studies based on classifiers used to identify the malignancy in the thyroid lesion are mentioned in [15,16]. Many machine learning techniques such as Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and Artificial Neural Network (ANN) have been studied for thyroid lesion classification. The classification accuracy of various classifiers for thyroid nodule in ultrasound images are summarized in Table 2.

Table 2. Accuracy of Thyroid Classifiers

Publication year	Ref.	Method	Accuracy (%)
1984	[42]	FA FA+C4.5 (Pruned)	94.38
		FAFA+C 4.5 (Rules)	94.38
		Einstein	91.91
		FAF A+Einstein	93.34
1997	[40]	A Fuzzy Classifier with Ellipsoidal Regions	93.34
1997	[41]	MLP	36.74
		LVQ	81.86
		RBF	72.09
		PPFNN	78.14
1999	[39]	k-NN method	96.90
		EACH method	95.60
		RPA method	96.10
2002	[30]	3NN-Par	94.20
		FED IC-Plain	96.10
2006	[31]	EDA	98.06
		WEDA	98.00
2006	[32]	HMM method	87 .91
		SOM method	88.84
2006	[33]	LDA	93 .44
		SVM	94.44
		GPC-EP (s-soft)	96.75
		GP C-EP(m-soft)	97.23

Table 2. (*continued*)

2007	[29]	AIRS	81
		AIRS with fuzzy weighted pre-processing	85
2008	[12]	MLNN with LM	92.96
		PNN	94.43
		LVQ	89.79
2008	[27]	AIRS	94.82
		IG-AIRS	95.90
2009	[26]	PNN with GA Feature selection	96.8
		SVM with GA Feature selection	99.05
2009	[18]	BPA	92
		RBF	80
		LVQ	98
2011	[25]	GDA–WSVM Expert System	91.86
2011	[24]	SVM	84.78
		ELM	93.56
		Radon- based approach	90.9

FED IC-Plain: Feature Extraction for Dynamic Integration of Classifiers
MLP with bp: multi layer perceptron with back-propagation.
MLP with fbp: multi layer perceptron with fast back-propagation.
DIMLP: DIMLP with two hidden layers and default learning parameters.
PNN with GA: Probabilistic Neural Network with Genetic Algorithm
MLNN with LM: Multilayer neural network Levenberg–Marquardt
IG-AIRS : Information Gain based Artificial Immune Recognition System
GDA–WSVM : Generalized Discriminant Analysis and Wavelet Support Vector Machine System
PLS-QDA: Partial Least Squares Discriminant Analysis-Quadratic Discriminant Analysis
CSFNN: Adaptive Conic Section Function Neural Network
PPFNN: Probabilistic Potential Function Neural Network
ESTDD: Expert system for thyroid diseases diagnosis
HOFDA: High order Fisher discriminate analysis
C4.5-1: C4.5 with default learning parameters
C4.5-3: C4.5 with parameter c equal to 95.
EDA: Evolutionary discriminate analysis
SMC: Single-model multigroup classifiers
GPC-EP: Gaussian Process Classifier
LVQ: Learning Vector Quantizer
DPM : Decision pathway modeling
HMM: Hidden Markov Model
WEDA: Wrapped Evolutionary discriminate analysis
SOM: Self Organizing map
PWC : Pairwise Classification
RBF: Radial Basis Function
Par : Parametric Approach
C4.5-2: C4.5 with parameter c equal to 5.
FAFA: Function attribute finding algorithm
NEFCLASS-J: Neuro Fuzzy Classification
LDA: Linear Discriminant Analysis
RPA: Recursive Partition Averaging
BPA : Back propagation algorithm
ELM: Extreme Learning machine
OAC: One-Vs-All Classification

In [17] the thyroid disease are diagnosed by training a neural network on the basis of signs and symptoms that outperforms human physicians especially in the presence of noise. In [10], the potential of boundary descriptors for the assessment of thyroid nodules on US images is investigated according to malignancy risk. The diagnosis of thyroid producing thyroid disorders using ANNs is presented in [18]. The best accuracy of LVQ Network for diagnosis is 98%. In [19] RBF, Probabilistic Neural Network (PNN) and Linear Vector Quantizer (LVQ) and SVMs are used for diagnosing thyroid diseases. The overall accuracy of diagnosis system is range from near 96% to 99%. A comparative thyroid disease diagnosis realized by multilayer, probabilistic, and learning vector quantization neural networks is presented in [13]. The results showed that Probabilistic Neural Network has given the best classification accuracies for thyroid disease dataset. An approach for differentiating benign and malignant thyroid nodules based on SVM with biased penalties is presented by [20]. The results showed that the method is able to get 90.1% with the sensitivity of 93.8% and the specificity of 86.6%. In [21] a computer-based image analysis system is proposed employing the SVM classifier for the automatic characterization of 120 verified thyroid nodules. Here the accuracy of SVM in classifying the low and high risk nodules is 96.7% where QLSMD classifier is 92.5% and QB classifier is 92%. In [22], five support vector machines (SVM) were adopted to select the significant textural features and to classify the nodular lesions of thyroid.

In [8] the computational characterization of thyroid tissue is investigated using supervised classification of directionality patterns in thyroid US images. The overall classification accuracy obtained by the application of the proposed Radon-based approach was 89.4%. In [22], the thyroid disease with a new hybrid machine learning method was diagnosed. The classification accuracy of 81 % was obtained with AIRS classification system. In [14] feature selection is argued as an important problem via diagnosis and demonstrate that GAs (Genetic Algorithms) provide a simple, general and powerful framework for selecting good subsets of features leading to improved diagnosis rates. In [23], a biometric system based on features extracted from the thyroid tissue accessed through 2D US was proposed. Using leave-one-out cross-validation method the identification rate was up to 94%. In [24] an automatic system is developed that classified the thyroid images and segmented the thyroid gland using machine learning algorithms. In [25], a Generalized Discriminant Analysis and Wavelet Support Vector Machine System (GDA_WSVM) method is presented for diagnosis of thyroid diseases.

4 Conlusions

Thyroid nodules are categorized according to the pathology as the enlarged follicles, the follicular cells with follicles, the papillary cells with follicles, the follicular cells with fibrosis, the papillary cells with fibrosis and the fibrosis. Many physicians are confused about the nature of various echo patterns of thyroid nodules because of low resolution of ultrasound. Various techniques are applied by different researchers to process Thyroid US as many of the structures are hardly visible due to noise ambiguity, vagueness and uncertainty. Thus, the utilization of new and more efficient classifiers could improve the accuracy performance towards classifying thyroid nodule as benign/malignant. In order to detect the abnormal structure, intuitive ways must be found out to interpret and describe the inherent ambiguity and vagueness in the US image using (intuitionistic and neutrosophic) fuzzy set theory. Therefore, this research would definitely be an aid, even to experienced radiologists, by providing a second opinion for the characterization of nodules. Moreover, it could be used as a valuable tool in follow-up diagnosis (such as thyroid cancer) where the validity of conclusions drawn by radiologist depends on the classification accuracy. Such techniques will help to aid the diagnosis process by automatically detecting the nodules in thyroid images and consequently lead to reduction of false diagnosis related thyroid diseases. Thyroid volume estimation from the segmented thyroid region and classification of thyroid nodules based on malignancy risk factor in ultrasound images could also involve in the future work.

References

[1] Unnikrishnan, A.G., Menon, U.V.: Thyroid disorders in India: An epidemiological perspective. Indian Journal of Endocrinology and Metabolism 15, 78–81 (2011)
[2] Horvath, E., Majlis, S., Rossi, R., Franco, C., Niedmann, J.P., Castro, A.: An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management. J. Clin. Endocrinol Metab, 748–751 (2009)
[3] Baskin, H.J.: Thyroid Ultrasound and Ultrasound-Guided FNA, 2nd edn. Springer (2008)
[4] Sharma, N., Aggarwal, L.M.: Automated medical image segmentation techniques. Jnl. of Medical Physics / Association of Medical Physicists of India 35(1), 3–14 (2010)
[5] Pal, S.K.: A review on image segmentation techniques. Pattern Recg., 1277–1294 (1993)
[6] Noble, J.A., Boukerroui, D.: Ultrasound image segmentation: A survey. IEEE Trans. on Medical Imaging 25, 987–1010 (2006)
[7] Ma, J., Luo, S., Dighe, M., Lim, D., Kim, Y.: Differential Diagnosis of Thyroid Nodules with Ultrasound Elastography based on Support Vector Machines. In: IEEE Int. Ultrasonics Symp. Proc, pp. 1372–1375 (2010)
[8] Savelonas, M., Maroulis, D., Iakovidis, D., Karkanis, S.: A VBAC Model for Automatic Detection of Thyroid Nodules in Ultrasound Images, pp. 1–4. IEEE (2005)
[9] Savelonas, M.A., Iakovidis, D.K., Dimitropoulos, N., Maroulis, D.: Computational Characterization of Thyroid Tissue in the Radon Domain. In: IEEE International Symposium on Computer-Based Medical Systems, pp. 1–4 (2007)
[10] Savelonas, M.A., Maroulis, D.E., Iakovidis, D.K., Dimitropoulos, N.: Computer-Aided Malignancy Risk Assessment of Nodules in Thyroid US Images Utilizing Boundary Descriptors. In: Panhellenic Conf. on Informatics, pp. 156–160. IEEE (2008)
[11] Savelonas, M.A., Iakovidis, D.K., Legakis, I., Maroulis, D.: Active Contours Guided by Echogenicity and Texture for Delineation of Thyroid Nodules in Ultrasound Images. IEEE Transactions on Information Technology in Biomedicine 13, 519–527 (2009)
[12] Chang, C., Lei, Y., Tseng, C., Shih, S.: Thyroid Segmentation and Volume Estimation in Ultrasound Images. In: IEEE Int. Conf. on Systems, Man and Cybernetics, pp. 3442–3447 (2008)
[13] Temurtas, F.: A comparative study on thyroid disease diagnosis using neural networks. Expert Systems with Applications 36, 944–949 (2009)
[14] Saiti, F., Naini, A.A., Shoorehdeli, M.A., Teshnehlab, M.: Thyroid Disease Diagnosis Based on Genetic Algorithms using PNN and SVM, pp. 1–4. IEEE (2009)

[15] Cai, J., Liu, Z.Q.: Pattern recognition using Markov random field models. Patt. Recog., 725–733 (2002)

[16] Cheng, H.D., Shan, J., Ju, W., Guo, Y., Zhang, L.: Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recognition 43, 299–317 (2010)

[17] Zhang, G., Berardi, V.L.: An investigation of neural networks in thyroid function diagnosis. Health Care Management Science 1, 29–37 (1998)

[18] Shukla, A., Kaur, P., Tiwari, R., Janghel, R.R.: Diagnosis of Thyroid Disorders using Artificial Neural Networks. In: IEEE Int. Advance Computing Conf., pp. 1016–1020 (2009)

[19] Rouhani, M., Mansouri, K.: Comparison of several ANN architectures on the Thyroid diseases grades diagnosis. In: Int. Comp. Science and IT- Spring Conf., pp. 526–528. IEEE (2009)

[20] Ma, J., Luo, S., Dighe, M., Lim, D., Kim, Y.: Differential Diagnosis of Thyroid Nodules with Ultrasound Elastography based on Support Vector Machines. In: IEEE Int. Ultrasonics Symp. Proc., pp. 1372–1375 (2010)

[21] Tsantis, S., Dimitropoulos, N., Cavouras, D., Nikiforidis, G.: A hybrid multi-scale model for thyroid nodule boundary detection on ultrasound images. Computer Methods and Programs in Biomedicine, 86–98 (2006)

[22] Polat, K., Sahan, S., Gunes, S.: A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis. Expert Systems with Applications 32, 1141–1147 (2007)

[23] Seabra, J.C.R., Fred, A.L.N.: Towards the Development of a Thyroid Ultrasound Biometric Scheme Based on Tissue Echo-morphological Features. In: Fred, A., Filipe, J., Gamboa, H. (eds.) BIOSTEC 2009. CCIS, vol. 52, pp. 286–298. Springer, Heidelberg (2010)

[24] Selvathi, D., Sharnitha, V.S.: Thyroid Classification and Segmentation in Ultrasound Images Using Machine Learning Algorithms. In: Proc. of Int. Conf. on Signal Processing, Communication, Computing and Networking Technologies, pp. 836–841. IEEE (2011)

[25] Dogantekin, E., Dogantekin, A., Derya, A.: An expert system based on Generalized Discriminant Analysis and Wavelet Support Vector Machine for diagnosis of thyroid diseases. Expert Systems with Applications 38, 146–150 (2011)

[26] Kodaz, H., Seral, O., Arslan, A., Salih, G.: Medical application of information gain based AIRS: Diagnosis of thyroid disease. Expert Systems with Applications, 3086–3092 (2009)

[27] Keles, A.: ESTDD: Expert system for thyroid diseases diagnosis. Expert Systems with Applications, 242–246 (2008)

[28] Polat, K., Sahan, S., Gunes, S.: A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis. Expert Systems with Applications 32, 1141–1147 (2007)

[29] Pechenizkiy, M., Tsymbal, A., Puuronen, S., Patterson, D.W.: Feature extraction for dynamic integration of classifiers. Fundamenta Informaticae 77(3), 243–275 (2007)

[30] Sierra, A.: High order Fisher's discriminants. Pattern Recogn. 35, 1291–1302 (2002)

[31] Sierra, A., Echeverria, A.: Evolutionary Discriminant Analysis. IEEE Trans. Evolutionary Computation 10(1) (February 2006)

[32] Hassan, R., Nath, B., Kirley, M.: A data clustering algorithm based on single hidden markov model. In: Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 57–66 (2006)

[33] Kim, H.C., Ghahramani, Z.: Bayesian Gaussian process classification with the EM-EP algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12), 1948–1959 (2006)

[34] Pasi, L.: Similarity classifier applied to medical data sets. In: Int. Conf. on Soft Computing, 10 Sivua, Fuzziness in Finland 2004, Helsinki, Finland & Gulf of Finland & Tallinn, Estonia (2004)

[35] Myles, A.J., Brown, S.D.: Decision pathway modeling. Jnl. of Chemometrics, 286–293 (2004)

[36] Raymer, M.L.: Knowledge Discovery in Biological Dataset Using a Hybrid Bayes classifier/Evolutionary Algorithm. IEEE Trans. on Bioinformatics and Bioengineering (2003)

[37] Ozyılmaz, L., Yıldırım, T.: Diagnosis of thyroid disease using artificial neural network methods. In Proc.of Int. Conf. on Neural Information Processing 4, 2033–2036 (2002)

[38] Duch, W., Adamczack, R., Grabczewski, K.: A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. IEEE Trans. Neural Net. 12, 277–306 (2001)

[39] Cheong, T.S., Yoon, C.H.: A memory based class recursive partition averaging. IEEE Tencon, 1038–1041 (1999)

[40] Abe, S., Thawonmas, R.: A fuzzy classifier with Ellipsoidal regions. IEEE Trans. Fuzzy Syst. 5(3) (August 1997)

[41] Serpen, G., Jiang, H., Allred, L.: Performance analysis of probabilistic potential function neural network classifier. In: Proc. of Artificial Neural Netw. in Engg. Conf., vol. 7, pp. 471–476 (1997)

[42] Mailloux, G.C., Bertranti, M.: Texture Analysis Of Ultrasound B-Mode Images By Segmentation. Ultrasonic Imaging 6, 262–277 (1984)

[43] Morifuji, H.: Analysis of ultrasound B-mode histogram in thyroid tumors. Nippon Geka Gakkai Zasshi 90(2), 210–221 (1989)

[44] Hirning, T., Zuna, I., Schlaps, D.: Quantification and classification of echographic findings the thyroid gland by computerized b-mode texture analysis. Eur. J. Radiol. 9, 244–247 (1989)

[45] Smutek, D., Šara, R., Sucharda, P., Tjahjadi, T., Švec, M.: Image texture analysis of sonograms in chronic inflammations of thyroid gland. Ultrasound Med. Biol. 29, 1531–1543 (2003)

[46] Keramidas, E.G., Iakovidis, D.K., Maroulis, D., Dimitropoulos, N.: Thyroid Texture Representation via Noise Resistant Image Features. In: IEEE Int. Symp. on Comp. Based Med. Sys., pp. 560–565 (2008)

[47] Keramidas, E.G., Maroulis, D., Iakovidis, D.K.: TND: A Thyroid Nodule Detection System for Analysis of Ultrasound Images and Videos. J. Med. Syst., 1–11 (2010)

[48] Chen, Y., Hou, C., Lee, M., Chen, S., Tsai, Y., Hsu, T.: The Image Feature Analysis for Microscopic Thyroid Tissue Classification. In: 30th Annual Int. IEEE EMBS Conf., 4059–4062 (2008)

A Brief Review of Data Mining Application Involving Protein Sequence Classification

Suprativ Saha and Rituparna Chaki

Department of Computer Sc. and Engineering
West Bengal University of Technology
Saltlake, Kolkata, India
{reach2suprativ,rituchaki}@gmail.com

Abstract. Data mining techniques have been used by researchers for analyzing protein sequences. In protein analysis, especially in protein sequence classification, selection of feature is most important. Popular protein sequence classification techniques involve extraction of specific features from the sequences. Researchers apply some well-known classification techniques like neural networks, Genetic algorithm, Fuzzy ARTMAP, Rough Set Classifier etc for accurate classification. This paper presents a review is with three different classification models such as neural network model, fuzzy ARTMAP model and Rough set classifier model. A new technique for classifying protein sequences have been proposed in the end. The proposed technique tries to reduce the computational overheads encountered by earlier approaches and increase the accuracy of classification.

Keywords: Data Mining, Neural Network Model, Fuzzy ARTMAP Model, Rough Set Classifier.

1 Introduction

The introduction of new technologies such as computers, satellites and many others has lead to an exponential growth of collected data in many areas. Traditional data analysis techniques often fail to process large amounts of data efficiently. In this case data mining technology can be used to extract knowledge from large amount of data. Recently, the collection of biological data like protein sequences, DNA sequences etc. is increasing at explosive rate due to improvements of existing technologies and the introduction of new ones such as the microarrays. So Data mining technique is used to extract meaningful information from the huge amount of biological data sequences, such as the DNA, protein etc. One important area of research is to classify protein sequences into different families, classes or sub classes.

Classification is the most important technique to identify a particular character or a group of them. Different classification methods or algorithms have been proposed by different researchers to classify the protein sequences. The Protein sequence consists of twenty different amino acids which are arranged in some specific sequences. Popular protein sequence classification techniques involve extraction of specific features from the sequences. These features depend on the structural and functional properties

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 469–477.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

of amino acids. These features are compared with their predefined values. Using neural networks, Genetic algorithm, Fuzzy ARTMAP, Rough Set based Classifier etc till date; none of researchers have achieved 100% accuracy level. This paper presents a comprehensive study of the on-going research on protein sequence classification followed by a comparative analysis.

The rest of the paper is organized as follows: Section 2 presents a review of classification models; section 3 consists of a comparative analysis, followed by a proposed work in section 4. Section 5 presents the conclusion.

2 Review

Different classification techniques have been used to classify protein sequence into its particular class, sub class or family. All these methods aim to extract some features, match the value of these features and finally classify the protein sequence. This paper focuses on mainly three types of classification techniques, the (i) Neural network Model, (ii) the Fuzzy ARTMAP Model, and (iii) the Rough Set Classifier.

2.1 Neural Network Model

Generally there are different types of approaches available for classification, such as decision trees and neural networks. Extracted features of protein sequences are distributed in a high dimensional space with complex characteristics, which is difficult to satisfy model using some parameterized approaches. So neural network based classifier have been chosen to classify protein sequence. Decision tree based techniques fails to classify patterns with continuous features especially as the number of attributes is larger.

Neural network model [2] has been used to classify unknown protein sequences by extracting some features from it which can apply as input of this model. 2-gram encoding method and 6-letter exchange group methods were used to find global similarity. For local similarity, user defined variables Len, Mut, and occur were used. Minimum description length (MDL) principle was also used to calculate the significance of motif. Some predefined values of these features were used as intermediate layers or hidden layers of the neural network. This model produces 90% to 92% accuracy.

In [1] authors want to classify the protein sequences using neural network model. Here n-gram encoding method ($n = 2, 3\ldots N$ and $N =$ len. of the input sequence) was used to extract feature which was applied to construct the pattern matrix. At the end by using neural network model new pattern was matched with the predefine pattern of protein super families or families. N-gram encoding method includes all 2-gram, 3-gram, etc encoding method, so to form the pattern matrix of features extracted from n-gram encoding method, individual also needed. Therefore in case of large sequences computational overhead also be increase. The accuracy level remains 90% only.

[7] Proposes an advance technique of [1]. At first 2-gram encoding method is applied and using only this result pattern matrix is build. If this matrix is unable to classify the input protein sequence, result of 3-gram encoding method is added to the pattern matrix. The result is then matched using neural network. The performance of this technique is largely dependent on the number of encoding operations performed.

In case all the sub patterns are to be checked, performance deteriorates sharply. The average performance is slightly better than [1].

In [9] authors used a probabilistic neural network model. The paper uses self organized map (SOM) network. The SOM networks can be used to discover relationships within a set of protein sequences by clustering them into different groups. Different types of features like Amino acid distribution, 2-gram amino acid distribution, etc, were extracted from the input protein sequence to construct the pattern matrix. According to the unsupervised learning method of neural network input sequences are placed in the 1^{st} layer of neural network, then pattern matrix is presented in the hidden layer (2^{nd} layer) for matching with some predefine values. Different outcome results are summarized in the 3^{rd} layer. 4^{th} or final layer of the probabilistic neural network model produced the final result of classification. The technique failed to produce impressive results in case of unclassified_p and unclassified_n parameters. The use of SOM network also causes hindrance in interpreting the results.

The main limitations of SOM networks for protein sequence classification are its interpretability of the results, and the model selection. SOM is a straight forward method; there is no chance of back propagation. But to reach a particular goal and increase the accuracy level of the classification back propagation is most important technique. In back propagation based model, there is a chance to move to the previous steps.

The problems faced by the SOM based technique in [9] is overcome by back propagation neural network (BPNN) technique in [4]. Here authors use extreme learning machine to classify protein sequence. This extreme learning machine included the advancement of back propagation technique of neural network model. To evaluate the performance of this machine authors extracted some features like 2-gram encoding method and 6-letter exchange method from the input protein sequence. A pattern matrix was formed using those features and used in the extreme learning machine. Finally accuracy level also is measured.

The use of neural network technique normally neural network is good at handling non-linear data (noise data). The protein sequence being linear, use of neural network does not add up. It has been observed that sequences of 20 different amino acids (Protein sequences) were used as working data in this paper. The data being linear, the use of neural network modelling fails to add any extra benefit. The paper fails to take care of noise in protein sequence even through it uses neural network. The model failed to process the physical relationships which are most important in this purpose. Again regarding the accuracy issue, neural network model provide 90%-92% accuracy. Improvement of this accuracy is mostly needed.

2.2 Fuzzy Artmap Model

Generally Fuzzy ARTMAP model, a machine learning method is used to classify the protein sequence. The basic difference between neural network model and fuzzy model is that neural network model do not analysis the data individually, it only provide a knowledge based information. On the other hand Fuzzy model calculates the membership value of every feature using membership functions and implements it in the whole model.

This model [5] was implemented to classify the unknown protein sequence into different predefine protein families or protein sub families with 93% high accuracy. A cleaning process was also been conducted on the databases. After that different features were extracted from protein sequences, e.g. physic-chemical properties of the sequences. They calculated the molecular weight (W) and the isoelectric point (pI) of the protein sequences, followed by Amino acid composition of the sequences. The hydropathy composition (C), the hydropathy distribution (D) and the Hydropathy transmission (T) also calculated. After extracting all forty different features an unknown protein sequence was used as the input of the Fuzzy ARTMAP model. Some predefine features which were extracted or generated from known protein sequence also used as the unit of classification rules. This model generated the name of family or sub family of the unknown protein sequence as the output which was taken as the input of the Fuzzy ARTMAP model.

In [6] author wants to classify protein sequence using Fuzzy model. Calculating the membership value using the membership function is most important in fuzzy model. At first feature is extracted using 6-letter exchange group method. Then membership value is assigned and constructs the pattern matrix. Using a fuzzy rule pattern matrix was distributed into 3 small groups (i) small, (ii) medium and (iii) large. Now according to the target, choose a group and further distribution was done to reach to goal. At the end, the model is tested using uniport 11.0 dataset which contain globin, kinase, ras and trypsin super families of protein. In this paper number of antecedent variables is huge. It is right that increase of antecedent variable, increase the classification accuracy but it also increases the CPU time.

[8] Proposes an advancement of the techniques proposed in [6]. This technique tries to decrease the CPU time without changing the classification accuracy. Here features also extracted using 2-gram encoding method and 6-letter exchange group method and according to the membership value of features pattern matrix was distributed into 3 small groups. But executing the distribution method a new algorithm is applied on the value of features. This algorithm provides a rank on the value of the features using the feature ranking algorithm and according to the rank features is arranged in descending order. Now collect the top ranked features to construct the pattern matrix. In this way this technique can reduce the CPU overhead. This is a normal, easy, human understandable and alignment – independent method. As a result every biologist can easily understand this method and feel free to implement it. At the end this method is evaluated and compared to the non fuzzy technique (C 4.5). The computational complexity is reduced, but the accuracy level remains the same as the earlier method [6].

Fuzzy modeling helps in the data analysis although storage and time requirement are high. The construction of fuzzy sets, for every iteration adds up to the computational complexity as well. This model also failed to process the physical relationships which are most important in this purpose.

2.3 Rough Set Classifier

Generally machine learning methods such as the neural network model, Fuzzy ARTMAP model etc., are insufficient to handle large number of unnecessary features, extracted for rule discovery [11]. As a result they try to select the features to reduce the computation time. But these methods also degrade their performance. Accuracy level is not sufficient since every feature is equally important for proper classification. Rough set classifier is a new model to overcome this problem.

Rough sets theory is a machine learning method, which is introduced by Pawlak [3] in the early 1980s. It implements the concept of set theory to make some decision. The indiscernibility relation that induced minimal decision rules from training examples is the important notation in rough set model. To identify the minimal set of the features, if-else rule is used on the decision table.

This new classification model [10] can classify the voluminous protein data based on structural and functional properties of protein. This model is faster, accurate and efficient classification tool than the others. Rough Set Protein Classifier provides 97.7% accuracy. It is a hybridized tool comprising Sequence Arithmetic, Rough Set Theory and Concept Lattice. It reduces the domain search space to 9% without losing the potentiality of classification of proteins. An innovative technique viz., Sequence Arithmetic (SA) to identify family information and utilize it for reducing the domain search space is proposed. Rules are generated and stored in Sequence Arithmetic database. A new approach to compute predominant attributes (approximate reducts) and use them to construct decision tree called Reduce based Decision Tree (RDT) is proposed. Decision rules generated from the RDT are stored in RDT Rules Database (RDTRD). These rules are used to obtain class information. The infirmity of RDT is overcome by extracting spatial information by means of Neighbourhood Analysis (NA). Spatial information is converted into binary information using threshold. It is utilized for the construction of Concept Lattice (CL). The Associated Rules from the CL are stored in Concept lattice Association Rule Database (CARD). Further, the domain search space is confined to a set of sequences within a class by using these Association Rules. Time complexity of this model is $O(n) + O(f) + O(\log C) + O(2^r)$, where the unknown sequence y with size n. No of the families in the database is f. ‘C’ is the number of classes in a given family and ‘r’ is the number of proteins in classes then the CL will have 2^r nodes.

In [11] authors use rough set classifier to extract all the features necessary for classification. The feature set was built based on compositional percentages of the 20 amino acids properties. The authors had used Rosetta system for data mining and knowledge discovery. In the first phase, a method is implemented on the whole datasets in which all the subfamilies were included ignoring the small size of sequences. Rough set model generally use standard Genetic Algorithms. The Rough Set was further applied to classify the data and evaluate the performance of the seven subfamilies. This paper achieves a satisfactory accuracy level without increase the computational time.

The Rough Set Classifier model provides knowledge based information only without any analysis of data. For properly classifying protein sequences, both play an important role. Instead of classifying protein sequence into classes or sub classes, this model provides a small known sequence from a long unknown protein sequence. Thus it requires extra time and space for further classification of the output sequence into classes or sub classes. The accuracy level is 97.7%, which leaves scope for improvement.

3 Comparative Analysis

Techniques	Neural Network based Classifier [1,2,4,7,9]	Fuzzy ARTMAP based Classifier [5,6,8]	Rough Set based Classifier [3,10,11]
Database Uses	The Int. Protein Seq. Database Release 62	i) SCOP 1.69 ii) ASTRAL 1.69	NCBI (Blast)
Features Selection	***1) Global similarity*** i) 2-gram encoding method ii) 6-letter exchange group methods. ***2) Local similarity*** i) Len, Mut, and occur calculation. ii) Min. description length (MDL) principle	i) Molecular weight (W) ii) Isoelectric point (pI) iii) Hydropathy composition (C) iv) Hydropathy distribution (D) v) Hydropathy transmission (T)	i) Sequence Arithmetic ii) Reduce based Decision Tree (RDT) iii) Neighbourhood Analysis (NA) iv) Concept Lattice (CL)
Accuracy Level	90% to 92%	93%	97.7%
Drawbacks	i) Better for Nonlinear and Noisy data. ii) Does not handle Physical relationship.	i) Concerned only about the physical Structure of AA. ii) Does not handle Physical relationship.	i) No analytical output. ii) Need Extra Time and Space

4 Proposed Model

The main purpose of this model is to classify the unknown protein sequence in to different families, classes or sub classes with high accuracy level and low computational time. To implement this goal choose a unknown protein sequence as an input and extracts some features from it and match with predefine values to classify the sequence into classes, sub classes and families. To do this first of all, extraction of features which is used to classify, is very important.

The proposed technique consists of three phases. The first phase aims to reduce the input dataset. The second phase helps to increase accuracy level of classification and the third phase implies the association rule to classify the protein sequence. Figure 1 gives a pictorial representation of the different modules of the proposed technique.

4.1 Phase 1

2-gram encoding method and 6-letter exchange group method both are used to extract the global similarity of the protein sequence [2]. These two methods are directly related to the structure of a protein sequence. So, to extract the knowledge based information, we have to calculate the global similarity of protein sequence. In the first phase if we extract the knowledge based information using the back propagation technique of the neural network [4] then we are able to reduce the total dataset. So those two methods are performed at first to construct the pattern matrix. Now this pattern matrix acts as the input of the neural network model. Here we use only two techniques to extract the feature, so number of the features is very less. In this situation it is right that we do not reach the require accuracy but we can be able to reduce the total dataset for further classification within low computational time.

4.2 Phase 2

In the second phase, after reduce the dataset, Molecular weight, isoelectric point, Hydropathy composition, Hydropathy distribution, Hydropathy transmission will be calculated to extract the features.. After extracting those features, a feature ranking algorithm will be applied on it. This algorithm will be able to provide a rank to the feature values and arrange the features values according to the descending order [8]. Top rank means have an extra ability to classify the protein sequence. After that pattern matrix will be generated using top ranked in the second highest level of reduce based decision tree. Now this pattern matrix will be distributed within three groups [6] and applied to the Fuzzy model. Those features which are used here generally deal with the molecular structure of the amino acids. So it is possible to eliminate huge no of classes, sub classes and families of protein in which the input protein sequence does not belongs. In this case data by data analysis was implemented instead of extracting knowledge based information. If our dataset is huge then data by data analysis takes huge computational time. But here our data set is small because in the first stage we are able to reduce the data set. The main advantage of the data by data analysis is it provides the high accuracy level.

4.3 Phase 3

In the third and final phase, Neighbourhood Analysis (NA) will be used to classify the input sequence in the particular class or family. To use neighbourhood analysis we generally apply association rule. This rule has a power to extract the particular association between the protein sequence and classes, sub classes and families. So it is possible to eliminate all other classes, sub classes and families of protein in which this input protein sequence do not belongs.

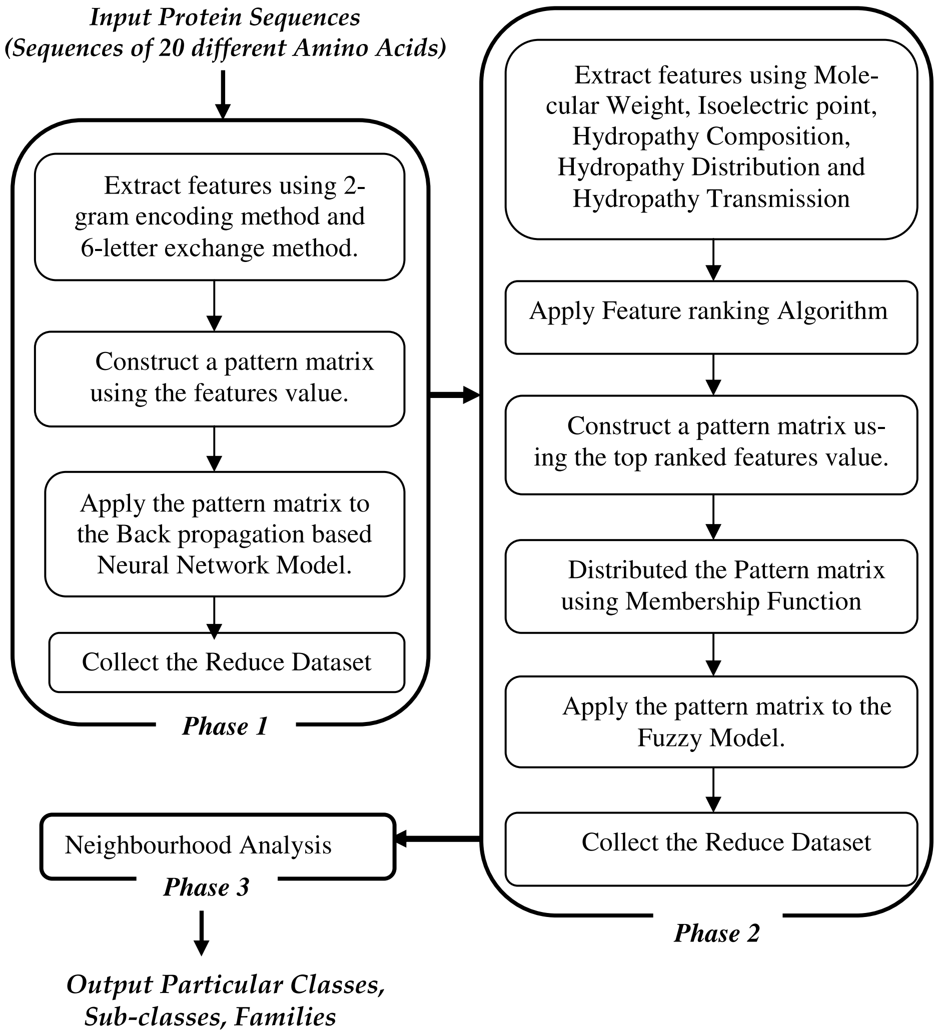

Input Protein Sequences
(Sequences of 20 different Amino Acids)
Extract features using 2-gram encoding method and 6-letter exchange method.
Construct a pattern matrix using the features value.
Apply the pattern matrix to the Back propagation based Neural Network Model.
Collect the Reduce Dataset
Phase 1
Extract features using Molecular Weight, Isoelectric point, Hydropathy Composition, Hydropathy Distribution and Hydropathy Transmission
Apply Feature ranking Algorithm
Construct a pattern matrix using the top ranked features value.
Distributed the Pattern matrix using Membership Function
Apply the pattern matrix to the Fuzzy Model.
Collect the Reduce Dataset
Phase 2
Neighbourhood Analysis
Phase 3
Output Particular Classes, Sub-classes, Families

Fig. 1. Pictorial representation of the different modules of the proposed technique

5 Conclusion

In the recent treads, analysis the large amount of biological data like protein sequences is very difficult using traditional database system. In this case data mining technique can be used to classify the unknown protein sequence. But the different models, which are used to classify the protein sequence is not perfect regarding the both accuracy level and computational time. This dissertation includes a detail review of ongoing research work involving three different techniques to classify the protein sequences. It has been observed that knowledge based and analysis of data form integral parts of protein sequence classification. The accuracy level of each proposed model has been studied. Finally, a new classification model have been proposed which can classify the unknown protein sequences into families, classes or sub classes, producing knowledge based information beside data analysis technique. In future different analysis will be done with this new proposed classification model.

References

1. Wu, C., Berry, M., Shivakumar, S., Mclarty, J.: Neural Networks for Full-Scale Protein Sequence Classification: Sequence Encoding with Singular Value Decomposition 21, 177–193 (1995)
2. Wang, J.T.L., Ma, Q.H., Shasha, D., Wu, C.H.: Application of Neural Networks to Biological Data Mining: A case study in Protein Sequence Classification. In: KDD, Boston, MA, USA, pp. 305–309 (2000)
3. Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z.: SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Research 31, 3692–3697 (2003)
4. Wang, D., Huang, G.-B.: Protein Sequence Classification Using Extreme Learning Machine. In: Proceedings of International Joint Conference on Neural Networks, IJCNN 2005, Montreal, Canada (2005)
5. Mohamed, S., Rubin, D., Marwala, T.: Multi-class Protein Sequence Classification Using Fuzzy ARTMAP. In: IEEE Conference, pp. 1676–1680 (2006)
6. Mansoori, E.G., Zolghadri, M.J., Katebi, S.D., Mohabatkar, H., Boostani, R., Sadreddini, M.H.: Generating Fuzzy Rules For Protein Classification. Iranian Journal of Fuzzy Systems 5(2), 21–33 (2008)
7. Zainuddin, Z., Kumar, M.: Radial Basic Function Neural Networks in Protein Sequence Classification. Malaysian Journal of Mathematical Science 2(2), 195–204 (2008)
8. Mansoori, E.G., Zolghadri, M.J., Katebi, S.D.: Protein Superfamily Classification Using Fuzzy Rule-Based Classifier. IEEE Transactions on Nanobioscience 8(1), 92–99 (2009)
9. Nageswara Rao, P.V., Uma Devi, T., Kaladhar, D., Sridhar, G., Rao, A.A.: A Probabilistic Neural Network Approach For Protein Super-family Classification. Journal of Theoretical and Applied Information Technology (2009)
10. Yellasiri, R., Rao, C.R.: Rough Set Protein Classifier. Journal of Theoretical and Applied Information Technology (2009)
11. Rahman, S.A., Bakar, A.A., Hussein, Z.A.M.: Feature Selection and Classification of Protein Subfamilies Using Rough Sets. In: International Conference on Electrical Engineering and Informatics, Selangor, Malaysia (2009)
12. Tzanis, G., Berberidis, C., Vlahavas, I.: Biological Data Mining

Hybrid Algorithm for Job Scheduling: Combining the Benefits of ACO and Cuckoo Search

R.G. Babukarthik[1], R. Raju[2], and P. Dhavachelvan[1]

[1] Department of Computer Science, Pondicherry University, Pondicherry, India
{r.g.babukarthik,dhavachelvan}@gmail.com
[2] Research Scholar, Bharatiyar University, Coimbatore, India rajupdy@gmail.com

Abstract. Job scheduling problem is a combinatorial optimization problem in computer science in which ideal jobs are assigned to resources at particular times. Our approach is based on heuristic principles and has the advantage of both ACO and Cuckoo search. In this paper, we present a Hybrid algorithm, based on ant colony optimization (ACO) and Cuckoo Search which efficiently solves the Job scheduling problem, which reduces the total execution time. In ACO, pheromone is chemical substances that are deposited by the real ants while they walk. When it comes to solving optimization problems it acts as if it lures the artificial ants. To perform a local search, we use Cuckoo Search where there is essentially only a single parameter apart from the population size and it is also very easy to implement.

Keywords: Job Scheduling, Ant Colony Optimization, Cuckoo Search.

1 Introduction

Scheduling is the process of allocation of limited resources to tasks over time. The main objective of the scheduling is optimization and hence it can be considered as a decision making process. Job scheduling problem is a combinatorial optimization problem, that validates the performance of heuristic algorithms and hence it is used in the manufacturing systems. The major difficulty encountered is that not many scheduling problems fit into a common description model. This makes it very difficult to define a common framework for scheduling problems and also to find algorithms which can be applied or adopted to tackle a great variety of problems. In fact, a well working algorithm for problem A might not work for problem B inspite of slight variation from problem A.

In this paper, we present a Hybrid algorithm, based on the Ant Colony Optimization (ACO) and Cuckoo Search which efficiently solves the Job scheduling problem, which reduces the total execution time. Our approach is based on heuristic principles and has the combined advantages of ACO and Cuckoo search. The main contributions of this paper can be summarized as follows:

- The proposal of a Hybrid algorithm which combines the advantage of ACO and Cuckoo Search.
- The evaluation of the Job creation time, Task creation time, Result retrieval time and destruction time.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 479–490.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

- The performance comparison of the Hybrid algorithm and Ant colony optimization.

2 Related Work

The main objective is to find a feasible plan which reduces the completion time (makespan) and the waiting time in such a manner that the processing order of jobs for each machine is scheduled [1]. ACO heuristic of the given model for the targeted architecture reduces the execution time of the application by exploring different solutions for mapping and scheduling of tasks and communications [2]. Multi-objective task scheduling optimization model is presented using ACO algorithm [3]. Improved Ant Colony Optimal algorithms are used for searching the Pareto optimal solutions [4]. A Mixed-Integer Linear Programming (MLIP) model for the job shop scheduling problem considers the sequence-dependent setup time and the arrival time constraints [5].

Agent and multi agent based techniques can be used for the job scheduling, but for combinatorial optimization problems it is not feasible [6][7][8]. A novel framework for the estimation of particle swarm distribution algorithm is applied for the selection of the local best solutions thereby obtaining further promising individual meant for model buildings [9]. It is a newly combined modeling method based on fuzzy system and evolutionary algorithm [10]. PSO has a drawback of premature convergence [11]. Feasible schedules in job shop problems are inadmissible, semi-active, active, non-delay [12]. Various optimization methods, such as canceling the history velocity, the double perturbation of the gBest and pBest of the particle need to escape from the local optimum. They are proposed by detecting the precise time and dimension of the double perturbation and also increase the performance of the rotary chaotic particle swarm optimization [13]. A new way of extending ACO to solving continuous optimization problems is by focusing on continuous variables sampling as a key in transforming ACO from discrete optimization to continuous optimization [14]. ACO algorithm belongs to constructive method, which can build a solution for combinatorial optimization problem in an incremental way step by step without backtracking. ACO algorithm provides a solution component until a complete solution is generated [15]. It formulates the aircraft arrival sequencing and scheduling problem in the form of permutation problem and proposes a new solution framework [16]. ACO and Cuckoo search algorithm are used in many application for the performance control [17][18][19][20][21].

3 Problem Definition

The most complex combinatorial optimization problem is job scheduling. Job scheduling can be described in the following manner. We have set of n jobs which need to be operated on m machines. Each activity will begin its execution only when the preceding activities in the ordering have finished their execution. Jobs which need to be completed, will visit the machine in different sequences. Each job can be performed on m machines. The following assumptions are additionally characterized.

- Job can be performed on any machine one at a time.
- Only a single operation can be performed on a machine.

- Once an operation has begun, it should not be interrupted.
- The starting time of the job are known in advance.

4 Proposed System

In this section, we present a hybrid algorithm for job scheduling which will have the advantage of Ant Colony Optimization and Cuckoo Search. We formulate the problems as follows. We have n jobs and m machines. Each job has their own ordering of execution that needs to be performed on m machines. Each job has its own starting time. The objective of the algorithm is to minimize the completion time and resources being consumed.

Table 1. Parameters

Parameters	
Index	***Notation***
No. of Jobs	N
No. of machines	M
No. of operation of job index	i
Job creation time	J_{ct}
Remote call	R_c
Disk space	D_s
Job submission time	J_{st}
Executing job	E_j
Job destruction time	J_{dt}
Job deletion time	t_{dmf}
Task deletion information	j_{dt}
Total time	T_t

4.1 Ant Colony Optimization (ACO)

For solving the computational problems ant colony optimization algorithm is a good technique, which can minimize and find the optimal path through graphs. The environment is used as a medium for the communication of Ants. With the help of the pheromone that have been deposited, exchange of information takes place, like the status of their “work”. Since the exchange of information has a local scope, the pheromones were left as a perception for the located ants. The ant finds its food source somehow and returns to its nest leaving behind a trail of pheromone. Ants follow four possible ways to reach its destination. Once the shortest route is found, the strengthening of the runway will be made in a more attractive way. Other ants take only this shortest route. Pheromone that has been deposited in other ways will be lost. Model explaining this behavior is described in the following manner.

An ant runs randomly around the colony, if the food sources have been discovered, it returns directly to the nest. Since the pheromones that have been deposited are attractive, nearby ants will be nurtured to follow this track. Returning to the colony, further these ants will deposit the pheromone to strengthen the route. For a particular time if there happens to be more than one route to attain the same food source then, the shorter route will be chosen by the ant to travel. The short route will be further enhanced and therefore becomes more attractive since the pheromones are volatile, the longer routes will be disappeared.

The following formulae show the most common algorithms of Ant Colony Optimization philosophy:

$$Pij= (\tau_{ij}) (\eta_{ij}) /\Sigma (\tau_{ij}) (\eta_{ij})$$

Where,

τij- is the amount of the pheromone on arc I j

α-is a parameter used to control the on fluency of τij

ηij-is the desirability of arc I j (a priori knowledge, typically 1/di,j)

β - is a parameter of control the influence of ηij

$$\mathbf{Tij = \rho\tau ij+\Delta\ Tij}$$

Where,

Tij is represented as the amount of pheromone lying on a given arc ij ρ is the pheromone evaporation rate.

Δ Tij is the amount pheromone deposited, typically given by

$$\mathbf{\Delta\ T\ k\ I\ j = \quad 1/Lk}$$

if ant k travels on arc I, j 0 otherwise

```
Initialization:
          The pheromone trails, the heuristic information and the parameters are
     initialized.
Iterative Loop:
          A colony of ants determines the starting job
          Construct a complete schedule for each ant.
Repeat:
       Apply state transition rule to select the next processing job
       Apply the local updating rule.
Until a complete schedule is constructed
         Apply the local search process.
         Apply the global updating rule.
Termination:
        If the maximum number of interactions is realized,
        Then STOP
        Else go to STEP 2
```

Fig. 1. ACO Algorithm

4.2 Cuckoo Search

Cuckoo Search (CS) is an optimization algorithm. The major advantage of Cuckoo Search algorithm is its simplicity. In reality, when we compare with other population or agent-based Meta heuristic algorithms such as particle swarm optimization and harmony search. In Cuckoo Search there is essentially only a single parameter $p_{a,}$ other than the population size n. Hence, it is very easy to implement. Cuckoo search is considered to be worthy for the imitation of the breeding behavior. Thus the Cuckoo Search can be applied for various combinatorial optimization problems. It is found that it can perform better than other Meta Heuristic algorithms. Cuckoo search uses the following representation.

In cuckoo search the eggs that are present in the nest is represented as the solution. The main objective is to use a new and better solution where the cuckoo replaces the solutions which are not so good. In other words, each nest has only one egg. Further the algorithm can be extended to more complex cases in which each nest containing more than one egg denotes a set of solution.

4.2.1 Basic Results of Cuckoo Search

i. Each cuckoo can lay only one egg at a time, and it dumps it in the randomly chosen nest.
ii. For the next generation the eggs of high quality in the optimum nest will be carried out.
iii. The number of available host's nest is fixed, and the probability of egg laid by a cuckoo is found by the host bird by p_a .
iv. The discovered set of worst nests needs further calculation to the obtain the solution

4.3 Random Walk Step Size

According to Yang and Deb discovery with the help of Levy's flights, random-walk style performance of the search is compared to simple random walk. The major concern is the applications of random walks and Levy's flights in the necessary equation for generating new solution is

$$Y_{t+1} = Y_t + sE_t$$

Where E_t is worn out from a standard normal distribution mean value as zero and unity standard deviation for random walks, or drained from Levy distribution for Levy flights. Apparently, it will be complicated where the random walks can also be connected with the similarity between a cuckoo's egg and the host's egg. The size of the step s determines how far a random walker can go for a fixed number of iterations. It will be tricky during the generation of Levy step.

If the value of s is too large, when compared to the old solution the new solution produced will be too far away. In the above scenario it is doubtful to be accepted. If the value of s is too small, then there will be a significant change and hence the search will not be efficient. Hence a proper step size is necessary to maintain the search efficiently.

```
Objective function: $f(x)$, x=($x_1$,$x_2$, ……. $x_d$)
Initialization:
Number of nests, random initial solution;
Iterative loop:
Get the current best nest
While (fmin>Max generation)
Get cuckoos by random walk and then replace it solution by Levy's flights
Evaluate the quality fitness ($Fn_i$)
Randomly choose nest among n, say j
If ( $Fn_i$>$Fn_j$ )
Replace j value by the new solution
End
Fractions of the nest which is worst are discarded and new ones are built;
Maintain the best solution and nests;
Rank the best solutions and nests, discover the current best;
To the next generation pass the current best solutions;
End while
```

Fig. 2. Cuckoo Search Algorithm

```
Initialization:
The pheromone trails, the Meta heuristic information and the number of nests, random initial solution
Iterative loop:
A colony of ants determines starting jobs Construct a complete schedule for each ant:
Repeat
Apply state transition rule to select the next processing job
Apply the local updating rule
Until a complete schedule is constructed
Apply cuckoo search process
Apply the global updating rule
Termination
If the maximum number of interactions is realized, the STOP Else go the step 2
Cuckoo search process
Objective function: $f(x)$, x=($x_1$,$x_2$, ……. $x_d$)
Iterative loop:
Get the current best nest While (fmin>Max generation)
Get cuckoos by random walk and then replace it solution by Levy's flights
Evaluate the quality fitness ($Fn_i$)
Randomly choose nest among n, say j
If ( $Fn_i$>$Fn_j$ )
Replace j value by the new solution
End
Fractions of the nest which is worst are discarded and new ones are built;
Maintain the best solution and nests;
Rank the best solutions and nests, discover the current best; To the next generation pass the current best solutions; End while
```

Fig. 3. Hybrid Algorithm

4.4 Hybrid Algorithm

In this algorithm we combine the advantages of both Ant Colony Optimization and Cuckoo Search. The major problem in ACO is that while the real ant walks, a chemical substance called Pheromone is deposited. While solving the optimization problems it lures the artificial ants and hence to perform a local search, we use Cuckoo Search where there is essentially only a single parameter apart from the population size.

Hence in ACO whenever a local search is performed, Cuckoo Search is applied because it is used for optimization problems. Comparing them with other population or agent-based Meta Heuristic algorithms such as particle swarm optimization and harmony search, there is basically a single parameter in Cuckoo Search apart from the population size *n*. Therefore, it is very easy to implement.

4.5 Flow Chart

Figure 1 shows the flow chart of the hybrid algorithm. For the job scheduling problems, when the jobs have been given as the input, the corresponding resources have to allocate to perform the operation. Hence matching of resource need to take care by algorithm, thus initialization of the job and the solution construction is performed. Random search operation is performed for the job that is matching of the resources to the job done, if the solution is not feasible the levy's flight has been performed. Updation of the search operation is performed globally and after execution of the job the process is terminated.

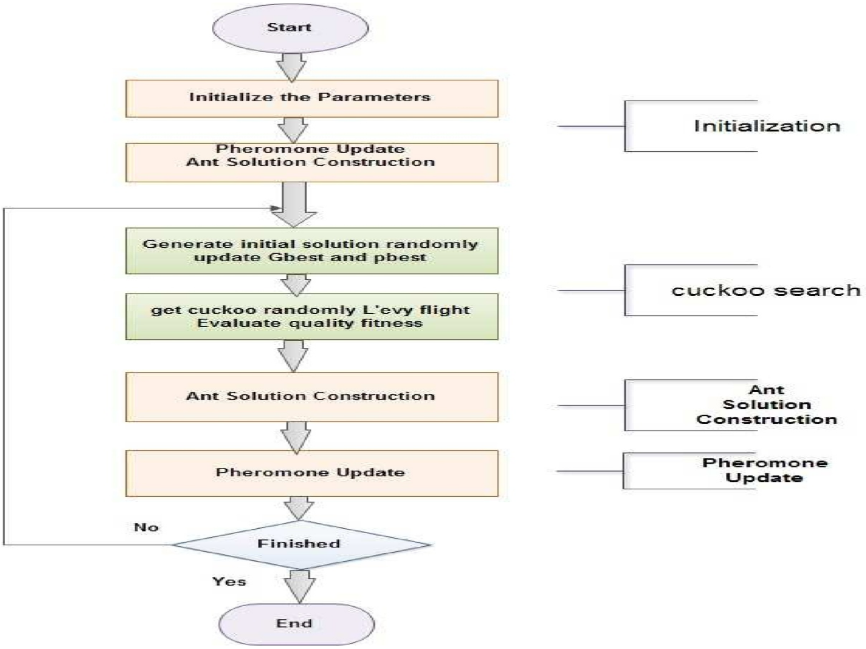


Fig. 4. Flow chart of Hybrid Algorithm

5 Experimental Results

Solving computationally demanding and data-demanding problem can be solved with the help of matlab more easily and quickly. Using parallel computing toolbox in matlab. We constructed the jobs and using workers we created 15 tasks for the corresponding jobs. And thus job creation time, tasks creation time, destruction time, result retrieval time and total execution time have been calculated. Using some of the parallel processing operation such as parallel for-loops and message-passing functions which allow us to implement task and data- parallel algorithms in matlab.

5.1 Job Creation Time

Job creation time can be stated as the time taken to create a new job. And thus for the job manager it includes the remote call and the time taken by the job manager to allocates space in its database for the operation. For the other types of schedulers, the job creation time includes writing some files to the disk.

$$J_{ct} = R_c + D_s \text{ or } J_{ct} = D_W$$

Table 2 shows the job creation time for the various tasks that has been given as the input, for the tasks 1 five job have been created; similarly for the tasks 2 and tasks 5 the corresponding jobs have been created. We took the average for the tasks1, tasks 2 and tasks 4. Figure 5 show that when the number of tasks increases the job creation time also increases with respect to the time in seconds. Hybrid algorithm shows clearly that job creation time for each and every task is a slight increase only.

5.2 Job Submission Time

Job submission time is the time taken to submit the job. Job manager, we can say to start executing the job that it has in the database. For the schedulers the time taken to execute all the tasks that have been created.

$$J_{st} = s_t \text{ or } J_{st} = E_j$$

5.3 Job Destruction Time

Job destruction time is the time to taken to completely delete the job and task information. For the job manager this includes the time taken to delete job and its associate information from the database.

$$J_{dt} = J_{dt} + t_{dinf}$$

Table 2 shows the job destruction time for the various tasks that has been given for the job that have been created , for the tasks 1 five job have been created and 5 job destruction time has been evaluated based on the time ; similarly for the tasks 2 and tasks 4 the corresponding jobs destruction time have been found. We took the average for the tasks1, tasks 2 and tasks 4. Figure 6 show that when the number of tasks increases the job destruction time also increases with respect to the time in seconds for the job that have been created. Hybrid algorithm shows clearly that job destruction time for each and every task increases gradually.

5.4 Total Time

Total time is the time taken to perform the Job creation time, Task creation time, Job submission time, Job waiting time, Task execution time, Result retrieval time, Job destruction time.

$$T_t = J_{ct} + T_{ct} + J_{st} + J_{wt} + T_{Et} + R_{rt} + J_{dt}$$

Table 3 shows the total time for the various tasks that has been given as the input, for the tasks 1 five total time have been calculated; similarly for the tasks 2 and tasks 5 the corresponding total time has been found. We took the average for the tasks1, tasks 2 and tasks 4. Figure 7 shows the performance of the Hybrid algorithm, when the number of tasks increases the total time also increases with respect to the time in seconds. Hybrid algorithm shows clearly that total time for each and every task increase, when compared with ACO algorithm. Table 3 shows that Hybrid algorithm search is performed more quickly and fastly, than ACO. There is tremendous increase in the performance when more and more task as been executed.

Table 2. Hybrid Job creation time and Hybrid Job destruction time

JOB CREATION TIME			
Sl.No	Task 1	Task 2	Task 4
1	0.0260	0.0290	0.0294
2	0.0285	0.0289	0.0265
3	0.0261	0.0254	0.0258
4	0.0269	0.0272	0.0281
5	0.0265	0.0251	0.0269
Total	0.134	0.1356	0.1367
Average	0.0268	0.02712	0.02734
JOB DESTRUCTION TIME			
Sl.No	Task 1	Task 2	Task 4
1	0.0256	0.0269	0.0262
2	0.0266	0.0263	0.0268
3	0.0255	0.0256	0.0261
4	0.0239	0.0256	0.0262
5	0.0255	0.0257	0.0260
Total	0.1271	0.1301	0.1313
Average	0.02542	0.02602	0.02626

Table 3. Hybrid Total time and Comparison of Hybrid & ACO

TOTAL TIME			
Sl.No	Task 1	Task 2	Task 4
1	4.2042	4.2083	4.2113
2	4.2021	4.2036	4.2052
3	4.2030	4.2034	4.2036
4	4.2142	4.2125	4.2380
5	4.2112	4.2122	4.2133
Total	21.0347	21.0400	21.0714
Average	4.20694	4.2080	4.21428

Hybrid and ACO algorithm			
Sl.No	Number of Tasks	Hybrid	ACO
1	1	4.20694	4.20691
2	2	4.208	4.206
3	4	4.21428	4.211
4	8	4.22055	4.2132
5	16	4.25946	4.2432
6	32	4.26580	4.2487
7	64	4.27216	4.251
8	128	4.27364	4.2538
9	256	4.2766	4.2551

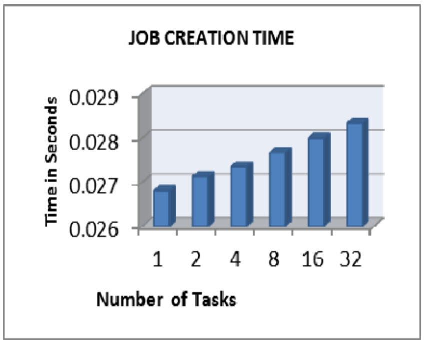


Fig. 5. Hybrid Job Creation Time

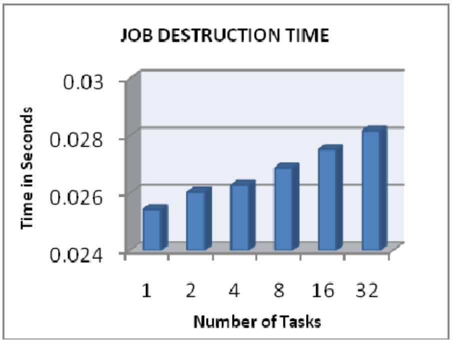


Fig. 6. Hybrid Job Destruction time

Using Hybrid algorithm the time taken for the job creation and job destruction is very minimum as the number of tasks increases, and more over it is clearly seen that for more and more tasks the execution speed of the Hybrid algorithm increases than the Ant Colony Optimization algorithm.

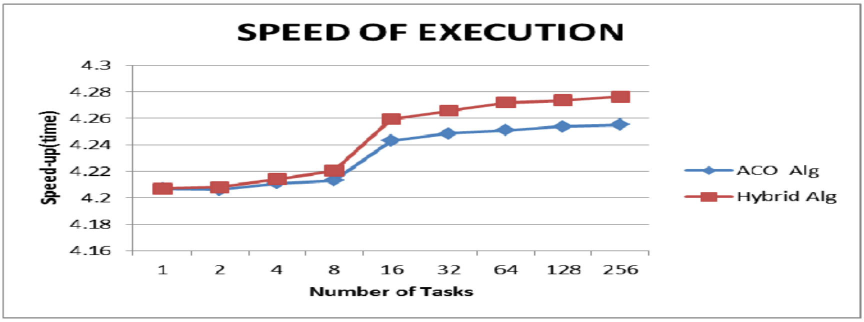


Fig. 7. Speed of Execution

6 Conclusion

Job scheduling problem is a combinatorial optimization problem, used to validate the performance of heuristic algorithms. In this paper we have proposed a new Hybrid algorithm for the job scheduling which combines the advantages of both Ant Colony Optimization and Cuckoo Search. The experimental analysis show that the performance of the algorithm is considerably increased as the number of the task increases. In future we plan to apply the Hybrid algorithm for job scheduling in the cloud computing environment for the analysis of the performance and to reduce the resources consumed.

References

1. Surekha, S.: PSO and ACO based approach for solving combinatorial Fuzzy Job Shop Scheduling. Int. J. Comp. Tech. Appl. 2(1), 112–120 (2010)
2. Ferrandi, F.: Ant Colony Heuristic for Mapping and Scheduling Tasks and Communications on Heterogeneous Embedded Systems. IEEE Transactions on Computer-Aided Design of Intergrated Circuits and Systems 29(6) (2010)
3. Guo, S., Huang, H.-Z.: Grid Service Reliability Modeling and Optimal Task Scheduling Considering Fault Recovery. IEEE Transactions on Realiability 60(1) (2011)
4. Tan, Q., Chen, H.-P.: Two-agent scheduling on a single batch processing machine with non-identical job sizes. Artificial Intelligence, Management Science and Electronic Commerce, AIMSEC (2011)
5. Azarkish, T.-M.: A new hybrid mutli-objective Pareto archive PSO algorithm for a bi-objective job shop scheduling problem. Elsevier Expert Systems with Applications (2011)
6. Dhavachelvan, P., Uma, G.V.: Multi-agent based Framework for Intra-Class Testing of Object-Oriented Software. International Journal on Applied Soft Computing 5(2), 205–222 (2005)
7. Dhavachelvan, P., Uma, G.V.: Multi-agent Based Integrated Framework for Intra-class Testing of Object-Oriented Software. In: Yazıcı, A., Şener, C. (eds.) ISCIS 2003. LNCS, vol. 2869, pp. 992–999. Springer, Heidelberg (2003)
8. Dhavachelvan, P., Uma, G.V.: Reliability Enhancement in Software Testing – An Agent-Based Approach for Complex Systems. In: Das, G., Gulati, V.P. (eds.) CIT 2004. LNCS, vol. 3356, pp. 282–291. Springer, Heidelberg (2004)

9. Ahn, C.W., An, J.: Estimation of particle swarm distribution algorithms: Combining the benefits of PSO and EDAs. Elsevier Information Sciences (2010)
10. Yang, X., Yuan, J.: An improved WM method based on PSO for electric load forecasting. Elsevier Expert Systems with Applications (2010)
11. Bae, C., Yeh, W.-C.: Elsevier Expert Expert Systems with Applications. Feature Selection with Intelligent Dynamic Swarm and Rough Set (2010)
12. Sha, Lin, H.-H.: A Multi-objective PSO for job-shop scheduling problems. Elsevier Expert Systems with Applications (2010)
13. Tao, Q., Chang, H.-Y.: A rotary Chaotic PSO algorithm for trustworthy scheduling of a grid workflow. Elsevier Computers & Operations Research (2011)
14. Hu, X.-M., Zhang, J.: SamACO: Variable Sampling Ant Colony Optimization Algorithm for Continuous Optimization. IEEE Transactions on Systems, Man, and Cybernetics 40(6) (2010)
15. Zhang, Z., Zhang, J., Li, S.: A Modified Ant Colony Algorithm for the Job Shop Scheduling Problem to Minimize Makespan. IEEE Explore (2010)
16. Zhan, Z.-H., Zhang, J.: An Efficient Ant Colony System Based on Receding Horizon Control for the Aircraft Arrival Sequencing and Scheduling Problem. IEEE Transactions on Intelligent Transaction Systems 11(2) (2010)
17. Manicassamy, J., Dhavachelvan, P.: Metrics Based Performance control Over Text Mining Tools in Bio-Informatics. In: ACM International Conference on Advances in Computing, Communication and Control, ICAC3 2009, India, pp. 171–176 (2009) ISSN: 978-1-60558-351-8
18. Manicassamy, J., Dhavachelvan, P.: Automating diseases diagnosis in human: A Time Series Analysis. In: Proceedings of International Conference and Workshop on Emerging Trends in Technology, ICWET 2010, India, pp. 798–800 (2010) ISSN: 978-1-60558-351-8
19. Victer Paul, P., Saravanan, N., Jayakumar, S.K.V., Dhavachelvan, P., Baskaran, R.: QoS enhancements for global replication management in peer to peer networks. Future Generation Computer Systems 28(3), 573–582 (2012)
20. Vengattaraman, T., Abiramy, S., Dhavachelvan, P., Baskaran, R.: An Application Perspective Evaluation of Multi-Agent System in Versatile Environments. International Journal on Expert Systems with Applications 38(3), 1405–1416 (2011)
21. Abirami, S., Baskaran, R., Dhavachelvan, P.: A survey of Keyword spotting techniques for Printed Document Images. Artificial Intelligence Review 35(2), 119–136 (2010)

Hybrid Ant Colony Optimization and Cuckoo Search Algorithm for Job Scheduling

R. Raju[1], R.G. Babukarthik[2], and P. Dhavachelvan[2]

[1] Research Scholar, Bharatiyar University, Tamil Nadu, India
rajupdy@gmail.com
[2] Department of Computer Science, Pondicherry University, Pondicherry, India
{r.g.babukarthik,dhavachelvan}@gmail.com

Abstract. Job scheduling is a type of combinatorial optimization problem. In this paper, we propose a Hybrid algorithm which combines the merits of ACO and Cuckoo Search. The major problem in the ACO is that, the ant will walk through the path where the chemical substances called pheromone is deposited. This acts as if it lures the artificial ants. Cuckoo search can perform the local search more efficiently and there is only a single parameter apart from the population size. It minimizes the makespan and the scheduling can be used in scientific computing and high power computing.

Keywords: Job Scheduling, Ant Colony Optimization, Cuckoo Search.

1 Introduction

Job scheduling problem has a combinatorial optimization problem. Job scheduling is used in compound equipment manufacturing system for authenticating the performance of heuristic algorithms. The major problem in job scheduling is that many scheduling do not fit into a common description model. Hence for scheduling problems it is too difficult to define a common frame work.

In this paper, we have proposed a Hybrid algorithm which combines the advantage of ACO and Cuckoo Search so as to solve the job scheduling problems. Job Scheduling can be used in scientific computing and high power computing for solving all the combinatorial optimizations problems. Our approach is based on heuristic principles which have the advantage of minimizing the makespan.

The main contribution of this paper is,

- The proposal of a Hybrid algorithm which combines the advantage of ACO and Cuckoo Search.
- The evaluation of the Job creation time, Task creation time, Result retrieval time and destruction time.
- The performance comparison of the Hybrid algorithm and Ant colony optimization.

2 Related Work

The predefined collection of tasks can be performed by JSSP with the allocation of required resources within a given time period of time. Combining the PSO and ACO

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 491–501.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

an approach for computational intelligence technique for solving JSSP is framed. The main objective is to minimize the makespan and the waiting time, in such a way the processing order of the job each and every machine is scheduled [1]. Power consumption of the system can be reduced using ACO with the proper allocation of tasks. When the size of the problem increases, a problem-specific optimization is used. By exploring the different solution for mapping and scheduling of the tasks and communication, using ACO it reduces the execution time of the application. Renewable and non-renewable resources are considered, for the heterogeneous embedded systems no formulation of ACO for concurrent mapping and scheduling. It is problem specific optimization.

Optimized performance is not achieved [2]. Using Local Node Fault Recovery (LNFR) the failed subtask execution can be resumed from the interrupted point which is used for fault recovery strategies in local situations. Using ACO multi objective task scheduling can be performed. LNFR can be used in life time of subtasks, number of recoveries to be performed, and grid service reliability models. Proposed a optimization model to minimize the total cost and to maximize the grid service reliability for task scheduling. Since gird has one RMS, it can be viewed as star topology which is very difficult for the complexity of grid [3]. Mobile agent based techniques can be used to avoid the failure, and further it can be improved by using the co-operative problem solving techniques [4][5]. Considered a scheduling model with two agent and batch processing machines with non-identical job. Consideration of the two agents is to minimize the completion time. Product allocation and resources distribution can be performed better with the help of improved ACO algorithm. In batch processing jobs from the same agent have their priority to scheduled, not focused on dynamic arrival of jobs [6].

Mixed-Integer Linear Programming (MILP) model for job shop scheduling solves the problem in mean weighted flow time, the sum of the weighted tardiness and the earliness costs. To find the pareto optimal solution for a given problem, they used a character of scatter search (SS) to select new swarm in each iteration. PSO has the good performance and computational cost and it is also very easy for implementation. The major disadvantage is that they used genetic operators and character of scatter search [7]. Local search can be performed efficiently with the help of PSO and global search is performed using genetic algorithm. Extended compact Particle Swarm Optimization (EcPSO) is developed by combining the benefits of genetic algorithm with PSO. Without losing the unique features they combined the PSO and Estimation distribution algorithm. Algorithm solves the deceptive and symmetric problems. Algorithm does not focus on continuous problems [8]. Wang-Mendel based on PSO is proposed. High-dimension non-linear optimization problems can be solved using PSO. A complete fuzzy rule set has been obtained through extrapolating using modified PSO algorithm to optimize the fuzzy rule centroid of the data covered area. [9]. The technique which has been used in the global replication management and in keyword matching can be used [10][11]. The major disadvantage of PSO is the premature convergence which can be solved by Intelligent Dynamic Swarm (IDS). It uses the feature selection technologies [12]. Performance of PSO is better in search quality and efficient than the traditional evolutionary heuristics [13].

Rotary Chaotic Particle Swarm Optimization (RCPSO) algorithm optimizes the scheduling performance in a multi-dimensional complex space. To optimize the grid workflow scheduling in discrete space a novel RD rule is presented to help the PSO.

In multi-dimensional complex space the scheduling performance is optimized [14]. Agent based techniques can be used for the combinatorial optimization problems [15][16][17]. SamACO is a way to sample promising variable values in the continuous domain and to use pheromones to guide ant's construction behavior. This has been developed to extend ACO to a continuous optimization [18].ACO algorithm belongs to the constructive method and hence solution for the combinatorial optimization problem is built step by step without backtracking. Further ACO algorithm adds solution components until a complete solution is generated [19]. Due to stochastic optimization nature of the ACS algorithm, only statistical conclusions can be made [20]. Efficiency of the algorithm to which it can be used in many application [21][22][23].

3 Problem Definition

For solving combinatorial optimization problem, an efficient algorithm is necessary. In this section, we have proposed a hybrid algorithm for job scheduling which will combine the advantages of ACO and Cuckoo search.

Problem can be defined as follows. We have N jobs and M machines. Each and every job has its own order of execution that has to be performed on M machines. Each job has its own starting time. The objective of this algorithm is to minimize the makespan and it can also be used for job scheduling in scientific and high power computing.

Some of the assumptions for the job scheduling problem are,

1. Jobs should be finite set.
2. Each and every job contains a series of operations that needs to be performed.
3. Machines should be finite set.
4. All the machines are capable of handling only one operation at a time.

Table 1. Parameters

	Parameters
Index	***Notation***
Input Variables	Tasks
Output variables	Processing time
No. of Tasks	32
No. of Machines	4
Task creation time	T_{ct}
Time to save disk information	t_{sinf}
Creation time	C_t

Some of the constraints are,

1. No job should visit the same machines twice.
2. No condition among operation of various jobs.
3. Preemption type of operation is not allowed.
4. A single machine is capable of handling individual job at a time.
5. No machines fail during its operation.

4 Proposed System

4.1 Hybrid Algorithm

It combines the advantages of Ant Colony Optimization and Cuckoo search. The major disadvantage in the ACO is that while trying to solve the combinatorial optimization problems the search has to performed much faster, but in ACO ant will walk through the path where the chemical substance called pheromone has been deposited. This acts as if it lures the artificial ants. Hence local search will be performing at the faster rate than in the ACO. In order to overcome the above drawback, Cuckoo search is used. Moreover in the Cuckoo search there is only a single parameter apart from the population size.

1) Scheduling Algorithm

Step 1: *Initialization - Job creation time, starting time.*
Step 2: *Find out the number of task T that need to be scheduled.*
Step 3: *Schedule the task T using a Hybrid algorithm.*
Step 4: *Assign the task T to the scheduler, using parallel computation to compute the tasks with the help of the workers.*
Step 5: *Find out the Job creation time* J_{ct}, *Task creation time* T_{ct}, *Result retrieval time* R_{rt}, *Job Destruction time* J_{dt} *and Total time* T_t.

Step 6: *Termination check - When the entire task* T_t *has been assigned to the scheduler, the algorithm terminates. Else go to step 2 for scheduling the tasks.In the former steps, Step 3 is the main process of the algorithm.*

Fig. 1. Scheduling Algorithm

4.2 Flow Chart

Figure 4 shows the flow chart of the Hybrid algorithm, thus the intialization of the parameter that need to be scheduled is performed. Then the number of task that has to be scheduled is identified. Scheduling of resource to the task is carried out with the help of Hybrid algorithm. Once the resource are assigned, the scheduler will assign the task to the worker and the execution of task time is calculated. Hybrid algorithm uses the ACO and Cuckoo search.

2) Ant colony Algorithm

```
Step 1: Initialization
Initialize the pheromone trails, Ant solution construction.
Step 2: Construction
For each ant in each step, choose the number of task that is needed to be scheduled.
Step 3: Local search
Perform the local search using cuckoo search and then update the $path_j$.
Step 4: Pheromone update
For each $path_j$ compute fitness value and update pheromone.
Step 5: Termination condition
If total_ iteration < max_iteration go to step 2 otherwise terminate.
```

Fig. 2. Ant colony Algorithm

3) Algorithm for cuckoo search

```
Step 1: Initialization
Initialization of nests and random initial solution.
Step 2: Evaluation
Get the current best nest.
Step 3: Loop construction
While (fmin > Max generation)
Get the cuckoo value by random walk, if not replace it by Levy's flights.
Step 4: Evaluation
Evaluate the quality fitness $Fn_j$.
Randomly choose nest among n, say j.
Step 5: Condition
If($Fn_i > Fn_j$)
Replace j value by new solution.
End
Step 6: Solution construction
Retain the best solution and nests.
Rank the solution and nests to choose the best.
Pass to next generation.
End while, else go to step 2.
```

Fig. 3. Cuckoo search

ACO Flow chart description: Initialization of the pheromone trails, and ant solution construction is performed. The task that need to be scheduled is choosen, assigning of the resource to the task is performed by Cuckoo search. Global upadation of the resource to the task is performed using the pheromone global update. Termination of task is carried out till the execution of the task is completed.

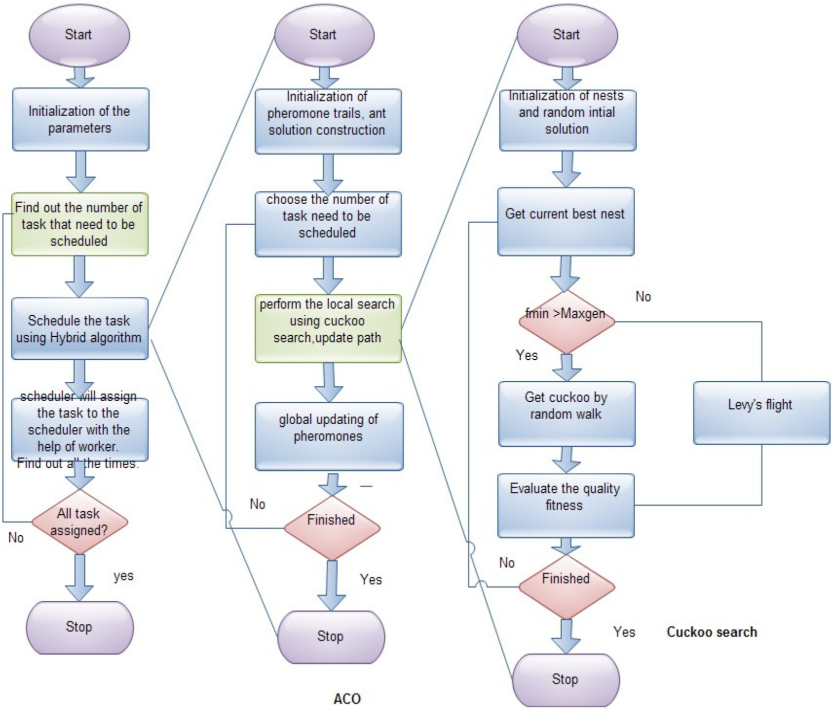

Start
Initialization of the parameters
Find out the number of task that need to be scheduled
Schedule the task using Hybrid algorithm
scheduler will assign the task to the scheduler with the help of worker. Find out all the times.
All task assigned?
No
yes
Stop
Start
Initialization of pheromone trails, ant solution construction
choose the number of task need to be scheduled
perform the local search using cuckoo search,update path
global updating of pheromones
No
Finished
Yes
Stop
ACO
Start
Initialization of nests and random intial solution
Get current best nest
fmin >Maxgen
No
Yes
Levy's flight
Get cuckoo by random walk
Evaluate the quality fitness
No
Finished
Yes
Stop
Cuckoo search

Fig. 4. Scheduling flow chart

Cuckoo search Flow chart description: Initilization of the nests and the random initial solution is performed. The current best nest is choosen by the random walk and then the evalution of the qulaity is fitness is performed. Else apply levy's flight for the evluation. Execution is carried out till all the solution is constructed.

5 Experimental Results

The performance of the Hybrid algorithm is stimulated using the parallel computing toolbox in matlab with the help of workers. The input value for the scheduling is given in form of number of tasks and the various values are Task creation time and the Result retrieval time. They are obtained in terms of time (seconds). Thus when the number of tasks is increased, the time taken for the creation of tasks is also increased. Result retrieval time for the number of tasks is also increased when the number of tasks is increased. It is obvious from the above result that in other types of algorithm such as PSO and ACO, when the number tasks is increased the time taken for the creation of each and every task is also considerably more, similarly for the result retrieval time will also be more. Where as in hybrid algorithm time taken for the creation of tasks is slightly increased.

5.1 Task Creation Time

Task creation time can be defined as the time taken to create and save the disk information. Thus job manager saves the task information in its database, for the other types of schedulers, saves the task information in files on the file system.

$$T_{ct} = c_t + t_{sinf}$$

Table 2 shows the Task creation time for the various tasks that has been given as the input, for the tasks 1 five task have been created; similarly for the tasks 2 and tasks 5 the corresponding tasks have been created. We took the average for the tasks1, tasks 2 and tasks 4. Figure 5 show that when the number of tasks increases the task creation time also increases with respect to the time in seconds. Hybrid algorithm shows clearly that task creation time for each and every task increase.

5.2 Result Retrieval Time

Result Retrieval time is the time taken to display the result to the client. Generally for the job manager this includes the time taken to obtain the results from the database. For the other types of schedulers is the time taken to read from the file system.

$$R_{rt} = D_{jrt}$$

Table 3 shows the result retrieval time for the various tasks that has been created , for the tasks 1 five job have been created and 5 result retrieval time has been evaluated based on the time for the task ; similarly for the tasks 2 and tasks 4 the corresponding result retrieval time have been found. We took the average for the tasks1, tasks 2 and tasks 4. Figure 6 show that when the number of tasks increases the result retrieval time also increases with respect to the time in seconds for the job that have been created. Hybrid algorithm shows clearly that result retrieval time for each and every task.

5.3 Total Time

Total time is the time taken to perform the Job creation time, Task creation time, Job submission time, Job waiting time, Task execution time, Result retrieval time, Job destruction time.

$$T_t = J_{ct} + T_{ct} + J_{st} + J_{wt} + T_{Et} + R_{rt} + J_{dt}$$

Table 2. Task creation time

TASK CREATION TIME			
Sl.No	Task 1	Task 2	Task 4
1	0.0330	0.0334	0.0340
2	0.0337	0.0357	0.0369
3	0.0320	0.0321	0.0322
4	0.0331	0.0346	0.0351
5	0.0338	0.0301	0.0341
Total	0.1656	0.1659	0.1729
Average	0.03312	0.03318	0.0348

Table 3. Result Retrieval time

RESULT RETRIEVAL TIME			
Sl.No	Task 1	Task 2	Task 4
1	0.0100	0.0104	0.0105
2	0.0102	0.0103	0.0107
3	0.0104	0.0110	0.0105
4	0.0103	0.0104	0.0122
5	0.0101	0.0105	0.0108
Total	0.0510	0.0526	0.0547
Average	0.0102	0.01052	0.01094

Table 4. Total time

TOTAL TIME			
Sl.No	Task 1	Task 2	Task 4
1	4.2042	4.2083	4.2113
2	4.2021	4.2036	4.2052
3	4.2030	4.2034	4.2036
4	4.2142	4.2125	4.2380
5	4.2112	4.2122	4.2133
Total	21.0347	21.0400	21.0714
Average	4.20694	4.2080	4.21428

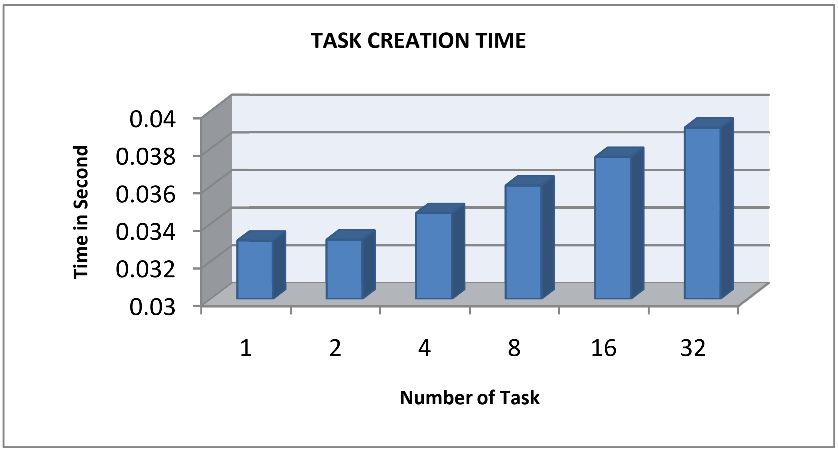


Fig. 5. Task creation time

The table 4 clearly shows the total time taken for the execution of tasks for the job that have been given has the input, thus for 64 task that have been created the corresponding execution time have been evaluated, moreover for the task 1,we have created 5 task and the corresponding execution time is founded. Similarly for task 2 and task4 average value of the execution time is taken.

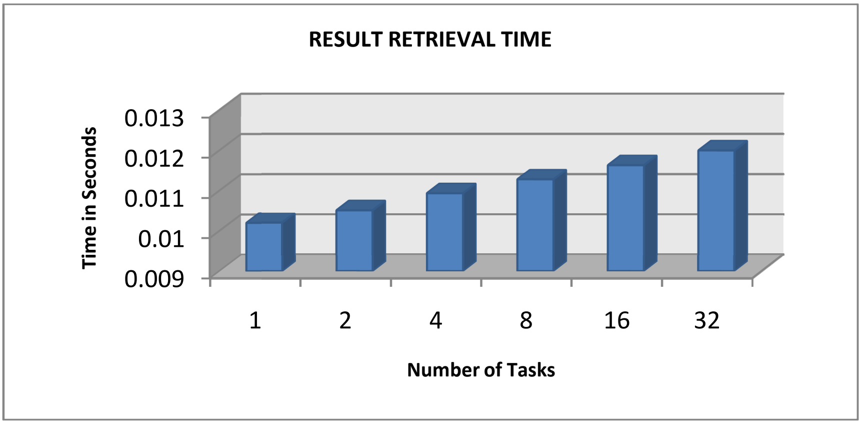


Fig. 6. Result Retrieval time

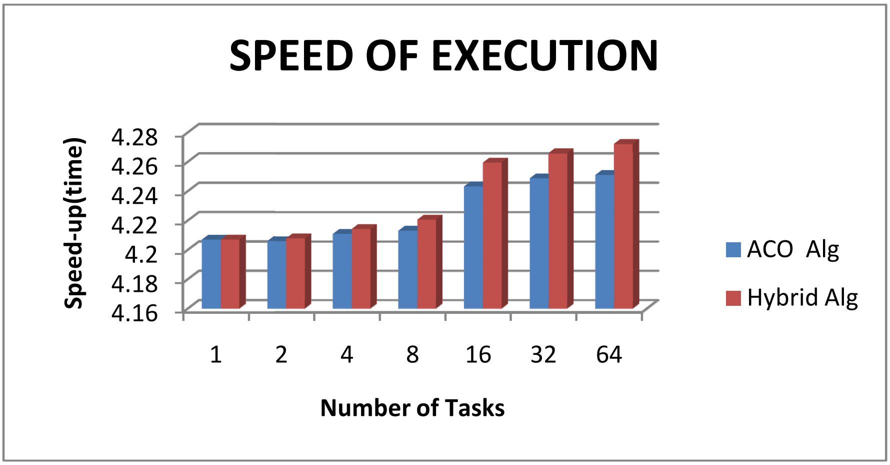


Fig. 7. Speed of Execution

The speed-up of the Hybrid Algorithm is shown in the Figure 7. The speed-up of the Hybrid algorithm increases when the number of task increased based on the time. Comparison is made with the ACO algorithm and the speed-up of ACO is evaluated based on the taks which has been given as the input. It shows clearly that the speed-up of the hybrid algorithm and ACO is increasing steadily for the number of tasks and when the more and tasks have been given as the input the speed-up is more in Hybrid Algorithm than in ACO. This is because the search operation is performed in the Hybrid algorithm by the cuckoo search.

6 Conclusion

Job scheduling is a type of combinatorial optimization problem, which can be used to validate the heuristic algorithms. In this paper, we have presented a Hybrid algorithm for job scheduling by combining the merits of ACO and Cuckoo search. The experimental analysis shows that as the number of tasks is increased, the time taken for the creation of tasks and result retrieval is also increased. In future we plan to apply the hybrid algorithm for job scheduling in scientific and high power computing to minimize the completion time and resource consumed.

References

1. Surekha, S.: PSO and ACO based approach for solving combinatorial Fuzzy Job Shop Scheduling. Int. J. Comp. Tech. Appl. 2(1), 112–120 (2010)
2. Ferrandi, F.: Ant Colony Heuristic for Mapping and Scheduling Tasks and Communications on Heterogeneous Embedded Systems. IEEE Transactions on Computer-Aided Design of Intergrated Circuits and Systems 29(6) (2010)
3. Guo, S., Huang, H.-Z.: Grid Service Reliability Modeling and Optimal Task Scheduling Considering Fault Recovery. IEEE Transactions on Realiability 60(1) (2011)
4. Venkatesan, S., Dhavachelvan, P., Chellapan, C.: Performance analysis of mobile agent failure recovery in e-service applications. International Journal of Computer Standards and Interfaces 32(1-2), 38–43 (2005) ISSN:0920-5489
5. Venkatesan, S., Chellapan, C., Vengattaraman, T., Dhavachelvan, P., Vaish, A.: Advanced Mobile Agent Security Models for Code Integrity and Malicious Availability Check. International Journal of Network and Computer Applications 33(6), 661–671 (2010)
6. Tan, Q., Chen, H.-P.: Two-agent scheduling on a single batch processing machine with non-identical job sizes. In: Artificial Intelligence, Management Science and Electronic Commerce, AIMSEC (2011)
7. Tavakkoli-Moghaddam, Azarkish: A new hybrid mutli-objective Pareto archive PSO algorithm for a bi-objective job shop scheduling problem. Elsevier Expert Systems with Applications (2011)
8. Ahn, C.W., An, J.: Estimation of particle swarm distribution algorithms: Combining the benefits of PSO and EDAs. Elsevier Information Sciences (2010)
9. Yang, X., Yuan, J.: An improved WM method based on PSO for electric load forecasting. Elsevier Expert Systems with Applications (2010)
10. Victer Paul, P., Saravanan, N., Jayakumar, S.K.V., Dhavachelvan, P., Baskaran, R.: QoS enhancements for global replication management in peer to peer networks. Future Generation Computer Systems 28(3), 573–582 (2012)
11. Abirami, S., Baskaran, R., Dhavachelvan, P.: A survey of Keyword spotting techniques for Printed Document Images. Artificial Intelligence Review 35(2), 119–136 (2010)
12. Bae, C., Yeh, W.-C.: Elsevier Expert Expert Systems with Applications. Feature selection with Intelligent Dynamic Swarm and Rough Set (2010)
13. Sha, Lin, H.-H.: A Multi-objective PSO for job-shop scheduling problems. Elsevier Expert Systems with Applications (2010)
14. Tao, Q., Chang, H.-Y.: A rotary Chaotic PSO algorithm for trustworthy scheduling of a grid workflow. Elsevier Computers & Operations Research (2011)

15. Dhavachelvan, P., Uma, G.V.: Complexity Measures For Software Systems: Towards Multi-Agent Based Software Testing. In: Proceedings-2005 International Conference on Intelligent Sensing and Information Processing, ICISIP 2005, Art. no. 1529476, pp. 359–364 (2005)
16. Vengattaraman, T., Abiramy, S., Dhavachelvan, P., Baskaran, R.: An Application Perspective Evaluation of Multi-Agent System in Versatile Environments. International Journal on Expert Systems with Applications 38(3), 1405–1416 (2011)
17. Vengattaraman, T., Dhavachelvan, P.: An Agent-Based Personalized E-Learning Environment: Effort Prediction Perspective. In: IEEE International Conference on Intelligent Agent & Multi-Agent Systems, IAMA 2009 (2009) ISBN: 978 1-4 244-4710-7
18. Hu, X.-M., Zhang, J.: SamACO: Variable Sampling Ant Colony Optimization Algorithm for Continuous Optimization. IEEE Transactions on Systems, Man, and Cybernetics 40(6) (2010)
19. Zhang, Z., Zhang, J., Li, S.: A Modified Ant Colony Algorithm for the Job Shop Scheduling Problem to Minimize Makespan. IEEE Explore (2010)
20. Zhan, Z.-H., Zhang, J.: An Efficient Ant Colony System Based on Receding Horizon Control for the Aircraft Arrival Sequencing and Scheduling Problem. IEEE Transactions on Intelligent Transaction Systems 11(2) (2010)
21. Victer Paul, P., Vengattaraman, T., Dhavachelvan, P.: Improving efficiency of Peer Network Applications by formulating Distributed Spanning Tree. In: Proceedings - 3rd International Conference on Emerging Trends in Engineering and Technology, ICETET 2010, Art. no. 5698439, pp. 813–818 (2010)
22. Saleem Basha, M.S., Dhavachelvan, P.: Web Service Based Secure E-Learning Management System - EWeMS. International Journal of Convergence Information Technology 5(7), 57–69 (2010) ISSN: 1975 9320
23. Dhavachelvan, P., Uma, G.V., Venkatachalapathy, V.S.K.: A New Approach in Development of Distributed Framework for Automated Software Testing Using Agents. International Journal on Knowledge–Based Systems 19(4), 235–247 (2006)

Particle Based Fluid Animation Using CUDA

Uday A. Nuli[1] and P.J. Kulkarni[2]

[1] Textile and Engineering Institute, Ichalkaranji
uanuli@yahoo.com
[2] Walchand College of Engineering, Sangli
pjk_walchand@rediffmail.com

Abstract. Particle based animation employing physical model is a highly compute-intensive technique for realistic animation of fluids. It has been used since its inception for the offline production of high quality special effects of fluids in the movies. However due to intense computational cost, it could not be adapted for real-time animations. This paper primarily focuses on formulation of parallel algorithms for particle based fluid animation using Smoothed Particle Hydrodynamics(SPH) approach employing CUDA enabled GPU to make it near real-time. SPH technique is highly suitable for SIMT architecture of CUDA enabled GPU promising better speedup than CPU based approaches. The most important hurdle in parallelization using CUDA is the existing parallel algorithms do not map efficiently to CUDA. In this paper we have employed parallel sorting based particle grid construction approach to reduce computational cost of SPH density and force computation from O (N2) to O (N).

Keywords: CUDA, particle animation, smoothed particle hydrodynamics.

1 Introduction

Animation is changing its trend from basic key-framed, non-realistic, and offline to realistic, model driven, and real time animation. Physically based animation is a technique that incorporate physical behavioral model of the object to estimate motion of the object. Since the motion is based on physical laws, the motion produced in the animation is more realistic. However such animation needs huge computational power to solve the equations governing the motion. Hence real-time physically based animation is possible through involvement of suitable parallel architecture system such as multi-core or computer cluster. In Particle base fluid animation using Smoothed Particle Hydrodynamics (SPH), fluid is treated as collection of particles. Motion of particles is governed by set of equations defined by SPH technique. Hence particle based animations is appropriate for implementation on SIMD parallel architecture.

Over past few years, Graphics Processing Unit (GPU) has evolved from a fixed function graphics pipeline to a general purpose, many core SIMD architecture. Although GPU architecture is inherently parallel since its invention, it was specifically limited to graphics functionality till NVIDIA introduced it as a Compute Unified Device Architecture (CUDA)[1][2].One of key fact about any parallel platform is the algorithms optimized for one platform may not run with same efficiency on other platform. Hence restructuring and optimization of algorithms before implementation

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 503–510.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

on a specific parallel platform is essential. This paper is devoted to restructuring and optimization of parallel algorithm for implementation of SPH based particle animation on NVIDIA CUDA platform.

2 Related Works

Fluid animations have a very long history. In past it has been implemented with various physical models, and hardware platforms. Jos Stam[3] has developed a Mesh-oriented solution for fluid animation. Nick Foster and Ron Fedkiw[4] derived full 3D solution for Navier-Stokes equations that produces realistic animation results. In addition to the basic method, the Lagrange equations of motion are used to place buoyant dynamic objects into a scene, and track the position of spray and foam during the animation process. Jos Stam[5] extended the basic Eulerian approach by approximating the flow equations in order to achieve near real time performance. He has also demonstrated various special effects of fluid with simple “C” code at typical frame rate of 4 to 7 minutes per frame.

Although Eulerian approach was a popular scheme for fluid animation, it has few important drawbacks such as; it needs global pressure correction and has poor scalability[6]. Due to these drawbacks such schemes are unable to take benefits of parallel architectures available today. Particle based methods are free from these limitations and hence are becoming more popular in fluid animation.

Reeves [7] introduced the particle system which is then widely used to model the deformable bodies, clothes and water. His paper has demonstrated animation of fire and multicolored fireworks. Particle System based animations are created with two approaches, one with motion defined by certain physical model and other by simple use of Newton’s basic laws. R. A. Gingold and J.J Monaghan proposed “Smoothed Particle Hydrodynamics”, a particle based model to simulate astrophysical phenomena [8] and later extended to simulate free-surface incompressible fluid flows[9]. This model was created for scientific analysis of fluid flow, carried out with few particles. Matthias Müller extended the basic SPH method for fluid simulation for interactive application [10] and designed a new SPH kernel.

The first implementation of the SPH method totally on GPU was realized by T. Harada [11] using OpenGL APIs. T. Harada has demonstrated 60,000 particles fluid animation at 17 frames per second which is much faster as compared to CPU based SPH fluid animation.

These papers clearly highlight the aptness of Smoothed Particle Hydrodynamics technique for SIMD parallel architecture to achieve realistic Particle based fluid animation.

3 Smoothed Particle Hydrodynamics (SPH)

Smoothed Particle Hydrodynamics [8][9][10] is a mesh-free, Lagrangian, particle method for modelling Fluid flow. This technique is introduced by Lucy and Monaghan in 1977 for modelling astrophysics phenomena and later on extended for modeling fluid phenomena.

SPH integrates the hydrodynamic equations of motion on each particle in the Lagrangian formalism. Relevant physical quantities are computed for each particle as an interpolation of the values of the nearest neighbouring particles, and then particles move according to those values. The basic Navier-Stokes equations are transformed to equivalent particle equations.

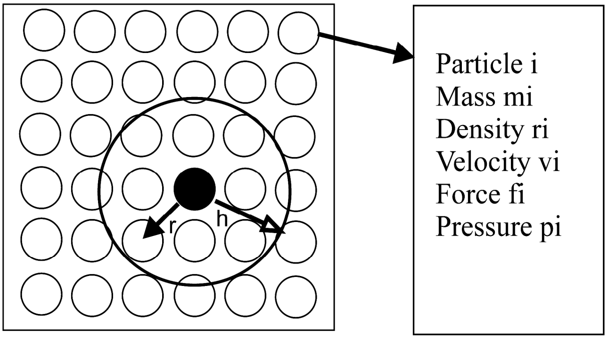


Fig. 1. Representation of fluid as collection of particles

According to SPH, a scalar quantity $As(r)$ is interpolated at location r by a weighted sum of contributions from all particles. The basic interpolation formula used is:

$$As(r) = \sum_j m_j \frac{A_j}{\rho_j} W(r - r_j, h) \tag{1}$$

Where, $As(r)$ is the scalar property of Particle at position r, m_j the mass of Jth particle at distance r_j from particle at r, ρ_j the mass-density of particle at location r_j, A_j the scalar property of particle at location r_j and W the Kernel Function[10] or the smoothing kernel. The kernel function has a cut of radius **h**. The cut-of radius sets W = 0 for all |**r** - **r**$_j$| > h. Mass-density of a particle is calculated by substituting density term in place of generic term $As(r)$. The equation for density ρs(r) terms is as follows [10]:

$$\rho_s(r) = \sum_j m_j W(r - r_j, h) \tag{2}$$

The pressure exerted on a particle due to other particles is derived from ideal gas law. The pressure is computed using following equation:

$$P = k(\rho - \rho_0) \tag{3}$$

Where P is the pressure exerted on the particle, k the stiffness constant of gas, ρ the mass density of the particle at time t in simulation and ρ_0 the mass density of the particle at rest condition. Every particle is influenced by viscous and pressure forces. These forces are computed using SPH formulations as:

$$f_i^{pressure} = -\sum_j m_j \frac{p_i + p_j}{2\rho_j} \nabla W\left(r_i - r_j, h\right) \tag{4}$$

$$f_i^{viscosity} = \mu \sum_j m_j \frac{v_j - v_i}{\rho_j} \nabla^2 W\left(r_i - r_j, h\right) \tag{5}$$

Where $f_i^{pressure}$ is the force due to pressure and $f_i^{viscosity}$ the force due to viscosity on ith particle exerted by other particles; p_i and p_j are the pressure, v_i and v_j the velocities, r_i and r_j the position of ith and jth particle; m_j is the mass and ρ_j is the density of the jth particle;

Velocity and particle position updates are carried out using equations specified by the Takahiro Harada [11].

4 System Architecture

Primary task in any physically based particle animation model is to estimate the motion of each particle. This is basically a physical simulation approach and comprises of calculation of particle spatial position in every time step. Motion estimation based on SPH involves computation of every particle's mass density, pressure and force exerted due to neighbouring particle within a distance of smoothing radius. Hence determination of neighbour particles has complexity of O(N^2) for all particles. This complexity can be reduced down to O(N), if spatial grid based neighbour search technique is employed. These huge periodic computations justify the need of SIMD parallel architecture in order to complete it in real time. However certain part of animation needs to be carried out on CPU due to execution constraints of CUDA. The computational part in animation is executed completely on CUDA as parallel CUDA Kernels whereas launching of CUDA Kernel takes place from CPU. The basic steps for Particle System Animation using SPH are as follows:

a) Initialize and setup particle system.
b) Render Particles.
c) Construct spatial grid for particles.
d) For each particle:
    i. Calculate density.
    ii. Calculate Pressure exerted on the particle due to its neighbours.
    iii. Calculate net Force on Particle due to inter particle pressure.
    iv. Add external force to Particle.
    v. Find Particle Acceleration.
    vi. Find next particle position.
    vii. Update Particle position.
e) Invoke OpenGL to Display the Particles at updated position.
f) Repeat from step d.

The animation process needs to be partitioned into two phases. During first phase the initialization of particle system, SPH parameters and GPU memory allocation is carried out. During second phase animation loop comprising particle motion estimation is carried out. First phase execution entirely takes place on CPU and second phase on GPU co-ordinated by CPU. Separate parallel kernels are written for spatial grid construction, Density, pressure, force and displacement computation. These kernels are called sequentially from CPU as indicated in Figure 3. Particle rendering can be implemented on entirely CPU or GPU based on sophistication requirements of animated result.

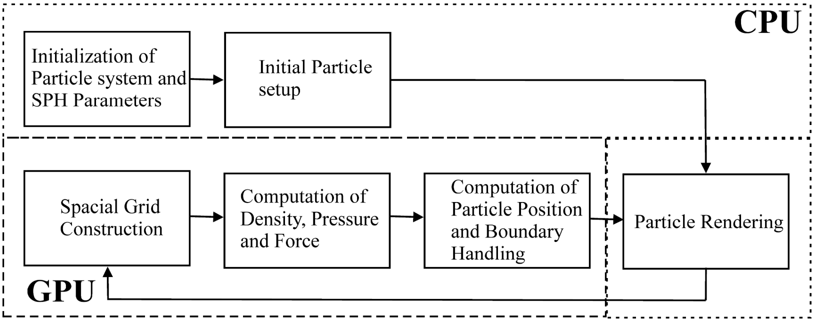

CPU
Initialization of Particle system and SPH Parameters
Initial Particle setup
Spacial Grid Construction
Computation of Density, Pressure and Force
Computation of Particle Position and Boundary Handling
Particle Rendering
GPU

Fig. 2. System Block Diagram

5 Particle Motion Estimation Using CUDA

Motion estimation computations, described in section 3 and 4, can be significantly speedup through implementation on parallel architecture like CUDA. The operations implemented as separate CUDA kernel to estimate particle position at every simulation step are as listed below:

1. Spatial Grid Construction.
2. Particle density and force computation.
3. Particle Velocity and displacement computation.

Particles are organized in a 3-dimensional spatial grid to reduce parameter computational cost from O (n^2) to O (n). Dynamic grid construction with variable number of particles per grid cells is the key requirements for grid construction approach. Both requirements can be satisfied by sorting based grid construction approach. The steps for spatial grid construction are as listed below.

1) Computation of Cell ID for each particle based on its location.
2) Grouping Particles according to grid cell.
3) Prepare cell table consisting Cell ID, Particle index for cell header, and particle count.
4) Prepare cell neighbour table consisting 26 neighbours for each cell.

The Cell ID is the identity of each cell in the dynamic grid. CUDA based implementation of Cell ID computation reduces its complexity from O (n) to O (1) by employing a CUDA thread per particle. Particles are grouped in cell by employing sorting technique on Cell ID. The radix sort algorithm is used to carry out sorting since it is the most efficient algorithm on CUDA [12]. Sorted Cell IDs are used to create a cell table comprising Cell ID, Particles per Cell, and Index for first particle in the cell. The modified parallel prefix scan algorithm is used to identify first cell index and Particles per cell count. Cell neighbour table stores information about 26 neighbours of cell and is computed form cell table performing parallel binary search on cell IDs. This is essential to reduced search complexity from O (n) to O (1) during iterations of animation.

Particle density and force are computed using equations as described in section 3. The force computed is used compute acceleration and velocity of each particle which is further used to compute displacement. The nature of computations makes it appropriate for data parallel architecture of CUDA. Number of threads launched on CUDA is decided by the Particle count used in animation. A single thread per particle strategy is employed for thread assignment to each particle. A simple cubical boundary is assumed for demonstration of animation hence it is possible complete these computations along with displacement in same kernel. A complicated boundary handling may require separate computation kernel.

6 Particle Rendering

Particles are rendered using OpenGL functions. A simplest approach is to render each particle as solid sphere. This approach is suitable for demonstration of fluid animation technique and not for professional animation. For Professional animation, Marching Cube like algorithm along with ray-tracing can be applied for fluid surface reconstructed from particle.

7 Results

Main motive behind employment of CUDA enable GPU for fluid animation is to achieve real-time performance. The result shown here clearly demonstrates considerable speedup in execution of every stage of animation. This particle animation has been carried out on a computer with Intel® Core™2 Duo CPU E7500 @ 2.93GHz with 2GB of RAM and NVIDIA GTX 280 GPU card. Timings are measured with NVIDIA CUDA Profiler and averaged. The result shows speedup for both animations with and without spatial grid. The optimized speed up is for animation employing spatial grid.

Table 1. Computation Time for 50000 Particles

Operations	CPU(ms)	GPU(ms)	Optimized GPU(ms)	Speed up	Optimized speedup
Pressure Computation	12725.74	71.26	2.653	178.58	4796.74
Force Computation	14672.61	190.66	5.092	76.96	2881.50
Displacement Computation	1.41	1.09	1.091	1.29	1.91
Total	27399.76	263.34	9.148	104.05	2995.16

Table 2. Computation Time for 100000 Particles

Operations	CPU(ms)	GPU(ms)	Optimized GPU(ms)	Speed up	Optimized speedup
Pressure Computation	58249.08	269.22	4.825	216.36	12072.35
Force Computation	65184.26	939.86	12.743	69.36	5115.29
Displacement Computation	2.94	1.14	1.132	2.58	2.59
Total	123436.3	1211.99	20.49	101.85	6024.22

8 Conclusion and Future Work

This Paper is focused on SPH, a relatively new Fluid Dynamics technique to simulate motion of fluid particles. SPH is a particle based parallelizable technique hence more suitable for CUDA. Equations of SPH expose more arithmetic intensity in computation; hence the animation's computational part can be executed on CUDA in real time. However the major hurdle in SPH based animation is the tuning of SPH parameters like smoothing radius and rest density. A small change in any of these parameter results into almost explosion of fluid.

Parallel Algorithm Design is the most critical issue in any application development for CUDA. Even though most of the algorithms discussed in the paper are already designed for some of earlier parallel architectures, it is not possible to adopt them without modification for CUDA. Also the typical asymptotic complexity equations are not appropriate for indicating performance of algorithm on CUDA. The asymptotic definition of complexity comments on the number of discrete steps executed by algorithm and not on the arithmetic intensity of computations per step. Hence asymptotic complexity equations do not represent true performance of an algorithm on CUDA.

This work can be extended for huge quantity fluid animation that demands more memory space than available on present CUDA devices. For animation of large quantity of fluid, the employment of alone CUDA is also insufficient. Hence development of framework for animation on cluster of CUDA can be considered as next possible approach. Also Surface quality of fluid can still be improved by optimizing ray tracing for interactive rate rendering on GPU.

References

1. NVIDIA, NVIDIA CUDA C, Programming Guide, version 4.2, `http://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf` (visited on May 9, 2012)
2. NVIDIA, NVIDIA CUDA C Best Practices Guide, version 3.1, `http://develper.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Best_Practices_Guide.pdf` (visited on May 9, 2012)

3. Jos, S.: Stable fluids. In: ACM SIGGRAPH 1999: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 121–128. ACM Press/Addison-Wesley Publishing Co., New York (1999)
4. Foster, N., Fedkiw, R.: Practical Animation of Liquids. In: ACM SIGGRAPH 2001, pp. 21–30 (2001)
5. Jos, S.: Real-Time Fluid Dynamics for Games. In: Proceedings of the Game Developer Conference (2003)
6. Tan, J., Yang, X.: Physically-based fluid animation: A survey. Science in China Series F: Information Sciences, vol. 52, pp. 723–740. Science China Press, co-published with SpringerLink (May 2009)
7. Reeves, W.T.: Particle Systems-A technique for modeling a class of fuzzy objects. ACM Transactions on Graphics 2(2), 91–108 (1983)
8. Gingold, R.A., Monaghan, J.J.: Smoothed particle hydrodynamics: theory and application to non-spherical stars. Monthly Notices of the Royal Astronomical Society 181, 375–398 (1977)
9. Monaghan, J.J.: Simulating free surface flows with SPH. Journal of Computational Physics 110(2), 399–406 (1994)
10. Müller, M., Charypar, D., Gross, M.: Particle-Based Fluid Simulation for Interactive Applications. In: Proceedings of the 2003 ACM SIGGRAPH, pp. 154–159 (2003)
11. Harada, T., Koshizuka, S., Kawaguchi, Y.: Smoothed Particle Hydrodynamics on GPUs. In: Proceedings of Computer Graphics International, pp. 63–70 (2007)
12. Satish, N., Harris, M., Garland, M.: Designing Efficient Sorting Algorithms for Manycore GPUs. In: Proceedings of 23rd IEEE International Parallel and Distributed Processing Symposium, pp. 1–10 (2009)

A Novel Model for Encryption of Telugu Text Using Visual Cryptography Scheme

G. Lakshmeeswari*, D. Rajya Lakshmi, Y. Srinivas, and G. Hima Bindu

GIT, GITAM University, Visakhapatnam, Andhra Pradesh
{lak_pr,rdavuluri}@yahoo.com, ysrinivasit@rediffmail.com,
nadiminti19@gmail.com

Abstract. This article presents a novel methodology of a visual cryptographic system based on Telugu Text. Visual Cryptography concerns with the methodology of hiding a secret message in *n* shares such that each individual receives one share and only the authorized person can identify the secret messages after superimposing one share upon the other. Many techniques are proposed in literature where in Rotation Visual Cryptography the underlying message of hiding is done by rotating the shares at different angles or directions. The proposed sliding scheme reveals the information by horizontal sliding of a share in downward direction by an angle of 180^0.

Keywords: Visual cryptography, Encoded Telugu text, Human Visual System, Sliding Scheme.

1 Introduction

An increased dependency on networks for information transmission has provided a large scope for hackers to utilize the leaks during transmission. To provide a secured transmission, senders had to increase the computational complexity which resulted in a higher degree of encryption and decryption process. But this lead to a complicated and time taking process. With the advent of Visual Cryptography this complexity is reduced to maximum extent.

Many efforts have been made for secured transmission of data using different methodologies. Amongst them, visual cryptography has evolved as one of the most successful method due to its simplicity and minimized computational complexities. Visual Cryptography Schemes can decode concealed images based purely on human visual systems, without any aid from cryptographic computation. This property has given rise a wide range of encryption applications[1]. Ever since the concept has been introduced in the year 1994 by "Shamir and Naor", many techniques have evolved, each of which refines a different aspect of this methodology.

[2] Naor and Shamir analyzed (k, n)-threshold visual cryptography schemes. In these schemes, a subset is qualified if and only if it consists of at least k participants. Most work concerning this subject focuses on two aspects, either the pixel expansion, i.e. the number of sub-pixels which is needed on the different levels to represent a

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 511–520.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

white or a black pixel, or the contrast, i.e. the difference of sub-pixels representing a white or a black pixel.

As a further generalization, the existence of a secret image can be concealed by displaying a different image on each transparency. Naor and Shamir [2] solved this problem for the (2, 2)-threshold scheme. By stacking the transparencies of each participant one upon another, a secret image is recovered, and this is in fact the only way to recover it[3].

The optimality of VCS is determined mostly by its pixel expansion m and the relative contrast. Pixel expansion m represents the loss in resolution from the original image to the decoded one. Therefore m needs to be as small as possible.

The theme of Rotation Visual Cryptography Schemes is to reveal the underlying message by rotation of a share at different angles[4]. Rotation schemes facilitated storage of multiple secrets in a single share.

Many methods are available in literature [1,3,5], which are mostly focused in hiding the text in the form of English language, but very little work is projected towards the usage of Telugu text for the purpose of secret transmission. Among the south central languages, Telugu is mostly preferred language, and it is being used in the government sectors as well as private sectors of Andhra Pradesh, hence in this paper, we propose a model based on Telugu Text. The Telugu coding scheme helps to minimize the bits required for hiding each character along with its diacritic and compound. We can hide a larger message with minimal effort[5]. This methodology using the (k,n) scheme , which can be very useful for secret transmission of messages in areas such as Banks, Home Department, Intelligence and other public services for the state of Andhra Pradesh, where Telugu language is commonly used. We can either transmit a continuous message or a fragmented password using the proposed system. The next section of the paper deals with the encryption model, section-3 deals with the decryption methodology, section-4 presents the proposed methodology, section-5 describes the experimental results of the proposed system and finally the conclusions are given in section-6.

2 Encryption Process

Step 1: Convert the Telugu text input into its equalent code[5].

Step 2: Select k number of shares randomly from a total of n shares, where k+1<n. The value of k and n are dynamic.

Step 3: Split the coded message obtained from step 1 in to k slices and generate k shares, where each share contains one slice of the coded message.

The random selection of shares in step 2 assures that the sequence of shares in which the information would be stored would not be the same always. A single share would contain only one portion of information.

3 The Process Adopted for Decryption

Step 1: Retrieve the k out of n shares which contain the embedded message with inverse random process.

Step 2: From the remaining *n-k* shares select a share which has to be slided over the selected k shares to reveal the information. Let the selected share be x.

Step 3: Slide the share x on each of the k shares horizontally downwards to reveal the message in each share sequentially.

4 Proposed Scheme

We propose the technique of inter-slicing in order to store text in a share. The number of slices is dynamic depending upon the value of k. For example consider 4 out of 7 shares to contain the secret message (k=4 and n=7). Let the numbers of the selected share numbers be 2,3,5 and 7. Each of the *k* shares are divided into k slices as shown in Figure 1.

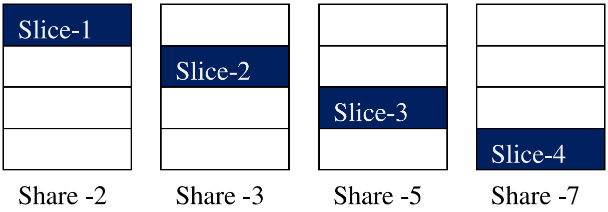


Fig. 1. Inter-sliced arrangement of shares

The shaded regions of the shares contain the embedded text. All shares must be of equal size.

For example consider the Telugu text "ఇక్కడ రెక్కి జరిగినది". The equalent encoded string for the given text is "0002401000 1A01E84110 0016012801 120021011F" [5]. Embed 10 characters each of this code in 4 shares(2,3,5 and 7) as shown in Figure 2.

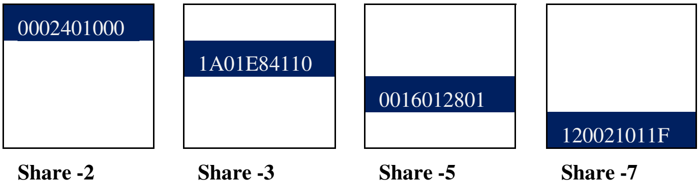


Fig. 2. Images with secret message

The original image would contain the complete code generated at step1 on the sender side and would appear as in Figure 3.

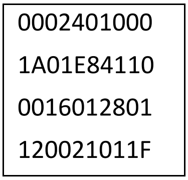
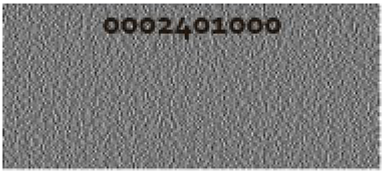

0002401000
1A01E84110
0016012801
120021011F

Fig. 3. Original Image with secret message

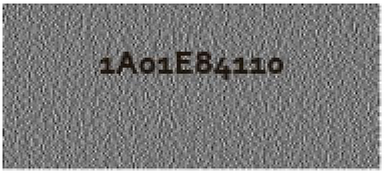
5 Experimental Results

The system is implemented using matlab. It considers the image containing the original encrypted message as its base image. To make the proposed system dynamic the values of *k* and *n* are not taken as constants instead are given as input on execution of the code.

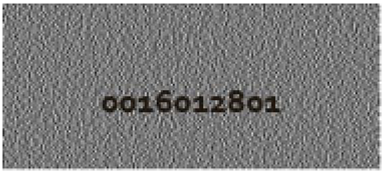
Resultant shares after superimposing the share '*x*' on each and every share to reveal the underlying text is shown in Figure 4.



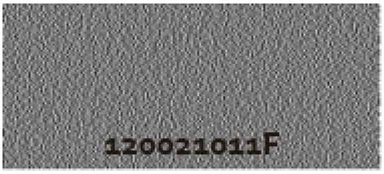
Share -2



Share -3



Share -5



Share -7

Fig. 4. Shares with disclosed information

The final share after combining the information of each share would appear as shown in figure 5.

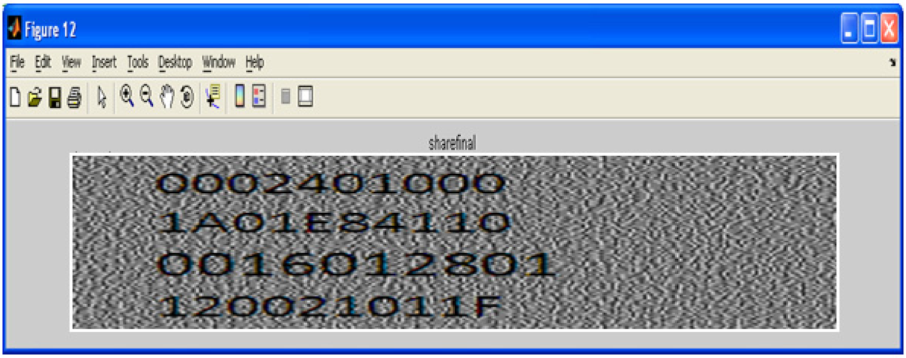


Fig. 5. Final share with complete secret message

The final Telugu text “ఇక్కడ రెక్కి జరిగినది” is obtained on decrypting the information n the final share[1].

Another application of the same concept can be demonstrated as follows:

The Figure 6 shows an image with a message in it.

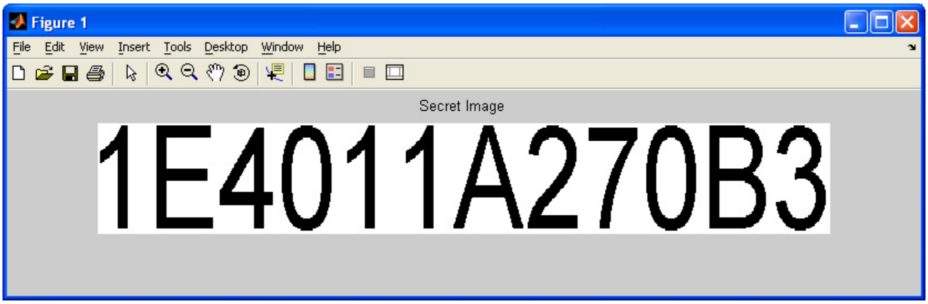


Fig. 6. Original Image with secret message

Apply the horizontal slicing technique and put the information in 5 shares, which are selected randomly. The shares would appear as shown in figures 7, 8, 9, 10 and 11.

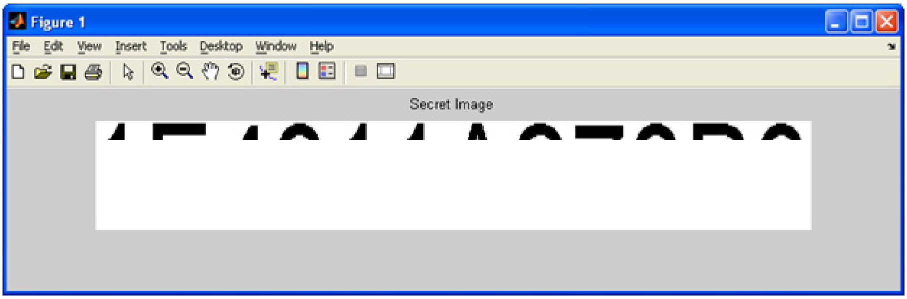


Fig. 7. Slice1 of the Original Image

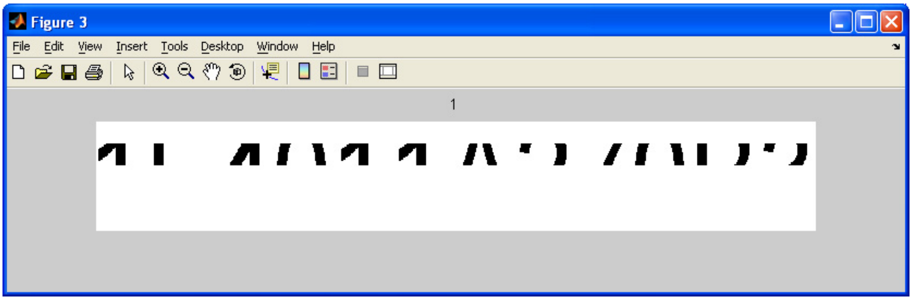


Fig. 8. Slice2 of the Original Image

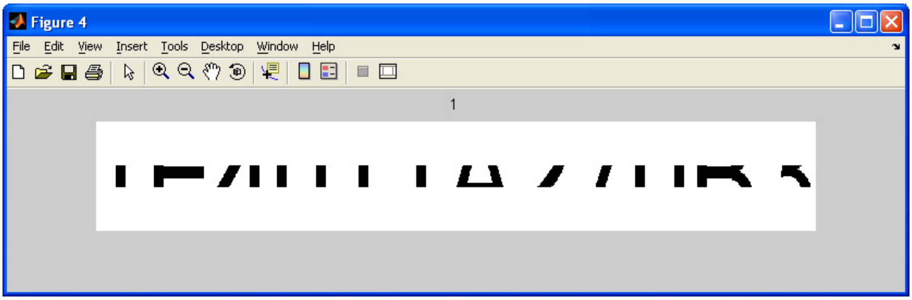


Fig. 9. Slice3 of the Original Image

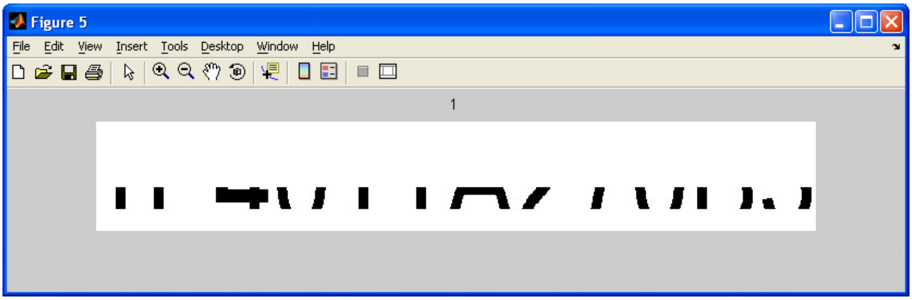


Fig. 10. Slice4 of the Original Image

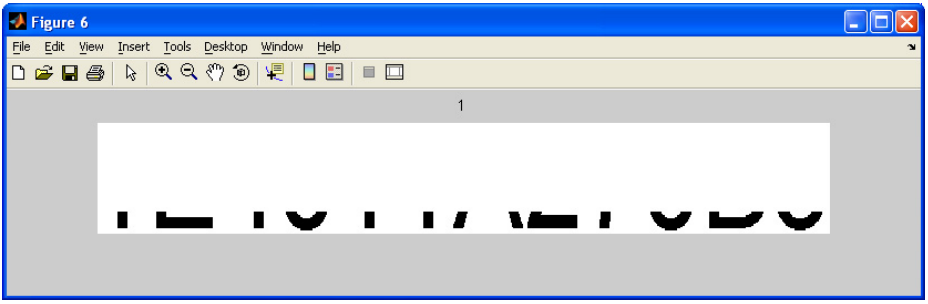


Fig. 11. Slice5 of the Original Image

Add noise to the shares, and use the share ‘x’ to disclose the content in the shares as specified in the decryption procedure. Resultant shares after superimposing the share ‘*x’* on each and every share to reveal the underlying text is shown in Figures 12, 13, 14, 15 and 16.

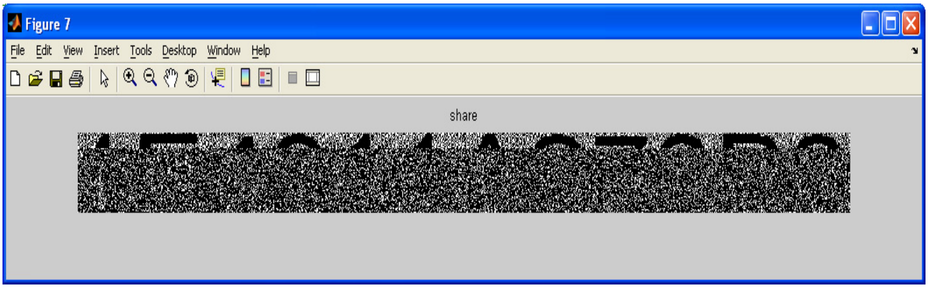


Fig. 12. Information of Slice1 disclosed from the first share

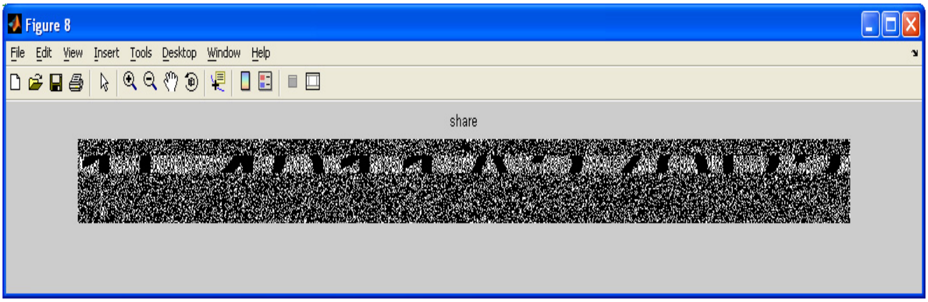


Fig. 13. Information of slice2 disclosed from the second share

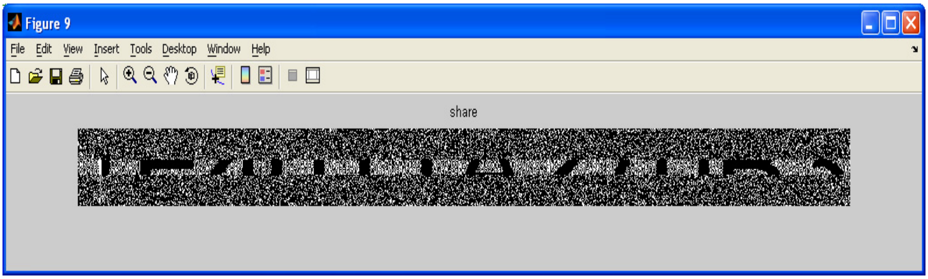


Fig. 14. Information of slice3 disclosed from the third share

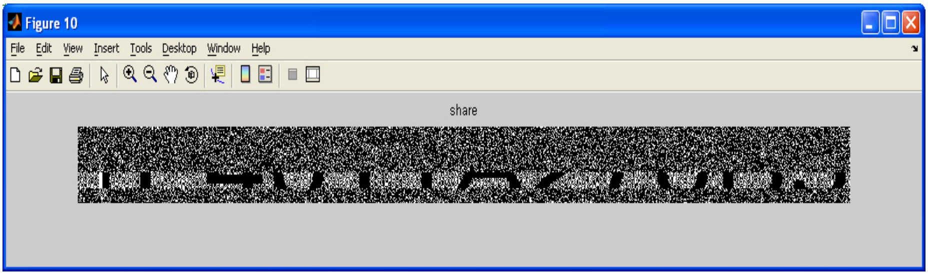


Fig. 15. Information of slice4 disclosed from the fourth share

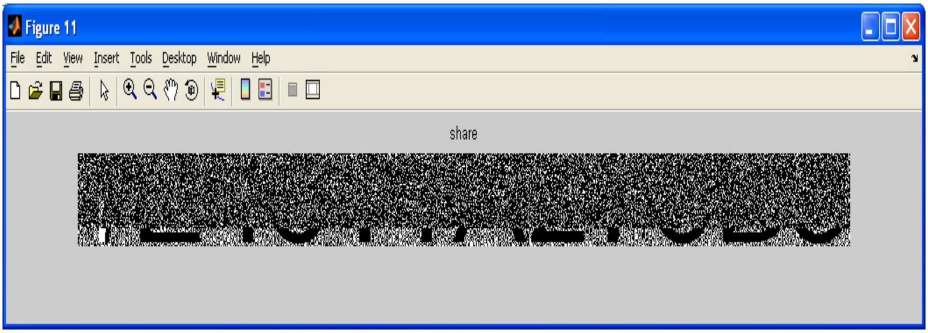


Fig. 16. Information of slice5 disclosed from the fifth share

The final share on combining all the above 5 shares is shown in figure 17.

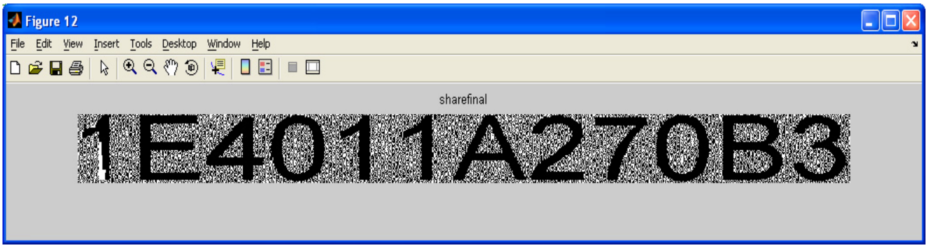


Fig. 17. Final share with complete disclosed message

6 Conclusion

The proposed scheme sliding Technique in Visual Cryptography is designed for encrypting a message or a short length password by slicing into k parts and storing each part in a share based on Telugu character code.

The decryption of the original message is possible only with the assistance of a share ‘x’ which has to be commonly stacked on all the shares with secret message. The resultant output is combined into a single share to disclose the complete secret text.

The proposed methodology not only works for the experimented Telugu language encryptions, but can be used as a generalized concept applicable for transmitting any secret information of any language in the world. It can also be used for password transmission, where each share can be sent to a person and on recombining we get back the complete password. The probability of guessing the password is much less when compared to the other methodologies as no single alphabet is completely displayed in any of the share.

References

1. Dinesh Reddy, B., Valli Kumari, V., Raju, K.V.S.V.N., Prassanna Raju, Y.H.: Rotation Visual Cryptography. Using Basic (2, 2) Scheme. IJCS & CT 3(2) (January 2011) (ISSN 0974-3375)
2. Naor, M., Shamir, A.: Visual Cryptography. In: De Santis, A. (ed.) EUROCRYPT 1994. LNCS, vol. 950, pp. 1–12. Springer, Heidelberg (1995)
3. Tsai, P.-F., Wang, M.-S.: An (3, 3)-Visual Secret Sharing Scheme for Handling Three Secret Data
4. Hsu, H.C., Chen, T.S., Lin, Y.H.: The ringed shadow image technology of Visual Cryptography by applying diverse rotating angles to hide the secret sharing. Networking, Sensing and Control 2, 996–1001 (2004)
5. Lakshmeeswari, G., Rajya Lakshmi, D., Lalitha Bhaskari, D.: Extended Encoding of Telugu Text for Hiding Compatibility. IJCA 30(5) (September 2011)
6. Weir, J., Yan, W.Q.: Sharing Multiple Secrets Using Visual Cryptography. In: IEEE International Symposium on Circuits and Systems, ISCAS 2009 (2009), doi: 10.1109/ISCAS.2009.5117797
7. Ge, L., Tang, S.: Sharing “Multi-Secret Based on Circle Properties”. In: 2008 International Conference on Computational Intelligence and Security (2008)
8. Ateniese, G., Blundo, C., Santis, A.D., Stinson, D.R.: Extended capabilities for Visual Cryptography. Theoretical Computer Science 250(1-2), 143–161 (2001)
9. Yu, B., Xu, X., Fang, L.: Multi-Secret sharing threshold Visual Cryptography Scheme. International Conference on Computational Intelligence and Security (2007)
10. Lakshmeeswari, G., Rajya Lakshmi, D., Srinivas, Y.: A New Encoding Scheme of Telugu Text for Information Hiding. International Journal of Computational Intelligence Techniques 2(1), 26–28 (2011) ISSN: 0976–20466 & E-ISSN: 0976–0474
11. Wu, H., Chang, C.C.: Sharing Visual Multiple Secrets using Circle Shares. Computer Standards & Interfaces 28, 123–135 (2005)

12. Zhou, Z., Arce, G.R., Crescenzo, G.D.: Halftone Visual Cryptography, vol. 15(8), pp. 2441–2453. IEEE (2006)
13. Fu, Z., Yu, B.: Research on Rotation Visual Cryptography Scheme. In: International Symposium on Engineering and Electronic Commerce (2009)
14. Myodo, E., Sakazawa, S., Takishima, Y.: Visual Cryptography based on void-and-cluster Half toning technique. In: ICIP, pp. 97–100 (2006)

Frequent Queries Selection for View Materialization

T.V. Vijay Kumar[1], Gaurav Dubey[3], and Archana Singh[2]

[1] School of Computer and Systems Sciences, Jawaharlal Nehru University,
New Delhi-110067, India
[2] Amity School of Computer Sciences, Amity Campus,
Sector 44, Noida, UP-201303, India
[3] Amity Institute of Information Technology, Amity Campus,
Sector 125, Noida, UP-201301, India

Abstract. A data warehouse stores historical data for answering analytical queries. These analytical queries are long, complex and exploratory in nature and, when processed against a large data warehouse, consume a lot of time for processing. As a result the query response time is high. This time can be reduced by materializing views over a data warehouse. These views aim to improve the query response time. For this, they are required to contain relevant information for answering future queries. In this paper, an approach is presented that identifies such relevant information, obtained from previously posed queries on the data warehouse. The approach first identifies subject specific queries and then, from amongst such subject specific queries, frequent queries are selected. These selected frequent queries contain information that has been accessed frequently in the past and therefore has high likelihood of being accessed by future queries. This would result in an improvement in query response time and thereby result in efficient decision making.

1 Introduction

The World Wide Web has allowed the accumulation of large amounts of data from across the globe. Organizations aspire to exploit this data for their benefit in order to stay competitive in the market. This would entail accessing data in a manner that would facilitate decision making. There are two ways to access this data namely the lazy (on-demand) approach or the eager (in-advance) approach[31]. In the former approach, the data is accessed based on the user query whereas in the latter approach, the data is accumulated and stored aprior and queries are posed against this stored data. Data warehousing is based on the latter approach[31] where data is extracted from disparate data sources, transformed and then loaded into a central repository, called a data warehouse[12]. A data warehouse stores historical data with the purpose of supporting decision making[12]. The queries associated with decision making are analytical and exploratory in nature. These queries, when processed against a large data warehouse, consume a lot of time for processing. As a result, the response time is high. Several effective solutions exist for organizing data for analytical queries, but the issue of poor query performance still needs to be addressed [18]. Traditional query optimization [5, 9] and indexing techniques [17], despite attempting to address this problem, could not scale up for longer and complex analytical queries on large data sets [16]. An alternative approach, which has been widely used to address the poor query response time of

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 521–530.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

analytical queries, is to materialize views in a data warehouse. Materialized views are constructed with the purpose of improving the query performance of the system [10, 16, 20]. To accomplish this, these should contain information that is relevant for answering future queries. This information cannot be arbitrarily identified, as it may result in materialized views that are not capable of answering future queries and thereby result in an unnecessary space overhead. Selecting such information, from the large amount of data available in the data warehouse, is referred to as the view selection problem [6]. View selection deals with selecting an appropriate set of views for materialization that can improve the query response time even while conforming to the resource constraints like storage space, memory etc [6, 8].

All possible views cannot be materialized, as the number of possible views grows exponentially with the number of dimensions and it becomes infeasible to materialize views for higher dimensional data sets within the storage space constraint. One way to construct materialized views is by materializing all possible views. Further, optimal selection of subsets of views, from among all possible views, is shown to be NP-Complete [10]. Alternative approaches exist in literature, that select views heuristically or empirically. Heuristic based approaches are greedy [8, 10, 22, 23, 24, 27, 28, 30] or evolutionary [11, 14]. Empirically, views are selected by assessing and monitoring queries on various factors like frequency, data size etc. [15, 19]. This paper focuses on constructing views empirically from queries posed on the data warehouse in the past.

Most query based approaches presented in literature [1, 2, 3, 21, 32] for constructing materialized views are workload driven, and based on the assumption that past queries are indicative of queries likely to be posed in future. In this paper, an approach is presented that first identifies subject area specific queries, from amongst previously posed queries. This is followed by selecting frequent queries from amongst subject specific queries. The materialized views constructed using these frequent queries are capable of answering most future queries in reduced query response times.

The paper is organized as follows: The frequent queries selection approach is given in section 2 and an example based on it in section 3. Section 4 is the conclusion.

2 Approach

The approach aims to select queries, from amongst all queries posed on the data warehouse in the past, containing relevant and required information for answering future queries. The approach selects such queries by first grouping similar queries posed in the past into groups or subject areas. This is followed by selecting queries that access frequently accessed data, referred to as frequent queries, from amongst the queries in each subject area. These selected frequent queries are subject specific and contain information that has high likelihood of answering future queries. The approach is similar to the one given in [25, 26, 29]. It differs in the similarity measure used, and the technique used for subject area identification and frequent query selection. The two phases used in the approach, as mentioned above, are discussed next.

2.1 Subject Area Identification

The approach considers the fact that most queries posed on the data warehouse happen to be subject specific and only few queries are posed a cross subject lines. The subject specific data, accessed in the past, needs to be consolidated. Accordingly, it is appropriate to identify subject specific queries, from amongst all the queries posed in the past. The approach identifies such subject specific queries by grouping the previously posed queries using the Nearest Neighbor Clustering technique [13]. The similarity between queries is computed using the DICE coefficient [7] according to which, the Similarity between a pair of queries Q_i and Q_j, i.e. $\mathrm{Sim}(Q_i, Q_j)$, based on Dice Coefficient measure [7], is given by

$$\mathrm{Sim}(Q_i\,,\,Q_j) = \frac{2\,|\,R(Q_i) \cap R(Q_j)\,|}{|\,R(Q_i)\,| + |\,R(Q_j)\,|}$$

where $R(Q_i)$ and $R(Q_j)$ are the relations accessed by queries Q_i and Q_j respectively.

Using the above similarity measure, the similarity between the previously posed queries is computed and a similarity matrix depicting this similarity is constructed. The Nearest Neighbor clustering technique uses the similarity matrix to group closely related queries into clusters. Each such cluster of queries, so identified, represents a subject area. The algorithm SubjectAreaIdentification, based on nearest neighbor clustering technique[13], used to identify subject areas, is given in Fig. 1. This algorithm takes the previously posed queries, the similarity matrix and a minimum query similarity threshold as input and produces the cluster of queries as output.

The algorithm first intitalizes the query count QC and the cluster count CC to 1. It then assigns the first query Q_{QC}, from amongst previously posed queries Q_p, into cluster C_{CC}. The next query in Q_p is then considered and its nearest neighbor, i.e. in terms of having maximum similarity, is identified from queries that are already assigned to clusters. If this similarity is greater than or equal to the minimum similarity threshold ε, then the query is assigned to the corresponding cluster. Otherwise, a new cluster is created and the query is assigned to it. This continues till all queries have been considered. The identified clusters specify the various subject areas.

```
ALGORITHM SubjectAreaIdentification
Inputs:  $Q_P$ : Previously posed Queries queries,
         $\varepsilon$ : Minimum query similarity threshold,
         SimMat  : Similarity Matrix showing similarity between queries
Output:     Cluster of Queries $C_Q$
Method:
STEP 1  Set query count QC = 1 and cluster count CC = 1.
STEP 2  Assign query $Q_{QC}$ in $Q_P$ to cluster $C_{CC}$
STEP 3  Increment QC by 1.
STEP 4  Find nearest neighbour of $Q_{QC}$ among the queries in $Q_P$ already assigned to clusters.
STEP 5  Using the SimMat, let MaxSim denote the similarity between $Q_{QC}$ and it's nearest
         Neighbor query in the existing clusters. Suppose the nearest is in cluster K
STEP 6  If MaxSim is greater than or equal to $\varepsilon$, then assign $Q_{QC}$ to $C_K$ otherwise increment
           CC by one and assign $Q_{QC}$ to $C_{CC}$
STEP 7  If every query has been considered then STOP else go to STEP 3.
```

Fig. 1. Algorithm SubjectAreaIdentification based on Nearest Neighbor[13]

There can be large numbers of queries in each subject area. Most of these may access similar information with others accessing dissimilar information. The queries accessing similar information are indicative of the information more likely to be accessed in future. Thus selecting such queries is beneficial, as the information accessed by them is more likely to be accessed by most future queries. Selection of such queries is discussed next.

2.2 Frequent Query Selection

As mentioned above, materialized views can reduce the query response time if they are capable of providing answers to most future queries. This necessitates that materialized views contain relevant and required information. This information cannot be arbitrarily identified, as it is required to contain information that is capable of providing answers to

```
INPUT:
    QS: Set of Queries
    $T_Q$: Query Transaction Table
    $\theta$: Minimum Query Support Threshold
    M: Minimum Transaction per Scan
OUTPUT:
    FQS: Frequent Query Set
METHOD:
Initially for each RelationSet
 Solid Square SS = NULL (Frequent)
 Solid Circle SC= NULL   (Infrequent)
 Dashed Square DS = NULL (Suspected Frequent)
 Dashed Circle DC = {All 1-Relation Sets} (Suspected Infrequent)
 WHILE (DS!=0) OR (DC!=0)
 BEGIN
     Read M Queries from QS into T
     FOR all queries in T
     BEGIN
         FOR each relation set R in DS and DC
         BEGIN
         IF R is in T
             Increment count of R, i.e. $R_C$, by 1
         FOR each relation set R in DC
         IF ($R_C$ is greater than or equal to $\theta$) THEN
             Move R from DC to DS
             IF (any immediate superset SR of R has all its subsets in SS and DS) THEN
                 Add a new relation set SR in DC
         END
         FOR each relation set R in DS
             IF (R has been counted through all queries) THEN
                 Move R to SS
         FOR each relation set R in DC
             IF (R has been counted through all queries) THEN
                 Move R to SC
     END
 Increment M by M
END
Frequent Relations Sets FRS contains relation sets in SS
Compute FQS as the queries in QS that contains atleast one frequent relation set in FRS.
```

Fig. 2. Algorithm FrequentQuerySelection based on DIC[4]

future queries. The approach identifies such relevant and required information by selecting queries that seek frequently accessed information. These queries, referred to as frequent queries, provide information that have high likelihood of answering future queries and therefore can appropriately be used for constructing the materialized view for the corresponding subject area. The approach uses the association rule mining technique DIC(Dynamic Itemset Counting) [4]. The frequent queries selection algorithm, based on DIC [4], is given in Fig. 2. The algorithm takes queries in a subject area along with the minimum query support threshold as input, and produces a set of frequent queries in the corresponding subject area as output.

The algorithm first marks the empty relation set NULL with a solid square and marks 1-relation set with dashed circles. Remaining relations sets are left unmarked. It then reads M queries from the query set QS into a transaction File. This is followed by incrementing the respective counters for relation sets in transaction files and marked with dashes. If a dashed circle's count for a relation set R exceeds the minimum support threshold θ, turn the dashed circle into a dashed square for R. If any immediate superset SR of R has all of its subsets as solid or dashed squares, mark a new relation set SR with a dashed circle. After the dashed relation sets have been counted through all the queries, make dashes as solid. The number of scans is incremented by M and the process is repeated. The process continues till there are no more dashed square and circle relation sets. The relation sets marked as solid square are the frequent relation sets. The queries containing any of the frequent relation sets are then selected as the frequent queries. In a similar manner, frequent queries in each subject area are selected. Materialized views can thereafter be constructed using the frequent queries selected in the respective subject area. These materialized views contain information more likely to be accessed by most future queries. This in turn would reduce the query response time and lead to efficient decision making.

3 Example

Consider the relations accessed by previously posed queries on the data warehouse given in Fig. 3.

Q1	Employee, Company, BranchDetails	Q11	Company, Module, Project
Q2	Hotel, Room, Guest	Q12	Room, Guest, Payment
Q3	Hotel, Room, Booking	Q13	Booking, Guest, Payment
Q4	Hotel, Room, Payment	Q14	Employee, Hotel, TouringDetails
Q5	Employee, Department, BranchDetails	Q15	Company, Project, BranchDetails
Q6	Hotel, Guest, Payment	Q16	Employee, Department, Company
Q7	Hotel, Room, Guest	Q17	Department, Module, Project
Q8	Company, Department, BranchDetails	Q18	Hotel, Room, Payment
Q9	Employee, Department, Project	Q19	Employee, Department, Project
Q10	Hotel, Room, Booking	Q20	Empoyee, Company, BranchDetails

Fig. 3. Relations accessed by Previously Posed Queries on a Data Warehouse

The similarity between the queries in Fig. 3 is computed using the DICE Coefficient[7]. These similarities are then used to construct a similarity matrix, given in Fig. 4.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
Q1	1	0	0	0	0.666	0	0	0.666	0.333	0	0.333	0	0	0.333	0.666	0.666	0	0	0.333	1
Q2	0	1	0.666	0.666	0	0.666	1	0	0	0.666	0	0.666	0.333	0.333	0	0	0	0.666	0	0
Q3	0	0.666	1	0.666	0	0.666	0.66	0	0	0.666	0	0.333	..333	0.333	0	0	0	0.666	0	0
Q4	0	0.666	0.666	1	0	0.666	0.666	0	0	0.666	0	0.666	0.333	0.333	0	0	0	1	0	0
Q5	0.666	0	0	0	1	0	0	0.666	0.666	0	0.333	0	0	0.333	0.333	0.666	0.333	0	0.666	0.666
Q6	0	0.666	0.666	0.666	0	1	0.666	0	0	0.333	0	0.666	0.666	0.333	0	0	0	0.666	0	0
Q7	0	1	0.666	0.666	0	0.666	1	0	0	0.666	0	0.666	0.333	0.333	0	0	0	0.666	0	0
Q8	0.666	0	0	0	0.666	0	0	1	0.333	0	0	0	0	0	0.666	0.666	0.333	0	0.333	0.666
Q9	0.333	0	0	0	0.666	0	0	0.333	1	0.333	0.333	0	0	0.333	0.333	0.666	0.666	0	1	0.666
Q10	0	0.666	0.666	0.666	0	0.333	0.666	0	0.333	1	0	0.333	0	0.333	0.333	0	0.333	0.666	0.333	0
Q11	0.333	0	0	0	0.333	0	0	0	0.333	0.333	1	0	0	0.333	0.333	0.333	0.666	0	0.666	0.333
Q12	0	0.666	0.333	0.666	0	0.666	666	0	0	0.333	0	1	0.666	0	0	0	0	0.666	0	0
Q13	0	0.333	0.333	0.333	0	0.666	0.333	0	0	0	0	0.666	1	0	0	0	0	0.333	0	0
Q14	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0	0.333	0.333	0.333	0	0	1	0	0.333	0	0.333	0.333	0.333
Q15	0.666	0	0	0	0.333	0	0	0.666	0.333	0.333	0.333	0	0	0	1	0.333	0.333	0	0.333	0.333
Q16	0.666	0	0	0	0.666	0	0	0.666	0.666	0	0.333	0	0	0.333	0.333	1	0.333	0	0.666	0.666
Q17	0	0	0	0	0.333	0	0	0.333	0.333	0.333	0.333	0	0	0	0.333	0.333	1	0	0.666	0
Q18	0	0.666	0.666	1	0	0.666	0.666	0	0	0.666	0	0.666	0.333	0.333	0	0	0	1	0	0
Q19	0.333	0	0	0	0.666	0	0	0.333	1	0.333	0.666	0	0	0.333	0.333	0.666	0.666	0	1	0.333
Q20	0.666	0	0	0	0.666	0	0	0.666	0.666	0	0.333	0	0	0.333	0.333	0.666	0	0	0.333	1

Fig. 4. Similarity Matrix showing similarity between queries Q1 . . Q20

Using the similarity matrix in Fig. 4, the queries in Fig. 3, minimum query similarity threshold ε=0.5, the subject areas are identified as given in Fig. 5.

Initialize QC=1 and CC=1
Assign Q_{QC} i.e. Q_1 to cluster C_{CC} i.e. C_1
Now $C_1 = \{Q_1\}$
Set QC=QC+1 i.e. QC=2
MaxSim of Q_{QC} i.e. Q_2 is 0 with nearest neighbor query Q_1
Since MaxSim < ε, set CC=CC+1 i.e. CC=2 and C_{CC} i.e. $C_2=\{Q_2\}$
Set QC=QC+1 i.e QC=3
MaxSim of Q_{QC} i.e Q_3 is 0.666 with its nearest neighbor query Q_2 in C_2
Since MaxSim > ε, Assign Q_3 to C_2
$C_2=\{Q_2, Q_3\}$
Set QC=QC+1 i.e. QC=4
MaxSim of Q_{QC} i.e Q_4 is 0.666 with its nearest neighbor query Q_2 and Q_3
Since MaxSim > ε, Assign Q_4 to C_2
$C_2=\{Q_2, Q_3, Q_4\}$
Set QC=QC+1 i.e. QC=5
MaxSim of Q_{QC} i.e. Q_5 is 0.666 with its nearest neighbor Q_1
Since MaxSim > ε, Assign Q_5 to C_1
$C_1=\{Q_1, Q_5\}$
Set QC=QC+1 i.e. QC=6
MaxSim of Q_{QC} i.e. Q_6 is 0.666 with its nearest neighbor Q_2, Q_3 and Q_4
Since MaxSim > ε, Assign Q_6 to C_2
$C_2=\{Q_2, Q_3, Q_4, Q_6\}$
The above steps are carried out in the similar manner to identify cluster of queries. The cluster of queries C_1, C_2, C_3, C_4 and C_5 identified represent the five subject areas S_1, S_2, S_3, S_4 and S_5 respectively as given below:
S_1={Q1, Q5, Q8, Q9, Q15, Q16, Q19, Q20}
S_2= {Q2, Q3, Q4, Q6, Q7, Q10, Q12, Q13, Q18}
$S_3=\{Q_{11}\}$, $S_4=\{Q_{14}\}$, S_5={Q17}

Fig. 5. Subject Area Identification using previously posed queries Q1 . . Q20

Next, the frequent queries are selected in each subject area using the FrequentQuerySelection algorithm given in Fig. 2. Consider subject area S1. The queries, along with the relation accessed by them, are given in Fig. 6.

Q1	Employee	Company	BranchDetails
Q5	Employee	Department	BranchDetails
Q8	Company	Department	BranchDetails
Q9	Employee	Department	Project
Q15	Company	Project	BranchDetails
Q16	Employee	Department	Company
Q19	Employee	Department	Project
Q20	Employee	Company	BranchDetails

Fig. 6. Queries along with relations accessed by them

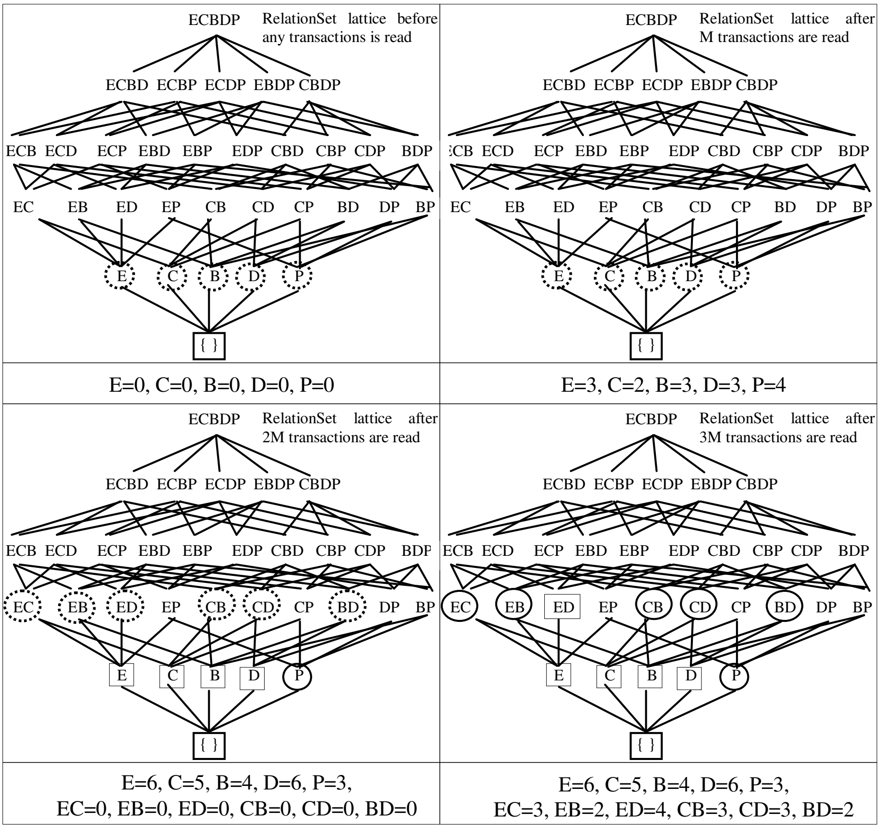

ECBDP
RelationSet lattice before any transactions is read
ECBD ECBP ECDP EBDP CBDP
ECB ECD ECP EBD EBP EDP CBD CBP CDP BDP
EC EB ED EP CB CD CP BD DP BP
E C B D P
{ }
E=0, C=0, B=0, D=0, P=0
ECBDP
RelationSet lattice after M transactions are read
ECBD ECBP ECDP EBDP CBDP
ECB ECD ECP EBD EBP EDP CBD CBP CDP BDP
EC EB ED EP CB CD CP BD DP BP
E C B D P
{ }
E=3, C=2, B=3, D=3, P=4
ECBDP
RelationSet lattice after 2M transactions are read
ECBD ECBP ECDP EBDP CBDP
ECB ECD ECP EBD EBP EDP CBD CBP CDP BDP
EC EB ED EP CB CD CP BD DP BP
E C B D P
{ }
E=6, C=5, B=4, D=6, P=3,
EC=0, EB=0, ED=0, CB=0, CD=0, BD=0
ECBDP
RelationSet lattice after 3M transactions are read
ECBD ECBP ECDP EBDP CBDP
ECB ECD ECP EBD EBP EDP CBD CBP CDP BDP
EC EB ED EP CB CD CP BD DP BP
E C B D P
{ }
E=6, C=5, B=4, D=6, P=3,
EC=3, EB=2, ED=4, CB=3, CD=3, BD=2

Fig. 7. Identification of frequent RelationSet in subject area S1 E: Employee, C: Company, B: BranchDetails, D: Department, P: Project

Using the query information in Fig. 6, the frequent relation set in S1 for minimum query support threshold θ=0.5 and M=4 is identified to be {Employee, Department} as given in Fig. 7. Thus, the frequent queries selected are Q5, Q9, Q16 and Q19 as they contain {Employee, Department} in their FROM clause. In a similar manner, the frequent relation set identified in S2 is {Hotel, Room} and thus the frequent queries are {Q2, Q3, Q4, Q7, Q10, Q18}. Such selected frequent queries would be used to construct the materialized views for the respective subject areas.

4 Conclusion

In this paper an approach, that selects frequent queries from previously posed queries on the data warehouse, is presented. The approach first identifies clusters of closely related previously posed queries. Each such cluster defines a subject area. The approach then selects frequent queries, from amongst all queries in each subject area. The selected frequent queries in each subject area identify the data accessed frequently in the past and therefore have high likelihood of being accessed by future queries. Materialized views constructed using these queries would improve the query response time. Further, materialized views being subject specific, as they are constructed for a specific subject area and most future queries posed on the data warehouse are subject specific, fewer numbers of views are required to answer future queries. As a result, the response time would be further reduced thereby aiding in efficient decision making.

References

1. Agrawal, S., Chaudhari, S., Narasayya, V.: 'Automated Selection of Materialized Views and Indexes in SQL databases'. In: 26th International Conference on Very Large Data Bases, VLDB 2000, Cairo, Egypt, pp. 495–505 (2000)
2. Aouiche, K., Darmont, J.: Data mining-based materialized view and index selection in data warehouse. Journal of Intelligent Information Systems, 65–93 (2009)
3. Baralis, E., Paraboschi, S., Teniente, E.: 'Materialized View Selection in a Multidimensional Database'. In: 23rd International Conference on Very Large Data Bases, VLDB 1997, Athens, Greece, pp. 156–165 (1997)
4. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic Itemset Counting and Implication Rules for Market Basket Data. In: SIGMOD Record, New York, vol. 6(2), pp. 255–264 (June 1997)
5. Chaudhuri, S., Shim, K.: Including Groupby in Query Optimization. In: Proceedings of the International Conference on Very Large Database Systems (1994)
6. Chirkova, R., Halevy, A.Y., Suciu, D.: A Formal Perspective on the View Selection Problem. In: Proceedings of VLDB, pp. 59–68 (2001)
7. Frakes, W.B., Baeza-Yates, R.: Information Retrieval, Data Structure and Algorithms. Prentice-Hall (1992)
8. Gupta, H., Mumick, I.S.: Selection of Views to Materialize in a Data warehouse. IEEE Transactions on Knowledge & Data Engineering 17(1), 24–43 (2005)

9. Gupta, A., Harinarayan, V., Quass, D.: Generalized Projections: A Powerful Approach to Aggregation. In: Proceedings of the International Conference of Very Large Database Systems (1995)
10. Harinarayan, V., Rajaraman, A., Ullman, J.D.: Implementing Data Cubes Efficiently. In: ACM SIGMOD, Montreal, Canada, pp. 205–216 (1996)
11. Horng, J.T., Chang, Y.J., Liu, B.J., Kao, C.Y.: Materialized View Selection Using Genetic Algorithms in a Data warehouse System. In: Proceedings of the 1999 Congress on Evolutionary Computation, Washington D.C., USA, vol. 3 (1999)
12. Inmon, W.H.: Building the Data Warehouse, 3rd edn. Wiley Dreamtech India Pvt. Ltd. (2003)
13. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
14. Lawrence, M.: Multiobjective Genetic Algorithms for Materialized View Selection in OLAP Data Warehouses. In: GECCO 2006, Seattle Washington, USA, July 8-12 (2006)
15. Lehner, W., Ruf, T., Teschke, M.: Improving Query Response Time in Scientific Databases Using Data Aggregation. In: Thoma, H., Wagner, R.R. (eds.) DEXA 1996. LNCS, vol. 1134. Springer, Heidelberg (1996)
16. Mohania, M., Samtani, S., Roddick, J., Kambayashi, Y.: Advances and Research Directions in Data Warehousing Technology. Australian Journal of Information Systems (1998)
17. O’Neil, P., Graefe, G.: Multi-Table joins through Bitmapped Join Indices. SIGMOD Record 24(3), 8–11 (1995)
18. Shah, B., Ramachandran, K., Raghavan, V.: A Hybrid Approach for Data Warehouse View Selection. International Journal of Data Warehousing and Mining 2(2), 1–37 (2006)
19. Teschke, M., Ulbrich, A.: Using Materialized Views to Speed Up Data Warehousing. Technical Report, IMMD 6, Universität Erlangen-Nümberg (1997)
20. Theodoratos, D., Sellis, T.: Data Warehouse Configuration. In: Proceeding of VLDB, Athens, Greece, pp. 126–135 (1997)
21. Theodoratos, D., Xu, W.: Constructing Search Spaces for Materialized View Selection’. In: 7th ACM Internatioanl Workshop on Data Warehousing and OLAP, DOLAP 2004, Washington, USA (2004)
22. Vijay Kumar, T.V., Ghoshal, A.: A Reduced Lattice Greedy Algorithm for Selecting Materialized Views. In: Prasad, S.K., Routray, S., Khurana, R., Sahni, S. (eds.) ICISTM 2009. CCIS, vol. 31, pp. 6–18. Springer, Heidelberg (2009)
23. Vijay Kumar, T.V., Haider, M., Kumar, S.: Proposing Candidate Views for Materialization. In: Prasad, S.K., Vin, H.M., Sahni, S., Jaiswal, M.P., Thipakorn, B. (eds.) ICISTM 2010. CCIS, vol. 54, pp. 89–98. Springer, Heidelberg (2010)
24. Vijay Kumar, T.V., Haider, M.: A Query Answering Greedy Algorithm for Selecting Materialized Views. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010. LNCS (LNAI), vol. 6422, pp. 153–162. Springer, Heidelberg (2010)
25. Vijay Kumar, T.V., Jain, N.: Selection of Frequent Queries for Constructing Materialized Views in Data Warehouse. The IUP Journal of Systems Management 8(2), 46–64 (2010)
26. Vijay Kumar, T.V., Goel, A., Jain, N.: Mining Information for Constructing Materialised Views. International Journal of Information and Communication Technology 2(4), 386–405 (2010)
27. Vijay Kumar, T.V., Haider, M.: Greedy Views Selection Using Size and Query Frequency. In: Unnikrishnan, S., Surve, S., Bhoir, D. (eds.) ICAC3 2011. CCIS, vol. 125, pp. 11–17. Springer, Heidelberg (2011)

28. Vijay Kumar, T.V., Haider, M., Kumar, S.: A View Recommendation Greedy Algorithm for Materialized Views Selection. In: Dua, S., Sahni, S., Goyal, D.P. (eds.) ICISTM 2011. CCIS, vol. 141, pp. 61–70. Springer, Heidelberg (2011)
29. Vijay Kumar, T.V., Devi, K.: Frequent Queries Identification for Constructing Materialized Views. In: The Proceedings of the International Conference on Electronics Computer Technology, ICECT 2011, Kanyakumari, Tamil Nadu, April 8-10, vol. 6, pp. 177–181. IEEE (2011)
30. Vijay Kumar, T.V., Haider, M.: Selection of Views for Materialization Using Size and Query Frequency. In: Das, V.V., Thomas, G., Lumban Gaol, F. (eds.) AIM 2011. CCIS, vol. 147, pp. 150–155. Springer, Heidelberg (2011)
31. Widom, J.: Research Problems in Data Warehousing. In: 4th International Conference on Information and Knowledge Management, Baltimore, Maryland, pp. 25–30 (1995)
32. Yang, J., Karlapalem, K., Li, Q.: Algorithms for Materialized View Design in Data Warehousing Environment. The Very Large databases (VLDB) Journal, 136–145 (1997)

HURI – A Novel Algorithm for Mining High Utility Rare Itemsets

Jyothi Pillai[1], O.P. Vyas[2], and Maybin Muyeba[3]

[1] Associate Professor, Bhilai Institute of Technology, Durg-491001, Chhattisgarh, India
jyothi_rpillai@rediffmail.com
[2] Professor, Indian Institute of Information Technology, Allahabad, Uttar Pradesh, India
dropvyas@gmail.com
[3] Senior Lecturer, Deptt. of Computing and Mathematics,
Manchester Metropolitan Univ., U.K.
M.Muyeba@mmu.ac.uk

Abstract. In Data mining field, the primary task is to mine frequent itemsets from a transaction database using Association Rule Mining (ARM). Utility Mining aims to identify itemsets with high utilities by considering profit, quantity, cost or other user preferences. In market basket analysis, high consideration should be given to utility of item in a transaction, since items having low selling frequencies may have high profits. As a result, High Utility Itemset Mining emerged as a revolutionary field in Data Mining. Rare itemsets provide useful information in different decision-making domains. High Utility Rare Itemset Mining, HURI algorithm proposed in [12], generate high utility rare itemsets of users' interest. HURI is a two-phase algorithm, phase 1 generates rare itemsets and phase 2 generates high utility rare itemsets, according to users' interest. In this paper, performance evaluation and complexity analysis of HURI algorithm, based on different parameters have been discussed which indicates the efficiency of HURI.

Keywords: Association Rule Mining, Utility Mining, Rare itemset, High Utility Rare itemset Mining.

1 Introduction

The most important assets of any corporation might be data. Data Mining extracts diamonds of knowledge from historical data and predicts outcomes of future situations, in the form of patterns and associations from large databases. One of the most widely used areas of data mining for retail industry is marketing. 'Market basket analysis' is a marketing method used by many retailers to promote products. Data Mining uses information about products purchased by customers to predict which products they would buy if given special offers [13]. Apriori algorithm, the first ARM algorithm developed by Rakesh Agrawal [10], firstly identifies those itemsets from sales dataset, having frequencies above a specified threshold and then generates association rules. Itemset Mining is done using Association Rule Mining (ARM). Most classical association rule mining algorithms consider utilities of the itemsets to be

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 531–540.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

equal [15]. In real life retail marketing or business other utility factors such as quantity, cost, revenue or profit of an item should also be considered. Yao et al defined Utility as a measure of how useful or profitable an itemset is [15]. In many practical situations rare ones are of higher interest (e.g., in medical databases, rare combinations of symptoms might provide useful insights for the physicians)[6]. Rare itemset mining is a challenging topic; the rare combinations of items in the itemset with high utilities provide very useful insights to the user. For example, a sales manager may not be interested in frequent itemsets that do not generate significant profit. In many real-life applications, high-utility itemsets consist of rare items, i.e. itemsets that occur infrequently in the transaction data set but may contribute a large portion of the profit; for eg, customers purchase microwave ovens or LEDs rarely as compared to bread, butter, etc. The former may yield more profit for the supermarket than the latter.

Jyothi et al proposed High Utility Rare Itemset Mining (HURI) algorithm [12], which finds high profitable rare itemsets according to user's perspective in two phases. In first phase, rare itemsets having support value less than the maximum support threshold are generated. Second phase finds high utility rare itemsets having utility value greater than the minimum utility threshold. The novel contribution of HURI is to effectively find rare itemsets, which are of high utility according to users' preferences. In this paper, performance evaluation of HURI based on different values of parameters such as support, utility, etc. have been described. The rest of paper is organized as follows. In section 2, we discuss some related works: section 3 presents the HURI algorithm. Section 4 performs the evaluation of HURI and section 5 presents conclusion and future work.

2 Related Work

The basic bottleneck in association rule mining is the rare itemset problem. In many applications, some items appear more frequently in the data, while others rarely appear. If frequencies of items vary, two problems may be encountered – (1) If minsup is set too high, then rules of rare items will not be found (2) To find rules that involve both frequent and rare items, minsup has to be set very low, where minsup is the minimum support of an item. This may cause combinatorial explosion in the number of itemsets.

Utility mining is now an important association rule-mining paradigm. Yao et al focuses on the measures used for utility-based itemset mining [8]. A unified framework is proposed for incorporating utility based measures into the data mining process via a unified utility function. Ying et al proposed a Two-Phase algorithm [14] that discovers high utility itemsets highly efficiently. Rare itemsets provide very useful information in real-life applications such as security, business strategies, biology, medicine and super market shelf management. For example [5] shows that normal behavior is very frequent whereas abnormal or suspicious behavior is less frequent. For example, from a marketing strategy perspective, it is important to identify product combinations that have a significant impact on company's bottom line, having highest revenue generating power [7].

S. Shankar et al presents a novel algorithm Fast Utility Mining (FUM) in [4], which finds all high utility itemsets within the given utility constraint threshold. The authors also suggest a novel method of generating different types of itemsets such as High Utility and High Frequency itemsets (HUHF), High Utility and Low Frequency itemsets (HULF), Low Utility and High Frequency itemsets (LUHF) and Low Utility and Low Frequency itemsets (LULF) using a combination of FUM and Fast Utility Frequent mining (FUFM) algorithms.

A different approach known as Apriori Inverse [9], involves use of maximum support measure to generate candidate itemsets, i.e., only items with a lower support than a given threshold are considered. Then rules are generated by an Apriori approach. In [6], L. Szathmary et al presented a novel algorithm for computing all rare itemsets by splitting the rare itemset mining task into two algorithms , (i) a naïve one that relies on an Apriori-style enumeration, Apriori-rare and (ii) an optimized method that limits the exploration to frequent generators only. Apriori-rare generates a set of all minimal rare generators, also called MRM. To retrieve all rare itemsets from minimal rare itemset, a prototype algorithm called "A Rare Itemset Miner Algorithm (ARIMA)" was proposed. ARIMA generates the set of all rare itemsets, splits into two sets: the set of rare itemsets having a zero support and the set of rare itemsets with non-zero support.

A totally different approach to all these algorithms presented demands developing new algorithms to tackle new challenges. Apriori-inverse [9], is a more intricate variation of the traditional Apriori algorithm. The main idea is that given a user-specified maximum support threshold, MaxSup and a derived MinAbsSup value, a rule X is rare if Sup(X)<MaxSup and Sup(X)>MinAbsSup. M. Adda et al [5] proposed a framework to represent different categories of interesting patterns and then instantiate it to the specific case of rare patterns. A generic framework called AfRIM for Apriori Rare itemset was presented to mine patterns based on the Apriori approach. The generalized Apriori framework was instantiated to mine rare itemsets. In [11], Lan et al proposed rare-utility itemset, by considering profit and quantity of each item in a transaction and an algorithm TP-RUIMD (Two-Phase Algorithm for Mining Rare Utility Itemsets in Multiple Databases), to find rare-utility itemsets in a multi-database environment. HURI algorithm proposed in [12] generates high-utility rare-itemsets, by considering the utility of itemsets other than the frequency of items in the transaction set. The utility of items is decided by considering factors such as profit, sale, temporal aspects, etc. of items. By using HURI, high-utility rare itemsets can be generated based on minimum threshold values and user preferences.

3 HURI Algorithm

3.1 Definitions

In this section, HURI algorithm is presented with formal definitions and examples to illustrate the approach.

Utility Mining. Let D1 (Table 1) be a given transaction database with a set of transactions {T1,T2,…,Tn} and a set of quantities of items I={i1, i2, i3,… , im} where each item i ε I has a set of utilities defined as U={u1, u2, u3,… , uk} (Table II). For example in transaction T29, the quantities of items A001, B002, C003, D004, E005… are 1,3, 0,1,1… respectively. The utility of an itemset X, i.e., u(X), is the sum of utilities of itemset X in all transactions containing X. An itemset X is called a high utility itemset if and only if u(X) >= min_utility, where min_utility is a user-defined minimum utility threshold [8]. Identification of the itemsets with high utilities is called as Utility Mining [5].

Utility Table. A utility Table UT (Table 2) contains items and their utility values where each item i has some utility value uj in U={u1, u2, u3,… , uk } for some k > 0. For example utility of item A001 from D1 is u(A001) = 4 in (Table 2).

Internal Utility. Internal utility value of item ip in a transaction T_q, denoted o(i_p, T_q) is the value of an item ip in a transaction T_q (Table 2), reflecting the occurrence of item. For eg., internal utility of item A0001 in transaction T1 is o(A001, T1) = 1, while internal utility of item A0001 in D1 is o(A001, D1) = 21(Table 1).

External Utility. External utility value of an item is a numerical value s(ip) associated with an item ip such that s(i_p)=u(i_p), where u is a utility function, a function relating specific values in a domain according to user preferences (Table 2). From Table 3, external utility of item A0001 in D1 is s(A0001) = u(A0001) = 4.

Item Utility. The utility of an item i_p in a transaction T_q, denoted U(i_p, T_q) is product of o(i_p, T_q) and s(i_p), i.e. internal and external utility respectively . For eg., total utility of item A0001 in D1 is U(A001) = s(A001)*o(A001)=4*21=84.

Transaction Utility. The transaction utility value of a transaction, denoted as U(Tq) is the sum of utility values of all items in a transaction Tq (Table I, Table II). From Table I and Table 3, the transaction utility of the transaction T1 from D1, U(T1) = U(A001)+U(B002)+U(C003)+ … + U(T020)=39.

Rare Itemset. Mining Rare itemsets are itemsets that occur infrequently in the transaction data set.

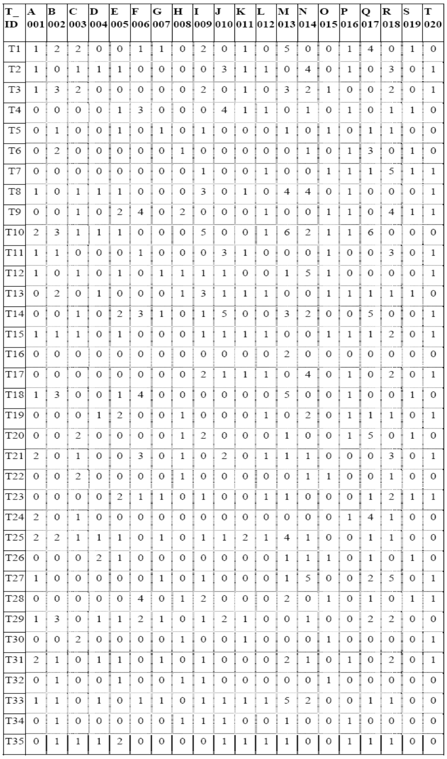
The HURI algorithm(Fig. 1) can be best understood by transactional dataset D1 (Table 1) and Item utility table (Table 2). Given a user-specified maximum support threshold maxsup, and a generated minabssup value, we are interested in a rule X if sup(X) < maxsup and sup(X) > minabssup. Rare rules are generated in the same manner as in apriori rule generation. Apriori-Inverse produces rare rules that do not consider any itemsets above maxsup. In Apriori inverse algorithm, rare itemsets are itemsets which fall below maxsup value. In HURI Algorithm (Fig. 1), high utility rare itemsets are generated in two phases:-

(i)In first phase, rare itemsets are generated by considering those itemsets which have support value less than the maximum support threshold (using apriori-inverse concept). Table III lists rare itemsets generated from dataset D1(Table 1).

(ii)In second phase, high utility rare itemsets having utility value greater than the minimum utility threshold are generated (Table 4).

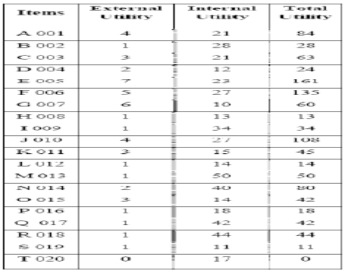
By applying HURI algorithm [12] on Transaction dataset described in Table 1 and by setting the value of maximum support threshold to 40%, the rare itemsets generated are listed in Table 3. Then HURI algorithm generates high utility rare itemsets which fall below a maximum support value but above a user provided high utility threshold. For example by setting high utility threshold as 45, the high utility rare itemsets generated from D1 are listed in Table 4. Both HURI and apriori inverse algorithm considers utility values of all items in transaction set in addition to frequency. But apriori inverse produces only rare itemsets whereas HURI produces high utility rare itemsets according to users' interest.

Table 1. Transaction Dataset D



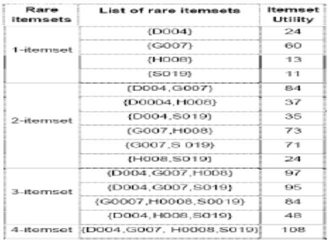
T_ ID	A 001	B 002	C 003	D 004	E 005	F 006	G 007	H 008	I 009	J 010	K 011	L 012	M 013	N 014	O 015	P 016	Q 017	R 018	S 019	T 020
T1	1	2	2	0	0	1	1	0	2	0	1	0	5	0	0	1	4	0	1	0
T2	1	0	1	1	1	0	0	0	0	3	1	1	0	4	0	1	0	3	0	1
T3	1	3	2	0	0	0	0	0	2	0	1	0	3	2	1	0	0	2	0	1
T4	0	0	0	0	1	3	0	0	0	4	1	1	0	1	0	1	0	1	1	0
T5	0	1	0	0	1	0	1	0	1	0	0	0	1	0	1	0	1	1	0	0
T6	0	2	0	0	0	0	0	1	0	0	0	0	0	1	0	1	3	0	1	0
T7	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	1	1	5	1	1
T8	1	0	1	1	1	0	0	0	3	0	1	0	4	4	0	1	0	0	0	1
T9	0	0	1	0	2	4	0	2	0	0	0	1	0	0	1	1	0	4	1	1
T10	2	3	1	1	1	0	0	0	5	0	0	1	6	2	1	1	6	0	0	0
T11	1	1	0	0	0	1	0	0	0	3	1	0	0	0	1	0	0	3	0	1
T12	1	0	1	0	1	0	1	1	1	1	0	0	1	5	1	0	0	0	0	1
T13	0	2	0	1	0	0	0	1	3	1	1	1	0	0	1	1	1	1	1	0
T14	0	0	1	0	2	3	1	0	1	5	0	0	3	2	0	0	5	0	0	1
T15	1	1	1	0	1	0	0	0	1	1	1	1	0	0	1	1	1	2	0	1
T16	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
T17	0	0	0	0	0	0	0	0	2	1	1	1	0	4	0	1	0	2	0	1
T18	1	3	0	0	1	4	0	0	0	0	0	0	5	0	0	1	0	0	1	0
T19	0	0	0	1	2	0	0	1	0	0	0	1	0	2	0	1	1	1	0	1
T20	0	0	2	0	0	0	0	1	2	0	0	0	1	0	0	1	5	0	1	0
T21	2	0	1	0	0	3	0	1	0	2	0	1	1	1	0	0	0	3	0	1
T22	0	0	2	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	0
T23	0	0	0	0	2	1	1	0	1	0	0	1	1	0	0	0	1	2	1	1
T24	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	4	1	0	0
T25	2	2	1	1	1	0	1	0	1	1	2	1	4	1	0	0	1	1	0	0
T26	0	0	0	2	1	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0
T27	1	0	0	0	0	0	1	0	1	0	0	0	1	5	0	0	2	5	0	1
T28	0	0	0	0	0	4	0	1	2	0	0	0	2	0	1	0	1	0	1	1
T29	1	3	0	1	1	2	1	0	1	2	1	0	0	1	0	0	2	2	0	0
T30	0	0	2	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1
T31	2	1	0	1	1	0	1	0	1	0	0	0	2	1	0	1	0	2	0	1
T32	0	1	0	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0
T33	1	1	0	1	0	1	1	0	1	1	1	1	5	2	0	0	1	1	0	0
T34	0	1	0	0	0	0	0	1	1	1	0	0	1	0	0	1	0	0	0	0
T35	0	1	1	1	2	0	0	0	0	1	1	1	1	0	0	1	1	1	0	0

Table 2. Item Utility Table



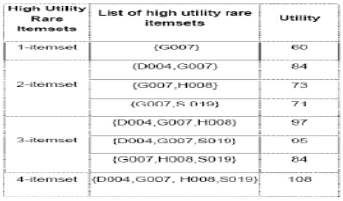
Items	External Utility	Internal Utility	Total Utility
A 001	4	21	84
B 002	1	28	28
C 003	3	21	63
D 004	2	12	24
E 005	7	23	161
F 006	5	27	135
G 007	6	10	60
H 008	1	13	13
I 009	1	34	34
J 010	4	27	108
K 011	3	15	45
L 012	1	14	14
M 013	1	50	50
N 014	2	40	80
O 015	3	14	42
P 016	1	18	18
Q 017	1	42	42
R 018	1	44	44
S 019	1	11	11
T 020	0	17	0

Table 3. Rare Itemset Table



Rare itemsets	List of rare itemsets	Itemset Utility
1-itemset	{D004}	24
	{G007}	60
	{H008}	13
	{S019}	11
2-itemset	{D004,G007}	84
	{D0004,H008}	37
	{D004,S019}	35
	{G007,H008}	73
	{G007,S 019}	71
	{H008,S019}	24
3-itemset	{D004,G007,H008}	97
	{D004,G007,S019}	95
	{G0007,H0008,S0019}	84
	{D004,H008,S019}	48
4-itemset	{D004,G007, H0008,S019}	108

Table 4. High Utility Rare Itemset Table



High Utility Rare Itemsets	List of high utility rare itemsets	Utility
1-itemset	{G007}	60
2-itemset	{D004,G007}	84
	{G007,H008}	73
	{G007,S 019}	71
3-itemset	{D004,G007,H008}	97
	{D004,G007,S019}	95
	{G007,H008,S019}	84
4-itemset	{D004,G007, H008,S019}	108

3.2 Algorithm HURI

```
Description: Finding High Utility Rare Itemsets of users interest
Ck: Candidate itemset of size k          Lk: Rare itemset of size k
For each transaction t in database
begin
        increment support for each item i present in t
End
L1= {Rare 1-itemset with support less than user provided max_sup}
for(k= 1; Lk!=Ø; k++)
begin
  C k+1= candidates generated from Lk;
      //loop to calculate total utility of each item
  For each transaction t in database
```

```
    begin
    Calculate total quantity of each item i in t
    Find total utility for item i using following formula:-
        u(i,t) = quantity[i] * user_provided_utility for i
    End
    //loop to find rare itemsets and their utility
    For each transaction t in database
    begin
    increment the count of all candidates in Ck+1 that are contained in t
        Lk+1 = candidates in Ck+1 less than min_support
      Add Lk+1 to the Itemset_Utility Table by calculating rare itemset utility using formula:
            Utility(R,t) = Σfor each individual item i in R (u(i,t));
    End
    //loop to find high utility rare itemset
    For each itemset iset in rare itemset Table R
    begin
    If (Utility(iset) > user_provided_threshold_for_high_utility_rare_itemset)
    then iset is a rare_itemset that is of user interest i.e.high_utility_rare_itemset
    else iset is a rare itemset but is not of user interest
    End
    Return high_utility_rare_itemsets
END
```

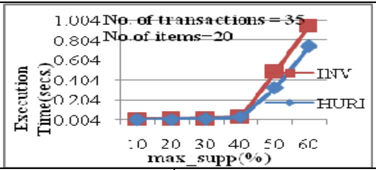
Fig. 1. Pseudo Code for HURI

4 Performance Evaluation of HURI

4.1 Comparative Analysis

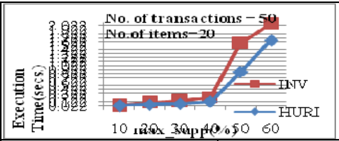
Apriori Inverse and HURI algorithms were compared using a transactional datasets of different sizes. Java as front end and MS Access as backend tool were used to evaluate HURI algorithm. We study the impact of different values of minimum support, number of transactions, number of items, on processing time etc. for comparing Apriori Inverse and HURI. The different comparative parameters are:-

(i) Number of rare itemsets generated.
(ii) Total execution time taken for generation of rare itemsets.



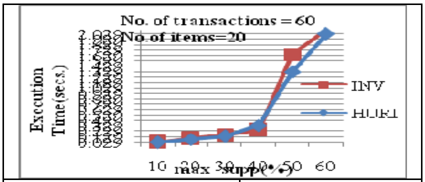
max_supp	Time Taken(secs.)	
	HURI	INV
10	0.004	0.006
20	0.009	0.011
30	0.014	0.015
40	0.031	0.031
50	0.325	0.484
60	0.741	0.954

Fig. 2. Execution Time on Database D1



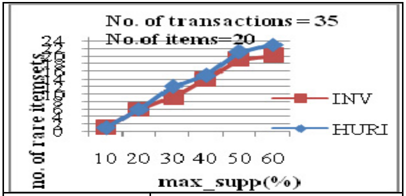
max_supp	Time Taken(secs.)	
	HURI	INV
10	0.022	0.024
20	0.053	0.099
30	0.076	0.143
40	0.123	0.234
50	0.867	1.597
60	1.65	2.053

Fig. 3. ExecutionTime on Database D2



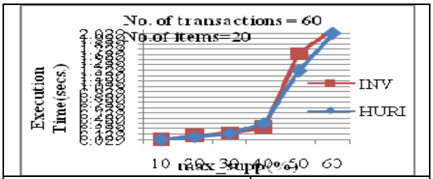
max_supp	Time Taken(secs.)	
	HURI	INV
10	0.029	0.031
20	0.089	0.099
30	0.145	0.168
40	0.342	0.255
50	1.324	1.654
60	2.031	2.112

Fig. 4. Execution Time on Database D3



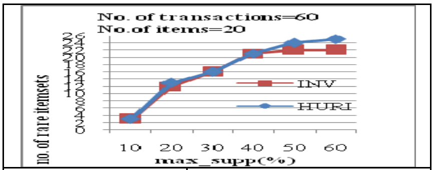
max_supp	No. of rare itemsets	
	HURI	INV
10	1	1
20	6	6
30	12	9
40	15	14
50	21	19
60	23	20

Fig. 5. Number of rare itemsets from D1



max_supp	Time Taken(secs.)	
	HURI	INV
10	0.029	0.031
20	0.089	0.099
30	0.145	0.168
40	0.342	0.255
50	1.324	1.654
60	2.031	2.112

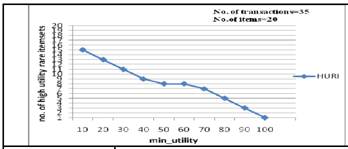
Fig. 6. Number of rare itemsets from D2



max_supp	No. of rare itemsets	
	HURI	INV
10	3	3
20	13	12
30	16	16
40	21	21
50	24	22
60	25	22

Fig. 7. Number of rare itemsets from D3

In item utility Table (Table 2), each item is assigned an external utility and internal utility is calculated from database D1. We considered three transaction sets, D1 (number of transactions is 50, number of items is 20), D2 (number of transactions is 50, number of items is 20) and D3 (number of transactions is 60, number of items is 20). Fig.2, Fig.3 and Fig.4 shows execution time of algorithms to generate rare itemsets from datasets D1, D2 and D3 respectively, by varying the support threshold. Fig.5, Fig.6 and Fig.7 shows number of rare itemsets generated from datasets D1, D2 and D3. The results show that HURI algorithm yields more rare itemsets with less execution time as compared to Apriori inverse. Apriori inverse produces only



Min.utility	No. of high utility rare itesets
	HURI
10	15
20	13
30	11
40	9
50	8
60	8
70	7
80	5
90	3
100	1
110	0

Fig. 8. Effect of utility threshold on high utility rare itemsets from dataset D1

rare itemsets whereas HURI produces high utility rare itemsets according to users' interest.

HURI was evaluated on dataset D1 under varied minimum utility thresholds, for generation of high utility rare itemsets (Fig.8). The experimental result shows that number of high utility rare itemsets decreases as minimum utility threshold increases, as desired, which indicates effectiveness of HURI.

4.2 Computational Complexity

The computational complexity of HURI can be affected by following factors –

4.2.1 Support Threshold

Increasing the support threshold results in more itemsets declared as rare. This has an adverse effect on the computational complexity of the algorithm because more candidate itemsets must be generated and counted as shown in Figure VIII, Figure IX and Figure I0. The maximum size of rare itemsets also tends to increase with higher support thresholds.

4.2.2 Number of Items (Dimensionality)

As number of items increases, more space will be needed to store support counts of items. If number of rare items also grows with dimensionality of the data, the computation and I/O costs will increase because of the larger number of candidate itemsets generated by the algorithm.

4.2.3 Number of Transactions

Since HURI algorithm makes repeated passes over the dataset, its run time increases with a large number of transactions.

4.2.4 Average Transaction Width

For dense datasets, the average transaction width can be very large. The maximum size of rare itemsets tends to increase as the average transaction width increases. As a result, more candidate itemsets must be examined during candidate generation and support counting.

A detailed analysis of time complexity for HURI algorithm is presented –

4.2.5 Generation of Rare 1-Itemsets

For each transaction, support count of every item presented in transaction is updated. If n is total number of transactions and m is the average transaction width, the time required for this operation is O(n*m).

4.2.6 Candidate Generation

To generate candidate k-itemsets, pairs of rare (k-1)-itemsets are merged. In the best-case scenario, every merging step produces a viable k-itemset. In the worst-case scenario, the algorithm must merge every pair o0f rare (k-1)-itemsets found in the previous iteration.

Therefore the overall cost of merging rare itemsets is $O(\sum_{k=2}^{m} (k-2) \, | C_k |)$

4.2.7 Support Counting

Each transaction of length x produces tCk itemsets of size k. If m is the maximum transaction width and αk is cost for updating support count of candidate k-itemset then cost for support counting is $O(n \cdot \sum_k ({}^{m}C_k)\, \alpha_k)$

4.2.8 High Utility Rare Itemsets

If m is maximum transaction width then time required to generate high utility rare itemsets is O(m2), which is negligible as compared to time taken for other operations for finding rare itemsets. Hence time taken in calculating Utilities like Internal utility, total item utility, Itemset Utility, Dataset Utility, etc. does not affect time taken by HURI algorithm to generate high utility rare itemsets. This proves the effectiveness and efficiency of HURI.

5 Conclusions and Future Work

Our paper proposes an innovative algorithm, HURI, for making business data mining more realistic and usable to business analyst. Data mining can identify products that are often purchased together, which can help product bundles that are more likely to be successful [13]. Frequency of item is not sufficient to answer a product combination i.e. whether it is highly profitable or whether it has a strong impact. Marketers are interested in knowing how various marketing programs affect the discovery of subtle relationships. The novelty of HURI is the ability to discover high utility rare itemsets. HURI algorithm may have practical meaning to real-world marketing strategies.The high utility rare itemsets are generated according to the users' preference.

HURI may be more beneficial on application to transactional data set. The high utility rare itemsets are generated based on transactional database information and external information about utilities. HURI uses the concept of Apriori inverse which produces only rare itemsets having support less than maximum support value where as HURI can produce high utility rare itemsets based on users' interest, support utility thresholds. The outcome of HURI would enable the top management or business analyst in crucial decision-making such as providing credit facility, finalizing discount policy, analyzing consumers' buying behaviour, organizing shelf space, quality improvement in supermarket scenario. Also the time complexity of HURI is almost the same as other algorithms. The future work includes the incorporation of temporal and fuzzy concept in HURI.

References

[1] Pillai, J., Vyas, O.P.: User centric approach to itemset utility mining in Market Basket Analysis. IJCSE 3(1), 393–400 (2011) ISSN : 0975-3397

[2] Pillai, J., Vyas, O.P.: Overview of Itemset Utility Mining and its Applications. IJCA (0975–8887) 5(1), 9–13 (2010)

[3] Pillai, J., Vyas, O.P., Soni, S., Muyeba, M.: A Conceptual Approach to Temporal Weighted Itemset Utility Mining. IJCA (0975-8887) 1(28) (2010)

[4] Shankar, S., Purusothoman, T.P., Jayanthi, S., Babu, N.: A Fast Algorithm for Mining High Utility Itemsets. In: Proceedings of IEEE IACC 2009, India, pp. 1459–1464 (2009)

[5] Adda, M., Wu, L., Feng, Y.: Rare Itemset Mining. In: Sixth International Conference on Machine Learning and Applications, pp. 73–80 (2007)

[6] Szathmary, L., Napoli, A., Valtchev, P.: Towards Rare Itemset Mining. In: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, vol. 1, pp. 305–312 (2007) ISSN: 1082-3409 , 0-7695-3015-X

[7] Erwin, A., Gopalan, R.P., Achuthan, N.R.: A Bottom-up Projection based Algorithm for mining high utility itemsets. In: Proceedings of AIDM 2007, Australia, Conferences in Research and Practice in Information Technology, CRPIT, vol. 84 (2007)

[8] Yao, H., Hamilton, H., Geng, L.: A Unified Framework for Utilty-Based Measures for Mining Itemsets. In: Proceedings of UBDM, pp. 28–37 (2006)

[9] Sun, X., Orlowska, M.E., Li, X.: Finding Temporal Features of Event-Oriented Patterns. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 778–784. Springer, Heidelberg (2005)

[10] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, pp. 207–216 (1993)

[11] Lan, G.C., Hong, T.P., Tseng, V.S.: A Novel Algorithm for Mining Rare-Utility Itemsets in a Multi-Database Environment. In: 26th Workshop on Combinatorial Mathematics and Computation Theory, pp. 293–302

[12] Pillai, J., Vyas, O.P.: High Utility Rare Itemset Mining (HURI): An approach for extracting high-utility rare itemsets. Journal on Future Engineering and Technology 7(1) (October 1, 2011)

[13] Data Mining: A Competitive Tool in the Bankingand Retail Industries, `http://icai.org/resource_file/9935588-594.pdf`

[14] Liu, Y., Liao, W.-k., Choudhary, A.K.: A Two-Phase Algorithm for Fast Discovery of High Utility Itemsets. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 689–695. Springer, Heidelberg (2005)

SMOTE Based Protein Fold Prediction Classification

K. Suvarna Vani[1] and S. Durga Bhavani[2]

[1] Department of Computer Science and Engineering,
V.R. Siddhartha Engineering College,
Vijayawada, A.P., India
`suvarnavanik@gmail.com`
[2] Department of Computer and Information Sciences,
University of Hyderabad,
Hyderabad, A.P., India
`sdb_dcis@yahoo.com`

Abstract. Protein contact maps are two dimensional representations of protein structures. It is well known that specific patterns occuring within contact maps correspond to configurations of protein secondary structures. This paper addresses the problem of protein fold prediction which is a multi-class problem having unbalanced classes. A simple and computationally inexpensive algortihm called *Eight-Neighbour* algortihm is proposed to extract novel features from the contact map. It is found that of Support Vector Machine (SVM) which can be effectively extended from a binary to a multi-class classifier does not perform well on this problem. Hence in order to boost the performance, boosting algorithm called SMOTE is applied to rebalance the data set and then a decision tree classifier is used to classify "folds" from the features of contact map. The classification is performed across the four major protein structural classes as well as among the different folds within the classes. The results obtained are promising validating the simple methodology of boosting to obtain improved performance on the fold classification problem using features derived from the contact map alone.

1 Introduction

Multi-class classification problem with imbalanced data is a challenging problem. The majority of multi-class pattern classification techniques are proposed for learning from balanced datasets [1]. There exist datasets having imbalanced data distribution, where some classes may have fewer training samples compared to other classes. Protein fold data set contains proteins from 27 most populated folds (classes) representing all major structural classes: all alpha, all beta, alpha + beta and alpha / beta. Protein fold prediction is a multi-class classification problem having highly imbalanced data amongst its classes. In literature, it can be seen that most of the datsets having insufficient instances among classes challenge the machine learning algorithms. The main objective of this paper is to find a solution for the multi-class protein fold prediction problem with the classes being highly imbalanced. Multi-class pattern classification techniques are principally proposed for learning from balanced dataset. Many datasets have imbalanced data distribution, where some classes of data may have fewer training samples compared to other classes. Support Vector Machines are believed to be less prone to the

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 541–550.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

class imbalance problem than other classification learning algorithms [3], since boundaries between classes are calculated with respect to only a few support vectors and the class sizes may not effect the class boundary too much. The underlining reason for this phenomenon is that as the training dataset gets more imbalanced, the support vector ratio between the high sample class and the small sample class also becomes more imbalanced. The small amount of cumulative error on the small class instances count for very little in the tradeoff between maximizing the width of the margin and minimizing the training error. Support Vector Machines simply learn to classify everything as the high sample class in order to make the margin the largest and the error the minimum. Boosting is a very popular technique for improving the accuracy of any given machine learning algorithm. Literature reports that AdaBoost and SMOTE algorithm have been successfully applied to most popular classifiers [4,5,6,7]. In literature, different boosting algorithms have been applied to unbalanced datasets consisting of a certain class of interest having very small size. Among the boosting algorithms AdaBoost[5,6] is shown to decrease the prediction error of minority classes. Significantly with increasing the prediction error of majority class a little bit, it can also produce higher values of margin which indicates a better classification. On the other hand, the combination of boosting and bagging, construct multiple classifiers by resampling the data space, weighting samples by boosting and replacing samples by bagging. The improvement in performance arising from ensemble combinations is usually the result of a reduction in variance. Variance measures how much a learning algorithm's guess bounces around for different training sets. Variance is therefore associated with overfitting: if a method overfits the data, the predictions for a single instance will vary between samples. Both boosting and bagging are capable of reducing variance, and hence are immune to the model overfitting problem. With an imbalanced data set, small class samples occurring infrequently, models that describe the rare classes have to be highly specialized. Standard learning methods pay less attention to the rare samples as they try to extract the regularities from the dataset. Such a model performs poorly on the rare class due to the introduced bias error. Bagging is believed to be effective for variance reduction, but not for bias reduction. AdaBoost and SMOTE, however are stated to be capable of both bias and variance reduction.

1.1 Boosting Algorithms

There are several boosting algorithms for classification of imbalanced data in the literature. All boosting algorithms can be divided into two categories The first category represents those that can be applied to most classifier learning algorithms directly, such as AdaCost[9], CSB1 and CSB2 [10], and RareBoost [11], the second category includes those that are based on a combination of the data synthesis algorithm and the boosting procedures, such as SMOTEBoost [13], and DataBoost-IM [13]. Synthesizing data may be application-dependent and hence involves extra learning cost. We only consider boosting algorithms that can be applied directly to most classification learning algorithms. In this category, Ada-Cost, CSB1 and CSB2 employ cost items to bias the boosting towards the small class, and RareBoost. The AdaBoost algorithm weights each sample to reflect its importance and places the greatest weights on those samples which are most often misclassified by the preceeding classifiers. The sample weighting strategy is equivalent to re-sampling the data space combining both up-sampling and down-sampling. Boosting attempts to reduce the bias error as it focuses on misclassified samples [9]. Such a focus may cause the learner to produce an ensemble function that difers significantly from the single learning algorithm. When the AdaBoost algorithm is adopted to tackle the class imbalance problem. Hence some of the advantages of AdaBoost for learning imbalanced data: 1.Applicable to most classifier learning algorithms. 2.AdaBoost algorithm is equivalent to resampling the dta space combining both up-sampling and down-sampling. 3.As a resampling method, Adaboost updates the data space automatically eliminating the extra learning cost for exploring the optimal class distribution.

1.2 SMOTE: Sythetic Minority Oversampling Technique

Synthetic Minority Oversampling Technique involves the process in which the minority class is oversampled by creating synthetic examples rather that by oversampling with replacement. Firstly, the minority class is oversampled by taking samples from the minority class and then introducing synthetic examples along the nearest neighbours of the minority class. Secondly, the majority class is under-sampled by randomly removing samples from majority class until the minority class becomes some specified percentage of the majority class.Thus a combination of under-sampling and over-sampling leads to the initial bias of the learner towards the majority (negative) class being reversed in favor of the minority (positive) class. The classifiers are generally taken to be SVM and NN learned on the datset by ”SMOTING” the minority class and under-sampling the majority class[12].

1.3 Protein Contact Map

Figure 1 shows 3D structure of protein (2IGD,Sequence length:61) and the corresponding cartoon topology of protein structure. The two dimensional representation of this

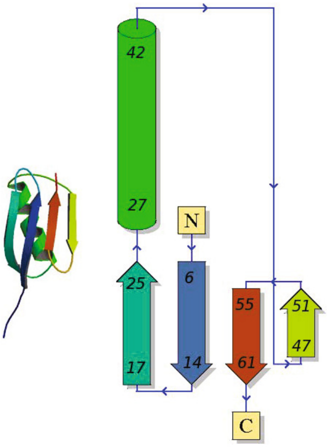


Fig. 1. 3D Structure and Cartoon topology for Protein :2IGD

structure is called a contact network which is a symmetric square matrix. The contact matrix C is obtained by taking a map from the amino acid (aa) sequence of the protein to itself with $C(a,b) = 1$ if aminoacid residue a is in proximity with b, say within 7 Angstrom units, otherwise zero. Protein contact maps are intermediary representations which are found to be useful to analyze the structural properties of proteins. [25,22]. Our work involves extracting useful features from the contact matrices and utilize them effectively for the protein fold prediction problem. Without loss of generality consider the lower triangular matrix for this study.

2 Methodology

Ding et al. consider features from sequences and construct a feature vector of dimension 20. Shamim et al. consider some additional features totalling upto 100 and also consider both sequence and secondary structural features. More recently, Chmielnicki and et al. apply the multi-class support vector machine classifier for protein fold recognition problem[18]. To the best of our knowledge the boosting techniques have not been applied to this problem.

2.1 Data Set

The data set of Ding et al is considered as a bench mark data in the literature [23],for protein folds. This data set contains proteins with less than 35-40% sequence identity representing all classes and further contains proteins belonging to 27 most populated folds, representing all major structural classes of all alpha, all beta, alpha + beta and alpha / beta. Along with the data set of Ding et al we use the recently updated database that holds consensus view of fold space based on SCOP, CATH and DALI [2,20,17]. This database is designed such that proteins domains are classified as a protein fold if they agree in at least two of three classification systems. These proteins are then downloaded from the protein data bank [21] for which the contact map is computed. The contact matrices of the map form the input data set for the classification experiments. Useful features are to be computed for these proteins. In the literature, amino acid composition, physico-chemical properties and secondary structural properties have been considered as features for classification. Ding [23]et. al consider 20 dimensional feature vectors and Nagarajaram et al use 100 such features for the protein fold prediction problem [24]. We propose novel contact map features along with the features proposed in the earlier paper [14] to form an eleven dimensional feature vector representation. We choose training set to have 244 sequences and testing set with 296 sequences.

2.2 Feature Extraction

If one looks at regions within contact maps in figure 2, it shows how the secondary structure interactions are nicely embedded within a contact map. The challenge then is to distinguish the differences in the patterns and automate the procedures such that these can be run on huge data sets. The secondary structures are visible along the diagonal in

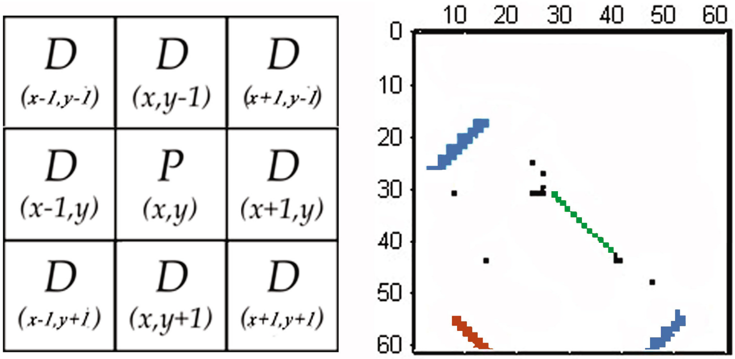


Fig. 2. Eight Directional Bit Positions and Protein 2IGD diagonal masked contact map

the contact map and the clusters seen away from the diagonal represent interactions between different secondary structures due to the folding of the protein chain. As proteins belonging to the same fold have highly similar arrangement of the secondary structures, we hypothesize that their corresponding contact maps exhibit similarity.The figure 1 shows Left:3D structure of protein (PDB file 2IGD,Sequence length:61) Right: The cartoon topology of protein structure. In this paper eleven features are derived : Number of helices, minimum helix length, maximum helix length, number of betas, minimum beta length, maximum beta length, number of clusters, minimum cluster density, maximum cluster density, number of parallel sheets and number of orthogonal sheets. The first six features are the secondary structure features derived from the diagonal proposed in an earlier paper [14]. Here we propose a novel yet simple method, the *EightNeighbour Algorithm*, to extract patterns in the off-diagonal region to compare similarity of contact maps. The remaining five features from the off-diagonal contact maps are derived from the Eight Neighbour Algorithm. EightNeighbour Algorithm In this algorithm off diagonal contact map pattern cannot be splitted. The figure 2 shows Left: Eight directional bit positions. If bit position is '1' then mask the corresponding bit and read the neighbour pixel. This process will be repeated in eight directions until the whole contineous pattern is completed. The minimum density of the pattern is five and the maximum density depends on the size of the protein.

2.3 Direction Features

The interactions are labled using the density of the cluster it means that how many pairs of amino acid contacts are present in 2IGD protein contact map. The covariance of each pattern in the contact map is computed in order to predict the direction of the pattern whether the pattern is along or away from the diagonal. If the covariance value is negative, the pattern is labelled as antiparallel sheet. If the covariance value is positive the pattern is identified as parallel sheet. If it is zero the pattern can be either parallel or antiparallel. Figure 2 shows Right:the contact map of the protein 2IGD. The cyan colour patterns represent the antiparallel sheets, orange colour pattern represents the parallel sheet and the green colour dots represent the helix information along the diagonal. Feature Sets Different feature sets are constituted to analyze the performance of the classifier with respect to the features

Feature set 1: secondary structural features from the diagonal region of the contact map: number of helices, minimum and maximum helix length, number of betas, minimum and maximum beta length (6).
Feature set 2: Cluster features : number of clusters, minimum cluster density, maximum cluster density, number of parallel and anti-parallel sheets (5).
Feature set 3: Diagonal + Cluster : number of helices, minimum and maximum helix length, number of betas, minimum and maximum beta length, number of clusters, minimum cluster density, maximum cluster density (9).
Feature set 4: Diagonal + Cluster + Direction : number of helices, minimum and maximum helix length, number of betas, minimum and maximum beta length, number of clusters, minimum cluster density, maximum cluster density, number of parallel and number of anti-parallel sheets (11).

2.4 Related Literature

Shamim et al. developed a new method for protein fold recognition using structural information of amino acid residues and amino acid residue pairs. They developed a SVM based classifiers that combines secondary structural state and solvent accessibility state frequncies of amino acids and amino acid pairs as feature vectors. They worked on an extended data set of Ding et al. which consists of 2554 proteins.Ding et al. consider the six feature sets extracted independently from protein sequences. One may apply machine-learning techniques based on a single parameter set for protein fold prediction. We found that using multiple parameter sets and applying majority voting scheme leads to increase better prediction accuracy. Alternatively, one may combine different parameter sets into one dataset so that each protein is represented by a 125-dimensional feature vector. Ding and Dubchak et. al consider features from protein sequences and their dimensions are 20. Nagarajaram and Shamim et al.consider some additional features counting to 100 and also increase the dataset size. The authors consider both sequence and secondary structural features. More recently, Wieslaw Chmielnicki and et. al apply the multi-class support vector machine classifier for protein fold recognition problem[18]. In this paper we intend to improve the accuracy of protein fold prediction classification by using SMOTE (Synthetic Minority Over-sampling TEchnique).

3 Classification Results

Machine learning methods like Support Vector Machines and Neural Networks are very powerful classifiers and have been used for protein structure prediction and fold prediction considering features from amino acid sequence[24,16]. On the other hand, these classifiers are not suitable for unbalanced data like protein classification data. The datasets have imbalanced data distribution, where some classes of data may have a few training samples compared to other classes. For our dataset, we found that radial basis

Table 1. Classifcation results of SVM Support Vector Machine with different Kernels

Different Features	LKernel	PKernel	RBF
Featureset 1	21.62	25.33	25
Featureset 2	23.98	25.33	21.95
Featureset 3	24.32	27.36	19.25
Featureset 4	**29.72**	27.70	19.93

Table 2. Classification of SCOP structural classification of a protein into the 4-major structural classes

Class	Shamim et al.	Ding et al.	SMOTE	SVM
All-Alpha	78	55	75.06	31.5
All-Beta	56	33.18	69	50.02
Alpha+Beta	35.8	31.4	63.3	25.33
Alpha/Beta	45.9	42.26	54	29.6
Average	**53.925**	**40.46**	**65.34**	**34.15**

Table 3. Ten fold classification of Alpha+Beta Class-Average True Positive:54%

TP	FP	Precision	Recall	F-Measure	ROC Area	Class
0.708	0.019	0.607	0.708	0.654	0.886	72
0.500	0.013	0.579	0.500	0.537	0.825	87
0.423	0.022	0.458	0.423	0.440	0.769	110

Table 4. Ten fold classification of Alpha/Beta Class-Average True Positive:63%

TP	FP	Precision	Recall	F-Measure	ROC Area	Class
0.190	0.027	0.200	0.190	0.195	0.63	46
0.810	0.020	0.586	0.810	0.680	0.941	47
0.905	0.008	0.792	0.905	0.844	0.949	48
0.591	0.015	0.591	0.591	0.591	0.828	51
0.500	0.002	0.909	0.500	0.645	0.868	54
0.667	0.008	0.762	0.667	0.711	0.870	57
0.625	0.012	0.682	0.625	0.652	0.845	58
0.708	0.020	0.586	0.708	0.642	0.927	62
0.450	0.007	0.692	0.450	0.545	0.840	69

function does not work well, the linear kernel works better and the polynomial kernel gives the best results. In Table 1 it is seen that the featureset 4 with linear kernel gives the best results. Now we apply SMOTE based technique to increase the accuracy of classification performance. The SMOTE algorithm rebalances the inadequate classes. As an intial step take the maximum instance class in the training set and cope-up the remaining class samples into that ratio maintained by the smote parameter k. Among the proteins present in SCOP families, we consider in the top six folds of All-alpha, top

Table 5. Ten fold classification of All Alpha Class-Average True Positive:74%

TP	FP	Precision	Recall	F-Measure	ROC Area	Class
0.87	0.005	0.875	0.875	0.875	0.93	1
0.64	0.014	0.667	0.640	0.653	0.85	3
0.85	0.013	0.680	0.850	0.756	0.96	4
0.66	0.007	0.778	0.667	0.718	0.82	7
0.75	0.020	0.600	0.750	0.667	0.90	9
0.72	0.008	0.722	0.722	0.722	0.88	11

Table 6. Ten fold classification of All Beta Class-Average True Positive:69%

TP	FP	Precision	Recall	F-Measure	ROC Area	Class
0.208	0.027	0.238	0.208	0.222	0.607	20
0.833	0.008	0.800	0.833	0.816	0.954	23
0.815	0.007	0.846	0.815	0.830	0942	26
0.708	0.013	0.680	0.708	0.694	0.909	30
0.875	0.003	0.913	0.875	0.894	0.956	31
0.667	0.013	0.667	0.667	0.667	0.862	32
0.571	0.010	0.667	0.571	0.615	0.803	33
0.667	0.020	0.571	0.667	0.615	0.841	35
0.750	0.013	0.692	0.750	0.720	0.911	39

nine folds of All- beta, top two folds of All alpha + beta and and All top nine folds of alpha / beta. SMOTE increases the training set to 617 instances. Now 80% of the data set is used for training the model and remaining 20% is used for testing. A ten fold cross validation is carried out in all the experiments to remove any bias that a protein of the dataset may impose on the classifier. A decision tree with J48 learning algorithm that is available in the open source software system WEKA is used for the classification task [19]. Classification experiments are conducted for different combinations of features for the data set.

4 Conclusion

In this paper, we studied several important issues in protein fold recognition and prediction in the context of a large number of folds using different techniques. We studied one-against-one,one-against-others and all-versus-all methods. These mehtods are not enough to improve prediction accuracy because of inadequate classes in dataset. As per literature multiclass classification for unbalanced data is the important ongoing research problem. The SMOTE algortihm is used to rebalance the data and to increase the accuracy also. SMOTE proves to be useful for the classification problem of protein fold prediction.It can be seen that some of the structural classes like Alpha/Beta have very low performance accuracy inspite of applying the boosting algorithm. It is still desirable to extract relevant features which is part of the ongoing study.

Acknowledgements. We thank Dr Somdatta Sinha of CCMB, Hyderabad who introduced us to the problem of analysis of protein contact map.

References

1. Ghanem, A.S., Venkatesh, S., West, G.: Multi-class Pattern Classification in Imbalanced Data. In: ICPR, pp. 2881–2884 (2010)
2. Day, R., Beck, D.A.C., Armen, R.S., Daggett, V.: A consensus view of fold space: Combining SCOP, CATH, and the Dali domain dictionary. Protein Science 12, 2150–2160 (2003)
3. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis Journal 6(5), 429–450 (2002)
4. Elkan, C.: Boosting and naive bayesian learning. Technical Report CS97-557, Department of Computer Science and Engneering, University of California,Sam Diego, CA (September 1997)
5. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 148–156. Morgan Kaufmann, The Mit Press (1996)
6. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning 37(3), 297–336 (1999)
7. Schwenk, H., Bengio, Y.: Boosting neural networks. Neural Computation 12(8), 1869–1887 (2000)
8. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Annals of Statistics 28(2), 337–374 (2000)
9. Fan, W., Stolfo, S.J., Zhang, J., Chan, P.K.: Adacost:misclasification cost-sensitive boosting. In: Proceedings of Sixth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, pp. 97–105 (1999)
10. Ting, K.M.: A comparative study of cost-sensitive boosting algorithms. In: Proceedings of the 17th International Conference on Machine Learning, Stanford University, CA, pp. 983–990 (2000)
11. Joshi, M.V., Kumar, V., Agarwal, R.C.: Evalating boosting algorithms to classify rare classes: Comparison and improvements. In: Proceeding of the First IEEE International Conference on Data Mining, ICDM 2001 (2001)
12. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: Improving prediction of the minority class in boosting. In: Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Databass, Dubrovnik, Croatia, pp. 107–119 (2003)
13. Guo, H., Viktor, H.L.: Learning from imbalanced data sets with boosting and data generation: The databoost-IM approach. SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets 6(1), 30–39 (2004)
14. Bhavani, S.D., Suvarnavani, K., Sinha, S.: Mining of protein contact maps for protein fold prediction. In: WIREs Data Mining and Knowledge Discovery, vol. 1, pp. 362–368. John Wiley & Sons (July/August 2011)
15. Hsu, C., Lin, C.J.: A comparision of methods for multi-class Support Vector Machines. IEEE Transactions on Neural Networks 13, 415–425 (2002)

16. Barah, P., Sinha, S.: Analysis of protein folds using protein contact networks. Pramana 71(2), 369–378 (2008)
17. Shi, J.-Y., Zhang, Y.-N.: Fast SCOP Classification of Structural Class and Fold Using Secondary Structure Mining in Distance Matrix. In: Kadirkamanathan, V., Sanguinetti, G., Girolami, M., Niranjan, M., Noirel, J. (eds.) PRIB 2009. LNCS (LNBI), vol. 5780, pp. 344–353. Springer, Heidelberg (2009)
18. Chmeilnicki, W., Stapor, K.: An efficient multi-class support vector machine classifier for protein fold recognition. In: IWPACBB, pp. 77–84 (2010)
19. `http://www.cs.waikato.ac.nz/ml/weka/`
20. `http://www.dynameomics.org/`
21. `http://www.rcsb.org/pdb/home/home.do`
22. Fraser, R., Glasgow, J.: A Demonstration of Clustering in Protein Contact Maps for Alpha Helix Pairs. In: Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B. (eds.) ICANNGA 2007. LNCS, vol. 4431, pp. 758–766. Springer, Heidelberg (2007)
23. Ding, C.H.Q., Dubchak, I.: Multi-class proteing fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358 (2001)
24. Shamim, M.T.A., Anwaruddin, M., Nagarajaram, H.: Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. Bioinformatics 23:24, 3320–3327 (2007)
25. Zaki, M.J., Nadimpally, V., Bardhan, D., Bystroff, C.: Predicting Protein Folding Pathways. In: Datamining in Bioinformatics. Springer (2004)

Conceptual Application of List Theory to Data Structures

B.K. Tripathy[1,*] and S.S. Gantayat[2]

[1] School of Computing Sciences & Engineering
VIT University, Vellore – 632 014, Tamilnadu, India
[2] Department of Computer Science & Engineering
GMR Institute of Technology, Rajam – 532 127, Andhra Pradesh, India
tripathybk@vit.ac.in, sasankosekhar.g@gmrit.org

Abstract. Following the approach of defining a set through its characteristic function and a multiset (bag) through its count function, Tripathy, Ghosh and Jena ([3]) introduced the concept of position function to define lists. The new definition has much rigor than the earlier one used in computer science in general and functional programming ([2]) in particular. Several of the concepts in the form of operations, operators and properties have been established in a sequence of papers by Tripathy and his coauthors ([3, 6, 7, 8]. Also, the concepts of fuzzy lists ([4]) and that of intuitionistic fuzzy lists ([5]) have been defined and studied by them. Recently an application to develop list theoretic relational databases and operations on them has been put forth by Tripathy and Gantayat ([9]). In the present article we provide another application of this approach in defining data structures like Stack, Queue and Array. One of the major advantages of this approach is the ease in extending all the concepts for basic lists to the context of fuzzy lists and intuitionistic fuzzy lists. We also illustrate this approach in the present paper.

Keywords: Lists, Stack, Queue, Array, Fuzzy lists, Intuitionistic Fuzzy lists.

1 Introduction

It is well known that the notion of a set and its characteristic function are interchangeable concepts. Similarly, the notion of a bag and its count function ([10]) are interchangeable concepts. So, it is natural to think of some such characteristic for lists, one of the most widely used data structures in the field of Computer Science. This could be realized by Tripathy, Ghosh and Jena ([3]), where they introduced the notion of position function for lists. As a continuation of their effort, they defined many concepts associated with lists through this new definition and also established several theorems which reveal the properties of lists. This study was further carried out in [6] and [7]. Just to recall, both the number of times elements occur and the order of their occurrence are important in a list. In a bag only the numbers of times elements occur are important. In the definition of a set the order of occurrence of elements is unimportant and the repetition of elements is considered as illegal.

---

[*] Corresponding author.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 551–560.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

The notion of fuzzy sets introduced by Zadeh ([11]) is an important model to capture impreciseness in data. The notions of fuzzy bags and fuzzy lists were not considered in literature before these were defined by Tripathy et al [4]. The concept of intuitionistic fuzzy sets introduced by Atanassov ([1]) is an extension of the notion of fuzzy sets and is a better model than fuzzy sets to model impreciseness in data. As it is natural, the concept of intuitionistic fuzzy lists was defined and their properties were studied by Tripathy et al ([5]).

Recently, as an application of the new definition of lists, a framework of relational data model and some operations on it have been defined by Tripathy et al ([9]). In this article we study another application of lists in realizing data structures. A list is defined as a linearly ordered collection of elements of similar data types with single or multiple occurrences. Since stack, queue and array are linear data structures; we establish here as how these data structures can be implemented using lists and operations on them. The operations on these data structures are, generally, insertion or deletion of an element, searching an element or finding its position and replacement of an element at a particular position by another element.

Also, we shall define these data structures in their fuzzy and intuitionistic fuzzy versions which can be used in some real time applications in the forth coming papers. Perhaps these concepts are used by us for the first time in the sequel. To be precise, in this article we have shown how the operators and operations of lists, fuzzy lists and intuitionistic fuzzy list can be used in defining data structures like stack, queue, array and operations on them. This creates a background basing upon which packages on lists and their applications can be developed dealing with data structures. The important factor is that once the basic operations are developed using functional programming the development of packages becomes easy.

The organization of the paper from this point onwards is as follows. We introduce some of the definitions of concepts associated with the new definition of list, which has been developed in different papers as mentioned above in section 2. We state two theorems to be used in the sequel. In section 3 we introduce several data structures using lists and operations on them. Also, we illustrate the execution of these operations through suitable examples. In section 4 we provide some concluding remarks. We end up the paper with a bibliography of referred works.

2 Definitions and Properties

Definition 2.1: A *list* L drawn from a set X is represented by a position function P_L, defined as P_L: X → P(N), where P(N) denotes power set of the set of non–negative integers N.

Definition 2.2: For any finite list L drawn from X, the *cardinality* of L is denoted by #L and is defined as

$$\# L = \sum_{x \in X} | P_L(X) | ,$$

whenever the right hand side exists. In this case L is said to be finite, otherwise L is said to be infinite.

Definition 2.3: For any finite list L drawn from X, the *reverse* of L, denoted by *rev*(L) is a list on X, and is given by the position function,

$$P_{rev(L)}(x) = \{\# L - 1 - t : t \in P_L(x)\} \text{ for each } x \in X.$$

Definition 2.4: For any finite list L drawn from X, the *head* of L is an element denoted by *hd*(L), where hd(L) = x if $0 \in P_L(x)$.

Definition 2.5: Let L be a list. We define the *tail* of L denoted by *tl*(L) as

$$P_{tl(L)}(x) = \{t - 1 : t \in P_L(x), t \neq 0\}.$$

Definition 2.6: For any list L we define the functions *init* and *last* as:

init(L) = rev(tl(rev(L))) and
last(L) = x; if $(\# L - 1) \in P_L(x)$.

Definition 2.7: For any $x \in X$ and a list L drawn from X, we denote the list which is obtained by adding an element x at the beginning of the list L by *cons(x, L)* and define it by its membership function,

$$P_{cons(x,L)}(y) = \begin{cases} \{r - 1 : r \in P_L(y), \text{ if } y \neq x; \\ \{0\} \cup \{r + 1 : r \in P_L(x), \text{ if } y = x. \end{cases}$$

Definition 2.8: A list L is *empty* if $P_L(x) = \phi$ for each $x \in X$ and denoted as L = [ ].

Definition 2.9: Let L_1 and L_2 be two finite lists drawn from X. Then we define the concatenation of L_1 and L_2 denoted by $L_1 \mathbin{+\!\!\!+} L_2$ and is given by the position function,

$$P_{(L_1 \# L_2)}(x) = P_{L_1}(x) \cup \{\# L_1 + t : t \in P_{L_2}(x)\} \ \forall x \in X.$$

Definition 2.10: Let L be finite list drawn from X and [x] be a singleton list. Then the position function of the list *L – [x]* is defined as follows.

Case-I: $P_L(x) \neq \phi$

$$P_{L - [x]}(y) = \begin{cases} \{r : r \in P_L(y), r < \min P_L(x)\} \cup \{r - 1 : r \in P_L(y), r > \min P_L(x)\}, & \text{if } y \neq x; \\ \{r - 1 : r \in P_L(x), r > \min P_L(x)\}, & \text{if } y = x. \end{cases}$$

Case-II: $P_L(x) = \phi$ $\qquad P_{L\text{-}[x]}(y) = P_L(y)$

Definition.2.11: Let L_1 and L_2 be two lists drawn from X, then the zip of L_1 and L_2 is given as

$$P_{zip(L_1, L_2)}(x, y) = P_{L_1}(x) \cap P_{L_2}(y)$$

Definition.2.12: For any two finite lists L and L′ drawn from X, the difference of the lists L – L′ is given by its position function, defined recursively by

$$P_{L-L'}(x) = \begin{cases} P_L(x), & \text{if } L' \text{ is an empty list}; \\ P_{\{L-[hd(l')]\}-tl(L')}(x), & \text{otherwise.} \end{cases}$$

Definition 2.13: Let L be a finite list drawn from X. Then the take operator on L, for any given n ∈ N, is denoted by *take*(n, L) and it is a list whose position function is given by

$$P_{take(n,L)}(x)=\begin{cases}\phi,\ if\ n\le \min P_L(x)\ or\ P_L(x)=\phi;\\ \{r: r<n\ and\ r\in P_L(x)\},\ if\ n>P_L(x).\end{cases}$$

Definition 2.14: Let L be a finite list drawn from X. Then, for any n ∈ N the drop operator on L is denoted by *drop*(n, L) and it is a list whose position function is given by

$$P_{drop(n,L)}(x)=\begin{cases}\phi, & if\ n>\max P_L(x);\\ \{r-n: r\ge n\ and\ r\in P_L(x)\}, & if\ n\le \max P_L(x).\end{cases}$$

Definition 2.15: For any list L drawn from X and any natural number n, we define the element index(L, n) as x ∈ X, such that

x = hd(take(n+1,L) – take(n, L)).

The following two properties of lists, which are Theorems established in ([6]) are to be used in the coming sections:

Theorem 2.1: If L_1 and L_2 are the lists drawn from X and f is a mapping from X into X, then for any n ∈ N,

a) $$take(n, L_1 \mathbin{+\!\!\!+} L_2)=\begin{cases}take(n,L_1), & if\ n\le \# L_1;\\ L_1 \mathbin{+\!\!\!+} take(n-\# L_1, L_2), & if\ n>\# L_1.\end{cases}$$

b) $$drop\left(n,\ L_1 \mathbin{+\!\!\!+} L_2\right)=\begin{cases}drop\left(n,\ L_1\right) \mathbin{+\!\!\!+} L_2\ , & if\ n\le \# L_1;\\ drop\left(n\ - L_1,\ L_2\right)\ , & if\ n\ >\# L_1.\end{cases}$$

Theorem 2.2: For any n ∈ N and for lists L_1 and L_2 drawn from X,

$$index\left(L_1 \mathbin{+\!\!\!+} L_2,\ n\right)=\begin{cases}index\left(L_1, n\right), & if\ n<\# L_1;\\ index\left(L_2,\ n\ -\ \# L_1\right), & if\ n\ge \# L_1.\end{cases}$$

3 Data Structures through Lists

A list is a linearly ordered collection of elements of similar data types with single or multiple occurrences. Since stack, queue and array are linear data structures; we shall establish how these data structures can be implemented using lists and operations on them can be realized through operations on lists established mentioned in section 2 and by other authors. The operations on these data structures are, generally, insertion or deletion of an element, searching an element, or finding its position and replacement of an element at a particular position by another element.

We shall establish the applications of lists and properties on them by showing the representation and operations on these data structures in the following.

3.1 Implementation of a STACK and Its Operations

A stack is a linearly ordered set of elements and the operation on elements in it follows the LIFO (Last In First Out) principle. The insertion and the deletion of an element is done from one end of this linear structure with the other end being fixed. A stack has a fixed size, which is the maximum number of elements, it can store in it. The beginning of the structure, which is a location in computer memory, is called the 'BOTTOM' of the Stack. At any instant of time there is a variable called STACK pointer, which contains the position of the latest element inserted into the stack. This position is called the 'TOP' of the stack. The process of inserting a new element into the stack is called PUSH and the process of deletion of an element from a stack is called POP, which is done from the TOP of the stack.

Suppose the size of a stack S at any point of time is #S and M is its maximum size. In the beginning we take S = [ ].

3.1.1 Insertion or Addition of an Element in S

$$\mathrm{PUSH}(x, S) = \begin{cases} S \mathbin{+\!\!\!+} [x], & \text{if } \#S < M; \\ \text{Overflow}, & \text{if } \#S = M. \end{cases}$$

3.1.2 Deletion or Removal of an Element from S

$$\mathrm{POP}(S) = \begin{cases} \mathrm{init}(S), & \text{if } \#S > 0; \\ [\ ], & \text{if } \#S = 0. \end{cases}$$

3.1.3 To Get the Top of the Elements from S

$$\mathrm{TOP}(S) = \mathrm{last}(S), \qquad \text{if } \#S \neq 0.$$

3.1.4 To Find an Element at the *i*th Position from the Top of S

$$\mathrm{PEEP}(S, i) = \mathrm{hd}(\mathrm{drop}\,(\#S - i, S)) \qquad \textit{or}$$

$$\mathrm{last}(\mathrm{take}(\#S - i+1, S)); \text{ where } 0 \leq i \leq \#S.$$

3.1.5 To Check whether the Stack S is Empty or Not

$$\mathrm{EMPTY}(S) = \text{True}, \text{ if } \#S = 0.$$

We shall illustrate the above operations in the following example.

Example 3.1: Consider the following list of five names.

S = ["John", "Sonu", "Rama", "Samir", "Tom"].

Here #S = 5. Let the list size M = 10. Then we can insert at most 5 names in the list.

(a) Suppose we want to insert a name x = "Anu" in the stack.

$$\begin{aligned} \mathrm{PUSH}(\text{“Anu”}, S) &= S \mathbin{+\!\!\!+} [\text{“Anu”}] \\ &= [\text{“John”},\text{“Sonu”},\text{“Rama”},\text{“Samir”},\text{“Tom”},\text{“Anu”}]. \end{aligned}$$

(b)

$$\begin{aligned} \mathrm{POP}(S) &= \text{delete the last element of the list S} \\ &= \mathrm{init}[\text{“John”}, \text{“Sonu”}, \text{“Rama”}, \text{“Samir”}, \text{“Tom”}] \\ &= [\text{“John”}, \text{“Sonu”}, \text{“Rama”}, \text{“Samir”}] \end{aligned}$$

(c) $\mathrm{PEEP}(S,2) = \mathrm{hd}(\mathrm{drop}(3,S)) = \mathrm{hd}([\text{“Samir”},\text{“Tom”}]) = [\text{“Samir”}]$.

3.2 Implementation of a QUEUE and Its Operations

A Queue is a linearly ordered set of elements which works on the FIFO (First In First Out) principle, that is, insertion or addition of elements take place at the end of the queue and deletion or removal of elements takes place from the beginning of the queue.

Suppose that a queue Q with the maximum size M, where M > 0 and Q contains #Q number of elements. Suppose x is an element of X, used in the operations.

3.2.1 Insertion or Addition of an Element into Q

$$\text{Q-INSERT}(x, Q) = \begin{cases} Q \mathbin{+\!\!\!+} [x], & \text{if } \#Q < M\,; \\ \text{Overflow}, & \text{if } \#Q \geq M \textit{ or size of } Q \text{ is fixed.} \end{cases}$$

3.2.2 Deletion or Removal of an Element from Q

$$\text{Q-DELETE}(Q) = \begin{cases} \text{tl}(Q)\quad, & \text{if } \#Q > 0; \\ \text{underflow}, & \text{otherwise.} \end{cases}$$

3.2.3 To Get an Element from Q

a) Front element of Q:

$$\text{FRONT}(Q) = \begin{cases} \text{hd}(Q)\quad, & \text{if } \#Q \neq 0; \\ [\,]\quad, & \text{otherwise.} \end{cases}$$

b) Rear element of Q:

$$\text{REAR}(Q) = \begin{cases} \text{last}(Q)\quad, & \text{if } \#Q \neq 0; \\ [\,]\quad, & \text{otherwise.} \end{cases}$$

3.2.4 To Check the Q is Empty or Not

$$\text{EMPTY}(Q) = \text{True}\ , \qquad \text{if } \#Q = 0.$$

Example 3.2: Consider the same example as given in Example 2.1., where the given list is a queue.

That is, Q = [“John”, “Sonu”, “Rama”, “Samir”, “Tom”].

Here #Q = 5. Suppose we want to insert a name x = “Anu” in a queue.

a) Q-INSERT(“Anu”, Q)
$= Q \mathbin{+\!\!\!+}$ [“Anu”] = [“John”, “Sonu”, “Rama”, “Samir”, “Tom”, “Anu”].

b) Q-DELETE(Q) = tl(Q)
= tl([“John”, “Sonu”, “Rama”, “Samir”, “Tom”])
= [“Sonu”, “Rama”, “Samir”, “Tom”].

c) FRONT(Q) = hd([“John”, “Sonu”, “Rama”, “Samir”, “Tom”]) = “John”

d) REAR(Q) = last([“John”, “Sonu”, “Rama”, “Samir”, “Tom”]) = “Tom”

3.3 Implementation of ARRAYS and Operations on Them

An array is a finite linearly ordered set such that an element can be inserted at any position and deleted from any position of the array.

An array is a finite list. Suppose A is an array of size M. The number of elements in an array A is #A. Suppose x is an element used in the operation. All the operations on arrays are basically operations on finite lists.

3.3.1 Insertion of an Element

Insertion of an element can take place at the following positions.

(i) at the beginning of an array (ii) at the end of an array
(iii) at any position inside an array

i. At the beginning of an array:

$$\text{B-INSERT}(x, A) = \text{cons}(x, A), \text{ if } \#A < M.$$

ii. At the end of an array: The operation is same as in the case of a Stack and a Queue.

$$\text{E-INSERT}(x, A) = A \mathbin{+\!\!+} [x], \text{ if } \#A < M.$$

iii. At any position of an array:

$$\text{INSERT}(x, i, A) = \text{take}(i - 1, A) \mathbin{+\!\!+} [x] \mathbin{+\!\!+} \text{drop}(i - 1, A), \text{ if } 1 \le i \le \#A < M.$$

3.3.2 Deletion of an Element

Deletion can be take place from the following positions.

(i) from the beginning of an array (ii) from the end of an array
(iii) from any position of an array

i. From the beginning of the array:

$$\text{B-DELETE}(A) = \begin{cases} \text{init}(A), & \text{if } 1 \le \#A; \\ [\,]\quad, & \text{if } \#A = 0. \end{cases}$$

ii. From the end of the array:

$$\text{E-DELETE}(A) = \text{tl}(A), \text{ if } \#A > 0.$$

iii. From any position of the array:

$$\text{DELETE}(i, A) = \text{take}(i - 1, A) \mathbin{+\!\!+} \text{drop}(i, A), \text{ if } i \ge 1;$$

where i is the location or position in the array.

3.3.3 Retrieval of an Element from an Array

To retrieve an element from an array we can use the GET operation. This operator can be implemented in the following cases.

a) from the beginning of a array b) from the end of an array
c) from any position of an array

a) From the beginning of an array:

$$\text{B-GET}(A) = \text{hd}(A), \text{ if } \#A > 0.$$

b) From the end of an array:

$$\text{E-GET}(A) = \text{last}(A), \text{ if } \#A > 0 \text{ and } \#A - 1 < M.$$

c) From any position of an array:

$$\text{GET}(i, A) = \text{hd}(\text{take}(i, A) - \text{take}(i-1, A)), \text{ if } 0 < i < \#A.$$

3.3.4 To Find the Next Element of an Array At the Position i

$$\text{NEXT}(i, A) = \text{hd}(\text{take}(i + 1, A) - \text{take}(i, A)), \text{ if } 0 \leq i \leq \#A.$$

3.3.5 To Find the Previous Element of an Array at the Position i

$$\text{PREVIOUS}(i, A) = \text{hd}(\text{take}(i - 1, A) - \text{take}(i - 2, A), \text{ if } 2 \leq i < \#A.$$

Example 3.3.1: Consider the array of numbers A, given by

$$A = [1, 5, 3, 1, 6, 8, 9]. \text{ Here } \#A = 7.$$

Suppose the maximum array size is 10. Now we use the above operations for this array, i.e. INSERT, DELETE and GET.

3.3.1.1 Insertion Operation

Suppose we want to insert an element x = 15 in the array.

i. Insertion at the beginning:

$$\text{B-INSERT}(15, A) = \text{cons}(15, A) = [15, 1, 5, 3, 1, 6, 8, 9]$$

ii. Insertion at the end:

$$\text{E-INSERT}(15, A) = A \mathbin{+\!\!+} [15] = [1, 5, 3, 1, 6, 8, 9] \mathbin{+\!\!+} [15]$$
$$= [1, 5, 3, 1, 6, 8, 9, 15]$$

iii. Insertion at any position:

Suppose i = 4, and x = 15.

$$\text{INSERT}(15, 4, A) = \text{take}(4 - 1, A) \mathbin{+\!\!+} [15] \mathbin{+\!\!+} \text{drop}(4 - 1, A)$$
$$= \text{take}(3, A) \mathbin{+\!\!+} [15] \mathbin{+\!\!+} \text{drop}(3, A) = [1, 5, 3] \mathbin{+\!\!+} [15] \mathbin{+\!\!+} [1, 6, 8, 9]$$
$$= [1, 5, 3, 15, 1, 6, 8, 9]$$

(i) To find the next and previous element at the position i = 4:

$$\text{NEXT}(4, A) = \text{hd}(\text{take}(5, A) - \text{take}(4, A))$$
$$= \text{hd}([1, 5, 3, 1, 6] - [1, 5, 3, 1]) = \text{hd}([6]) = 6$$

$$\text{PREVIOUS}(4, A) = \text{hd}(\text{take}(3, A) - \text{take}(2, A))$$
$$= \text{hd}([1, 5, 3] - [1, 5]) = \text{hd}([3]) = 3$$

3.3.1.2 Deletion Operation

i. Deletion from the beginning:

$$\text{B-DELETE}(A) = \text{tl}(A) = \text{tl}\,[1, 5, 3, 1, 6, 8, 9] = [5, 3, 1, 6, 8, 9]$$

ii. Deletion from the end:

E-DELETE(A) = init(A) = init [1, 5, 3, 1, 6, 8, 9] = [1, 5, 3, 1, 6, 8]

iii. Deletion from any position:

Suppose we want to delete an element from the position i = 3.

DELETE(3, A) = take(2, A) ⧺ drop(3, A) = [1, 5] ⧺ [1, 6, 8, 9]
= [1, 5, 1, 6, 8, 9]

3.3.1.3 GET Operation

i. To get an element from the beginning:

B-GET(A) = hd(A) = 1.

ii. To get an element from the end:

E-GET(A) = last(A) = 9.

iii. To get an element from any position:
Suppose the position of the element i = 5.

GET(5, A) = hd(take(5, A) – take (4, A)) = hd ([1, 5, 3, 1, 6] – [1, 5, 3, 1])
= hd([6]) = 6.

4 Conclusion

The definition of lists using the notion of position function as defined by Tripathy et al in 2001 has been further extended to define the notions of Fuzzy lists [4] and intuitionistic fuzzy lists [5]. Application of this new approach has made it simple to define these extended notions and define their properties. Recently one application of these new approaches has been made to define the relational model. In this paper we provided the realization of data structures like queue, stack, and array using the modified definition. Many other data structures can also be implemented in a similar manner. Also, we have extended these data structures to define their fuzzy as well as intuitionistic fuzzy versions. A package on lists and operations on them can be developed and then used for developing the above data structures following the methods provided in this article. The applications of these data structures can be practically implemented in job scheduling in operating systems, memory allocations, and similar types of some real life operations, which we propose as a direction for further research.

References

[1] Atanassov, K.T.: Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems 20, 87–96 (1986)

[2] Bird, R., Walder, P.: Introduction to Functional Programming. Prentice Hall International Series in Comp. Sc. (1988)

[3] Tripathy, B.K., Jena, S.P., Ghosh, S.K.: On the Theory of Bags and Lists. Information Sciences (USA) 132, 241–254 (2001)

[4] Tripathy, B.K., Jena, S.P., Ghosh, S.K.: On the Theory of Fuzzy Bags and Fuzzy Lists. Int. J. Fuzzy Maths 9(4), 1209–1220 (2001)
[5] Tripathy, B.K., Choudhury, P.K.: Intuitionistic Fuzzy Lists. Notes on Intuitionistic Fuzzy Sets 9(2), 61–73 (2003)
[6] Tripathy, B.K., Gantayat, S.S.: Some More Properties of Lists and Fuzzy Lists. Information Sciences (USA) 166, 167–179 (2004)
[7] Tripathy, B.K., Pattnaik, G.P.: On Some Properties of Lists and Fuzzy Lists. Information Sciences (USA) 168, 9–23 (2004)
[8] Tripathy, B.K., Gantayat, S.S.: Some New Properties of Lists and Fuzzy Lists. Communicated to Information Sciences (USA) (2007)
[9] Tripathy, B.K., Gantayat, S.S.: Some New Properties of Lists and a Framework of a List Theoretic Relational Model. Communicated to International Journal of Technology and Engineering Sciences (IJTES) (2012)
[10] Yager, R.R.: On the Theory of Bags. Intl. Jour. of General Systems 13, 23–37 (1986)
[11] Zadeh, L.A.: Fuzzy Sets. Information and Control 8, 338–353 (1965)

A Comprehensive Study on Multifactor Authentication Schemes

Kumar Abhishek[1], Sahana Roshan[2], Prabhat Kumar[3], and Rajeev Ranjan[3]

[1] Department of Computer Science and Engineering, N.I.T. Patna-800005
kumar.abhishek@nitp.ac.in
[2] Department of Information Technology, Al Musanna College of Technology, Oman
sahana@act.edu.om
[3] Department of Information Technology, N.I.T. Patna-800005
{prabhat8,rrnitp}@gmail.com

Abstract. In recent years, the development in Information technology has unleashed new challenges and opportunities for new authentication systems and protocols. Authentication ensures that a user is who they claim to be. The trust of authenticity increases exponentially when more factors are involved in the verification process. When security infrastructure makes use of two or more distinct and different category of authentication mechanisms to increase the protection for valid authentication, It is referred to as Strong Authentication or Multifactor Authentication. Multifactor authentication uses combinations of "Something you know," "Something you have,", "Something you are" and “Somewhere you are”/“Someone you know”, to provide stronger remote authentication than traditional, unreliable single-factor username and password authentication. In this paper we do a survey on the different aspects of multifactor authentication, its need, its techniques and its impact.

Keywords: Multi Factor Authentication, One Time Passwords, OTP, Biometrics.

1 Introduction

An authentication system determines how a user is identified and verified to the computer. Verification of user’s identity is the main goal behind an authentication system, that is the user is actually who they say they are.

E-commerce has endured a huge hit in the last years due to authentication issues. Traditional reusable passwords have reached their breaking point and organizations and security conscious individuals are searching for an adaptive, flexible, secure and affordable solution to meet the increasing and ever changing business and user needs. In the traditional methods, a system that requires authentication challenges the user for a secret, typically a pair of username and password. The entry of the correct pair grants access on the system’s services or resources. Unfortunately, this approach is susceptible to several vulnerabilities and drawbacks. These shortcomings range from user selected weak or easily guessable passwords to more sophisticated threats such as malware and keyboard sniffers.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 561–568.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Existing authentication methodologies involve three basic "factors":

- Something the user knows: Knowledge Factor (e.g., password, PIN)
- Something the user has :Inherence Factor (e.g., ATM card, smart card)
- Something the user is: Ownership factor (e.g., biometric characteristic) [1]

One more recent inclusion is the fourth factor that is associated with "Someone you know": Social Factor.

Multi-factor authentication methods are difficult to compromise than single-factor methods. Let us take for example, the use of a logon ID/password is single-factor authentication (i.e., something the user knows); whereas, an ATM transaction requires multifactor authentication: something the user possesses (i.e., the card) combined with something the user knows (i.e., PIN). "Out-of-band" controls for risk mitigation can also be included in multifactor authentication methodology. An effective authentication method should have customer acceptance, reliable performance, scalability to accommodate growth, and interoperability with existing systems and future plans [1].

Multifactor authentication methods, properly designed and implemented, are more reliable and stronger fraud deterrents [1].

Another aspect when considering the kind of authentication to employ is the clients required security and risk assessment. For an E-commerce environment, The risk should be evaluated in light of the type of customer; the customer transactional capabilities; the sensitivity of customer information being communicated to both the institution and the customer; the ease of using the communication method; and the volume of transactions. The adequacy of such authentication techniques should be assessed by Institutions in light of new or changing risks such as phishing, pharming, malware, and the evolving sophistication of compromise techniques.

The main emphasis should be on building a good Risk assessment models to calculate the mitigated risks. The risk assessment process should:

Identify all transactions and levels of access associated with Internet-based customer products and services;

- Identify and assess the risk mitigation techniques, including authentication methodologies, employed for each transaction type and level of access; and
- Include the ability to gauge the effectiveness of risk mitigation techniques for current and changing risk factors for each transaction type and level of access. [1]

In this paper, we first explore the simple authentication mechanisms that use a single factor to authenticate users and discuss its pros and cons. We also discuss the need for multifactor authentication and the circumstances that lead to its emergence. Then the paper focuses on multiple factors of authentication and its growth over the past few years.

2 Single Factor Authentication

The most prevalent authentication type in use is single-factor authentication. The features of this traditional (Username-Password) scheme can be listed as:

- Easy to implement
- Requires no special equipment
- Easy to forget
- Can be susceptible to shoulder surfing
- Security based on password strength
- Lack of identity check
- Cost of support increases

Many problems exist within the world of single factor authentication. The first name of username/password combination, the username, may seem non-threatening in a security sense. By knowing the username or even the current naming convention of the usernames in a single-factor authentication site, gives the potential hacker 50% of the information to gain access to vital information.

The Microsoft Corporation has attempted to mitigate one of the inherit problems with the username/password combination. Microsoft has been a proponent of the idea of using "Strong Passwords." Instead of having people use common names for password, Microsoft has detailed the use of using a combination of letters, numbers, and special characters for passwords. While guessing someone's password will be more difficult if the password is "#$#rU78!" the use of strong passwords will still not deter individuals from writing their passwords down. In fact, the use of "strong passwords" will likely increase the number of times someone jots down their password, just so they do not forget it. [2]

Possessing an ID card and swiping it to gain access into a facility is another basic example for single factor authentication.

3 Multi-factor Authentication

According to the FFIEC,

"By definition true multifactor authentication requires the use of solutions from two or more of the three categories of factors. Using multiple solutions from the same category … would not constitute multifactor authentication.[1]"

For applications that require greater security, it may be advisable to implement more than one type of the above mentioned schemes. The implementation thus gets termed as Multifactor authentication.

One problem with implementing multifactor authentication generally is the lack of understanding of "true" multifactor authentication. Supplying a username and password (Both being Knowledge factors) is single factor authentication despite being multiple pieces of distinct information. Supplying additional information in the form of answers to challenge questions (Again Knowledge factor) is still single-factor authentication. Adding a visual image for identification would still be single-factor authentication. An example of true Multi-factor Authentication is requiring the user to insert something (Ownership factor) or requiring a valid fingerprint via biometric reader (Inherence factor). [3][4]

3.1 The Second Factor

In the most basic form, two-factor Authentication has been in the market for quite some time now. The classical example for this would be to possess an ATM card or a credit card (Ownership Factor) and using it along with a secret pin (Knowledge Factor) . This is how two-factor authentication has been implemented in the banking sector. In E-banking or investment sectors, there is a password (Knowledge Factor) and a token (Ownership Factor). The tokens can be delivered in many methods that will be discussed later. More on deployment of these tokens using USBs are discussed in [6].

Either the Ownership or the Inherence factor can be used alongside the Knowledge factor to implement two factor authentications. But the implementation ultimately depends on factors like cost, convenience, flexibility. Multi-factor authentications products make are a need for the day today. A solution is required that should be easy to use and administer and when combined with username/password scheme should provide reasonable security. Costs are the biggest issues when it comes to multifactor authentication, but there are a variety of options at varying price points.

The advantage of implementing Two-factor Authentication using tokens can be listed as follows:

- Reducing the Window of Opportunity
- Eliminating Passive Attacks
- Limited validity for the data obtained
- Mitigating the Risk of Active Attacks
- Increasing the Cost to Implement Fraud [7]

The use of token along with the traditional username-password scheme for two-factor authentication is the concept of One Time Passwords. A One-Time Password (OTP) is valid for only one login session. The static password is usually changed either it has expired or when the user has forgotten it and needs to reset it. The static password are susceptible to cracking as they are cached on computer hard drives and stored on servers. By use of one-time password, password changes each time the user logs in. A one-time password system uses a different password every time you want to authenticate yourself. Each password is used only once; thus, the term “one-time”.

OTP’s can be implemented through hardware or software mechanisms. OTPs are not vulnerable to replay attacks, this means that if a potential intruder manages to record an OTP that was already used to log into a service or to conduct a transaction; he or she will not be able to abuse it since it will be no longer valid. For human beings its very difficult to memorize OTP so there is a requirement of additional technology in order to work.

Randomness is employed by OTP generation algorithms as it is necessary because otherwise it would be easy to predict future OTPs from observing previous ones. There are variation in concrete OTP algorithms vary greatly with respect to their details. Various approaches for the generation of OTPs are:

a. OTP can be generated on time-synchronization between the authentication server and the client. More on Time Synchronized OTPs is discussed in [8]
b. A new OTP can be generated based on the previous password by using a mathematical algorithm. An example of this type of algorithm, credited to Leslie Lamport, uses a one-way function as discussed in [9].
c. Challenge-response one-time passwords will require a user to provide a response to a challenge. For example, this can be done by inputting the value that the token has generated into the token itself. Duplicates can be avoided by using an additional counter is usually involved, so if one happens to get the same challenge twice, this still results in different one-time passwords. However, the computation does not usually involve the previous one-time password; i.e. usually this or another algorithm is used, rather than using both algorithms [9]. In another flavor of the challenge based OTP, The authentication server displays a challenge (a random number) to the user when he attempts to authenticate. When the user enters the challenge number into the token, it executes a special algorithm to generate a password. This is done as the authentication server has the same algorithm, it can also generate the password. Thus the authentication of the user can be completed if the passwords match.

The user should be made aware of the next use of OTP by different ways. Electronic tokens are being used by some systems that the user carries and that generate OTPs and show them using a small display. There are some systems which consist of software that runs on the user's mobile phone. Yet other systems generate OTPs on the server-side and send them to the user using an out-of-band channel such as SMS messaging. Also there is a technique employed by some system in which OTPs are printed on paper that the user is required to carry with them.

One-time password systems can be easy to deploy. Some systems use one-time passwords generated on a hardware device that is communicated directly to the computer, say through a USB port. One-time password systems are generally acceptable to customers, due to their similarity to password systems.

The verifier will need special software and/or hardware. Protected storage and management of the base secrets is required. When a one-time password is used it can be used with any of the other systems if an attacker obtains it. Shorter windows reduce the scope of such attacks. Also, these attacks may be countered by protecting the communication channel.

The table provides the general details of OTP from a security, cost and usability perspective. The generalization may differ affected by the difference in implementation.

The other mechanism that can be used as the second factor of authentication include Terminal Profiling, TAN(Transaction Authentication Number) Lists, SMS tokens which can be either instant SMS tokens or Batch SMS tokens, Smart cards and Chip readers, Chip enabled USBs, Virtual Keypads and Biometrics. More on each of these is discussed in [5].

Table 1. Different aspects of OTP [5]

Protection against Passive Attacks	HIGH
Protection against Active Attacks	LOW to MEDIUM
Initial Cost Involved	MEDIUM
Support and Usage Costs	LOW to NONE
Ease of Use	MEDIUM
Portability	HIGH
Special Software for client required	MOSTLY NO

3.2 The Third Factor

Biometric authentication allows for a physical characteristic of the user to be analyzed by an automated process in making an access control decision. Biometrics may be based on unchanging and unalterable physiological characteristics such as a fingerprint or retinal scan. A biometric system can provide either: Identification or Verification. [13]

Various biometric techniques and identifiers are being developed and tested, these include:

- fingerprint recognition;
- face recognition;
- voice recognition;
- keystroke recognition;
- handwriting recognition;
- finger and hand geometry;
- retinal scan; and
- iris scan.

Generally, Biometrics is the preferred third factor to authenticate users in a system. Most systems can be combined with a biometric scanner such as a fingerprint scanner or an iris scanner. Use of Biometrics authenticates the user to the device rather than the authentication server. While easy to use, futuristic and fashionable, biometric sensors are still relatively expensive and increase the cost of the authentication device. [5].

Biometric authentication provides some inherent advantages as compared to other non-biometric identifiers since biometrics correspond to a direct evidence of the personal identity versus possession of secrets which can be potentially stolen. Moreover, most of the times biometric enrollment is executed in-person and in controlled environments making it very reliable for future use. Biometric authentication poses however several non-trivial security challenges because of the inherent features of the biometric data itself. Addressing these challenges is crucial for the large scale adoption of biometric authentication and its integration with other authentication techniques and with access control systems. [14]

A privacy preserving multifactor authentication protocol with biometrics has been proposed in [15]. Also, [16] proposes a truly three-factor authentication scheme. The three factors are of three different data types, where smart cards show what you have, passwords represent what you know, and biometrics mean what you are, and they are all verified in the server. A generic framework for three-factor authentication in distributed systems has been described in [17].

3.3 The Fourth Factor

A factor "Someone you know" is categorized as the fourth factor of authentication associated with the system of knowledge of some person. This principle of identification by an entrusted person is being used from the beginning of mankind. In the electronic environment, this principle is used to verify and identify by e-mail or phone call. [6].

One practical implementation of this system involves a list of system of guarantees. Authentication mechanisms a are applied on a group of users where one of the users with appropriate rights – a guarantee, uses his authentication devices for emergency authentication of another user –an applicant. If a user loses or forgets his authenticator, A new guarantee who has appropriate rights, can provide a temporary access. One flavor of this mechanism has been implemented in RSA SecurID system as described in [10].

Kevin Mitnick describes several methods of obtaining unauthorized access by social engineering in [11]. Authentication by e-mail or phone is not reliable and sufficient to identify a user to an authentication system. User identification and the guarantee of identifying the user correctly are the most important factor here is the user identity verification process [5].

A proposal for the fourth factor authentication is described in [12].

4 Conclusion

Reusable password protection systems are considered to be at one end of the authentication spectrum and are weak and vulnerable against the most prevalent and easily executed attacks. Strong Multifactor Authentication is absolutely necessary to provide a secure and reliable way of identifying users who need access to a system. There are many different ways of achieving strong authentication, all increase security by magnitudes compared to simple password protection.

Strong Authentication is an integral part of every security strategy and the right authentication system needs to be designed considering the different factors discussed in this paper. More economical and secure options need to be made available which should be flexible and easy to adopt.

References

[1] Authentication in an Internet Banking Environment, Federal Financial Institutions Examination Council, `http://www.ffiec.gov/pdf/authentication_guidance.pdf`

[2] Wildstrom, Stephen: Securing Your PC: You're on Your Own. Business Week, p. 26 (May 26, 2003)

[3] Analysis & Review of FFIEC Multi-Factor Authentication Guidance, Data Risk Governance, `http://datariskgovernance.com/risk-assessment/multi-factor-authentication-in-banking/`

[4] Multifactor Authentication, Tevora Business Solution Blog, `http://blog.tevora.com/authentication/multifactor-authentication-2/`

[5] Gpayments, Two-Factor Authentication: An essential guide in the fight against Internet fraud (February 2006)

[6] Sobotka, J., Doležel, R.: Multifactor authentication systems, vol. 1(4) (December 2010) ISSN 1213-1539

[7] Kemshall, A.: Phil Underwood, Options for Two Factor Authentication, White paper by Securenvoy

[8] `http://www.otptoken.com/`

[9] `http://www.ndkey.com.cn/otp_tech_introduction.html`

[10] RSA White paper, RSA SecurID Authenticators (2009), `http://www.rsa.com`

[11] Mitnick, K.D., Simon, W.L.: The Art of Deception: Controlling the Human Element of Security. Wiley (2002)

[12] Brianard, John, Ari, J., Rivest, Ronald: Fourth Factor Authentication: Some-body you know (2006), `http://www.rsa.com`

[13] Davis, J., Orr, T.: A Case for Multi-factor Authentication in Public Key Infrastructure (Spring 2006)

[14] Bhargav, A., Squicciarini, A., Bertino, E.: Privacy Preserving Multi-Factor Authentication with Biometrics

[15] Bhargav-Spantzel, A., Squicciarini, A.C., Bertino, E., Modi, S., Young, M., Elliott, S.J.: Privacy preserving multi-factor authentication with biometrics. J. Comput. Security 15(5), 529–560 (2007)

[16] Fan, C.-I., Lin, Y.-H.: Provably Secure Remote Truly Three-Factor Authenti-cation Scheme with Privacy Protection on Biometrics. IEEE Transactions on Information Forensics and Security 4(4) (December 2009)

[17] Huang, X., Xiang, Y., Chonka, A., Zhou, J., Deng, R.H.: A Generic Framework for Three-Factor Authentication: Preserving Security and Privacy in Distributed Systems. IEEE Transactions on Parallel and Distributed Systems 22(8) (August 2011)

Quality Assessment Based Fingerprint Segmentation

Kumud Arora[1] and Poonam Garg[2]

[1] Assistant Professor,
Inderprastha Engg. College,
Ghaziabad
`Kum_arora1@yahoo.com`
[2] Professor,
IMT,
Ghaziabad
`pgarg@imt.edu`

Abstract. Lack of robust segmentation against degraded quality image is one of the open issues in fingerprint segmentation. Good fingerprint segmentation effectively reduces the processing time in automatic fingerprint recognition systems. Poor segmentation result in spurious and missing features thus degrading performance of overall system. Segmentation will be more effective if done in accordance to the quality of image. Fingerprint images with high quality have wide range of features which can be used for segmentation than the low quality image, where the fingerprint features are not clearly visible. This paper focus on the two folded segmentation process comprising of quality evaluation and segmentation based on it. Various global and local features are used for assessing quality of image and thereby using them for segmenting ridge area from plain background. The segmented images are compared using percentage of foreground area to total area, genuine number of minutiae points extracted from segmented area. The time taken for image segmentation is also used as a performance parameter. The proposed approach has been tested with images of different qualities from NIST and FVC data sets and the results are proven to be better than the conventional segmentation approaches.

Keywords: OCL (Orientation Certainty Level), Quality Index, CM (Consistency Measure).

1 Introduction

Fingerprint segmentation is one of the first and most integral pre processing steps for any fingerprint verification/identification system. With the onset of the use of fingerprint recognition at large scale, there is rising demand to increase the reliability of fingerprint identification in non ideal conditions. A captured fingerprint image usually consists of two components-foreground region and background region .The poor quality images pose difficulty in detecting features from the image and hence may decrease the recognition performance. The aim of fingerprint segmentation is to identify and exclude un-interested regions and unrecoverable poor quality fuzzy regions from

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 569–579.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

the captured fingerprint image and keeps ridge area as foreground. After segmentation subsequent processing will be focused only on foreground of fingerprint image.

Effective fingerprint segmentation can not only reduce the computation amount for post processing steps in the system, but also improve the reliability of extracted features notably [1]. Fingerprint image is segmented according to different features between ridge area and non ridge area. Previous studies demonstrate that the performance of a fingerprint recognition system is heavily affected by the quality of fingerprint images [6], [7]. A number of factors can affect the quality of fingerprint images [11]: occupation, motivation/collaboration of users, age, temporal or permanent cuts, dryness/wetness conditions, temperature, dirt, residual prints on the sensor surface, etc. Unfortunately, many of these factors cannot be controlled and/or avoided. For this reason, assessing the quality of captured fingerprints is important for a fingerprint recognition system. When fingerprint images include a noisy background, feature extraction algorithms extract a lot of false features.

In this paper, segmentation is based on the quality estimation from the multiple features (local and global features both) captured from fingerprint image. In section 2 the proposed algorithm is illustrated. Proposed algorithm consists of four phases: Pre-processing phase, Computing Quality index from various features, adaptive segmentation in accordance to the quality index is applied and finally post processing phase is applied to merge the isolated blocks. Pre-processing phase consists of application of Gaussian filter to weaken the noise effect captured from the sensor surface and quality estimation based on direction field is done. According to quality index of the input fingerprint image three approaches are used. For good quality index, dominant ridge score and orientation certainty level is used for segmentation .For average quality input image, the adaptive gradient variance approach is combined with consistency measure of fingerprint is used for segmentation .To increase the segmentation reliability for low quality images gradient based segmentation is used to approximate the foreground region of fingerprint images along with orientation reliability score. After the segmentation stage, post processing is done by taking into consideration the continuity of ridges in the neighboring blocks. The experimental results based on proposed method is displayed in section 4 and in section 5, the conclusion is presented.

2 Methodology: Problem Formulation and Proposed Approach

The problem considered here is to extract foreground region from fingerprint image. For good quality image, a simple technique of tracing flow of ridges is enough to segmentation. In average or poor quality images the orientation flow is smudged by noise or some other factor, so tracing ridge flow alone may result in improper segmentation. For segmentation of poor quality image where ridges are corrupted over fairly large portion of captured area, the reliability of orientation flow analysis needs to be assured before segmentation.

The philosophy of the presented quality assessment based segmentation scheme lies in its multidimensional approach of analyzing and estimating quality indices of the fingerprint images. This multilayered technique acquires varied estimation factors, providing a broad base to the quality assessment system thereby improving segmentation robustness. The techniques in this arrangement complement each other, while minimizing their individual weaknesses and reinforcing the overall system strength. The approach consists of:

2.1 Preprocessing

The image pre-processing process consists of image normalization [3] and Gaussian smoothing. We used normalization to normalize the gray-level variations. After that, a 5x5 Gaussian filter was applied to reduce the amount of noise in the image.

2.2 Image Quality Assessment Using Multi-level Factors

First of all, we consider which factors are important when a person evaluate the quality of a fingerprint image using human visual system, which can help us establish better evaluation system. It's obviously that the first impression is whether the foreground of the fingerprint image can supply the enough information. So the size of foreground is very important for image quality assessment. Second, viewers would use some global features to judge, such as image contrast and frequency information, which reflect the image quality basically. And then, viewers need to watch the images carefully on purpose. And they would mostly likely to watch structure of ridge lines .If all these factors have better performance, the image will have better quality.

2.2.1 Effective Area Estimation

Quality assessment process is initiated by the foreground area calculation, which is defined as the percentage of the foreground blocks to the total area [19].A smaller value of effective area would mean a smaller area of fingerprint has been captured. Image requires re-capturing if the effective area value is less than set threshold value.

2.2.2 Main Energy Ratio

Because of noise, we define main energy ratio as:

$$\text{Main Energy Ratio} = \frac{Ep1+Ep2}{E-(Ep1+Ep2)} \tag{1}$$

Where $E=\sum F(i,j)$, if F(i,j)> E_{p1} x 30%

Where $F(i,j)$ is the value of image in the frequency domain after FFT, s is a circle with radius of r, r is the distance from the peak to the DC component and is round to the nearest integer. Ep1 is the value of one peak. Ep2 is the value of the other peak. Ep1 is approximately equal to Ep2. The quality values for the low- and high-quality image are 0.35 and 0.88, respectively.

2.2.3 Image Contrast

Image contrast can be defined as the normalization of variance. Two options for the image contrast are available: Michelson Contrast, Weber Contrast [20].

2.2.4 Consistency Measure

To measure the consistency of the image, firstly the 255 gray levels image are changed into binary image. Secondly, the 3*3 neighbourhood is used, whose centre moved from pixel (2,2) to (255,255), to measure the consistency of ridge flow lines. Following equations are used to determine consistency and consistency measure:

$$\mathrm{consistency(i,j)} = \begin{cases} 0.2 * \mathrm{cen(i,j)} * (9 - \mathrm{sum(i,j)}) + \mathrm{cen(i,j)}, 4 < sum\mathrm{(i,j)} < 9 \\ (1 - \mathrm{cen(i,j)}) + (0.2 * \mathrm{sum(i,j)} * \mathrm{cen(i,j)}), \mathrm{sum(i,j)} \leq 4 \end{cases} \tag{2}$$

$$\text{Consistency measure} = \sum_{i=2}^{r-1} \sum_{j=2}^{c-1} \mathrm{consistency(i,j)} \tag{3}$$

Where cen(i,j) being the value of the pixel (i,j), sum(i,j)is the summation of values of 3*3 neighbourhood cantered at pixel (i,j),r & c be the rows and columns present in image.

Rule based system for calculating Quality Index using multiple parameters extracted from global analysis and local analysis of image:

If (Aeff and Contrast > Th1) then
    If Consistency Measure (CM)>Th3 & Energy>Th4 then
        Classify blocks to “good”, average, poor, very poor

Threshold values are determined for valid local and global features taken into account. It maps quality of image blocks and overall structure to “good”, “average”, “bad” or“very noisy “identity.

2.3 Segmentation

Features for segmentation are selected according to assessment factor calculated. The range of features used for segmentation varies from simple variance, gradient variance, Energy concentration along dominant orientations, Dominant Ridge Score.

2.3.1 Energy Strength along Ridge Valley Orientations

The grey level gradient **(dx,** *dy)* at a pixel exhibits the orientation and the orientation strength of the image at this pixel. By performing Principal Component Analysis on the image gradients in an image block, an orthogonal basis for an image block can be formed by finding its eigen values and eigenvectors. The ratio between the two eigen values gives an indication of how strong the energy is concentrated along the dominant direction with two vectors pointing to the normal and tangential direction of the average ridge flow respectively. The covariance matrix C of the gradient vector for a N units image block is given by

$$C = \frac{1}{N}\sum\left\{\begin{bmatrix} dx \\ dy \end{bmatrix} \begin{bmatrix} dx & dy \end{bmatrix}\right\} = \begin{bmatrix} dx^2 & dxdy \\ dy & dy^2 \end{bmatrix} \quad \begin{bmatrix} c1 & c3 \\ c3 & c2 \end{bmatrix} \tag{4}$$

Where, dx and dy exhibit the intensity gradient of each pixel calculated by Sobel operator of 3 by 3 windows.

$$\mathrm{MIN} = \frac{(c1 + c2) - \sqrt{(c_1 - c_2)^2 + 4c_3^2}}{2} \tag{5}$$

$$\lambda\,\mathrm{MAX} = \frac{(c1 + c2) + \sqrt{(c_1 - c_2)^2 + 4c_3^2}}{2} \quad (6), \qquad \mathrm{OCL} = \frac{\lambda\,\mathrm{MIN}}{\lambda\,\mathrm{MAX}} \quad (7)$$

OCL values distribute between the range of [0, 1].OCL=0 means that ridges and valleys in a block change consistently in the same direction. While OCL=1 means that they are not consistent at all. These blocks may belong to the background with no ridges and valleys. For a small OCL value, ridges and valleys are very clear with good orientation consistency and, as the OCL value increases, they change irregularly. In the OCL images, the gray level of blocks indicates fingerprint images orientation consistency level.OCL value for image is obtained by summation of all the block values.

2.3.2 Local Gradient and Consistency Measure Based Method

In order to segment the foreground blocks effectively, the threshold values are defined for global and local gradient values along X values and Y values. The underlying idea for block segmentation is same as used in Junetal.[23] Foreground threshold is set by the following steps

1) Divide the input image I into non-overlapping blocks, size of blk* blk

2) Compute the gradients $\partial x(i,j)$, and $\partial y(i,j)$,at each pixel (i,j) which is the center of the block.

3) Calculate each block mean and variance value for x and y component of the gradient using the following equations:

$$Mx = \frac{1}{blk^2}\sum_{i=-\frac{blk}{2}}^{\frac{blk}{2}}\sum_{j=-\frac{blk}{2}}^{\frac{blk}{2}} \partial x(i,j) \ \ \& \ My = \frac{1}{blk^2}\sum_{i=-\frac{blk}{2}}^{\frac{blk}{2}}\sum_{j=-\frac{blk}{2}}^{\frac{blk}{2}} \partial y(i,j)\,.$$

where blk is the size of block.In our experiment block size is set to 8.

4) Compute deviation for both M_x and M_y using the equations:

$$Vx = \frac{1}{blk^2}\sum_{i=-\frac{blk}{2}}^{\frac{blk}{2}}\sum_{j=-\frac{blk}{2}}^{\frac{blk}{2}} (\partial x(i,j) - Mx)^2 \ \&$$

$$Vy = \frac{1}{blk^2}\sum_{i=-\frac{blk}{2}}^{\frac{blk}{2}}\sum_{j=-\frac{blk}{2}}^{\frac{blk}{2}} (\partial y(i,j) - My)^2 \tag{8}$$

5) Compute the Gradient Variance's mean:

$$VMx = \frac{1}{r*c}\sum_{i=1}^{r}\sum_{j=1}^{c} Vx \ \ , \ VMy = \frac{1}{r*c}\sum_{i=1}^{r}\sum_{j=1}^{c} Vy \tag{9}$$

6) Calculate the foreground regional variance estimate for x and y component of the gradient as follows:

$$Vfx = \frac{Vsx}{Nsx}, Vfy = \frac{Vsy}{Nsy}\ , \tag{10}$$

where VSx, Vsy and NSx, Nsy are defined respectively as blocks gradient sum and blocks gradient number along X axis and Y axis satisfying the condition, Vx>=VM_x and Vy>=VMy

7) Calculating background regional variance estimate for x and y component of gradients:

$$Vbx = \frac{Vsfx}{Nbx}\; Vby = \frac{Vsfy}{Nby}\;, \tag{11}$$

Where VSfx, Vsfy and Nbx, Nby are defined respectively as blocks gradient sum and blocks gradient number along X axis and Y axis satisfying the condition Vx <=vfx and Vy<=Vfy

8) If Vx<Vbx and Vy<Vby, the block is considered as background otherwise it belongs to foreground.

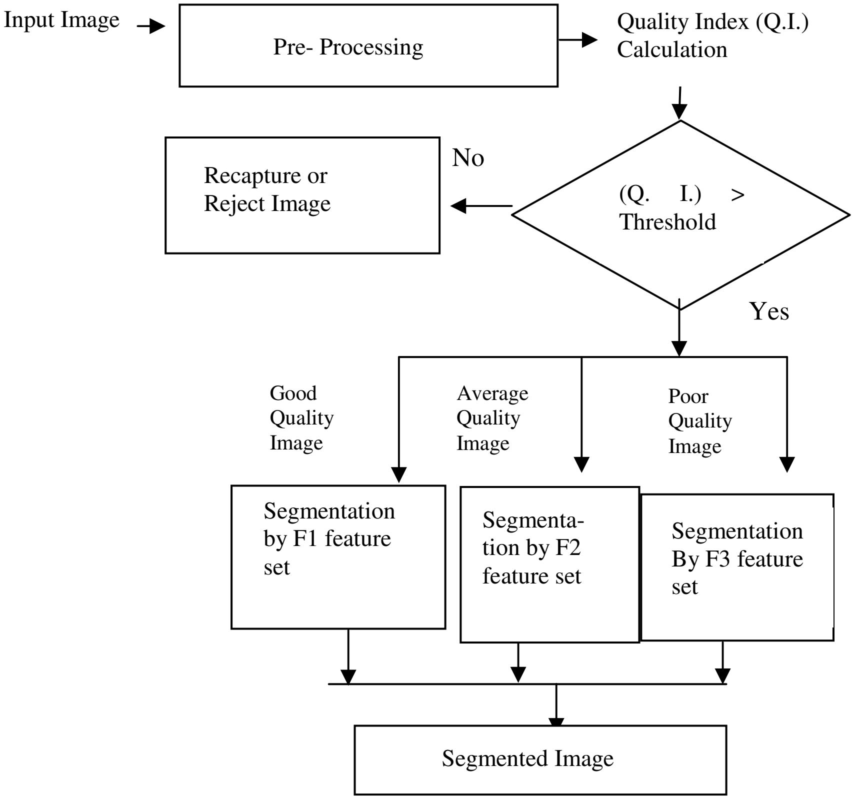

Input Image
Pre- Processing
Quality Index (Q.I.) Calculation
(Q. I.) > Threshold
No
Recapture or Reject Image
Yes
Good Quality Image
Average Quality Image
Poor Quality Image
Segmentation by F1 feature set
Segmentation by F2 feature set
Segmentation By F3 feature set
Segmented Image

Fig. 1. FlowChart of Proposed Approach

Feature sets used in Approach F1: OCL and dominant Score; F2: Local Gradient and Consistency Measure; F3: Gray Variance Degree and consistency Measure

2.3.3 Dominant Ridge Score

The dominant ridge score, orientation of each pixel is obtained as:

$$\theta ij = 90° + \frac{1}{2} arctan\left\{\frac{2Gxy}{Gxx - Gyy}\right\} \tag{12}$$

where the gradients $Gx(i,j)$ is the sum of gradients along X axis in the block ,and $Gy(i,j)$,is the sum of gradients calculated along Y direction in the block. Gxy is the sum of the values of $Gx(i,j) * Gy(i,j)$ for a block .Gxx is the sum of square of the Gx values of the block .Gyy is the sum of the square of the Gy values of the block. Then orientation of each pixel is normalized to 8 normalized orientations (0, 22.5, 45, 67.5, 90, 112.5, 125 or 157.5). Then dominant ridge score is defined as:

$$n_{dom}/\ n_w \tag{13}$$

where n_{dom} and n_w are corresponding, the number of block orientation and the number of pixels in block.

3 Experimental Results and Discussions

In order to validate the actual performance of the proposed approach described in the previous section, fingerprint images from FVC2004 DB3 databases are selected, which fingerprint images quality is poor. We evaluated the proposed method in three ways: 1) the estimation ability of quality, 2) segmentation between genuine foreground blocks and background area, 3) verification performance. The foremost parameter required for fingerprint recognition is segmentation area and in case of sample images the value is much higher than the required threshold.

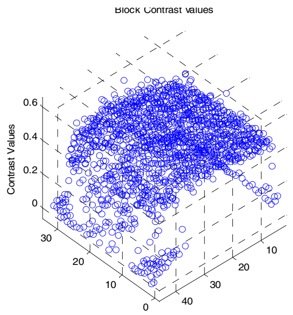


Fig. 2. Contrast measure of Good Quality Image

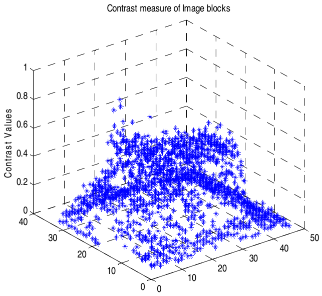


Fig. 3. Contrast measure of Poor Quality Image

In Fig.(2 & 3) those markers which have low contrast values are considered to be background .In Fig.2 large number of markers indicating blocks of image have contrast value greater than threshold value thereby predicting good quality of image. Fig.(4 &5) shows contrast measurement of the blocks of image. The markers which indicate higher consistency value represents the singular points present in the image..In fig.4 maximum number of blocks has good consistency measure except for the regions which are isolated or the regions which don’t contain any ridge area. The blocks which have poor orientation strength are considered to be background blocks.

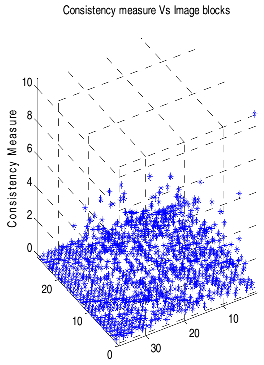


Fig. 4. Consistency Measure of Good Quality Image

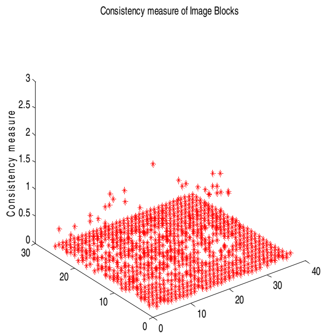


Fig. 5. Consistency Measure of Poor Quality Image

Table 1. Overall Quality Estimation of sample images from various attributes

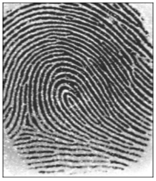
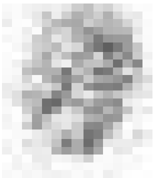
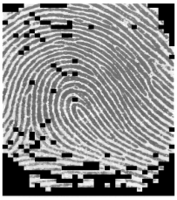
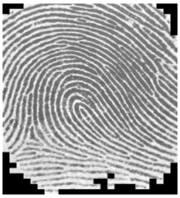
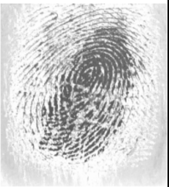
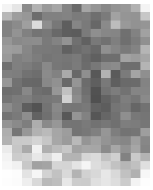
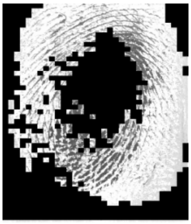
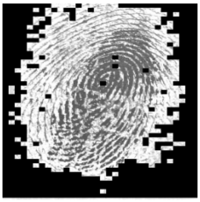
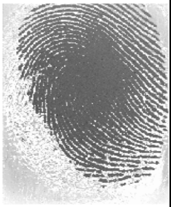
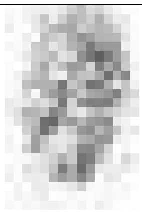
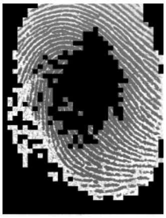
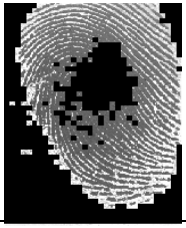
Image Id	**Total Number of Blocks in Image**	**Feature Set Under Evaluation**				**Quality Estimation**
		Effective Fore-ground	**Orienta-tion Strength**	**Consistency Measure**	**Energy Ratio**	
1	**1456.**	**83.17%**	**0.8856**	**89.91**	**79.98**	**Good**
2	**1456**	**74.13%**	**0.529**	**49.62**	**52.79**	**Poor**

Images in the first column are the original fingerprints. Second Column represents Quality map of the sample image. Third column fingerprints are segmented fingerprints using the traditional gradients, coherence, gray variance method. In fourth column segmented images with proposed approach are there. It can be seen that some fingerprints regions with low contrast or high noise still remain in the segmented fingerprints if we use traditional approaches of segmentation, which may generate spurious features during feature extraction.

Segmentation performance = T2/T1, where T2= Time to compute Minutiae points in original Image & T1= Time taken to compute minutiae points in Segmented Image

Segmentation performance for image (1)= T2/T1= 0.0356.
Segmentation performance for image 2 = T2/T1= 0.689(Poor).

Table 2. Overall Fingerprint Image Quality and Segmentation Analysis

Original Image	Quality Map	Segmentation using Coherence, Mean & Variance	Segmentation using Proposed Quality Based method
(1) 			
(2) 			
			

4 Conclusions and Future Work

In this paper, a new method for improving the robustness of the segmentation was proposed by defining image quality map. An attempt is made to correlate the image quality and segmentation performance. It is shown with the help of experiments that image features like gradients which produce very good results in case of good image may not produce good results in case of poor images which are inherently noisy and contain gray level fluctuations that generate wrong results.Future work will be concentrated to improve the combination of image features and hence image quality classification system by using genetic algorithm .Isolated low quality regions can be improved by reconstruction by Line adjacency information around blurred or broken areas.

References

[1] Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer (June 2003)

[2] U. I. A. of India, `http://www.uidai.gov.in/` (last accessed on December 25, 2011)

[3] Hong, L., Wan, Y., Jain, A.: Fingerprint image enhancement: algorithm and performance evaluation. IEEE Trans. Pattern Analysis and Machine Intelligence 20(8), 777–789 (1998)

[4] Chen, Y., Dass, S.C., Jain, A.K.: Fingerprint Quality Indices for Predicting Authentication Performance. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 160–170. Springer, Heidelberg (2005)

[5] Alonso-Fernandez, F., Fierrez, J., Ortega-Garcia, J., Gonzalez- Rodriguez, J., Fronthaler, H., Kollreider, K., Bigun, J.: A Comparative Study of Fingerprint Image-Quality Estimation Methods. IEEE Information Forensics and Security 2, 734–743 (2007)

[6] Simon-Zorita, D., Ortega-Garcia, J., et al.: Image quality and position varia-bility assessment in minutiae-based fingerprint verification. Proc. Inst. Elect. Eng., Vis. Image Signal Process. 150(6), 402–408 (2003)

[7] Fiérrez-Aguilar, J., Chen, Y., Ortega-Garcia, J., Jain, A.K.: Incorporating Image Quality in Multi-algorithm Fingerprint Verification. In: Zhang, D., Jain, A.K. (eds.) ICB 2005. LNCS, vol. 3832, pp. 213–220. Springer, Heidelberg (2005)

[8] Chen, Y., Dass, S.C., Jain, A.K.: Fingerprint Quality Indices for Predicting Authentication Performance. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 160–170. Springer, Heidelberg (2005)

[9] Otsu, N.: A threshold selection method from gray level histogram. IEEE Trans. Syst. Man Cybern. SMC-9, 62–66 (1979)

[10] Helfroush, M., Mohammadpour, M.: Fingerprint segmentation. presented at 3rd International Conference on Information and Communication Technologies: From Theory to Applications, Damascus, Syria (2008)

[11] Joun, S., Kim, H., Chung, Y., Ahn, D.: An experimental study on measur-ing image quality of infant fingerprints. In: Proc. KES 2003, pp. 1261–1269 (2003)

[12] Ko, T., Krishnan, R.: Monitoring and reporting of fingerprint image quality and match accuracy for a large user application. In: Proc. AIPR 2004, pp. 159–164 (2004)

[13] Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Target de-pendent score normalization techniques and their application to signature verification. IEEE Trans. Syst. Man. Cybern. C, Appl. Rev. 35(3), 418–425 (2005)

[14] Lee, S., Choi, H., Choi, K., Kim, J.: Finger-print-Quality Index Using Gradient Components. IEEE Transactions on Information Forensics and Security 3(4) (December 2008)

[15] Tabassi, E., Wilson, C., Watson, C.: Fingerprint image quality. NIST. Res. Rep. NISTIR7151 (August 2004)

[16] Xie, S.J., Yang, J.C., Yoon, S., Park, D.S.: An Optimal Orientation Certainty Level Approach for Fingerprint Quality Estimation. In: Second International Symposium on Intelligent Information Technology Application, vol. 3, pp. 722–726 (2008)

[17] Turroni, F., Maltoni, D., Cappelli, R., Member, D.M.: Improving Fingerprint Orientation Extraction. IEEE Transactions on Information Forensics and Security 6(3) (September 2011)

[18] Wang, S., Zhang, W., Wang, Y.: New features Extraction and Application in Fingerprint segmentation (2002)

[19] Saquib, Z., Soni, S.K., Vij, R.: 2010 International Conference on Computer Design And Appliations, ICCDA 2010. IEEE (2010)

[20] Drahanský, M.: Realization of Experiments with Image Quality of Fingerprints. International Journal of Advanced Science and Technology 6 (May 2009)

[21] Lim, E., Toh, K.A., Suganthan, P.N., Jiang, X.D., Yau, W.Y.: Fingerprint image quality analysis. In: ICIP 2004, vol. 2, pp. 1241–1244 (2004)

Improved Algorithms for Anonymization of Set-Valued Data

B.K. Tripathy[1], A. Jayaram Reddy[2], G.V. Manusha[2], and G.S. Mohisin[2]

[1] School of Computing Science and Engineering
VIT University, Vellore- 632014, TN, India
[2] School of Information Technology and Engineering
VIT University, Vellore-632014, TN, India
{tripathybk,ajayaramreddy,gshahidmohisin2008}@vit.ac.in,
manushagv@gmail.com

Abstract. Data anonymization techniques enable publication of detailed information, while providing the privacy of sensitive information in the data against a variety of attacks. Anonymized data describes a set of possible worlds that include the original data. Generalization and suppression have been the most commonly used techniques for achieving anonymization. Some algorithms to protect privacy in the publication of set-valued data were developed by Terrovitis *et al.*,[16]. The concept of k-anonymity was introduced by Samarati and Sweeny [15], so that every tuple has at least (k-1) tuples identical with it. This concept was modified in [16] in order to introduce k^m -anonymity, to limit the effects of the data dimensionality. This approach depends upon generalisation instead of suppression. To handle this problem two heuristic algorithms; namely the DA-algorithm and the AA-algorithm were developed by them.These alogorithms provide near optimal solutions in many cases.In this paper,we improve DA such that undesirable duplicates are not generated and using a FP-growth we display the anonymized data.We illustrate through suitable examples,the efficiency of our proposed algorithm.

Keywords: K-anonymization, K^m-anonymization, Direct anonymization, Apriori-based anonymization, set-valued data, count –tree.

1 Introduction

In [10] supermarket transactions were considered as the motivating example to describe the requirement of anonymization of set valued data. Suppose an adversary finds some of the items purchased by a customer. If the supermarket database is published later, even after removing the personal identities, there is a chance that the database contains only one transaction containing the items seen by him. Then the adversary can easily identify the other items purchased by the particular customer and get useful information out of it. Identifying the transaction details in this way is known as re-identification. Inorder to preserve the data from being re-identified data can be k-anonymized. According to Sweeney [16] the data is k-anonymized if the information for each person contained in the release cannot be distinguished from at

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 581–594.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

least k -1 individuals whose information also appears in the release. So, we need to transform the original database D to the anonymized database D'. Even after the data is k-anonymized the data cannot be completely protected from being re-identification. However, in this approach the set of attributes in a database are divided into two broad categories. These are the sensitive attributes and the non-sensitive attributes. But in [15], the attributes are considered to be alike and are not divided into such categories with the view that any subset of the set of attributes can be sensitive attributes and the others as non-sensitive attributes depending upon the application. So, assuming that an adversary has knowledge about at the most m items and we want to prevent him from distinguishing the transaction from a set of k published transactions in the database. Equivalently, for any set of m or less items, there should be at least k transactions, which contain this set, in the published database D'. So, making use of this concept a k^m -anonymization model was developed in [10] and algorithms were developed to deal with such type of set-valued data.

A subset of items in a transaction play the role of quasi-identifier.By which the data can be re-identified by linking techniques.The items in the transaction can be anonymized through various anonymization approaches. We describe some of these below. In Suppression information is removed from the data. For example, the gender attribute can be removed from a patient database. In Generalization the information is generalized from more specific to less specific or can be coarsened into sets. For example, in an employee database DOB can be generalized form (dd-mm-yy) format to only year. In Perturbation noise is added to the data. For example, salary can be added to an employee database replacing one of the sensitive attributes. In Permutation sensitive associations between entities are swapped. For example, in the purchase of medicines by a person, a medicine can be swapped with number of people purchased that particular medicine. Out of these techniques mostly Generalization or Suppression is used for the anonymization of data. It can be noted that when suppression is used for anonymization then there is greater loss of information than when Generalization is used. So, the Generalization technique is used for transactional database or "market basket" data analysis.

Three algorithms were introduced in [16] to achieve k^m -anonymization. However, the problem in these algorithms is the generation or redundant transactions while generating the additional tuples to achieve anonymization. In this paper we improve upon the two algorithms (DA and AA algorithms) in [16] by adding several steps so that the number of transactions generated is the exact number required.

2 Literature Survey

As mentioned in the introduction, one of the earliest attempts to anonymization of databases is the introduction of the notion of k-anonymity by Samarati [14] and Sweeny [15]. A table is k-anonymised if each record is indistinguishable from at least k-1 other records with respect to a set of quasi-identifier (QI) attributes. The QIs are than generalised and the records with identical QI values thus form an anonymised group. The process of transforming a database table D into a table D' after anonymization is called recoding. It has been established in [12] that the problem of optimal k-anonymity for multidimensional QI is NP-hard, under both generalisation and suppression models. Several approximate algorithms those minimise the number of suppressed values have been obtained. These are in [12] with a bound of O(k.logk), in [2] with a bound of O(k), and with a bound of O(logk) in [3]. In [9] an algorithm called Incognito is proposed, which uses dynamic programming approach to find an optimal solution. The problem here is the concept of full-domain recoding, which requires that all values in a dimension must be mapped to the same level of hierarchy. Inspired by Incognito, Terrovitis *et al* .,[16] proposed three algorithms, the optimal anonymization (OA) algorithm and two heuristic algorithms called the direct anonymization (DA) algorithm and the apriori anonymization (AA) algorithm. Here, the full-domain recoding is not assumed. Also, in k-anonymity the set of QI attributes are known beforehand. However, in case of [16], since any set of m items (which corresponds to the attributes) can be used by the adversary, no QI set can be predetermined.

The concept of k-anonymity has been extended to the notion of l-diversity by Machanavajjhala *et al.,* [11] and further to the notion of t-closeness by Li *et al* .,[10]. Several approaches to solve the l-diversity problem efficiently are provided in [5,21,22]. Many fast l-diversity algorithms have been developed by Tripathy *et al* .,[17,18,19,20]. Also, some of the l-diversity algorithms developed by Tripathy *et al.,* take care of uncertainty in databases by using rough set methods to achieve l-diversity. Some of these algorithms deal with hybrid databases.

3 Concepts and Existing Algorithms

In this section we introduce some concepts to be used in the paper and also introduce the existing algorithms proposed in [16] along with explanations.

3.1 Generalization

The data generalization concept explored from data mining as a way to hide detailed(more specific) information, rather than discover trends and patterns.Data mining is frequently described as the process of extracting valid, authentic and actionable information from large databases. In other words, data mining derives patterns and trends that exist in data. These patterns and trends can be collected together and defined as a mining model. Data masking is the process of obscuring (masking) specific data elements within data stores. It ensures that sensitive data is replaced with realistic but not real data. The goal is that sensitive customer information is not available outside of the authorized environment.Once the data is masked, not only it can discover useful patterns, but also mask private (personal) information. Here, we adapted an approach of bottom-up generalization of data mining to generalize the data. The generalized data remains useful to classification but becomes to link to other sources.The generalization space is specified by a hierarchial structure of generalization.

For example let us consider anonymity problem, a data holder wants to release a person-specific data R, but wants to prevent from linking the released data to an external source E through shared attributes $R \cap E$, called virtual identifier. One approach is to generalize specific values into less specific but semantically consistent values to create k-anonymity if one record r in R is linked to some external information,atleast k-1 other records are similarly linked by having the same virtual identifier (i.e. the information used by intruder to find the information of a person) value as r(i.e. one record in the person specific data). The idea is to make the inference ambiguous by creating extra-neous linkages. An example is generalizing "birth date" to "birth-year" so that everybody in the same year are linked to a medical record with birth year,but most of these linkages are non-existing in the real life. The key is identifying the "best"generalization.

k^m –anonymization model is proposed for transactional databases, where m is the atmost number of items known to an adversary.For an set of m or less items there should be atleast k transactions which contain m itemset in the published database D', to prevent an adversary from distinguishing the transaction from a set of k transactions in the database. But here there is no fixed,well defined set of quasi-identifier for the sensitive data.A subset of items in a transaction can act as quasi-identifier for the sensitive ones or vice versa. To solve the k^m –anonymization problem for a transactional database, generalization is in use.If original database D does not satisfy the k^m -anonymity then it is transformed to D' by replacing items with their generalized ones.Here in supermarket database while entering the item is provided with its respective generalization. Generalization replaces intial attribute with generalized attribute.

For example consider T={orange, goodday, mango, timepass}, in this {orange, mango} can be generalized to Fruits and {goodday, timepass}can be generalized to biscuits and total transaction to {Fruits, biscuits} .

Example 1: In the table below we present a set of items along with their generalisations, which form the components of any transaction in a super market.

Table 1. Database of items and their generalizations

Item ID	Items	Generalisation
1	Apple	Fruit
2	Orange	Fruit
3	Pineapple	Fruit
4	Clinicplus	Shampoo
5	Dove	Shampoo
6	Aswini	Oil
7	VVD	Oil
8	Margo	Soap
9	Lux	Soap
10	Chik	Shampoo

Table 2. Database of some transactions

ID	Items in Cart
1	Apple,Aswini,ClincPlus
2	Dove,ClincPlus,Aswini
3	Apple,Aswini,ClincPlus,Orange,Pineapple
4	ClinicPlus,Apple,Dove,Pineapple
5	Aswini,ClincPlus,Orange
6	Apple,Orange,Pineapple
7	Aswini,ClinicPlus,Orange

Table 3. Transformed Transactional Database

ID	Items in Cart
1	Fruit,Shampoo,Oil,
2	Oil,Shampoo
3	Fruit,Shampoo,Oil
4	Fruit,Shampoo
5	Fruit,Shampoo,Oil,
6	Fruit
7	Fruit,Shampoo,Oil

3.2 Count Tree

To find whether generalization applied provides K^m-anonymity, it is to count efficiently the support of all combinations of m-items that appear in the database. To avoid scanning the database each and every time generalization has to be checked. To acheive these two goals a datastructure was constructed which keep track of not only all combinations of m items from the generalized database but also it must know how each generalized value effects the database. The support value of each combination of all items in the transactions is calculated.Inorder to keep track of the support of all the transactions a count tree-data structure was constructed.

To count the support of all these combinations and to store them the count-tree is used, based on the count tree algorithm. The tree assumes an order of items and their generalizations, based on their frequencies(supports)in D.To compute this order, a database scan is required.The support of each itemset with upto m items can be computed by following a corresponding path in the tree and using the support value of the corresponding node[16].Count-tree follows the apriori principle which states that the support of an item set is always less than or equal to the support of its subsets.

Here in the database the items which are present and not present in the transaction are represented 1's and 0's respectively. Based on this the frequent item sets are generated. A frequent item set is an item set whose number of occurrences is above a threshold. For each combination of items in the transaction the support value is calculated and is displayed. The items which are having the support value less than the minimum support value then those items are neglected.

Based on the count tree two anonymization techniques can be performed i.e. Direct Anonymization (DA) and Apriori -based anonymization (AA).

Definition 1. Support: The support or utility or prevalence for an association rule X=>Y is the percentage of transactions in the database that contains both X and Y.

$$\text{Support}(X \rightarrow Y) = \frac{\text{No.of tuples containing both X and Y}}{\text{Total no.of tuples}} = P(X \cap Y).$$

Table 4. Algorithm for Creation of the tree for k^{m} anonymity([16])

```
Populate Tree (D, tree, m)
1: For all t in D do for each transaction
2:     expand t with the supported generalized items
3:     For all combination of c ≤m items in the expanded t do
4:         If $\neg\exists\ i, j \in c$ such that i generalizes j then
5:             insert c in tree
6:             increase the support counter of the final node
```

3.2.2 Output of Count Tree

Table 5. The following table provides an example of the output of the above algorithm

Transaction Table		
Item	**No. of Occurrences**	**Support**
Apple	1	0.5
PineApple	1	0.3571428571428715
Dove	1	0.21428571428571427
Margo	1	0.14285714285714285
Lux	1	0.07142857142857142
VVD	1	0.14285714285714285
Orange	1	0.5
Aswini	1	0.5714285714285714
ClincPlus	1	0.7142857142857143
Chik	1	0.07142857142857142

3.3 Optimal Anonymization(OA)

To find the optimal cut i.e. no generalization that satisfies k^{m} -anonymity and has the least information loss, we can examine systematically the generalizations in the cut hierarchy, in a bottom-up, breadth first fashion. Initially the cut C_{ng} which corresponds to no generalization is put to queue Q. While Q is not empty, we remove first cut from C from it and examine whether it satisfies k^{m} -anonymity [16]. If it satisfies then it becomes a candidate solution. If it does not satisfy k^{m} -anonymity, its immediate ancestors in the hierarchy, which do not have a descendant cut that satisfies k^{m} -anonymity are added to the queue.

Table 6. Optimal Anonymization Algorithm

```
OA(D,I,K,m)
$C_{opt}$ :=null; $C_{opt}$.cost := $\infty$        // initialize $C_{opt}$
add $C_{ng}$ to an initially empty queue Q
While (Q is not empty) do
    pop next cut C from Q
    if C does not provide $K^m$-anonymity to D then
        for all immediate ancestors $C_{ans}$ of C do
            if $C_{ans}$ does not appear in H then
                push $C_{ans}$ to Q
    else                        // C provide $K^m$-anonymity to D
        for all immediate ancestors $C_{ans}$ of C do
            add $C_{ans}$ to H
            if $C_{ans}$ in Q then
                delete $C_{ans}$ from Q
        if $C \cdot cost < C_{opt}$.cost then
            $C_{opt}$ := C
return $C_{opt}$.
```

3.4 Direct Anonymization

Direct anonymization is to scan the count tree once and then use the generalized combinations to find a solution that optimizes problem of re-identification. Optimal Anonymization method is based on pre-computation of complete count tree for sets consisting of up to m item sets [16]. Direct anonymization scans the tree to detect m-sized paths that have support less than K.For each such paths,it generates all possible generalization.In this direct anonymization, the database is scanned and a count tree is constructed.Once the count tree has been created; direct anonymization initializes the output generalization C_{out} as bottommost cut of the lattice (i.e. no generalization). Then performs preorder traversal of count tree. Based on the initial support count neglect the item sets whose support count is less than the initial. For every node encountered, if the item corresponding to that node has already been generalized in C_{out} , direct anonymization backtracks as all complete m-sized paths passing from there correspond to itemsets that will not appear in the generalized database based on C_{out} (and therefore their supports need not be checked).

Table 7. Direct Anonymization Algorithm

```
DA (D, I, k, m)
1. Scan D and create count-tree
2. Initialize $C_{out}$
3. For each node v in preorder count-tree tranversal do
    4. If the item of v has been generalized in $C_{out}$ then 5. backtrack
    6. If v is a leaf node and v.count<k then
        7. J:= itemset corresponding to v
        8. find generalization of items in J that make J k-anonymous
        9. merge generalization rules with $C_{out}$
        10. backtrack to longest prefix of path J,where no item has been generalized in C
11.Return $C_{out}$.
```

3.5 Apriori Algorithm

Apriori is a classical algorithm for learning association rules. Association rule mining is finding the frequent patterns, associations, correlations or casual structures among sets of items or objects in transaction databases, relational databases, and other information repositories. Apriori is designed to operate on databases containing transactions (for example, collection of items bought by customers). Apriori uses a "bottom up" approach where frequent subsets are extended one item at a time (a step known as a candidate generation and groups of candidates are tested against the data).The algorithm terminates when no further successful extensions are found.It uses a level wise search, where K-itemsets (an itemsets that contain K items is a K-itemset) are used to explore (K+1) itemsets, to mine frequent itemsets from transactional database for association rules. First, the set of frequent 1-itemsets is found. This set is denoted by L1, which is used to find L2, the set of frequent 2-itemsets, which are used to find L3 and so on until no more frequent K-item sets can be found.

Table 8. Apriori based Anonymization Algorithm

```
AA (D, $I$, $k$, m)
1: Initialize $c_{out}$
2: For i: = 1 to m do                          //for each item set length
3:   initialize a new count-tree
4:   For all t ϵ D do                          //scan D
5:           extend t according to $C_{out}$
6:           add all i-subsets of extended t to count-tree
7:       run DA on count-tree for $m = i$ and update $C_{out}$
```

4 The Proposed Algorithms

As mentioned earlier we improve the two algorithms DA and AA in order to reduce the generation of redundant transactions which makes the further analysis of the output efficient and simpler. First we present the improved DA algorithm below. We have added new steps from 12 to 25 in the existing algorithm.

4.1 Improved Direct Anonymization Algorithm

Table 9. Improved Direct Anonymization

```
DA(D,I,k,m)
1: Scan D and create count-tree
2: Initialize $C_{out}$
3: For each node v in preorder count-tree tranversal do
4:   If the item of v has been generalized in $C_{out}$ then
5:        backtrack
6:   If v is a leaf node and v.count<k then
7:       J: = item set corresponding to v
8:       find generalization of items in J that make J k-anonymous
9:       merge generalization rules with $C_{out}$
10:      backtrack to longest prefix of path J, wherein no item has been generalized
              in $C_{out}$
11: Return $C_{out}$.
12. For i: =1 to $C_{out}$ do
13.   Initialize count=0
14.   Scan each transaction in $C_{out}$
15.   Seperate each item in a transaction and store it in p
16.   Increment count.
17. For j: =1 to count do
18.    For all g belongs $C_{out}$ do
19.       Compare each item of p with that of $C_{out}$
20.         If all items of i equal to cout
21.         Increment r
22.   If $k_a$ equal to r then backtrack to i
23    Else if r is greater than $k_a$ then get the index position of the similar
        transactions
24.   make them NULL until $k_a$ equal to r
25      Else update the transactions in database, where $k_a$–anonymization value.
```

4.2 Improved Apriori Based Anonymization Algorithm

Table 10. Improved Apriori Based Anonymization

```
AA   (D, I, k, m)
1.Initialize cout
2. For i: = 1 to m do     // for each item set length
3.Initalize a new count-tree
4.For all t ϵ D do
5.      extend t according to $C_{out}$
6.      add all i-subsets of extended t to count tree
7.For i: =1 to $C_{out}$ do
8.          Initialize count=0
9.         Scan each transaction in $C_{out}$
10. Seperate each item in a transaction and store it
           in p
11.         Increment count.
12.  For j: =1 to count do
13.        For all g belongs $C_{out}$ do
14.            Compare each item of p with that of $C_{out}$
15.             If all items of i equal to cout
16.             Increment r
17.        If $k_a$ equal to r then backtrack to i
18. Else if r is greater than $k_a$ then get the
Index position of the similar Transactions.
19. make them NULL until $k_a$ equal to r
20. Else update the transactions in database,
where $k_a$-anonymization.
```

It may be noted that the steps 7 to 20 in this algorithm are same as steps 12 to 25 in the previous algorithm.

5 Comparison of the Algorithms

The execution time of both Direct anonymization and Apriori based anonymization is compared.Here apriori based anonymization takes less amount of time when compared to direct anonymization. In order to get the output generalization C_{out} we run Direct anonymization on count tree in Apriori based anonymization that's the reason why we get the similar output for both direct and apriori anonymization.

5.1 Experimental Analysis

We consider here a Supermarket database where there is a provision to add an item,add an transaction as well as to view the transactions.Here a database is created with limited number of transactions let us consider 10 transactions.

5.1.1 Database

We are considering the following items based on the items transactions are created.

Table 11. Items in the database and their generalization

Item ID	Items	Generalisation
1	Apple	Fruit
2	Orange	Fruit
3	Pineapple	Fruit
4	Clinicplus	Shampoo
5	Dove	Shampoo
6	Aswini	Oil
7	VVD	Oil
8	Margo	Soap
9	Lux	Soap
10	Chik	Shampoo

5.1.2 Transactions in Database

We are considering a small database which contains 15 transactions.

Table 12. Transactions in the database

Transaction ID	Transaction Items
1	Apple,Aswini,ClincPlus
2	Dove,ClincPlus,Aswini
3	Apple,Aswini,ClincPlus,Orange,Pineap
4	ClinicPlus,Apple,Dove,Pineapple
5	Aswini,ClincPlus,Orange
6	Apple,Orange,Pineapple
7	Aswini,ClinicPlus,Orange
8	Apple,Aswini,ClincPlus,Dove,Margo, Orange,PineApple,VVD

Table 12. (*continued*)

9	Aswini,ClincPlus,VVD
10	ClincPlus,Orange,Margo
11	ClinicPlus,Dove
12	Aswini
13	PineApple,Chik,Lux
14	Apple,Dove,Margo
15	ClinicPlus,Aswini

5.1.3 Output Using Earlier Algorithms

We implemented the following algorithms in NETBEANS IDE using JAVA SWINGS with backend technology as SQL SERVER. Outputs of both the algorithms are similar.

Table 13. Output of Direct Anonymization and Apriori Anonymization

Transaction id	Generalized transactions
1	Fruit,Shampoo,Oil,
2	Oil,Shampoo
3	Fruit,Shampoo,Oil
4	Fruit,Shampoo
5	Fruit,Shampoo,Oil,
6	Fruit
7	Fruit,Shampoo,Oil
8	Fruit,soap,Shampoo,Oil
9	Shampoo,Oil
10	Fruit,soap,Shampoo
11	Oil
12	Fruit,soap,Shampoo,
13	Fruit,soap,Shampoo,
14	Oil,Shampoo,
15	Fruit,Shampoo,Oil
16	Fruit,soap,Shampoo,Oil,
17	Fruit,Shampoo,Oil
18	Fruit,soap,Shampoo,
19	Fruit,
20	Oil,
21	Shampoo,Oil,
22	Fruit,Shampoo,
23	Oil,Shampoo,
24	Fruit,soap,Shampoo,
25	Oil,Shampoo,
26	Fruit,soap,Shampoo,

5.1.4 Output Based on New Approach

In this new approach, the number of duplicate transactions is decreased because the algorithm checks for all the conditions inorder to achieve K^m-anonymity completely. As said earlier the outputs are same for both algorithms.

Table 14. Output of Direct Anonymization and Apriori Anonymization

Transaction ID	Generalized Transactions
1	Fruit,Shampoo,Oil,
2	Oil,Shampoo
3	Fruit,Shampoo
4	Fruit
5	Fruit,soap,Shampoo,Oil,
6	Fruit,soap,Shampoo,
7	Fruit,soap,Shampoo,
8	Oil,Shampoo
9	Oil,
10	Oil,
11	Fruit,Shampoo,Oil,
12	Fruit,soap,Shampoo,Oil,
13	Fruit
14	Fruit,Shampoo

5.2 Theoretical Estimation

In this section we provide a theoretical estimation of the efficiency of the table generated from the earlier algorithm and our algorithms. Suppose the total number of transactions in the original table be n.Let there be m_i transactions of size r_i=1,2,...p. So that $\sum_{i=i}^{p} m_i = n$ .If out of m_i transactipns there are $m_{p(i)}$ transactions are repeated or permutations of each other then the number of duplicate transactions generated by the earlier algorithms is $\sum_{i=1}^{p} k.m_i$ .Where as our algorithms will generate only $\sum_{i=1}^{p} k(m_i - m_{p(i)})$ duplicate transactions.

5.2.1 Worst Case Analysis

If $m_{p(i)} = m_i \forall i$ then $k.m_i$ additional transactions are generated if $m_i>k$ .If $m_i \le k$ then we add $(k-m_i)$ transactions.This is true for all i.So, if $m_i>k$,$i=1,2,...p$ our algorithms do not generate any additional transactions where as the earlier algorithm generate $\sum_{i=1}^{p} k.m_i = k.n$ duplicate transactions.

5.2.2 Best Case Analysis

$m_{p(i)} = 1$ for all i, even then the earlier algorithms generate n,k items,where as our algorithms generate no additional transactions if $m_i > k$,$i=1,2,\ldots p$ Otherwise, our algorithms generate only those many transactions as are necessary to make up for k-anonymity.

6 Conclusion

In this paper we have improved the k^m -anonymity algorithms developed in [16] for anonymization of set-valued data. The algorithms in [10] generate many redundant transactions and it is very inconvenient for further analysis. The improved algorithms generate the exact number of tuples required for the generalisation. This reduces the size of the output table considerably and makes it simpler for further analysis. We provided an example to illustrate the efficiency of the new algorithms over the existing algorithms. Also, theoretically we have computed the extent of improvement in the results.

References

[1] Aggarwal, G., Feder, G., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., Zhu, A.: Achieving Anonymity via Clustering. In: Proc. of ACM PODS, pp. 153–162 (2006)

[2] Aggarwal, G., Feder, G., Kenthapadi, R., Motwani, R., Panigrahy, R., Thomas, D., Zhu, A.: Approximation Algorithms for k-Anonymity. Journal of Privacy Technology (2005)

[3] Atzori, M., Bonchi, F., Giannotti, F., Pedreschi, D.: Anonymity Preserving Pattern Discovery. VLDB Journal (2008) (accepted for publication)

[4] Bayardo, R.J., Agrawal, R.: Data Privacy through Optimal k-Anonymization. Proc. of ICDE, pp. 217–228 (2005)

[5] Ghinita, G., Karras, F.P., Kalnis, P., Mamoulis, N.: Fast Data Anonymization with Low Information Loss. In: VLDB, pp. 758–769 (2007)

[6] Ghinita, G., Tao, Y., Kalnis, P.: On the Anonymization of Sparse High-Dimensional Data. In: Proceedings of ICDE (2008)

[7] Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proc. of ACM SIGMOD, pp. 1–12 (2000)

[8] Iyengar, V.S.: Transforming Data to Satisfy Privacy Constraints. In: Proceedings of SIGKDD, pp. 279–288 (2002)

[9] LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: Efficient Full-domain k-anonymity. In: Proceedings of ACM SIGMOD, pp. 49–60 (2005)

[10] Li, N., Li, T., Venktasubramanian, S.: t-closeness Privacy Beyond k-anonymity and l-diversity. In: Proceedings of ICDE, pp. 106–115 (2007)

[11] Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, S.: l-diversity: Privacy Beyond k-Anonymity. In: Proceedings of ICDE (2006)

[12] Meyerson, A., Williams, R.: On the Complexity of Optimal k-Anonymity. In: Proceedings of ACM PODS, pp. 223–228 (2004)

[13] Park, H., Shim, K.: Approximate algorithms for k-Anonymity. In: Proceedings of the ACM SIGMOD, pp. 67–78 (2007)

[14] Samarati, P.: Protecting Respondents Identities in Microdata Release. IEEE TKDE 13(6), 1010–1027 (2001)

[15] Sweeney, L.: K-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty. Fuzziness and Knowledge-Based Systems 10(5), 557–570 (2002)

[16] Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy Preserving Anonymization of Set-Valued Data. In: PVLDB 2008, Auckland, New Zeland, pp. 115–125 (2008)

[17] Tripathy, B.K., Devineni, H., Jayasri, K.J., Bhargava, M.: An Efficient Clustering Algorithm for l-diversity. In: Proceedings of the International Conference on Advances and Emerging Trends in Computing Technologies, ICAET 2010, June 21-24, pp. 76–81. SRM university (2010)

[18] Tripathy, B.K., Panda, G.K., Kumaran, K.: A Rough Set Approach to develop an efficient l-diversity Algorithm based on Clustering. In: Proc. of the 2nd IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence, January 8-9, p. 34 (2011)

[19] Tripathy, B.K., Panda, G.K., Kumaran, K.: A Fast l - Diversity Anonymisation Algorithm. In: Proc. of the Third International Conference on Computer Modelling and Simulation, ICCMS 2011, Mumbai, January 7-9, pp. V2-648–652(2011)

[20] Tripathy, B.K., Maity, A., Ranajit, B., Chowdhuri, D.: A fast p-sensitive l-diversity Anonymisation algorithm. In: Proceedings of the RAICS IEEE Conference, Kerala, September 21-23, pp. 741–744 (2011)

[21] Xiao, X., Tao, Y.: Anatomy: Simple and Effective Privacy Preservation. In: Proceedings of VLDB, pp. 139–150 (2006)

[22] Zhang, Q., Koudas, N., Srivastava, D., Yu, T.: Aggregate Query Answering on Anonymised Tables. In: Proceedings of ICDE, pp. 116–125 (2007)

An Efficient Flash Crowd Attack Detection to Internet Threat Monitors (ITM) Using Honeypots

K. Munivara Prasad[1,3], M. Ganesh Karthik[3], and E.S. Phalguna Krishna[2,3]

[1] Assistant Professor (SL)
[2] Assistant Professor
[3] Department of Computer Science and Engineering,
Sree Vidyanikethan Engg. College, Tirupati
{prasadkmv27,ganeshkarthik16,phalgunakrishna}@gmail.com

Abstract. Now a days there is a rapid increase of traffic to a given web server within a short time as the number of Internet users increases, and such a phenomenon is called a flash crowd. Once flash crowds occurs a response rate decreases or the web server may crash as the load increases. In this paper we implement the Internet Threat Monitoring (ITM), is a globally scoped Internet monitoring system whose goal is to measure, detect characterize, and track threats such as distribute denial of service (DDoS) attacks and worms. To block the monitoring system in the internet the attackers are targeted the ITM system. In this paper we address flash crowd attack against ITM system in which the attacker attempt to exhaust the network and ITM's resources, such as network bandwidth, computing power, or operating system data structures by sending the malicious traffic. We propose an information-theoretic frame work that models the flash crowd attacks using Botnet on ITM. Based on this model we generalize the flash crowd attacks and propose an effective attack detection using Honeypots.

Keywords: Internet Threat Monitors (ITM), DDoS, flash crowd attack, Botnet and Honeypot.

1 Introduction

Internet is the global network which provides the various communications for the users; it also provides the better scalability and openness. This causes the unprotected and unauthorized transactions to the users. This feature of the internet is useful for the attackers to perform some attacks by sending malicious data through malicious and suspicious transaction with out bothering the security. The lack of authentication means that attackers can create a fake identity, and send malicious traffic with impunity. A denial-of-service (DoS) attack [2] is an explicit attempt by attackers to prevent an information service's legitimate users from using that service. Network bandwidth, computing power, or operating system data structures are the resources of the victim which has been exhaust by these attacks. Flood attack, Ping of Death attack, SYN attack, Teardrop attack, DDoS, and Smurf attack are the most common types of DoS attacks. The hackers who launch DDoS attacks typically target sites or services

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 595–610.
springerlink.com
© Springer-Verlag Berlin Heidelberg 2013

provided by high- profile organizations, such as government agencies, banks, credit-card payment gateways, and even root name servers.

A common attack used to attack a victim machine by sending a large amount of malicious traffic is a flash crowd-based Distributed Denial of Service (DDoS) attack. Network level congestion control can throttle peak traffic to protect the network. Network monitors are used to monitor the traffic in the networks to classify them as genuine or attack traffic and also these monitors gives the traffic as an input to several DDoS detection algorithms for detection of DDoS attacks. However, it cannot stop the quality of service (QoS) for legitimate traffic from going down because of attacks. Two features of DDoS attacks hinder the advancement of defense techniques. First, it is hard to distinguish between DDoS attack traffic and normal traffic. There is a lack of an effective differentiation mechanism that results in minimal collateral damage for legitimate traffic. Second, the sources of DDoS attacks are also difficult to find in a distributed environment.

Therefore, it is difficult to stop a DDoS attack effectively. The Internet Threat Monitoring (ITM) System basically has two main components one is centralized data center and another is the number of monitors which are distributed across the Internet. Each monitor covers the range of IP addresses and monitors the traffic to send the traffic logs to data center. The data center now collects the traffic logs from monitors and analyzes the collected traffic logs to publish reports to ITM system users.

The collected logs, as a random sample of the Internet traffic, can still provide critical insights for the public to measure, characterize, and track/detect Internet security threats. The idea of ITM systems dates back to DShield and CAIDA network telescope [4], [5][17], which have been successfully used to analyze the activities of worms and DDoS attacks [3], [6].

The reason is that if an attacker discovers the monitor locations, it can easily avoid detection (by ITM systems) by bypassing the monitored IP addresses and directing the attack to the much larger space of unmonitored IP addresses. Furthermore, such an attacker may even mislead the reports published by an ITM system by manipulating traffic to the identified monitors, generating highly skewed samples. Since ITM reports are trusted by the public as a random (unbiased) sample of Internet traffic, the confidentiality of monitor locations is vital for the usability of ITM systems.

The monitor locations of an ITM system can be compromised by introducing several attacks by the attackers which includes Localization attacks [1] and DDoS Attacks which exploits some vulnerability or implementation bug in the software implementation of a service to bring that down or that use up all the available resources at the target machine or that consume all the bandwidth available to the victim machine, this is called as Bandwidth attacks. There are several attacks such as Localization attacks[1] and DDoS Attacks are introduced by the attackers in order to compromised the monitor locations of an ITM system

In this paper we introduce an information theoretic frame work to model existing flash crowd attacks in ITM system monitors. In the flash crowd attack the attacker sends the large volume of unwanted traffic to the targeted monitor for this he uses the botnet. Based on the Information-theoretic model we propose an effective approach to detect flash crowd attacks using Honey pots.

2 Related Work

Probing traffic based Localization attack [7] [8] in which an attacker sends high rate short length port scan messages to the targeted network to compromise the monitor locations in ITM system. Then, attacker queries the data center to determine whether a short spike of high-rate traffic appears in the queried time-series data, for confirmation of the attack.

A steganographic localization attack [9] an attacker launches a stream of low-rate port-scan probing traffic which is marginally modulated by a secret Pseudonoise (PN) code. While the low-rate property prevents the exhibition of obvious regularity of the published traffic data at the data center, based on the carefully synchronized PN code, the attacker can still accurately identify the PN-code-modulated traffic in the retrieved published traffic data from the data center. Thereby, the existence of monitors in the targeted network can be compromised. To this end, the PN-code-based steganographic attack presented in our paper can be understood as a covert channel problem [10], because the attack traffic encoded by a signal blends into the background traffic and is only recognizable by the attacker which knows the secret pattern of the PN code.

In [1] introduced the information theoretic framework to evaluate the effectiveness of the localization attacks by using the minimum time length required by an attacker to achieve a predefined detection rate as the metric. But this frame work is defined in specific to the localization attacks only; they are not given any solution for other DDoS attacks. The frame work allows the ITMs which are registered within the data center given, and the access is restricted to that private region only. But public access of the ITMs and data center allows more scope to provide security against different attacks.

3 Proposed Work

In [1] the authors define a model in which the ITMs in the networks sends the traffic logs periodically to the data center and the data center collects the traffic logs and publishes the reports to ITM system users which are registered, that means it creates the private environment or region. In the private region the scope for DDoS attacks are very less, and they are restricted this model only for Localization attacks. In this section we have defined a model which will provide the following extensions.

Public accessing: Public accessing of the data center increases the network usage and provides better communication with the outside world rather than private environment. In this any user from outside the private region can get the communication with the private network, if the user is genuine he can get the status of the monitor before sending the data to internal monitors, to avoid the attacks. If the user is an attacker, then this status information can be misused to perform the attacks on the monitor. The data center sends the status information to any users (public or private) based on the request query, but the private (internal) users can get the highest priority.

Usage of Botnets for Flash crowd Attack: A denial-of-service (DoS) attack is an explicit attempt by attackers to prevent an information service's legitimate users from using that service. In a DDoS attack, these attempts come from a large number of distributed hosts that coordinate to flood the victim with an abundance of attack packets simultaneously. The attacker may use the botnets [11], [12] and other alternatives to launch the attack.

3.1 Flash Crowd

Launching a Flash crowd attack: Once the DDoS network has been set up and the infrastructure for communication between the agents and the handlers established, all that an attacker needs to do is to issue commands to the agents to start sending packets to the victim host. The agents try to send unusual data packets in order to maximize the possibility of causing disruption at the victim and the intermediate nodes.

Concept of flash crowds and its damages

A flash crowd is a surge in traffic to a particular Web site that causes the site to be virtually unreachable. Flash crowd this basically a sudden increase in the overall traffic to any specific web page or a website on the internet and you do any sudden occurrence of any event that triggers that particular massive traffic and people accessing that particular web page or web site[23]. Sudden surges of traffic, also known as flash crowds, present a significant problem to Web sites Current systems deal with flash crowds by offloading a portion of the Web site load to a content delivery network. However, it is hard to determine in advance when the offloading should start. As a result, most systems react only after detecting a flash crowd during its initial phase.

The continuous growth in the number of Internet users often results in popular Web sites becoming overloaded. In these cases, the number of requests received by the Web site grows rapidly, causing the server's capacity to soon become exceeded. Overloaded Web sites service as many requests as possible (usually with very low performance), and simply drop the remaining ones. Such events are often referred to as Slashdot effects, hot spots, or flash crowds.

A widely adopted solution consists of offloading a portion of the Web site load to some distributed infrastructure such as a content delivery network (CDN) [24]. In that case, the Web site replicates its content to a number of CDN servers, and starts redirecting the clients to these servers when the load becomes too high. This effectively increases the client-serving capacity of the Web site by that of the CDN, which enables the Web site to service all the clients with a good performance. The problem with offloading is that the Web site has to decide when exactly the offloading should begin. If it starts too early, then the CDN capacity will be utilized unnecessarily, resulting in higher maintenance costs of the Web site. On the other hand, if the Web site begins offloading too late, then some of the clients shall still experience the reduced Web site performance during the initial flash crowd phase. Worse yet, replicating Web site content to a CDN during a flash crowd puts additional stress on the Web site, thus reducing its performance even further and causing the replication itself to last much longer compared to replication done in advance. Ideally, a Web site would start offloading right before a flash crowd begins such that the CDN capacity is utilized optimally.

3.2 BOTs

The attacker uses the bots to generate huge number of packets to attack the victim by sending these huge packets as large traffic to generate flash crowd attack. The attacker first identifies the compromised servers in terms of security and controls the systems which are under the control of these compromised servers. The compromised systems under the servers known as the bot. Normally the attacker communicates the bots by using the Internet relay chat (IRC)[14] .IRC is the public network there the users can enter and communicate each other or with the groups openly.

The attacker launches the DDoS attacks through the bots by sending the commands using these IRC network. The DDoS attacks can be blocked or the detection can not be possible, but by identifying the IRC server one can block the packets to the victim.

3.3 Botnet

Internet is the globally established network where different users or systems exist and it provides the better scalability and openness to the users in terms of services. The open accessing of the internet allows different threats and one of the major threats is from large number of compromised computers also called as bots or Zombies and the group of these computers called as Botnet. By using these botnets the attackers performs the attacks on the victims by simply sitting in house, from offices or organizations and any private or public network around the world. Every botnet or the group of compromised bots is controlled by a master commonly called as attacker or hacker. These botnets conducts various attacks which includes DDoS, e-mail spamming, keylogging, click fraud, and spreading any malware to the victim. Compared to any attack the botnets consists of pool of compromised bots and these are capable to conduct or damage the victim tremendously with collective power or capacity than the individual attacker. Example for these type of attacks are flooding, flash crowd and ports scan attacks there the attacker uses the botnet power to generate the large number of traffic to blocks the victim resources.

3.4 IRC-Based Command and Control

A bot generally communicates with a controller to receive commands send by the attacker or send back information any to the attacker. It establishes the communication channel directly to the controller for transactions. The problem is that this connection could compromise the controller’s location. Instead, the bot controller can use a proxy such as public message drop point (e.g., a well- known message board). The websites and other drop points can introduce significant communication latency; a more active approach is attractive. The commonly used communication channel is through the IRC.[14]

IRC provides a common protocol that is widely used across the Internet and has simple textbased command syntax. There is also a large number of existing IRC networks that can be used as public exchange points. In addition, most IRC networks lack any strong authentication, and a number of tools to provide anonymity on IRC networks are available. Thus, IRC provides a simple, low-latency, widely available, and anonymous command and control channel for botnet communication. An IRC network is composed of one or more IRC servers as depicted in figure 1.

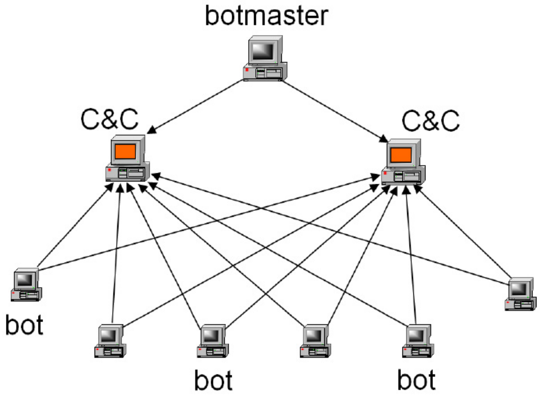


Fig. 1. Compromised computers. In a distributed denial- of-service attack (DDoS), these computers serve three major roles: master controller, command and control server, and bot.

In the botnet communication every bot connects to a public IRC network or a hidden IRC server on the compromised system. The bot receives the commands directly from the IRC controller by entering the named channel.

3.5 Honeypots

Honeypot is an effective and efficient tool for identifying and understanding intruder's toolkits, tactics, and motivations. A honeypot suspects every packet transmitted to/from it, giving it the ability to collect highly concentrated and less noisy datasets for network attack analysis. Honeypots are decoy computer resources set up for the purpose of monitoring and logging the activities of entities that probe, attack or compromise them. Activities on honeypots can be considered suspicious by definition, as there is no point for benign users to interact with these systems. Honeypots come in many shapes and sizes; examples include dummy items in a database, low-interaction network components like preconfigured traffic sinks, or full-interaction hosts with real operating systems and services.

Honeypots excel at detection, addressing many of the problems of traditional detection. Honeypots reduce false positives by capturing small data sets of high value, capture unknown attacks such as new exploits or polymorphic shell-code, and work in encrypted and IPv6 environments. In general, low- interaction honeypots make the best solutions for detection. They are easier to deploy and maintain.

4 Proposed Model

In this paper we divided the entire model into two regions namely private region and public region. The Internet Threat Monitors (ITM) are distributed across the Internet and each monitor records the traffic addressed to range of IP addresses and send the traffic logs periodically to the data center. The data center then analyzes the traffic logs collected from the monitors and publishes the reports to ITM system users. The collection of monitors under the data center forms the private region because the ITMs are registered before sending the logs to the data center. Any user can get the reports of the requested ITM by sending the query request to the data center and the data center is answerable to all the ITMs which are registered.

The public region of our model specifies the unregistered users of the data center who does not have any permission to access the data center, but they can get the traffic reports related to any ITM by sending the query request to the data center. The data center scope is extended to the public domain but it can only give the traffic reports to the public users. Allowing the public users or network accessing to the data center and monitors, causes decrease in the performance because of the overload of the data center. These can be balanced by introducing the priorities to the users; the internal or private region users have the highest priority than the public users .This priorities does not disturb the existing scenario but this can enhance the service to the public domain ,this will not be a over burden to the data center.

In this section we are constructing the botnet as the public user network without having any registration with data center and performing the flooding attack on the ITM which is local to the data center.

Generation of flash crowd attack with Botnet

A DDoS (Flash crowd) attack mechanism typically includes a network of several compromised computers [15]. These compromised computers serve three major role - master controller, command and control (C&C) server, and bot. An attacker prepares a DDoS attack by exploiting vulnerabilities in one computer system and making it the DDoS “master controller.” From here, the attacker identifies and communicates with other compromised systems. A C&C server is a compromised host with a special program running on it, this server distributes instructions from the attacker to the rest of the bots, which form a botnet[11]. (A bot is a compromised host that runs a special program.) Each C&C server is capable of controlling multiple bots, each of which is responsible for generating a stream of packets to the intended victim. Often, the bots employed to send the flood of requests are infected with a virus that lets attackers use them anonymously.

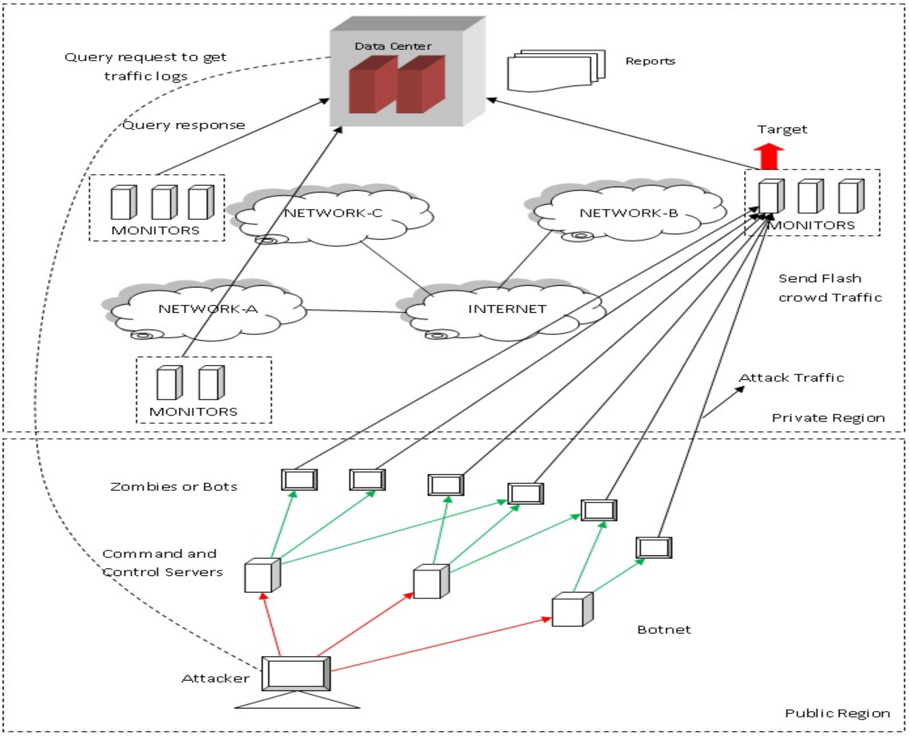


Fig. 2. Workflow of flash crowd attacks using botne

A Flash crowd attack happens in several phases

- *Discover vulnerable hosts*. To launch a DDoS attack, attackers first build a network of computers that they can use to produce the volume of traffic needed to deny services to legitimate users. To create this network, they first scan and identify vulnerable sites or hosts. Vulnerable hosts are usually those that run either no antivirus software or an out-of-date version, or those that aren't properly patched. Attackers use these compromised hosts for further scanning and compromises.
- *Establish a botnet*. After gaining access, attacker must then install attack tools on the compromised hosts to form a botnet.
- *Launch an attack*. In the next phase, attackers send commands to C&C servers for their bots to attack by sending hundreds of thousands of requests to the target simultaneously.
- *Flood a target*. In the final phase, monitor receives a flood of requests to the point where they can't operate effectively.

5 Prevention

Preventive mechanisms attempt either to reduce the possibility of DDoS attacks or enable potential victims to endure the attack without denying services to legitimate users.

- *System security mechanisms* increase a host's overall security posture and prevent it from becoming part of a botnet or a DDoS victim. Examples of system security mechanisms include reliable firewall filtering, proper system configuration, effective vulnerability management, timely patch installation, robust antivirus programs, controlled and monitored system access, and solid instruction detection.

- *Resource multiplication mechanisms* provide an abundance of resources to counter DDoS threats, such as increasing the capacity of network bandwidth, routers, firewalls, and servers. Additional examples include deploying information services at diverse network locations and establishing clusters of servers with load-balancing capabilities. Resource multiplication essentially raises the bar on how many bots must participate in an attack to be effective. While not providing perfect protection, this last approach has often proved sufficient for small- to mid-range DDoS attacks.

Preventing Flash crowd Attacks

In this section we introduce a general methodology to prevent flash crowd attacks. It is based on the following line of reasoning:

1. To mount a successful Flash crowd attack, a large number of compromised machines are necessary.
2. To coordinate a large number of machines, the attacker needs a remote control mechanism.
3. If the remote control mechanism is disabled, the Flash crowd attack is prevented.

Our methodology to mitigate flash crowd attacks aims at manipulating the root-cause of the attacks, i.e., influencing the remote control network. Our approach is based on three steps:

1. Infiltrating the remote control network.
2. Analyzing the network in detail.
3. Shutting down the remote control network.

In the first step, we have to find a way to smuggle an agent into the control network. In this context, the term agent describes a general procedure to mask as a valid member of the control network. This agent must thus be customized to the type of network we want to plant it in. The level of adaptation to a real member of the network depends on the target we want to infiltrate. For instance, to infiltrate a botnet we would try to simulate a valid bot, maybe even emulating some bot commands.

Once we are able to sneak an agent into the remote control network, it enables us to perform the second step, i.e., to observe the network in detail. So we can start to monitor all activity and analyze all information we have collected.

In the last step, we use the collected information to shut down the remote control network. Once this is done, we have deprived the attacker's control over the other machines and thus efficiently stopped the threat of a flash crowd attack with this network. Again, the particular way in which the network is shut down depends on the type of network.

6 Detection of Flash Crowd Attacks

In this section we present efficient way of detecting the attacks on the ITMs in the given information theoretic frame work. We divide the attack detection process into three phases, Firstly the primary detection of DDoS attacks [20] on the ITMs and the later is the detection of flash crowd attacks on the ITMs.

In the primary detection phases the system detects the attacks based on traffic information aggregated from all monitors in the ITM system. If the overall traffic rate (e.g., volume in a given time interval) exceeds a predetermined threshold, the defender issues an alarm. The threshold value can be maintained either at data center or the individual ITMs based on the type of schemes used [1] in the network. In the primary detection phase the system detects some attack was happened in the network. If the detection scheme is centralized, then whenever the aggregate traffic exceeds the threshold maintained at the data center then the data center finds the attack and that attacked monitor can be identified by verifying the individual traffic logs of each ITM from the report. Otherwise if the detection strategy is distributed then each monitor maintained an individual threshold and checked the aggregate traffic regularly. If the traffic exceeds the threshold then it find the attack was happened and sends the status as attacked to the data center. After getting the attacked status from the ITM the data center blocks the corresponding ITM and displays the status of the ITM as blocked in the status reports, which will avoids the further traffic to or from the attacked ITM with the rest of the networks.

The second stage of detection specifies the detection of the flash crowd attacks. Detecting a flash crowd forces the origin server to adjust its operation such that it can continue servicing clients. This typically requires using some distributed infrastructure of *mirror servers*, which provide the origin server with additional hosting capacity. Mirror servers can originate either from a CDN or from a community of contributed servers. The origin server typically exploits the capacity of mirror servers by using them to service some fraction of client requests. This section surveys various techniques for flash crowd detection and handling. Before starting to adapt its operation, the origin server needs to actually detect a flash crowd. The simplest technique is to monitor the request rate, and initiate adaptation once that rate exceeds a certain threshold. The intuition behind this technique is that the high request rate can be treated as the indication of an upcoming flash crowd[24]. In practice, however, using a single adaptation threshold is not enough, as request rates might oscillate around it, thereby causing frequent and unnecessary system adaptations. To address the oscillation problem, several algorithms use a simple watermarking technique. In that case, the origin server defines two watermarks, high and low, and determines its state as follows:

We say that a server enters the overloaded state, when the request rate reaches the high watermark. The server remains in overloaded state until the request rate drops below the low Watermark. When the request rate reaches the low watermark, the server returns to its normal operation mode.

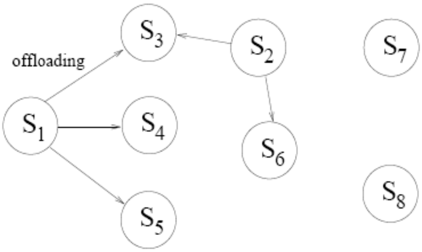


Fig. 3. Dot Slash architecture. S1 and S2 are origin servers. S3, S4, S5 and S6 are mirror servers.

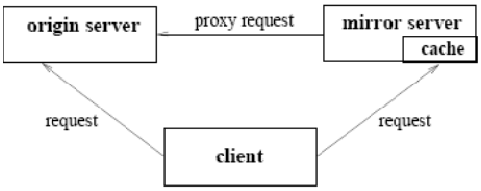


Fig. 4. Dot Slash HTTP redirection

The third stage of detection specifies the detection of the flash crowd attacks. Once the attack is onformed then the data center identifies the attacked monitor and the traffic logs will be handover to the flash crowd detection phase. In this paper the flash crowds are generated using botnet, so botnet tracking is required to detect and block the flooding attacks on the attacked ITM.

In this section we define the approaches for detecting the botnet. Once the botnet is successfully identified and blocked then automatically the flash crowd attacks can be avoided. In this connection the honeypots play the major role to block the botnet by identifying the command and control through the IRC server.

6.1 BOTNET Detection

Botnets are a very real and quickly evolving problem that is still not well understood. In this paper, we outline the problem and investigate methods of stopping bots. We identify three approaches for handling botnets:

(1) Prevent systems from being infected,

(2) Directly detect command and control communication among bots and between bots and controllers, and,

(3) Detect the secondary features of a bot infection such as propagation or attacks.

The first approach is to prevent the system from the attack, these can be done by using by introducing the anti-virus software, firewall or any security measures in the system.

The detection of command and control defines the second approach. The controlling of botnets are done in general with IRC and detection of IRC can be done by monitoring the TCP port 6667 which is used for IRC traffic. One could also look for non-human behavioural characteristics in traffic, or even build IRC server scanners to identify potential botnets.

The third approach used to detect the botnet is purely depends on the identification of secondary features of bot infection or attack behaviour. Finding the command and control directly is not possible in this approach this can be done based on the correlation of data from different sources to locate bots.

In this paper we explore the second and third approach for stopping botnets. The problem with the first approach is that preventing all systems on the Internet from being infected is nearly an impossible challenge. As a result, there will be large pools of vulnerable systems connected to the Internet for many years to come.

6.2 Detecting Command and Control

To avoid the damage of bots, we identified two approaches for detecting botnets: detect the command and control communication, or detect the secondary features of a bot infection.

IRC-based Botnet Detection

Today, most known bots use IRC as a communication protocol, and there are several characteristics of IRC that can be leveraged to detect bots. One of the simplest methods of detecting IRC- based botnets is to offramp traffic from a live network on known IRC ports (e.g., TCP port 6667) and then inspects the payloads for strings that match known botnet commands. Unfortunately, botnets can run on non-standard ports. Another method is to look for behavioral characteristics of bots. One study found that bots on IRC were idle most of the time and would respond faster than a human upon receiving a command. The system they designed looked for these characteristics in Netflow traffic and attempted to tag certain connections as potential bots [15].

The idle IRC activity was successfully detected using this method but it is unable to findhigh false positive rate, for this honeypots is used to minimize the false positives.

One attack pool set up a vulnerable system and waited for it to be infected with a bot. They then located outgoing connections to IRC networks and used their own bot to connect back and profile the IRC server [16].

Honeypots are used to connect the bots directly rather than connecting IRC server and these honeypot checks the characteristics of command and control in outgoing connections. We identified all successful outgoing TCP connections and verified that they were all directly related to command and control activity by checking the payloads. There were a wide range of interesting behaviors, including connections from the bot to search engines to locate and use bandwidth testers, downloading posts from popular message boards to get server addresses, and the transmission of comprehensive host profiles to other servers.

These profiles consists of detailed information about the operating system, host bandwidth, users, passwords of the users, file shares, filenames and permissions for all files in the system, and a number of other minute details about the infected host.

We then analyzed all successful outgoing connections for specific characteristics that could be used to identify botnet command and control traffic.

6.3 Collecting Malware with Honeypots

A honeypot is a network resource (computers, routers, switches, etc.) deployed to be probed, attacked, and compromised. A honeynet is a network of honeypots. Honeypot is a special software which periodically collects data about the system behaviour and provides automatic post-incident forensic analysis. The collected data enables us to determine the necessary information about an existing botnet.

The honeypots collets the data or attack traffic either from the data center or ITMs based on the detection scheme. Two types of detection schemes are defined in this paper based on the position of the honeypots.

Centralized scheme: In this approach only one honeypot is used for the detection and it is placed at the data center. Once the data center identifies the attacked ITM, then the traffic logs of the attacked ITM are send to the centralized honeypot to find the botnet.

Distributed Scheme: In the distributed approach the honeypots are placed at each ITM of the data center. Whenever the data center identifies the attacked monitor either by using centralized or distributed threshold detection approaches, then the attack traffic can be handover to the attached honeypot of that attacked ITM. The honeypot then identifies the botnet which causes the flash crowd attack.

The centralized scheme is more economical when compared to the distributed scheme because it uses only one honeypot at the data center instead of using individual honeypots for each ITM.But the efficiency of the system depends on the number of honeypots used in the network, If honeypots are more in the network then the detection of botnets is very simple and easy. Whenever more than one ITMs are attacked in the network, then the centralized scheme is less efficient than the distributed scheme.

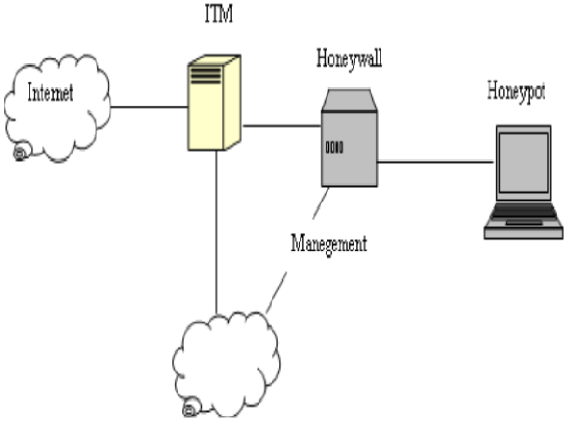


Fig. 5. Set up for tracking botnet using Honeypot

GenII Honetpot is a windows honeypoy and used to collect the necessary information about the attack. The Windows honeypot runs an unwatched version of Windows 2000 or Windows XP. This system is thus very vulnerable to perform the attacks. The average expected lifespan of the honeypot is less than ten minutes. The shortest compromise time was only a few seconds: Once we plugged the network cable in, a bot compromised the machine and installed itself on the machine.

As explained in the previous section, a bot tries to connect to the C&C server to obtain further commands once it successfully attacked the honeypot. This is where the Honeywall comes into play. The Honeywall is a transparent bridge that enables the two tasks Data Control and Data Capture. Due to the Data Control facilities, it is possible to control the outgoing traffic. Using available tools for Data Control we can replace all suspicious in- and outgoing messages. A message is suspicious if it contains typical IRC messages for command and control, for example “TOPIC”, “PRIVMSG”, or “NOTICE”. Thus we are able to reduce the bot from accepting valid commands from the master channel. It can therefore cause no harm to others and therefore we have caught a bot inside our Honeynet. In addition with the detection, we can also derive all necessary sensitive information for a botnet from the data we have obtained up to that point in time: The Data Capture capability of the Honeywall allows us to determine the DNS/IP address of the bot which wants to connect the IRC.

In addition, we can obtain from the Data Capture logs the nickname, the indent information, the server’s password, channel name, and the channel password as well. So we have collected all necessary information about the attack and the honeypot can catch further malware. Since we do not care about the captured malware for now, we rebuild the honeypot every 24 hours to have a “clean” system every day. This 10 has proven to be a good time span since after this amount of time the honeypot tends to become unstable due to installed malware.

7 Conclusion and Future Work

The approach integrates active real time flash crowd attack flow identification from botnet with determining required number of honeypots. The honeypot controller has been modeled at Data center or ITMs to trigger honeypot generation in response to suspected attacks and route the attack traffic to honeypots. The performance of the proposed scheme is independent of attack traffic due to presence of honeypots at data center or ITMs. It gives stable network functionality even in the presence of high attack load.

Some of the avenues for further extensions are with larger and heterogeneous networks. Back tracking can be applied on attack flows to reach the attack source. Both of them hold promise for evaluating and improving our DDoS detection and defense method and data center information protection. The data center load can be still minimized by used some distributed load sharing algorithms.

References

[1] Yu, W., Zhang, N., Fu, X., Bettati, R., Zhao, W.: Localization attacks to internet threat monitors: Modeling and counter measures. IEEE Transactions on Computers 59(12) (December 2010)

[2] Mirkovic, J., Reiher, P.: A Taxonomy of DDOS Attack and DDOS Defense Mechanisms. ACM SIGCOMM Computer Comm. Rev. 34(2), 39–53 (2004)

[3] SANS, Internet Storm Center (2010), `http://isc.sans.org/`

[4] Moore, D., Voelker, G.M., Savage, S.: Inferring Internet Deny-of-Service Activity. In: Proc. 10th USNIX Security Symp., SEC (August 2001)

[5] Yegneswaran, V., Barford, P., Jha, S.: Global Intrusion Detection in the Domino Overlay System. In: Proc. 11th IEEE Network and Distributed System Security Symp., NDSS (February 2004)

[6] Bailey, M., Cooke, E., Jahanian, F., Nazario, J., Watson, D.: The Internet Motion Sensor: A Distributed Blackhole Monitoring System. In: Proc. 12th Ann. Network and Distributed System Security Symp., NDSS (February 2005)

[7] Bethencourt, J., Frankin, J., Vernon, M.: Mapping Internet Sensors with Probe Response Attacks. In: Proc. 14th USNIX Security Symp., SEC (July/August 2005)

[8] Shinoda, Y., Ikai, K., Itoh, M.: Vulnerabilities of Passive Internet Threat Monitors. In: Proc. 14th USNIX Security Symp., SEC (July/August 2005)

[9] Wang, X., Yu, W., Fu, X., Xuan, D., Zhao, W.: Iloc: An Invisible Localization Attack to Internet Threat Monitoring Systems. In: Proc. IEEE INFOCOM, Mini-Conf. (April 2008)

[10] Cabuk, S., Brodley, C., Shields, C.: Ip Covert Timing Channels:Design and Detection. In: Proc.2004 ACM Conf. Computer and Comm. Security, CCS (October 2004)

[11] Cooke, E., Jahanian, F., McPherson, D.: The Zombie Roundup: Understanding, Detecting, and Disrupting Botnets. In: Proc. Steps to Reducing Unwanted Traffic on the Internet Workshop, SRUTI (July 2005)

[12] Freiling, F.C., Holz, T., Wicherski, G.: Botnet Tracking: Exploring a Root-Cause Methodology to Prevent Distributed Denial-of- Service Attacks. In: Proc. 10th European Symp. Research in Computer Security, ESORICS (September 2005)

[13] The mstream distributed denial of service attack tool, `http://staff.washington.edu/dittrich/misc/mstream.analysis.txt`

[14] Oikarinen, J., Reed, D.: RFC 1459: Internet Relay Chat Protocol (1993)

[15] Racine, S.: Analysis of Internet Relay Chat Usage by DDoS Zombies. Master's thesis, Swiss Federal Institute of Technology Zurich (April 2004)

[16] The Honeynet Project. Know your enemy: Tracking botnets (March 2005), `http://www.honeynet.org/papers/bots/`

[17] CAIDA, Telescope Analysis (2010), `http://www.caida.org/analysis/security/telescope`

[18] Provos, N.: A Virtual Honeypot Framework. In: Proc. 12th USENIX Security Symp., SEC (August 2004)

[19] Sachdeva, M., Kumar, K., Singh, G., Singh, K.: Performance Analysis of Web Service under DDoS Attacks. In: Proc. IACC 2009, pp. 1002–1007 (March 2009)

[20] Yokota, K., Takahashi, T., Asaka, T.: A Load Reduction System to Mitigate Flash Crowds on Web Server, pp. 503-508 (2011)

[21] Sachdeva, M., Kumar, K., Singh, G., Singh, K.: Performance Analysis of Web Service under DDoS Attacks. In: Proc. IACC 2009, pp. 1002–1007 (March 2009)

[22] Wang, J., Phan, R.C.W., Whitley, J.N., Parish, D.J.: Augmented Attack Tree Modeling of Distributed Denial of Services and Tree Based Attack Detection Method. In: Proc. CIT 2010, pp. 1009–1014 (June 2010)

[23] Ari, I., Ethan, B.H., Scott, L.M., Darrell, A.B., Long, D.E.: Managing Flash Crowds on the Internet. In: Proceedings of the 11th IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer Telecommunications Systems, MASCOTS 2003. IEEE (2003) 1526-7539/03 $ 17.00 © 2003 IEEE

[24] Chi, C.-H., Xu, S., Li, F., Lam, K.Y.: Selection Policy of Rescue Servers Based on Workload Characterization of Flash Crowd. In: 2010 Sixth International Conference on Semantics, Knowledge and Grids (2010)

[25] Yokota, K., Takahashi, T., Asaka, T.: A Load Reduction System to Mitigate Flash Crowds on Web Server. In: 2011 Tenth International Symposium on Autonomous Decentralized System (2011)

Hierarchical Directed Acyclic Graph (HDAG) Based Preprocessing Technique for Session Construction

S. Chitra[1] and B. Kalpana[2]

[1] Assistant Professor, Department of Computer Science,
Government Arts College, Coimbatore – 641 018, Tamilnadu, India
`chitra.sivakumar@ymail.com`
[2] Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore – 641 043, Tamilnadu, India
`kalpanabsekar@yahoo.com`

Abstract. Web access log analysis is to examine the patterns of web site usage and the features of user's behavior. Preprocessing of the log data is very essential for efficient web usage mining as the normal log data is very noisy. Session construction is very vital step in the preprocessing phase and recently various real world problems can be modeled as traversals on graph and mining from these traversals provides effective results. On the other hand, the traversals on unweighted graph have been taken into consideration in existing works. This paper oversimplifies this to the case where vertices of graph are given weights to reflect their significance. Patterns are closed frequent Directed Acyclic Graphs with page browsing time. The proposed method constructs sessions using an efficient Directed Acyclic Graph approach which contains pages with calculated weights. Hierarchical Directed Acyclic Graph (HDAG) Kernel approach is used for session construction. The HDAG directly accepts several levels of both chunks and their relations, and then efficiently computes the weighed sum of the number of common attribute sequences of the HDAGs. This will help site administrators to find the interesting pages for users and to redesign their web pages. After weighting each page according to browsing time a DAG structure is constructed for each user session.

Keywords: Web Usage Mining, Session Construction, Directed Acyclic Graph (DAG), Preprocessing, Robots Cleaning.

1 Introduction

World Wide Web is a playing a vital role in the present scenario. The process of efficiently utilizing the data in the internet has become an active area of research. Web usage mining is an active, technique used in this field of research. It is also called web log mining in which data mining techniques are applied to web access log. A web access log is a time series record of user's requests each of which is sent to a web server whenever a user sent a request. Due to different server setting parameters, there are many types of web logs, but typically the log files share the same basic information such as client IP address, request time, requested URL, HTTP status code, referrer etc.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 611–621.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Web usage mining extracts regularities of user access behavior as patterns, which are defined by combinations, orders or structures of the pages accessed by the internet. Web usage mining consists of three main steps:

- Data Preprocessing
- Knowledge Extraction
- Analysis of Extracted Results

Preprocessing is a significant step since the Web architecture is very complex in nature and 80% of the mining process is done at this phase.

Administrators of the web sites have to know about the users' background and their needs. For this statistical analysis such as Google Analytics are used to analyze the logs in terms of page views, page exit ratio, visit duration etc. With the help of this analysis administrators can know about frequently accessed page, average view time and so on. But there are few drawbacks in statistical analysis. It gives low level error report on unauthorized entry points, invalid urls are not found properly etc. Web usage mining enables administrators to provide complete analysis than statistical methods. It extracts a lot of patterns for administrators to analyze. This paper provides a method which analyses log files and extracts access patterns containing browsing time of each page using graphs [16].

Graph and traversal are extensively used to model a number of classes of real world problems. For example, the structure of Web site can be modeled as a graph in which the vertices represent Web pages, and the edges correspond to hyperlinks between the pages [7]. Mining using graphs turns out to be a center of interest. Traversals on the graphs are the models of User navigations on the Web site [14]. Once a graph and its traversals are specified, important information can be discovered. Frequent substructure pattern mining is an emerging data mining problem with many scientific and commercial applications [15]. This paper provides a new version to the previous works by considering weights attached to the vertices of graph. Such vertex weight may reflect the importance of vertex. For example, each Web page may have different consequence which reflects the value of its contents.

There are three phases in this method. First one is preprocessing phase which includes data cleaning, user identification, session identification, DAG construction. The second phase is pattern extraction phase using clustering and the third phase is pattern analysis phase. This paper discusses elaborately about the first phase and briefs about other phases.

A new graph-based approach, called Hierarchical Directed Acyclic Graph kernels (HDAG) [17] is used in this approach for the preprocessing step. The method is observed to be effective in session construction and includes characteristics of tree and sequence kernels [18]. The HDAG is a hierarchized graph-in-graph structure. This framework defines a kernel function between input objects by employing convolution "sub-kernels" that are the kernels for the decompositions (parts) of the objects.

2 Related Works

Various commercially available web server log analysis tools are not designed for high traffic web servers and provide less relationship analysis of data among accessed files which is essential to fully utilize the data gathered in the server logs [3]. The statistical analysis introduces a set of parameters to describe user's access behaviors. With those parameters it becomes easy for administrators to define concrete goals for organizing their web sites and improve the sites according to the goals. But the drawback in this analysis is that the results are independent from page to page. Since user's behavior is expected to be different dependent on length of browsing time, the calculation of accurate browsing time is more important [5].

A labeled graph is a tuple $G = (V, E, \varphi)$, where V is the set of vertices, E is the set of edges and $\varphi : V \rightarrow L$ is a labeling function with L a finite set of labels [9]. For an edge $(u, v) \in E, u$ is the parent of v and v is the child of u. If there is a set of vertices $\{u_1, \dots, u_n\} \subseteq$ V such that $(u_1, u_2) \in E, \dots, (u_{n-1}, u_n) \in E$, $\{u_1, \dots, u_n\}$ is called a path, u_1 is an ancestor of un and u_n is a descendant of u_1 . There is a cycle in the graph if a path can be found from a vertex to itself. An edge $(u, v) \in E$ of the graph is said to be a transitive edge if besides the edge (u, v), there also exists another path from u to v in G . A labeled DAG is a labeled graph without cycles. Let $D = \{D_1, \dots, D_n\}$ be a set of labeled DAGs and $\epsilon \geq 0$ be an absolute frequency threshold. DIGDAG algorithm specifies that a DAG P is a frequent embedded sub-DAG of D if it is embedded in at least ϵ DAGs of D.

Prediction of users interest is the most important aspect of web usage mining. For this frequency and order of visited pages are considered. But Time spent on web pages is more important factor which is estimated from the log information and it is used as an indicator in other fields such as information retrieval, human-computer interaction (HCI) and E-Learning [2].

Duration time is the time that a user spends on reading a page in a session. Let P_i and P_{i+1} are two adjacent pages in a session. The timestamp field of P_i is Ti, and of P_{i+1} is Ti+1. Suppose T_3 is the loading time of P_i , and T_4 is the loading time ancillary files. By subtracting the time required for loading P_i and the ancillary files from the time difference between the requests of P_i and that of P_{i+1} , the duration time of P_i can be calculated [4].

The browsing time of an accessed page equals the difference between the access time of the next and present page. But with a more careful analysis, this difference includes not only user's browsing time, but also the time consumed by transferring the data over internet, launching the applications to play the audio or video files on the web page and so on. The user's real browsing time is difficult to be determined; it depends on the content of the page, the real-time network transfer rate, user's actions and computer's specifications and so on [13].

All of these works attempt mainly to find the exact browsing time of users so that web administrators can understand the interest of their users in web pages. In the proposed method a more accurate browsing time is found and creation of sessions as graphs depending on the time accessed.

3 Preprocessing

The quality of session construction significantly affects the whole performance of a web usage mining system. To improve the quality log data should be reliable. Preprocessing is a vital phase before mining to select the reliable data. Data Cleaning, user identification, sessions construction are the steps in preprocessing.

3.1 Data Cleaning

Data Cleaning enables to filter out useless data which reduce the log file size to use less storage space and to facilitate upcoming tasks [8]. It is the first step in data preprocessing. The log format used in this method is Extended Common Log Format with the fields as follows: “ipaddress, username, password, date/timestamp, url, version, status-code, bytes-sent, referrer-url, user-agent”.

If a user needs a particular page from server entries like gif, JPEG, etc., are also downloaded which are not helpful for further investigation are eliminated. The records with failed status code are also eliminated from logs. Automated programs like web robots, spiders and crawlers are also to be eradicated from log files. Thus removal process includes elimination of irrelevant records as follows:

- If the status code of all record is fewer than 200 and better than 299 then those records are eradicated.
- The cs-stem-url field is verified for its extension filename. If the filename has gif, jpg, JPEG, CSS, and so on they are eradicated.
- The records which request robots.txt are eradicated and if the time taken is incredibly little like less than 2 seconds are considered as automated programs traversal and they are also eradicated [8].
- All the records which have the name “robots.txt” in the requested resource name (URL) are recognized and straightly eradicated.

3.2 User Identification

In this step users are identified from log files. Sites which need registration stores the user data in log records. But those sites are few and often neglected by users. IPaddress, referrer URL and user agent in the log record is considered for this task. Unique users are identified as follows:

- If two records has dissimilar IP address they are differentiated as two different users else if both IP address are similar then User agent field is verified.
- If the browser and operating system information in user agent field is dissimilar in two records then they are recognized as different users else if both are identical then referrer url field is checked.
- If URL in the referrer URL field in present record is not accessed before or if url field is blank then it is considered as a new user.

3.3 Session Identification

A user session is defined as a sequence of requests made by a single user over a certain navigation period and a user may have a single or multiple sessions during a period of time. The objective of session identification is to segregate the page accesses of each user into individual sessions. Reconstruction of precise user sessions from server access logs is a difficult task because the access log protocol (HTTP protocol) is status less and connectionless. There are two simple methods for session identification. One is based on total session time and other based on single page stay time. The set of pages visited by a specific user at a specific time is called page viewing time. It varies from 25.5 minutes [12] to 24 hours [8] at the same time as default time is 30 minutes by R. Cooley [4]. The second method depends on page stay time which is calculated with the difference between two timestamps. If it goes over 10 minutes the second entry is understood as a new session. The third method based on navigation of users through web pages. But this is accomplished by using site topology which is not used in our method.

4 Session HDAG Construction

In the proposed method sessions are modeled as a graph. Graph mining extracts user access patterns as a graph structure like the web sites link structure. To make efficient analysis when users handle more pages at the same time using tab browsers graph mining gives excellent results. Vertices are represented as web pages and edges are represented as hyperlink between pages. A graph is represented as a tuple of vertices, edges which connect the vertices [13]. User navigations are given as traversals in a graph. Each traversal can be represented as a sequence of vertices, or equivalently as a sequence of edges. DAG construction phase has following tasks.

4.1 Calculation of Browsing Time

1. The first task is to calculate browsing time of each page. For this the timestamp fields of the records are considered. Real Browsing time is very difficult to calculate since it depends on network transfer rate, user's actions, and computer specifications and so on. Browsing Time and Request Time recorded in log are abbreviated as BT and RT . Browsing time BT_p of page 'p' is equal to the period of time with the time difference between the RT_p of the request which include 'p' as a reference and another RT of the request which include 'P' as a requested page. In the log record one of the fields is bytes_sent which is the size of the web page. 'c' is the data transfer rate. So the real browsing time is assumed as

$$BT_p = BT_p' – \text{bytes\_sent} / c$$

where BT_p' is the difference between reference and request page of 'p'.

4.2 Calculation of Weight of Pages

The second task in this method is to fix minimum and maximum browsing time for each page as BT_{min} and BT_{max} is used to calculate the weighing function which is to be used as a label in the graph. They are assumed by the administrators. The next step is to discretise the browsing time and given to each page as the weight which denotes the length of browsing time. Weighting function is calculated as follows

$Wt\ (p, BT_p) = 0$ when $BT_p \neq$ null and $BT_p < BT_{min}$
$Wt\ (p, BT_p) = 1$ when $BT_p \neq$ null and
$BT_{min} \leq BT_p \leq BT_{max}$
$Wt\ (p, BT_p) = 2$ when $BT_p \neq$ null and $BT_{max} < BT_p$ $Wt\ (p, BT_p) = 3$ when $BT_p =$ null

If weight is ‘0’ it is assumed as the time to browse is too short and the user simply passed the page. If weight is ‘1’, administrators conclude it is a valid browsing time and user is interested in the content of the page. If weight is ‘2’ the time is too long and it is assumed as if the user left the page and if the weight is ‘3’ the page does not exists as reference page in that session. It is assumed as the end page and the user does not move from this page.

4.3 DAG Construction

Directed Acyclic Graph (DAG) is a tuple of Vertex, Edge and a label. This type of graph doesn’t have cyclic structures. After weighting all pages based on the browsing time a DAG structure is built for each user session. Vertex is labeled by a page and its weight. Each vertex is represented by a set of page and it’s weight as (p, wt (p, BTp). Edge connects reference page to request page for each request. Edges show users page transition and only the direction is considered. If any cyclic structure exists a new vertex is created and the graph structure is converted to acyclic. In this method DAGs which give user session information for mining is constructed. The advantage over other graph methods is use of numerical values like browsing time is considered. A simplest form of the weighting function is used depending on the browsing time which is longer or shorter than the threshold. The threshold is based on the content of the page.

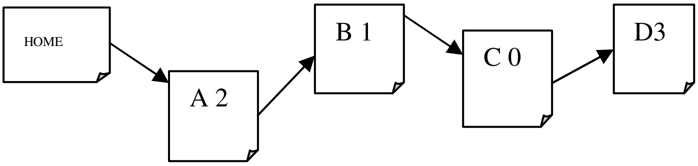


Fig. 1. Directed Acyclic Graph Construction

4.4 Proposed Hierarchical Directed Acyclic Graph (HDAG)

This paper defines HDAG as a Directed Acyclic Graph (DAG) with hierarchical structures [17]. That is, certain nodes contain DAGs within themselves.

4.4.1 Kernel Function

$x \in X$ is a composite structure and $x_1, \dots, x_D$ where $x_d \in X_d$. R is relation on the set $X_1 \times \cdots X_D \times X$ such that $R(\mathbf{x}, x$) is true if x are the parts of x . $R^{-1}(x)$ is defined as $R^{-1}(x) = \{\mathbf{x}: \mathrm{R}(\mathbf{x}, x)\}$.

Consider that $x, y \in X$, x be the parts of x with $x = x_1, \dots, x_D$, and y be the parts of y with $y = y_1, \dots, y_D$. Then, the similarity $K(x, y)$ between x and y is defined as the following generalized convolution:

$$K(x, y) = \sum_{x \in R^{-1}(x)} \sum_{y \in R^{-1}(y)} \prod_{d=1}^{D} K_d(x_d, y_d) \tag{1}$$

Convolution Kernels are abstract concepts, and that instances of them are determined by the definition of sub-kernel $K_d(x_d, y_d)$.

An explicit definition of both the Tree Kernel [18] and String Sub-sequence Kernel (SSK) $K(x, y)$ is written as:

$$K(x, y) = \langle \phi(x) \cdot \phi(y) \rangle = \sum_{i=1}^{M} \phi_i(x) \cdot \phi_i(y) \tag{2}$$

All sub-structures occurring in x and y are listed, where M represents the total number of possible sub-structures in the objects. ϕ , the feature mapping from the sample space to the feature space, is given $\phi(x) = (\phi_1(x), \dots, \phi_M(x))$.

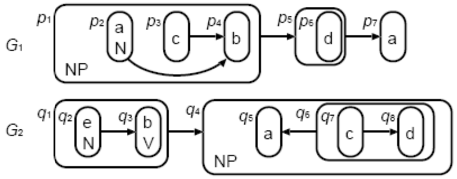


Fig. 2. Examples of HDAG Structure

There are several levels of chunks in the web usage mining such as web pages, web links, named entities and these are bound by relation structures. HDAG is designed to enable the representation of all of the structures in the web usage mining, hierarchical structures for chunks and DAG structures for the relations of chunks. This richer representation is extremely useful to enhance the performance of similarity measure between web pages, moreover, learning and clustering tasks in the application areas of web usage mining.

As shown in Figure 2, the nodes are allowed to have more than zero attributes, because nodes in texts usually have several kinds of attributes. For example, attributes include web pages, web links,etc.

The set of nodes in HDAGs G_1 and G_2 as P and Q, respectively, p and q represent nodes in the graph that are defined as $\{p|p_i \in P, i = 1, \ldots, |P|\}$ and $\{q|q_i \in Q, j = 1, \ldots, |Q|\}$ respectively. The expression $p_1 \rightarrow p_4 \rightarrow p_7$ to represent the path from p_1 to p_7 through p_4 .

"Attribute sequence" is defined as a sequence of attributes extracted from nodes included in a subpath. As a basic example of the extraction of attribute sequences from a sub-path, $q_1 \rightarrow q_3$ in figure 2 contains the four attribute sequences 'e-b', 'e-V', 'N-b', and 'N-V' which are the combinations of all attributes in q_2 and q_3 .

This framework makes similarity evaluation robust; the similar sub-structures can be computed in the value of similarity, on the contrary to exact matching that never evaluate the similar substructure. The similarity between HDAGs, which is the definition of the HDAG Kernel, follows equation (2) where input objects x and y are G_1 and G_2 respectively.

5 Pattern Extraction Phase

Once a graph and its traversals are specified, valuable information can be retrieved through graph mining. Normally they are in the form of patterns. Frequent patterns which are sub traversals occurred in a large ratio are considered for analysis. To discover DAG's i.e., sub graphs DIGDAG mining algorithm is used which derive closed frequent sets. It replaces closed frequent DAG mining problem with the problem of closed frequent item-set mining on edges with the restriction that all the labels of the vertices in a DAG must be distinct. By the reconstruction of DAG structures from the mined closed frequent edge set, closed frequent DAG's are obtained. DIGDAG extracts the embedded DAGs based on not only on parent-child relationship but also ancestor-descendant relationship of vertices. The input for DIGDAG are the user session DAG set and the minimum support $\sigma(\geq 0)$ as inputs. Access patterns are obtained as frequent DAGs.

6 Clustering Patterns

The last step is clustering of the mined patterns. The purpose of clustering is to group patterns which have similar page transitions. Each pattern is analyzed as different user behavior with browsing time. Weight of each page is not considered in clustering. The similarity of the patterns is to be estimated. Similarity of graphs is based on the labels of vertices and the edges. There are many clustering algorithms available to group the similar patterns. Administrators have to analyze the patterns respectively and it is time-consuming. They have to understand the meaning of each and every sub pattern to find out the problem of their web sites. If a content page has 0 weights then they have to redesign the page.

7 Experimental Results

To confirm the usefulness and effectiveness of the proposed methodology, an experiment is carried out with the web server log of the library of South-Central University for Nationalities. The preliminary data source of the experiment is from May 28, 2006 to June 3, 2006, which size is 129MB. Experiments were carried out on a 2.8GHz Pentium IV CPU, 512MB of main memory, Windows 2000 professional, MatLab 7.10. Table-1 is the obtained results from the experiment.

Table 1. The Processes and Results of Data Preprocessing in Web Usage Mining

Number of records in raw web log	Number of records after data cleaning	Number of users
747890	112783	55052

Table 1 show that after data cleaning, the number of log data diminished from 747890 to 112783.

Four samples from the same university are obtained to evaluate the cleaning phase. From Fig-3 it is confirmed that the unwanted and irrelevant records are cleaned.

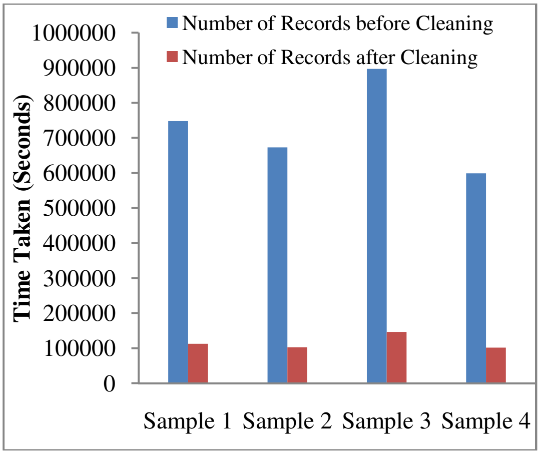


Fig. 3. Data Cleaning of Sample Records

Table 2. User Session Identification by using Directed Acyclic Graph (DAG)

Approaches	IP Address	User id	Session id	Path Completed
DAG	116.128.56.89	1	1	16-17-18-17-18-19-20
	116.128.56.89	1	2	25-26-30-35
HDAG	116.128.56.89	1	1	16-17-18-19-20
	116.128.56.89	1	2	25-26-35

Table 2 shows the comparison of the DAG and HDAG approaches. The path completed by both the approaches is given. It is observed from the table, that the only the best and the interested paths are chosen by the HDAG approach due to its hierarchical structure.

8 Conclusion

It is very difficult to extract the interesting patterns available in the web log data without preprocessing phase. Preprocessing phase helps to clean the records and discover the interesting user patterns and session construction. But understanding user's interest and their relationship in navigation is more important. In this paper, an efficient technique is proposed to analyze web logs in detail by constructing sessions as Hierarchical Directed Acyclic graphs (HDAG). The proposed method takes advantage of both statistical analysis and web usage mining. Patterns are reduced by closed frequent mining for efficient analysis. Web site administrators follow the results and improve their web sites more easily. From the experimental results it is obvious that the proposed method successfully cleans the web log data and helps in identifying the user session.

References

1. Mobasher, B.: Data Mining for Web Personalization. LCNS. Springer, Heidelberg (2007)
2. Catlegde, L., Pitkow, J.: Characterising browsing behaviours in the World Wide Web. Computer Networks and ISDN systems (1995)
3. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining World Wide Web browsing patterns. Knowledge and Information Systems (1999)
4. Cooley, R., Mobasher, B., Srivastava, J.: Web mining: Information and Pattern Discovery on the World Wide Web. In: International Conference on Tools with Artificial Intelligence, Newport Beach, pp. 558–567. IEEE (1997)
5. Mihara, K., Terabe, M., Hashimoto, K.: A Novel web usage mining method. Mining and Clustering of DAG Access Patterns Considering Page Browsing Time (2008)
6. Hofgesang, P.I.: Methodology for Preprocessing and Evaluating the Time Spent on Web Pages. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (2006)
7. Lee, S.D., Park, H.C.: Mining Weighted Frequent Patterns from Path Traversals on Weighted Graph. IJCSNS International Journal of Computer Science and Network Security 7(4) (2007)
8. Spilipoulou, M., Mobasher, B., Berendt, B.: A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis. Informs Journal on Computing Spring (2003)
9. Suresh, R.M., Padmajavalli, R.: An Overview of Data Preprocessing in Data and Web usage Mining. IEEE (2006)
10. Termier, A., Tamada, Y., Numata, K., Imoto, S., Washio, T., Higuchi, T.: DIGDAG, a first algorithm to mine closed frequent embedded sub-DAGs. In: The 5th International Workshop on Mining and Learning with Graphs, MLG 2007 (2007)
11. Wang, T., He, P.-L.: Find Duration Time Maximal Frequent Traversal Sequence on Web Sites. In: IEEE International Conference on Control and Automation (2007)

12. Li, Y., Feng, B., Mao, Q.: Research on Path Completion Technique in Web Usage Mining. In: International Symposium on Computer Science and Computational Technology. IEEE (2008)
13. Li, Y., Feng, B.: The Construction of Transactions for Web Usage Mining. In: International Conference on Computational Intelligence and Natural Computing. IEEE (2009)
14. Etminani, K., Delui, A.R., Yanehsari, N.R., Rouhani, M.: Web Usage Mining: Discovery of the Users' Navigational Patterns Using SOM. In: First International Conference on Networked Digital Technologies, pp. 224–249 (2009)
15. Nina, S.P., Rahman, M., Bhuiyan, K.I., Ahmed, K.: Pattern Discovery of Web Usage Mining. In: International Conference on Computer Technology and Development, vol. 1, pp. 499–503 (2009)
16. Lee, C.-H., Fu, Y.-H.: Web Usage Mining Based on Clustering of Browsing Features. In: Eighth International Conference on Intelligent Systems Design and Applications, vol. 1, pp. 281–286 (2008)
17. Suzuki, J., Hirao, T., Sasaki, Y., Maeda, E.: Hierarchical Directed Acyclic Graph Kernel: Methods for Structured Natural Language Data. Meeting of the Association for Computational Linguistics, pp. 32–39 (2003)
18. Collins, M., Duffy, N.: Parsing with a Single Neuron: Convolution Kernels for Natural Language Problems. Technical Report UCS-CRL-01-10, UC Santa Cruz (2001)

Extending Application of Non-verbal Communication to Effective Requirement Elicitation

Md. Rizwan Beg[1], Md. Muqeem[2], and Md. Faizan Farooqui[2]

[1] Professor & Head, Department of CS&E, Integral University, Lucknow 226026, India
rizwanbeg@gmail.com
[2] Research Scholar, Department of Computer Application, Integral University, Lucknow 226026, India
muqeem.79@gmail.com, faizan_farooqui2000@yahoo.com

Abstract. Requirements elicitation is the first stage in the process of developing a software product. The purpose of the requirements elicitation process is to build an understanding of the problem. It is fundamentally a communication process between a requirements elicitor and various Stakeholders. Interviewing stakeholder during Elicitation process is a communication intensive activity involving various techniques & tactics requiring communication skills apart from experience and knowledge. Interview involves Verbal and Non-verbal communication .Generally during Interview process Requirement is elicited via verbal Communication .A lot of work has been done to improve the Interview process and to observe and document verbal communication but the area of Non-verbal communication is still unexplored in the field of Requirement Elicitation. Interviewer should give emphasis to Non-verbal communication along with verbal communication so that he can elicit requirements more efficiently and effectively. In this paper we emphasize on Behavioral Aspects of Non-verbal communication i.e. the use of facial expressions, eye contact, gestures, Tone of voice, body posture, orientation, touch, and various cues and signals such as distance, amused, sleepy, pitch, sound, pacing, shaking, sweating are important in non-verbal communication during interviews for eliciting requirements. In this paper we have discussed our findings i.e. behavioral aspects, cues and signals during non verbal communication, these aspects of Non-Verbal Communication are to be followed to minimize problems encountered during requirement elicitation, so that the requirement are elicited and recorded properly.

1 Introduction

In requirements engineering, requirements elicitation is the practice of obtaining the requirements of a system from users, customers and other stakeholders. Requirement elicitation is considered to be a very vital activity in requirement engineering. It is a proven fact that poor elicitation of requirements leads to a project failure. So for the improvement in the software industry's success rate more attention is required in the elicitation process. Elicitation is all about determining the needs of stakeholders and discovering what the user wants. It is one of the most critical activities in software development life cycle. The failure of the software projects has been concerns of the software industry from many years. Many surveys have been conducted to investigate

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 623–630.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

the projects failure statistics. According to the Standish reports [1], success rate of software project is only 28%. A major contributing factor in such a low rate of success is said to be unclear and imprecise requirements [1][2].In 2006 C. J. Davis et al discovered that "Accurately capturing system requirements is the major challenge in large software projects"[3]. The work was reflection of Lindquist (2005) according to whom "poor requirements management can be attributed to 71% of software projects failure, the cost of this failure is enormous." [2]. It is a reasonably well documented fact that software requirements definition has a big impact on final product quality. Generally when Requirement are Elicited from Stakeholders elicitor is not able to elicit the requirements because stakeholder stops communicating the requirement this may be because of (1) Stakeholder is not aware about his own Requirements .(2)Stakeholder knows requirement but is unable to express. (3) Stakeholders Knows his requirement but doesn't want to reveal it. To above problems, formal RE methods and techniques were developed. Davis et al. [3] mention interviews as the most popular RE technique for gathering requirements. Frey and Oishi [4] define interviews which are used to gain an understanding of a particular topic as a means of communication and information exchange where one person asks prepared questions (interviewer) and another answers them (respondent). The other argument is that the interview has been rated as one of the most effective requirements elicitation techniques by practitioners .Interview is a communication intensive activity in which non verbal communication plays an important role. In this paper we emphasize on Non Verbal Communication to make Interview process more effective.

2 Interview a Better Elicitation Technique

Requirement elicitation is to extract & dig out the knowledge about requirements. Elicitation techniques like interviews, brainstorming, workshop, use case, focus groups, JAD/RAD, prototypes, etc are used. Overall these techniques have one thing common that is the elicitation team has to interface with the stakeholder(s) representative(s) or user groups using questions to enquire requirements [17]. Out of which Interviews [5][6]are probably the most traditional and commonly used technique for requirements elicitation by analysts. Because interviews are essentially human based social activities, they are inherently informal and their effectiveness depends greatly on the quality of interaction between the participants. Interviews provide an efficient way to collect large amounts of data quickly from groups or individuals. Interviews enable the elicitor to question the stakeholder directly about their thoughts and opinions, and allows the freedom to describe and reflect in detail on their views and beliefs. The interviews were specifically intended to explore not only what experts thought of the state of practice in requirements elicitation, but also what they believed would be the key components of a new and improved approach for requirements elicitation .Most Software Projects adopt the Face to Face interview i.e. verbal communication with clients and considers one of the effective method of requirement Elicitation .If the Elicitor along with the verbal communication gives emphasis to Non–verbal Communication then requirements may be recorded more effectively & efficiently.

3 Motivation: Problems in Interviews

The interview technique also faces with few problems apart from the problems which are the intrinsic part of requirement elicitation. Such as (1) the stakeholders may have difficulties in expressing the needs of the software system that is desired [12]. (2) Stakeholder also has limitations of memory and communication abilities [16]. (3) Whatever the people do the most they cannot describe it as they think it is usual without any importance "tacit knowledge"[14]. (4) Information may be inaccurate as suggested by cognitive sciences [13]. (5) Communication problems such as domain terminology, different view, biases, status, gender, and environment [14] [10] (6) Interview generate large amount of data. (7) It is hard to compare the different response. (8) It is difficult to analyze that a particular question is still unanswered [12]. (9) The personality, attitude, and manner of asking questions can affect information the elicitation teams receive [15] [11]. Interview involves Verbal and Non-verbal communication during requirement Elicitation process. Researchers suggest that 70% of the information conveyed by the stakeholder is Non-verbal. In this paper we discuss or findings during Non-Verbal Communication, observing and documenting these aspects lead to effective requirement Elicitation process.

4 Non-verbal Communication

Nonverbal communication (NVC) [8], or body language, is a vital form of communication, a natural, unconscious language that broadcasts our true feelings and intentions in any given moment, and clues us in to the feelings and intentions of those around us. When we interact with others, we continuously give and receive wordless signals. All of our nonverbal behaviors, the gestures we make, the way we sit, how fast or how loud we talk, how close we stand, how much eye contact we make send strong messages. These messages don't stop when you stop speaking either. Even when we are silent, we are still communicating nonverbally. Oftentimes, what we say and what we communicate through body language are two totally different things. When faced with these mixed signals, the listener has to choose whether to believe your verbal or nonverbal communication, and in most cases we choose NVC. NVC, such as gestures, does hold potential as a source of additional user behavior data in a cross-cultural testing situation [7]. When such user data collection happens across different cultures, data is often collected and analyzed, ignoring the rich qualitative cues embedded in non-verbal communications such as gestures. In cross-cultural situations, gestures can yield additional information from the user. NVC plays an important role while conducting interviews during Requirement Elicitation. Non -verbal behavior of the clients during Interview can be analyzed by the elicitor in order to record the requirement. If Interviewer has ability to observe and documenting Non-verbal communication during interview as from the findings and suggestions given in this paper the effective requirements may be recorded.

5 Proposed Work: Role of Non-verbal Communication during Interview Process

During Requirement Elicitation when Requirement are Elicited from Stakeholders elicitor is not able to elicit the requirements because stakeholder stops communicating the requirement this may be due toCase 1: Stakeholder is not aware about his own Requirements: In this case requirements are not known by the stakeholders .Stakeholders are confused about their own requirements, provides unnecessary technical details rather than specify their requirements. Case 2: Stakeholder Knows requirements but unable to express: In this case requirements are known by the stakeholders but unable to express it due to poor communication skills, domain terminology. Case 3: Stakeholders knows his requirement but doesn't want to reveal it: In this case stakeholder aware about his requirement, he can express them but don't want to reveal them. The above three cases lead to poor requirement elicitation. To handle these issues we emphasize on Behavioral Aspects of Non-verbal communication and various cues, signals that are important in Non-verbal communication during interview process for eliciting requirements. In Interview process of Requirement Elicitation, much of the non-verbal communication is exchanged using body movements and has up to 55% impact. Certain behaviors can be better understood only through the interpretation of non-verbal cues with body movements being the only way of their communication. We have observed and recorded the behavior of these three types of Stakeholders .Accordingly we present our findings and suggestions in tabular fashion table 1 & table 2.

Table 1. Behavioral Aspects in Non-Verbal Communication

Behavioral Aspects	Case 1	Case 2	Case 3
Eye Contact	• Widen eyes • Look here and there • Blank looks • Eyes rolled upwards	• Eyes moving all around • Downward glances	• Little eye contact • Look somewhere else • Glimpse from the corner of their eye. • Look down with shrugged shoulders
Facial expression	• Surprised expression • Lifted eyebrows • Shrunk and wrinkled hearts. • Curls outs lower lips. • Bite their lower lips.	• Confused expression	• Make a grimace • Disinterested in topic • Limited to mouth movement instead of whole face.

Table 1. (*continued*)

Way of speaking	• Lots of use of “ANDS” & “BUTS”.	• Stammer a lot	• Reluctant to talk • About the subject • Change subject • Use of our words to answer questions • Contradict us.
Tone	• Speak in high voice	• Speak in low voice	• Speak in monotonous tones
Reaction	• Nervousness	• Unable to express himself	
Body posture	• Immobile	• Shrink to one side	• May lean back indicate indifference or lack of Interest. • Limited & stiff physical expressions cross their arms. • Hunch their back. • Clench their fist. • Show defensiveness. • Feel hostile. • Stooped posture. • Shuffling feet’s. • Get angry. • Lean forward • Fight facial expression • Cover their mouth • While walking hide their hands.

Table 1. (*continued*)

Gestures	• Shrugs jerk up shoulders to say no in a response.	• Lot of hand gestures	• Few hand & arm movements. • Touch their face, throat, & mouth. • Scratch their nose or ears. • Touch their chest/heart with open hands. • Constantly swallow or clear their throat. • Stick their tongue out to moist their lips. • Closed, descending & insecure poses. • Tapping hands or feet.
Space	• Turn their body away from askers	• Place objects	• They will pull their tie or t-shirt.
Handshake	• Shake only fingers not hand	• Relaxed hand shaking but at distance from the front person	• A firm handshake
Orientation	• Present them back to back.	• Likely to sit side by side.	• Sit face to face

Observing Non-Verbal Communicational behavioral aspects, various cues and signals as discussed in the tables given above we can very easily decide whether the stakeholder is not aware of its requirements, unable to express it or doesn't want to reveal it. Once identified as case1, case 2, case3 discussed earlier we can take appropriate measures. For case 1 stakeholder should be made aware about his own requirements by counseling and providing domain knowledge. For case 2 stakeholder should be motivated to express his requirement by providing prototypes, front end visuals etc. For case 3 the matter could be reported to top management and fresh interviews should be conducted with different set of stakeholders. If these aspects of Non-Verbal Communication are observed and documented, problems encountered during requirement elicitation may be minimized and requirement may be elicited and recorded more efficiently and effectively.

Table 2. Cues & signals in Non-Verbal Communication

CUES & SIGNALS	CASE 1:	CASE 2:	CASE 3:
Distance	Moderate	Low	High
Amused	High	Moderate	Low
Anxious	High	Moderate	Low
Confused	Moderate	High	Low
Sleepy	High	Low	Moderate
Slouching	Low	Moderate	High
Stiff	Low	Moderate	High
Cringing	High	Moderate	Low
Towering	Low	Moderate	High
Crouching	Low	Moderate	High
Scratching head	Low	Moderate	High
Biting fingernails	Low	Moderate	High
Folding arms	Low	Moderate	High
Narrowing eyes	Low	Moderate	High
Voice Intensity	Low	Moderate	High
Timing and Pace	Moderate	Low	High
Sound	High	Moderate	Low
Pacing	Low	Moderate	High
Pitch	High	Low	Moderate
Shaking	High	Moderate	Low

Acknowledgments. First and foremost, our sincere thanks goes to Prof. Syed Wasim Akhtar, Honorable Vice Chancellor, Integral University, Lucknow. Prof Akhtar has given us unconditional support in many aspects, which enabled us to work on the exciting and challenging field of Software Engineering. We would also like to give our special thanks to Prof. T. Usmani, Pro Vice Chancellor, Integral University. His encouragement, help, and care were remarkable during the past few years. We are also grateful to Prof S. M. Iqbal Chief Academic Consultant, Integral University. Prof Iqbal provided us with valuable thoughts for our research work. My gratitude also goes to Dr. Irfan Ali Khan, Registrar, Integral University for his constructive comments and shared experience.

References

[1] Standish Group, The CHAOS Report (2005), `http://www.standishgroup.com/sample_research/PDFpages/chaos1994.pdf`

[2] Beg, M.R., Abbas, Q., Verma, R.P.: Analysis of various studies showing the impact of the requirements engineering on software project success (2008)

[3] Davis, A., Dieste, O., Hickey, A., Jurist, N., Moreno, A.M.: Effectiveness of requirements elicitation techniques: Empirical results derived from a systematic review. In: Proceedings of the 14th IEEE International Requirements Engineering Conference, pp. 176–185 (2006)

[4] Frey, H., Oishi, S.M.: How to Conduct Interviews by Telephone and in Person. Sage Publications, London (1995)

[5] Agarwal, R., Tanniru: Knowledge Acquisition Using Structured Interviewing: An Empirical Investigation. Journal of Management Information Systems 7(1), 123–140

[6] Holtzblatt, K., Beyer: Requirements Gathering: The Human Factor. Communications of ACM 38(5), 30–32

[7] Kendon, A.: Gestures and speech: how they interact, pp. 13–45. Sage, CA (1983)

[8] Mehrabian, A.: Nonverbal Communication (2007)

[9] Samovar, L., Porter, R., McDaniel, E.: Communication between Cultures (2009)

[10] Byrd, T.A., Cossick, K.L., Zmud, R.W.: A Synthesis of Research on Requirements Analysis and Knowledge Acquisition Techniques, pp. 117–138 (1992)

[11] Fey, Goldberg: Legal Interviewing from a Psychological Perspective: an Attorney's Handbook, vol. 14, p. 217 (1977)

[12] Levesque, G., Michel-Ange, Zamor, Shostak, B.: Requirements Management: from technical to managerial aspects

[13] Gilovich, T., Griffin, D., Kahneman, D. (eds.): Heuristics and biases: The psychology of intuitive judgment. Cambridge University Press (2002)

[14] Goguen, J., Linde, C.: Techniques for Requirements Elicitation. In: Proc. First IEEE International Symposium on Requirements Engineering, pp. 152–164 (1993)

[15] Goguen, J.A., Linde, C.: Techniques for requirements elicitation. In: Proceedings of the IEEE International Symposium on Requirements Engineering (1992)

[16] Kuberski, P.: The Persistence of Memory: organism, myth, text, California (1992)

[17] Beg, M.R., Abbas, Q., Verma, R.P.: Interview process model for requirement elicitation. International Journal of Computer Science and Applications 1(2) (August 2008)

A Critical Review of Migrating Parallel Web Crawler

Md. Faizan Farooqui[1], Md. Rizwan Beg[2], and Md. Qasim Rafiq[3]

[1] Research Scholar, Department of Computer Application, Integral University,
Lucknow 226026, India
faizan_farooqui2000@yahoo.com
[2] Professor & Head, Department of CS&E, Integral University, Lucknow 226026, India
rizwanbeg@gmail.com
[3] Chairman, Department of Computer Engineering, Aligarh Muslim University,
Aligarh, India
mqrafiq@hotmail.com

Abstract. The size of the internet is very large and it has grown enormously, search engines are the tools for World Wide Web navigation. In order to provide powerful search facilities, search engines maintain comprehensive indices for documents and their contents on the Web by continuously downloading Web pages for processing, known as web crawling. In this paper we reviewed various web crawlers and their performance attributes. We study mobile and parallel web crawling approach that makes web crawling system more effective and efficient. The major advantage of the mobile approach is that the analysis portion of the crawling process is done locally where the data resides rather than remotely inside the Web search engine. This can significantly reduce network load which, in turn, can improve the performance of the crawling process. The major advantage of parallel crawling is that as the size of the Web grows, it becomes imperative to parallelize a crawling process, in order to finish downloading pages in a reasonable amount of time. We identify fundamental issues related to migrating parallel crawling and also propose metrics to evaluate a migrating parallel crawler. Lastly, we summarize the web crawlers and their performance attributes that effects the process of web crawling.

1 Introduction

The Internet is a global system of interconnected computer networks. The searching and indexing tasks for the web are currently handled from specialized web applications called search engines. The modern search engines can be divided into three parts they are the publically available search engine, the data- base and the web crawling system. A web crawler is an automated program that browses the Web in a methodological manner. The process of traversing the web is called Web crawling. Web crawler starts with a queue of known URLs to visit. As it visits these pages, it scans them for links to other web pages. The Web Crawler consists of Crawler Manager, Robot.txt downloader, Spider and Link Extractor. A Crawl Manager takes a set of URLs from Link extractor and sends the next URLs to the DNS resolver to obtain its IP address. Robot.txt file are the means by which web authors express their wish as to

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 631–637.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

which pages they want the crawler to avoid. Link extractor extract URLs from the links in the downloaded pages and sends the URLs to the crawler manager for downloading afterwards.

2 Literature Survey

In [13], the author demonstrated an efficient approach to the “download-first process-later” strategy of existing search engines by using mobile crawlers. In [14] author has implemented UbiCrawler, a scalable distributed and fault-tolerant web crawler. In [15] author presented the architecture of PARCAHYD which is an ongoing project aimed at designing of a Parallel Crawler based on Augmented Hypertext Documents. In [16] the author studied how an effective parallel crawler is designed. As the size of the Web grows, it becomes imperative to parallelize a crawling process. In [17] the author proposes Mercator, which is a scalable, extensible crawler. [18] Presented Google, a prototype of a large scale search engine which makes heavy use of the structure present in hypertext. [19] Aims at designing and implementing a parallel migrating crawler in which the work of a crawler is divided amongst a number of independent.

3 Motivation: Problems in Generic Web Crawlers

According to [1], Web crawlers of big commercial search engines crawl up to 10 million pages per day. Assuming an average page size of 6K [2], the crawling activities of a single commercial search engine adds a daily load of 60GB to the Web.

Scaling Issues: One of the first Web search engines, the World Wide Web Worm [3], was introduced in 1994 and used an index of 110,000 Web pages. Big commercial search engines in 1998 claimed to index up to 110 million pages [1]. The Web is expected to grow further at an exponential speed, doubling its size (in terms of number of pages) in less than a year [4].

Efficiency Issues: Current Web crawlers download all these irrelevant pages because traditional crawling techniques cannot analyze the page content prior to page download [13].

Index Quality Issues: Current commercial search engines maintain Web indices of up to several hundred million pages[1].

4 Migrating Parallel Web Crawler

The Migrating Parallel Crawler system consists of Central Crawler, Crawl Frontiers, and Local Database of each Crawl Frontier and Centralized Database. It is responsibility of central crawler to receiving the URL input from the applications and forwards the URLs to the available migrating crawling process. Crawling process migrated to different machines to increase the system performance. Local database of each crawl frontier are buffers that locally collect the data. This data is transferred to the central crawler after compression and filtering which reduces the network bandwidth overhead. The central crawler has a centralized database which will contain the documents collected by the crawl frontiers independently. The main advantages of the migrating parallel crawler approach are Localized Data Access, Remote Page Selection, Remote Page Filtering, Remote Page Compression, Scalability, Network-load dispersion, Network-load reduction.

5 Issues in Migrating Parallel Web Crawler

The following issues are important in the study of a migrating parallel crawler interesting:

Communication bandwidth: Communication is really important so as to prevent overlap, or to improve the quality of the downloaded content, crawling processes need to communicate and coordinate with each other to improve quality thus consuming valuable communication bandwidth.

Quality: Prime objective of migrating parallel crawler is that to download the "important" pages first to improve the "quality" of the downloaded pages.

Overlap: When multiple processes run in parallel to download pages, it is possible that different processes download the same page multiple times. One process may download the same page that other process may have already downloaded, one process may not be aware that another process has downloaded the same page.

6 Coordination in Migrating Parallel Web Crawler

The coordination is done in order to prevent overlap and also crawling processes need to coordinate with each other so that the correct and quality pages are downloaded. This coordination can be achieved in following ways:

Independent: In this method Crawling processes may download pages independent of each other without any coordination. The downloaded pages may overlap in this case, but we assume that overlap is not significant enough.

Static assignment: In this method the Web is partitioned into sub partitions and each sub partition is assigned to each Crawling processes before the actual crawling process may begin. This type of arrangement is called static assignment.

Dynamic assignment: In this coordination scheme there is a central coordinator that divides the Web into small partitions and each partition is dynamically assigned to a Crawling processes for further downloading.

7 Crawling Modes for Static Assignment in Migrating Parallel Web Crawler

In this scheme of static assignment, each Crawling processes are responsible for a certain partition of the Web and have to download pages within the partition. There may be possibility that some pages in the partition may have links to pages in another partition. Such types of links are referred to as an inter-partition link.

Firewall mode: Each Crawling processes strictly download the pages within its partition and inter-partition links are not followed. In firewall mode all inter-partition links are either thrown away or ignored.

Cross-over mode: In this mode each Crawling processes download pages within its partition. It follows inter-partition links when links within its partitions are exhausted.

Exchange mode: In this mode Crawling processes periodically exchange inter-partition links. Crawling processes are not allowed to follow inter-partition links [16].

Table 1. Comparison of three Crawling Modes[16]

Mode	Coverage	overlap	Quality	Communication
Firewall	Bad	Good	Bad	Good
Cross-Over	Good	Bad	Bad	Good
Exchange	Good	Good	Good	Bad

8 Evaluation Models in Migrating Parallel Web Crawler

This section deals with metrics that we define to quantify the advantages or disadvantages of migrating parallel crawler.

Coverage: the coverage is defined as I/U, where U is the number of pages downloaded by migrating parallel crawler and I is number of unique pages downloaded by the migrating parallel crawler.

Overlap: The overlap is defined as (N−I) /I. Where, N is number of pages downloaded by the crawler, and I is the number of unique downloaded pages by migrating parallel crawler.

Quality: The quality of downloaded pages is defined as $|A_N \cap P_N|/|P_N|$. The migrating parallel crawler downloads the important N pages, and P_N is that set of N pages. A_N is set of N pages that an actual migrating parallel crawler would download, which is different from P_N.

Communication overhead: The communication overhead is defined as E/I. Where I in total pages downloaded and E represents exchanged inter-partition URLs.

Table 2 summarizes different crawling techniques with their performance attributes. Crawling process of migrating parallel crawlers move to the resource which needs to be crawled to take the advantage of localized data access. After accessing a resource, migrating parallel crawler move on to the next machine or server and send result of crawling to central database in compressed form. From the table it is evident that Migrating parallel crawlers are more efficient in terms of network load reduction, I/O performance, reducing communication bandwidth, network load dispersion etc.

Table 2. Performance Attributes for Different Web Crawlers

	High performance distributed web crawler	Incremental crawler	Parallel crawler	Agent based	Migrating parallel Crawler	Domain Specific Crawler	Mobile Crawler	Breadth First Crawling
Robustness	++	--	--	--	++	--	++	--
Flexibility	++	--	--	++	++	--	--	--
Manageability	++	--	--	++	++	--	--	--
I/O Performance	++	--	--	++	++	--	--	--
Network resources	++	--	--	--	++	--	++	--
OS limits	++	--	--	--	++	--	++	--
Higher performance	++	--	--	--	++	--	++	--
Extensibility	--	++	--	--	--	--	++	--
Incremental crawling	--	++	--	--	--	--	--	--
Reasonable cost	++	--	--	--	--	--	--	++
Overlapping	--	--	++	++	++	++	++	--
Communication bandwidth	--	--	++	--	++	++	++	--
Network load dispersion	--	--	++	--	++	++	++	--
Network load reduction	--	--	++	++	++	++	++	--
Freshness	--	++	--	--	--	--	--	--
Page rank	--	--	--	--	--	--	--	++
Scalability	--	--	--	--	--	++	--	--
Load sharing	--	--	--	--	--	++	++	--
High quality	--	--	--	--	--	--	--	++
Agent oriented	--	--	--	++	++	--	--	--

9 Conclusion

In this paper we reviewed various web crawlers and their performance attributes. We study mobile and parallel web crawling approach that makes web crawling system more effective and efficient. We identify fundamental issues related to migrating parallel web crawler and also propose metrics to evaluate a migrating parallel web crawler. Lastly, we summarize the web crawlers and their performance attributes that effects the process of web crawling. The research directions in migrating parallel crawler include:

- The crawling process should crawl in domain specific manner
- Crawling on the host/server should be done on breadth first manner
- Multi threaded server for central coordination
- Agent based Crawl workers

This future work will deal with the problem of quick searching and downloading the data. The parallel crawlers will be studied and a new crawler technique will be compared with the existing one. The data will be collected and analyzed with the help of tables and graphs.

Acknowledgments. First and foremost, our sincere thanks goes to Prof. Syed Wasim Akhtar, Honorable Vice Chancellor, Integral University, Lucknow. Prof Akhtar has given us unconditional support in many aspects, which enabled us to work on the exciting and challenging field of Software Engineering. We would also like to give our special thanks to Prof. T. Usmani, Pro Vice Chancellor, Integral University. His encouragement, help, and care were remarkable during the past few years. We are also grateful to Prof S. M. Iqbal Chief Academic Consultant, Integral University. Prof Iqbal provided us with valuable thoughts for our research work. My gratitude also goes to Dr. Irfan Ali Khan, Registrar, Integral University for his constructive comments and shared experience.

References

[1] Sullivan, D.: Search Engine Watch. Mecklermedia (1998)
[2] Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Stanford University, Stanford, CA, Technical Report (1997)
[3] McBryan, O.A.: GENVL and WWW: Tools for Taming the Web. In: Proceedings of the First International Conference on the World Wide Web, Geneva, Switzerland (1994)
[4] Kahle, B.: Archiving the Internet. Scientific American (1996)
[5] Gosling, J., McGilton, H.: The Java Language Environment. Sun Microsystems, Mountain View, CA, White Paper (April 1996)
[6] White, J.E.: Mobile Agents. MIT Press, Cambridge (1996)
[7] Harrison, C.G., Chess, D.M., Kershenbaum, A.: Mobile Agents: Are they a good idea? IBM Research Division, T.J. Watson Research Center, White Plains, NY, Research Report (September 1996)
[8] Nwana, H.S.: Software Agents: An Overview. Knowledge Engineering Review 11, 3 (1996)
[9] Wooldridge, M.: Intelligent Agents: Theory and Practice. Knowledge Engineering Review 10, 2 (1995)
[10] Maes, P.: Modeling Adaptive Autonomous Agents. MIT Media Laboratory, Cambridge, MA, Research Report (May 1994)
[11] Maes, P.: Intelligent Software. Scientific American 273, 3 (1995)
[12] Finin, T., Labrou, Y., Mayfield, J.: KQML as an agent communication language. University of Maryland Baltimore County, Baltimore, MD (September 1994)
[13] Hammer, J., Fiedler, J.: Using Mobile Crawlers to Search the Web Efficiently (2000)
[14] Boldi, P., Codenotti, B., Santini, M., Vigna, S.: UbiCrawler: A Scalable Fully Distributed Web Crawler (2002)

[15] Sharma, A.K., Gupta, J.P., Aggarwal, D.P.: PARCAHYDE: An Architecture of a Parallel Crawler based on Augmented HypertextDocuments (2010)
[16] Cho, J., Garcia-Molina, H.: Parallel crawlers. In: Proceedings of the Eleventh International World Wide Web Conference, pp. 124–135 (2002)
[17] Heydon, A., Najork, M.: Mercator: A scalable, extensible web crawler. World Wide Web 2(4), 219–229 (1999)
[18] Singh, A., Singh, K.K.: Faster and Efficient Web Crawling with Parallel Migrating Web Crawler (2010)
[19] Wu, M., Lai, J.: The Research and Implementation of parallel web crawler in cluster. In: International Conference on Computational and Information Sciences (2010)

Parallel Character Reconstruction Expending Compute Unified Device Architecture

Anita Pal, Kamal Kumar Srivastava, and Atul Kumar

Department of Computer Science & Engineering
Shri Ramswaroop Memorial College of Engineering and Management, Lucknow (U.P.), India
{anitapal13,2007.srivastava,atulverma16}@gmail.com

Abstract. Neural networks, or the artificial neural networks to be more precise, represents a technology that is rooted in many disciplines: neuroscience, mathematics, statistics, physics, computer science and engineering. Neural network finds applications in such fields as modeling, time series analysis, pattern recognition signal processing and control by virtue of an important property: the ability to learn from input data with or without a teacher .In a biological system, learning involves adjustments to the synaptic connections between neurons same for artificial neural networks (ANNs) works too that has made it applicable to valid applications. Neural Network architecture has the ability to learn for the things and then later on classify the things. Neural Network for Character Recognition is based over Multilayered Architecture having Back-propagation algorithm. First Network is been trained for the alphanumeric handwritten characters and then testing the network with the trained or untrained handwritten characters. We achieved a greater computation enhancement by using modified back- propagation algorithm having an added momentum term, which lowers the training time and speeds the system. The time is more reduced with its parallel implementation using CUDA.

Keywords: CUDA, GPU, Back-propagation algorithm.

1 Introduction to CUDA

Compute Unified Device Architecture is a general purpose programming model where programmer kicks off batches of threads over graphics processing units. Graphics processing units are dedicated super threaded, massively parallel data co-processor. CUDA is an extension towards the programming languages C/C++ which provides some libraries to utilize graphics processing units in massively data parallel architecture.

1.1 Device

In a matter of just a few years, the programmable graphics processor unit has evolved into an absolute computing workhorse. With multiple cores driven by very high memory bandwidth, today's GPUs offer incredible resources for both graphics and non-graphics processing. The main reason behind such an evolution is that the GPU is specialized for compute-intensive, highly parallel computation exactly what graphics

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 639–648.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

rendering is about and therefore is designed such that more transistors are devoted to data processing rather than data caching and flow control. More specifically, the GPU is especially well-suited to address problems that can be expressed as data-parallel computations – the same program is executed on many data elements in parallel with high arithmetic intensity, the ratio of arithmetic operations to memory operations. Because the same program is executed for each data element, there is a lower requirement for sophisticated flow control; and because it is executed on many data elements and has high arithmetic intensity, the memory access latency can be hidden with calculations instead of big data caches. Data-parallel processing maps data elements to parallel processing threads. Many applications that process large data sets such as arrays can use a data-parallel programming model to speed up the computations. In 3D rendering large sets of pixels and vertices are mapped to parallel threads. Similarly, image and media processing applications such as post-processing of rendered images, video encoding and decoding, image scaling, stereo vision, and pattern recognition can map image blocks and pixels to parallel processing threads. In fact, many algorithms outside the field of image rendering and processing are accelerated by data-parallel processing, from general signal processing or physics simulation to computational finance or computational biology.

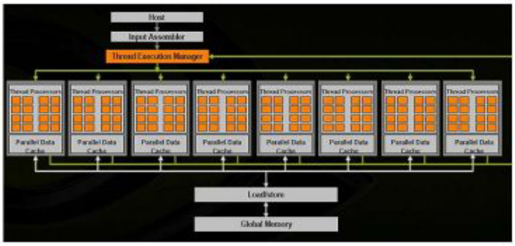


Fig. 1. Architecture of GPU

Up until now, however, accessing all that computational power packed into the GPU and efficiently leveraging it for non-graphics applications remained tricky:

The GPU DRAM could be read in a general way – GPU programs can gather data elements from any part of DRAM - but could not be written in a general way - GPU programs cannot scatter information to any part of DRAM, removing a lot of the programming flexibility readily available on the CPU.

Some applications were bottlenecked by the DRAM memory bandwidth, underutilizing the GPU's computational power. [14]

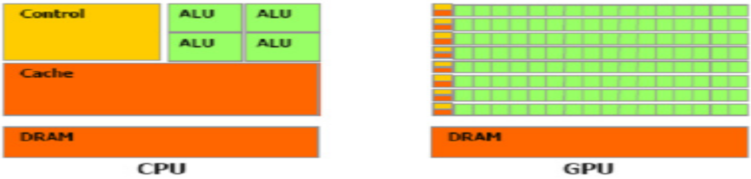


Fig. 2. Comparison between Number of Processsors in GPU and CPU

1.2 CUDA: A New Architecture for Computing on the GPU

CUDA stands for Compute Unified Device Architecture and is a new hardware and software architecture for issuing and managing computations on the GPU as a data-parallel computing device without the need of mapping them to a graphics API. Operating system's mechanism of multitasking which is responsible for managing the access to the GPU and many CUDA and graphics applications running concurrently. CUDA software stack is also collected by many layers: Which is hardware driver, application programming interface(API) and In this we use two higher-level mathematical libraries of common usage, which is CUFFT and CUBLAS. In this, the hardware which has been used to support lightweight driver and runtime layers. As a resultant its performance has been increased.

CUDA use the features of parallel data cache or on-chip shared memory for the fast general read and write access. In this all the threads use to share data with one another. Applications can take advantage of it by minimizing over-fetch and round-trips to DRAM and therefore becoming less dependent on DRAM memory bandwidth.

1.3 A Highly Multi-threaded Co-processor

When programmed through CUDA, the GPU is viewed as a compute device capable of executing a very high number of threads in parallel. In GPU it function as a coprocessor to the main CPU, or host: In this a sector of a part of application that is executed many times, but independently on different data, can be separated into a function that is executed on the device as many different threads. The effect, such a function is compiled to the instruction set of the device and the resulting program, known as a kernel, is downloaded to the device. Both the host and the device maintain their own DRAM, referred to as a host memory and device memory, respectively. One can also copy data from one DRAM to the other through optimized API calls the device's high-performance Direct Memory Access (DMA) engine.

1.3.1 Thread Batching

Threads that executes a kernel is organized as a grid of thread blocks is–

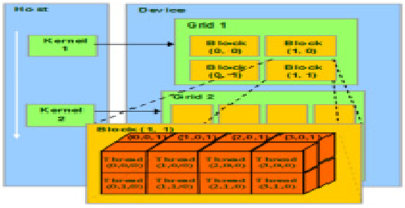


Fig. 3. Structure of Block and Grid on GPU

1.3.2 Thread Block

A thread block is a batch of threads that can cooperate together by efficiently sharing data through some fast shared memory and synchronizing their execution to coordinate memory accesses. In kernel we can specify the synchronization points normally, whenever threads in a block are suspended until they all reach the synchronization point. In this case each thread assigns a unique ID, which is the number of thread within the block. It also help in complex addressing based on the thread ID, the application can specify a block as a two- or three-dimensional array of architecture size and identify each thread using a two 2- or 3-component index instead.

1.3.3 Grid of Thread Blocks

A block has only limited maximum number of threads. In the thread, blocks of same dimensional and size that execute the same kernel would be batched together into a grid of blocks, so the total number of threads can be launched in a single kernel supplication is much larger. For the above, the expense of reduced thread cooperation, because threads in different thread blocks from the same grid and the grid cannot be communicate and synchronize with each other. The above model allows kernels to efficiently run without recompilation on various devices with different parallel capabilities: In this, a device may run all the blocks of a grid sequentially.

1.3.4 Cuda Memory Model

A thread that executes on the device has only access to the device's DRAM and on-chip memory through the following memory spaces as illustrated in the figure below.

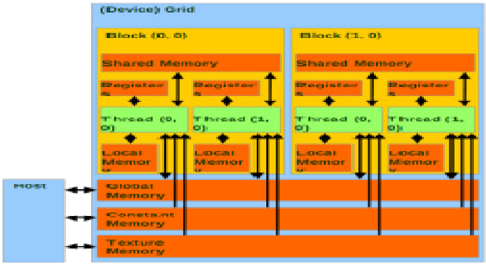
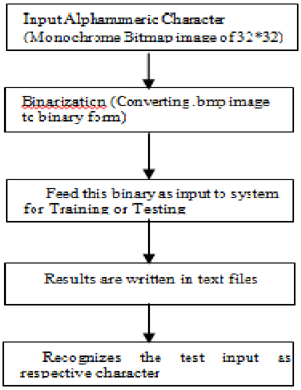


Fig. 4. CUDA Memory Model

2 Proposed Methodology


Input Alphanumeric Character (Monochrome Bitmap image of 32*32)
Binarization (Converting .bmp image to binary form)
Feed this binary as input to system for Training or Testing
Results are written in text files
Recognizes the test input as respective character

2.1 Training Methodology

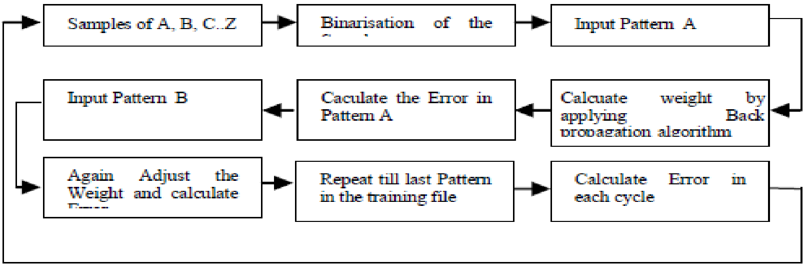

Samples of A, B, C..Z
Binarisation of the
Input Pattern A
Input Pattern B
Caculate the Error in Pattern A
Calcuate weight by applying Back propagation algorithm
Again Adjust the Weight and calculate
Repeat till last Pattern in the training file
Calculate Error in each cycle

Fig. 5. Training Methodology on Back-propagation Network

3 Implementation

3.1 Binarization

First performed binarization of the input image which is of monochrome bitmap image of size 32*32 and which is like-

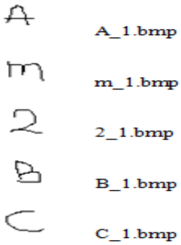


Fig. 6. Sample Input Images

And many more, we have collected about 100 sample of each character from different people and used these during the training of the system and later other test samples for testing the accuracy .And the binary form of the character A_5.bmp is:

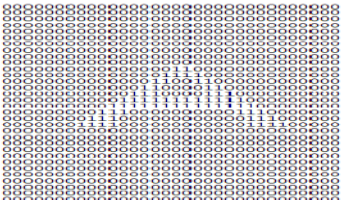


Fig. 7. Binararization of Input Image

3.2 Phases of Implementation

The network works in 2 modes: -

1. Training Mode
2. Testing Mode

3.2.1 Training Mode

A network in which learning is employed is said to be subjected to training .Training is an external pressure regimen. Learning is the desired process that takes place internal to the network. Here, the user provides a training file in the current directory called TRAINING.TXT. This file contains patterns. Each pattern has a set of inputs followed by a set of inputs followed by a set of outputs. Each value is separated by one or more spaces. Another file that is used in training is the weights file. Once the simulator reaches the error tolerance that was specified by the user, or the maximum number of iterations, the simulator saves the state of the network by saving all of its weights in a file called WEIGHTS.TXT. The file can then be used subsequently in another run of the simulator in testing mode.

3.2.2 Testing Mode

In this mode the user provides test data to the simulator in a file called TEST.TXT. This file contains only input pattern. When this file is applied to an already trained network, an OUTPUT.TXT file is generated, which contains the outputs from the network for all the input patterns The network goes through one cycle of operation in this mode, covering all the modes in the test data file to start up the network, the weights file ,WEIGHTS.TXT is read to initialize the state of the network .

3.3 Parallel Implementation on CUDA

Since the neural network has a large amount of computations in the TRAINING MODE as well as in the TESTING MODE and as per the model of Multilayer Neural Network it allows us to port the network on a parallel architecture. We ported our problem on NVIDIAs CUDA parallel architecture which is follows SIMD approach. The back propagation algorithm can be majorly divided into 3 steps namely

FORWARD PROPAGATE
BACKWARD PROPAGATE
WEIGHT UPDATION

The three steps in back propagation model can be ported in parallel. The model has three layer of neurons input layer, hidden layer and output layer, the computations in all the three layers are independent of each other. However, the result of the preceding layers is fed into the successive layers, this is a restriction and hinders parallelism. But in spite of this one can make the computations in a input layer, hidden layer or output layer in parallel as the neurons in each of these layer are independent of each other.

3.3.1 Forward Propagation

Forward pass is the first step in the back propagation algorithm in which we calculate the summation of inputs with the corresponding weight and is passed through an activation function.

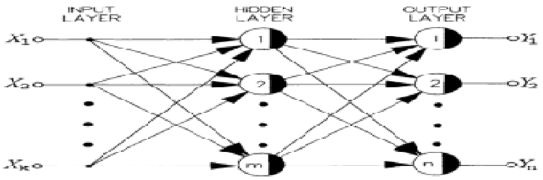
$$\mathrm{NET}= X1W1+X2W2+\text{--------}+ XnWn$$

$$\mathrm{OUT} = F(\mathrm{NET}) = 1/(1+e\text{-}\mathrm{NET})$$

Threads were launched as per the configuration of hidden layer and output layer i.e 500 threads for the hidden layer and 6 for output layer. Each thread computes the NET and OUT function as described above and sends the result back to CPU. Since the net input of the output layer is dependent on the hidden layer so these are launched on GPU one by one.

Input Layer, No. of Input neurons = 1024 (32*32), No threads launched as activation function not required, Hidden Layer - No. of Hidden neurons = 500, threads launched = 500, Output Layer- No. of output neurons = 6, No. of threads launched = 6, Output Bit Sequence

Like A=000001, B=000010, C=000011 and likely all character are assigned this desired output sequence for the training of characters.



Where k=1024, m=500 and n=6

Fig. 8. Model of Single Hidden Layered Neural Network

Now in parallel implementation of model each thread calculates the activation value for each hidden layer neuron hence the activated threads for hidden layer is 500, which greatly reduces computation time. Now these values are being transmitted for each output neuron, these values acts as input to the output layer. Now again parallel activation values for output neuron can be calculated by launching same number of threads, as the number of output neurons giving output bit sequence.

3.3.2 Weight Updating and Backward Propagation

After forward propagation the output bit sequence is checked by the desired output sequence as given in TRAINING.TXT .If these two bit sequence do not match then the next step is to calculate the backward errors and then weights are updated as to minimize the errors. The error calculation is done serially but weight updating is done in parallel. The updating of weights in hidden neurons and output neurons is independent of each other and hence it can be parallelized.

4 Testing

Both CPU and GPU versions of the program works on the handwritten input monochrome bitmap image and these programs first converts those images into binary form before making them actual training or testing data for the system. Now the system is given images like this-

Binary form of input image is-

```
    0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000 0000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000011000
00000000000000000000000000011110000000000000000000000000000000110011000000000000000000000000000001100001100000000000000000000000
000011100001110000000000000000000000000001111111111111000000000000000000000011111111111111000000000000000000000111000000000001110 0
0000000000000011100000000000001110000000000000000011110000000000000001111000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000 000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
```

Input image must be given in monochrome bitmap format otherwise the results will not be proper and size of the image is also fixed which is 32x32.

5 Result

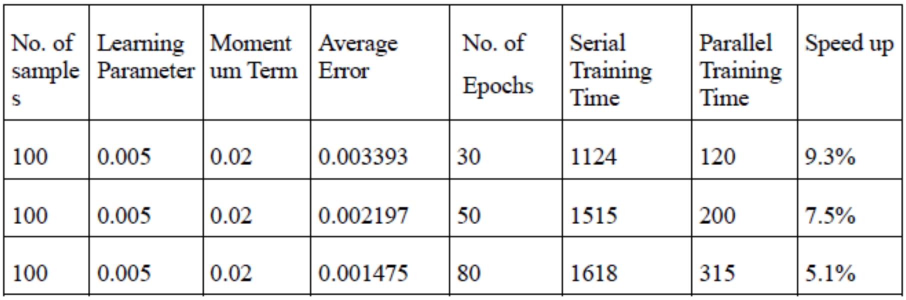
The system was simulated using a feed forward neural network which consisted of 1024 input neurons, 500 hidden layer neurons and 6 output neurons. The image was binarized in 32 * 32 pixel values so 1024 input neurons and the system will recognize 26 uppercase and lowercase English alphabets and numerals from 0 to 9 so output neurons like 000001 for 'A' and so on. Here we experiment the result by varying the number of Epochs i.e number of cycles for which the network is to be trained. Epochs were varied from 5X to 9X. The structure of the neural network was kept constant (Input neurons and Hidden neurons were kept constant). As the epochs were increased the average error per patter was reduced and the network accuracy rate was increased.

Table 1. Comparison of Time for Serial and Parallel Version of the System

No of characters in test Sample	No. of Epochs	Learning Parameter	Momentum Term	Character Correctly Recognized	%age Recognized
100	100	0.005	0.02	77	77%

The character recognition system is being tested for different samples of 'A, B' which were not supplied in the training file and the recognition accuracy is listed in the table shown below. There we have supplied some samples of images that were not trained before, so as to get accuracy of the system .That how much system being able to recognize untrained samples. Result of checking accuracy of system is –

Table 2. Testing the Accuracy of the System



No. of samples	Learning Parameter	Momentum Term	Average Error	No. of Epochs	Serial Training Time	Parallel Training Time	Speed up
100	0.005	0.02	0.003393	30	1124	120	9.3%
100	0.005	0.02	0.002197	50	1515	200	7.5%
100	0.005	0.02	0.001475	80	1618	315	5.1%

It is been observed that the network recognizes well when the epochs were increased. This is due to the fact that when we increase the cycles the average error is reduced and more accurate results are obtained.

6 Conclusion

The proposed method for the handwritten character recognition uses the back propagation learning algorithm and a momentum term for weight modification process which yielded remarkable results. The additional momentum term introduced in weight modification process helps in better convergence of weights and reduced error while training the samples. Due to the back-propagation of error element in Multilayer Perceptron (MLP), it frequently suffers from the problem of Local-Minima; hence the samples may not converge. The network may get trapped in local minima even though there is a much deeper minimum nearby. The network showed better accuracy as number of epoch were increased, so the network should be trained for a specific error tolerance rate to yield better results. The training time was reduced when we used Nvidia's CUDA architecture and GTX 260 having 260 cores. Since we know that time is a critical factor in the training of any neural network, at the same time we cannot compromise with the accuracy rate. So, utilizing the power of GPU proved to be a better option in reduction of training time for any neural network and a speed up of 9X was obtained using one hidden layer and 500 hidden neurons. A huge amount of speed up will be obtained when the hidden neurons will be increased as well as the hidden layers would be increased. Since the CUDA architecture is SCALABLE in nature we can also obtain a speed up on increasing the number of cores in a graphics card. More work need to be done especially on the test for more complex handwritten characters. This type of system saves a lot of time of users and helps in digitization of their important documents. These systems are not having only commercial users but also for personal uses as well. There are still many government organizations that still using conventional methods of working and pen downing important things, that consumes a lot time and human effort as well. These organizations are needed to be digitized as in this form, data would be more safe and would remain preserved. Future work that needed to be done on our current system will include especially-Reduction in training time of the system, Increasing the accuracy of the system, Extend our system to recognize words as well as our present system is just for single letter at a time, modifications in the current algorithm used for better training of system and overcomes the drawbacks of local minima and overtraining, develop a better user interface.

References

1. Liu, C.-L.: Normalization-Cooperated Gradient Feature Extraction for Handwritten Character Recognition. Transactions on Pattern Analysis and Machine Intelligence 29(8), 1465–1469 (2007)
2. Verma, B.K.: Handwritten Hindi Character Recognition Using Multilayer Perceptron and Radial Basis Function Neural Network. In: IEEE Conference on Neural Network, vol. 4, pp. 2111–2115 (1995)
3. Chung, Y.Y., Wong, M.T.: Handwritten Character Recognition by Fourier Descriptors and Neural Network. In: Proceedings of IEEE TENCON 1997. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications, vol. 1, pp. 391–394 (1997)
4. Starzyk, J.A., Ansari, N.: Feedforward neural network for handwritten character recognition. In: Proceedings of 1992 IEEE International Symposium on Circuits and Systems, ISCAS 1992, vol. 6, pp. 2884–2887 (1992)
5. Li, Y., Li, J., Meng, L.: Character Recognition Based on Hierarchical RBF Neural Networks. In: Sixth International Conference on Intelligent Systems Design and Applications, ISDA 2006, vol. 1, pp. 127–132 (2006)
6. Smith, S.J., Baurgoin, M.O.: Handwritten character classification using nearest neighbor in large database. IEEE Trans. On Pattern and Machine Intelligence 16(10), 915–919 (1994)
7. Pal, A., Singh, D.: Handwritten English Character Recognition using Neural Network. International Journal of Computer Science and Communication 1(2), 141–144 (2010)
8. Faaborg, A.J.: Using Neural Networks to Create an Adaptive Character Recognition System, Cornell University, Ithaca, NY (2002)
9. `http://neileshchorakhalikar.wordpress.com/cuda/`

A Novel Algorithm for Obstacle Aware RMST Construction during Routing in 3D ICs

Prasun Ghosal, Satrajit Das, and Arindam Das

Bengal Engineering and Science
University, Shibpur
Howrah 711103, WB, India
p_ghosal@it.becs.ac.in, satrajit_das@sifymail.com, rajonline05@gmail.com
http://it.becs.ac.in/p_ghosal/

Abstract. Three dimensional integrated circuits offer an attractive alternative to 2D planar ICs by providing increased system integration by either increasing functionality or combining different technologies. Routing phase during layout design of 3D ICs plays a critical role. The problem again becomes worse in presence of obstacles across the routing layers. This obstacle aware routing tree construction has become a challenging problem among the researchers recently. In this work, an efficient algorithm has been proposed for the construction of rectilinear minimum Steiner tree (RMST) in presence of obstacles across the routing layers using a shortest pair approach. Due to ever increasing design complexity issues, careful measures have been taken to reduce the time complexity of the proposed algorithm. The novelties of this work may be stated as follows (i) proposed algorithm helps to construct an RMST in presence of obstacles, (ii) time complexity of the proposed algorithm is very much competitive with available tools, (iii) proposed algorithm efficiently reduces the number of Steiner points during the construction of RMST in presence of obstacles in comparison to the standard solution available in absence of obstacles. Experimental results are quite encouraging.

Keywords: Obstacle aware routing, Routing in 3D ICs, Shortest pair routing.

1 Introduction

Three-dimensional (3D) technology can be broadly defined as any technology that stacks semiconductor elements on top of each other and utilizes vertical, as opposed to peripheral, interconnects between the wafers. 3D integration technology offers the promise of being a new way of increasing system performance even in the absence of scaling. This promise is due to a number of characteristic features of 3D integration, including (a) decreased total wiring length, and thus reduced interconnect delay times; (b) dramatically increased number of interconnects between chips; and (c) the ability to allow dissimilar materials,

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 649–658.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

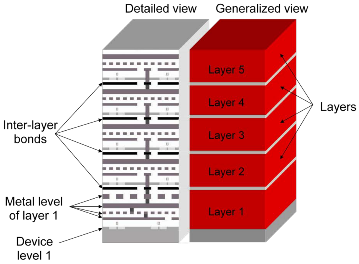


Fig. 1. 3D integration technique

process technologies, and functions to be integrated. 3D ICs have three main process components: (a) a vertical interconnect, (b) aligned bonding, and (c) wafer thinning with backside processing. The order of these steps depends on the integration approach chosen, which can depend strongly on the end application. 3D integration technique may be illustrated with figure 1.

2 Background and Motivation

During the routing phase of physical design cycle of chip design, building an RSMT i.e. rectilinear Steiner minimum tree is considerably a difficult problem. Initially, the RSMT issue did not assume any obstruction in the routing area. But today one has to consider Obstacle aware RSMT (OARSMT), as the current VLSI designs consists of several obstructions viz. pre-routed nets, macro cells, IP blocks, etc. Both RSMT and OARSMT are NP-complete issues. Researchers have been paying real attention to this problem. In [1] [2] [3] [4] researchers have worked in their own ways and have applied their own techniques to resolve this matter. From introducing the track graph and then followed by contributions like making FORst (hierarchical heuristic), An-OARSMan (non-deterministic heuristic), CDCTree, the connection graph, AMAZE (quick maze routing based algorithm), etc; their contributions have helped to move towards progress. RSMT for a given terminal set is the fundamental problem in integrated circuit computer aided design (IC CAD). Routing a net is an important issue in VLSI physical design. In 3D this problem becomes more complex due to close proximity of the modules and thereby increasing the congestion. In the presence of obstacles this routing process becomes more complex and challenging.

3 Problem Formulation

The description of the present problem of interest may be stated as follows.

3.1 Problem in 2D

Let us consider a net is given with number of terminal points with their coordinates. Obstacles are also given those are distributed across the routing region. We have to generate a rectilinear Steiner tree by connecting all the terminal points and bypassing the obstacles (in Manhattan manner) such that the total cost is minimum. When we connect the two terminals in Manhattan manner the obstacle may fall on the connecting path. Both the cases in the figure 2 indicate that how two points can be connected with minimum cost. Detailed heuristic for avoiding the regions are described in the following section.

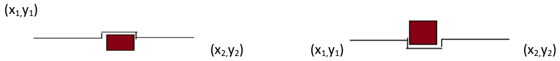


Fig. 2. Avoiding obstacles during routing in a layer

Cost function calculation: During the decision making in the proposed heuristic for selecting the shortest pair for joining initially the Euclidean distances between the pairs of points are computed. But during routing that path will be replaced by the Manhattan path and the total wirelength is also computed in that manner.

3.2 Problem in 3D

Let us consider m number of terminals and s number of obstacles distributed throughout the n layers (considering in 3D there are n layers available). It may so happen that n_i layer may contain m_1 terminals and s_1 obstacles. Also the generated rectilinear Steiner tree is also distributed across the layers. Obstacles are considered similarly as described in previous case in 2D and also treated as described previously. To connect the layer we need TSV (through silicon via), and connections are done in such a fashion that the TSV cost is minimum.

Cost function: Tree can be generated within the single layer or can be spread on different layers. Hence, total cost may be assumed to be divided into two parts viz. inter layer and intra layer. When cost is calculated within a single layer it is intra layer and when calculated in different layers it is inter layer. So the total cost is the summation of both, i.e.

$$Cost_{total} = \Sigma_{layeri=1}^{n}(\Sigma_{allpairs} cost_i) + \Sigma_{layeri=1}^{n}(\Sigma_{layerj=1,j\neq i}^{n} cost_{i,j}) \tag{1}$$

Two layers are given those are connected through TSV. Red portions represents obstacles. These are bypassed using the proposed heuristic. Cost is also calculated accordingly.

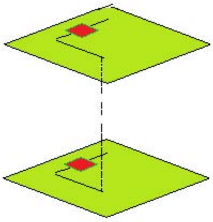


Fig. 3. Avoiding obstacles during routing in 3D

4 Proposed Solution

Let us consider the following table 1 generated from the connectivity information supplied for the net under consideration. Here x_i, y_i represents the coordinates of the terminals and $d_{i,j}$ represents the distance between a pair of ith and jth points. The basic scheme for the general flow of the algorithm may be described textually as follows.

Step 1: for $d_{i,j}$ if i=j, then set $d_{i,j} = 0, distance = 0$
Step 2: Set $d_{i,j} = d_{j,i}$
Step 3: Let $d_{2,4}$ is the minimum distance within the matrix. Then the coordinate corresponding to the distance $d_{2,4}$ is (x_2, y_2) and (x_4, y_4). Connect them through downward tree manner. If an obstacle is found in this path connect it in upward tree manner. If no obstacle is found connect both points in downward tree manner. Different cases are illustrated in the following figures. Set the flag for (x_2, y_2) and (x_4, y_4) to indicate the these points are visited (connected).
Step 4: After reaching (x_4, y_4) we consider the next shortest distance among (x_4, y_4) and rest of the terminals with checking the flag (so that already visited terminals are not calculated again). Now we get the next shortest distance say $d_{5,4}$. Hence the coordinate connecting this distance is (x_4, y_4) and (x_5, y_5). Set the flag as in step 3.
Step 5: Repeat the step 3 and 4 until all terminal point are reached.
Step 6: calculate the cost when we connect the terminals in appropriate manner described early.

Pseudo codes of the entire proposed algorithm with different procedures are described in Algorithms 1, 2, 3, 4, and 5.

Table 1. Table representing the distance matrix

	x_1, y_1	x_2, y_2	x_3, y_3	x_4, y_4	x_5, y_5
x_1, y_1	$d_{1,1}$	$d_{1,2}$	$d_{1,3}$	$d_{1,4}$	$d_{1,5}$
x_2, y_2	$d_{2,1}$	$d_{2,2}$	$d_{2,3}$	$d_{2,4}$	$d_{2,5}$
x_3, y_3	$d_{3,1}$	$d_{3,2}$	$d_{3,3}$	$d_{3,4}$	$d_{3,5}$
x_4, y_4	$d_{4,1}$	$d_{4,2}$	$d_{4,3}$	$d_{4,4}$	$d_{4,5}$
x_5, y_5	$d_{5,1}$	$d_{5,2}$	$d_{5,3}$	$d_{5,4}$	$d_{5,5}$

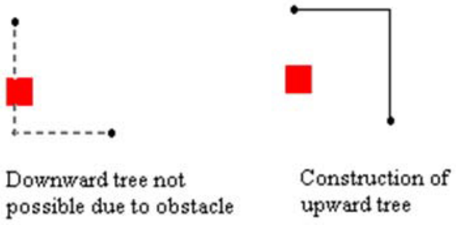

Downward tree not possible due to obstacle
Construction of upward tree

Fig. 4. Case I: Routing tree construction with obstacles

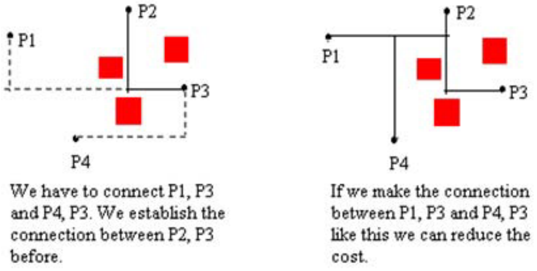

P2
P1
P3
P4
We have to connect P1, P3 and P4, P3. We establish the connection between P2, P3 before.
P2
P1
P3
P4
If we make the connection between P1, P3 and P4, P3 like this we can reduce the cost.

Fig. 5. Case II: Routing tree construction with obstacles

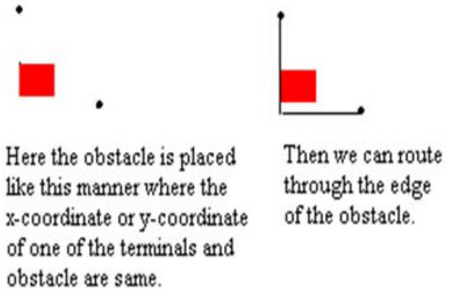

Here the obstacle is placed like this manner where the x-coordinate or y-coordinate of one of the terminals and obstacle are same.
Then we can route through the edge of the obstacle.

Fig. 6. Case III: Routing tree construction with obstacles

```
Input  : The adjacency matrix of 2D netlist, Obstacles, Terminals
Output: Optimum global routing path with RMST for the given netlist

Initialize();
/* Initialize matrix and all loop variables */ Generate_matrix();
Compute d;
Find_d_min ();
Expand_tree();
```

Algorithm 1. Algorithm for obstacle aware routing tree construction

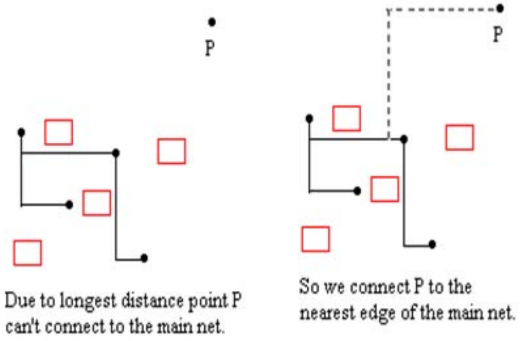

P
P
Due to longest distance point P can't connect to the main net.
So we connect P to the nearest edge of the main net.

Fig. 7. Case IV: Routing tree construction with obstacles

```
Find_d_min ();

Initialize loop variable and set minimum = 32000;
Find out the minimum distance from the matrix except 0;
Establish connection between these two points in Manhattan manner;
/* Use functions tree_downward() or tree_upward() */ Calculate_cost (pt1, pt2);
Set_flag();
/* Set a flag for used vertex.*/
```

Algorithm 2. Procedure for finding minimum distance

```
expand_tree(int point1);

From point1 successively calculate the minimum value from the matrix;
Establish connection between these two points in Manhattan manner;
/* Use functions tree_downward() or tree_upward() */ Calculate_cost (pt1, pt2);
Set_flag();
/* Set a flag for used vertex.*/
```

Algorithm 3. Procedure for expanding the routing tree

5 Experimental Results

Proposed algorithm has been implemented in C++ under Linux GNU GCC environment and executed using an Intel 3 GHz processor chip with 1 GHz memory. Experimental results have been summarized in tables 2 and 3. Total wire-length cost comprising of intra-layer as well inter-layer costs for all the layers and layer combinations has been represented as Total cost in tables. The fields $L < layer - no > cost$ denotes the total intra-layer wiring cost for that particular layer. $L < layer1 > \_ < layer2 > cost$ fields represents the total inter-layer cost of interconnects for that layer combinations. #SP denotes the

```
obs_check();

if(terminals and Steiner coordinates lies between obstacles coordinate by
tree_downwards()) Set flag_1 = 1;
Consider tree_upwards();
else set flag_1 = 0;
if(terminals and Steiner coordinates lies between obstacles coordinate by
tree_upwards()) Set flag_2 = 1;
else Set flag_2 = 0;
Case 1: if(flag_1==1 and flag_2==0)
Downward connection not possible
Take upward connection;
Case 2: if(flag_1==0 and flag_2==1)
Upward connection not possible
Take downward connection;
Case 3: if(flag_1==0 and flag_2==0)
Both connections are possible
Take downward connection;
Case 4: if(flag_1==1 and flag_2==1)
Bypass obstacle to connect;
```

Algorithm 4. Procedure for checking obstacles

```
Calculate_cost(pt1, pt2);

if(level of point1== level of point2)
Calculate Euclidean distance;
Store into intra_layer cost;
else
Calculate Euclidean distance;
Store into inter_layer cost;
Add total intra_layer cost and total inter_layer cost;
Store into total_cost;
```

Algorithm 5. Procedure for cost calculation

number of Steiner points generated by the algorithm for construction of the entire Steiner tree. CPU time is noted for each and every execution to get an idea of the computational complexity of the proposed algorithm. From a close inspection of the tables reported it is clearly seen that the proposed algorithm is pretty much efficient as far as the execution time is concerned. Execution time does not exceed 0.35 seconds even for Steiner routing among a net consisting of 100 number of terminals and 50 number of obstacles. Another important point noticeable from the experimental results is that number of Steiner points generated during the construction of Steiner tree has decreased for the same

Table 2. Experimental Results for Random benchmarks with no obstacles

#Terminals	Total Cost	L1 cost	L2 cost	L3 cost	L1_2 cost	L2_3 cost	L1_3 cost	#SP	CPU Time
10	85	21	23	0	39	2	0	9	0m0.002s
20	168	6	47	4	35	34	42	19	0m0.002s
30	194	5	27	14	35	60	53	29	0m0.007s
40	181	29	36	0	23	58	35	39	0m0.017s
50	202	33	25	23	24	47	50	49	0m0.046s
60	260	40	34	37	77	44	28	59	0m0.077s
70	256	22	25	33	59	66	51	69	0m0.106s
80	319	42	46	60	41	59	71	79	0m0.169s
90	320	20	30	52	111	36	71	89	0m0.237s
100	346	32	31	51	78	49	105	99	0m0.338s

Table 3. Experimental Results for Random benchmarks in presence of obstacles

#Terminals	#Obstacles	Total Cost	L1 cost	L2 cost	L3 cost	L1_2 cost	L2_3 cost	L1_3 cost	#SP	CPU Time
10	5	120	0	6	0	98	10	6	9	0m0.002s
	10	104	9	28	0	56	5	6	9	0m0.002s
20	5	117	17	23	8	28	35	6	19	0m0.003s
	10	111	4	31	0	27	19	30	19	0m0.005s
	15	163	0	15	0	109	21	18	18	0m0.003s
30	10	168	9	6	6	64	79	4	29	0m0.008s
	15	182	11	27	23	72	38	11	28	0m0.008s
	20	164	0	11	23	36	46	48	28	0m0.007s
40	10	249	13	18	3	82	117	16	38	0m0.022s
	20	243	12	58	16	67	62	28	38	0m0.021s
	30	190	12	28	12	35	74	29	37	0m0.024s
50	10	246	7	8	11	66	91	63	49	0m0.039s
	20	216	10	12	12	65	79	38	48	0m0.061s
	25	218	32	19	24	36	58	49	48	0m0.042s
	30	195	36	18	25	39	52	25	48	0m0.051s
60	10	245	29	18	26	69	55	48	59	0m0.066s
	20	241	4	16	8	74	69	70	59	0m0.075s
	30	247	43	32	29	48	60	35	58	0m0.077s
70	10	277	28	52	34	56	55	52	59	0m0.110s
	25	262	23	35	41	57	77	29	69	0m0.132s
	35	266	12	33	66	67	40	48	68	0m0.114s
80	10	254	16	21	31	73	44	69	74	0m0.159s
	30	313	22	12	50	129	46	54	75	0m0.184s
	40	312	39	21	43	41	71	97	69	0m0.159s
90	20	298	21	21	35	59	79	83	77	0m0.249s
	30	318	34	33	37	32	91	91	76	0m0.209s
	45	306	43	29	34	43	83	74	80	0m0.225s
100	20	226	24	12	24	72	40	54	72	0m0.246s
	35	329	31	19	31	86	59	103	95	0m0.315s
	50	330	28	35	65	70	72	60	87	0m0.327s

number of terminals in presence of obstacles in comparison to the situation when no obstacles were present. This again proves the efficiency and effectiveness of the proposed algorithm as the design cost rapidly increases with the increase in number of Steiner points in the layout.

6 Conclusions and Future Works

In this paper, we propose a method for routing of nets spanned across the entire 3D architecture in presence of obstacles, and perform empirical study of interconnect lengths for routing within the individual layers as well as inter-layer routing cost contributing towards the cost for TSV across different device layers. Major achievements of the study include (i) To construct a routing path starting from the smallest distance terminals and complete the tree from that terminals, (ii) To construct a collision free Steiner tree by considering the closest coordinates using an efficient cluster formation technique, (iii) For construction of the Steiner tree Manhattan routing is considered, which is, well adopted by the industry, and (iv) the proposed algorithm is very fast as far as time complexity is concerned, which is, an essential requirement in computer aided electronic design automation tools handling with ever increasing design complexity nowadays.

In the present work, we have assumed the obstacles to be of square shaped with a very small dimension which is further customizable by the user and the routing path is Manhattan i.e. we can route the terminals only either in 0 or 90 degree orientation. But obstacles may be irregular in shape in practical life. For that purpose also, a large non-uniform obstacle may be considered as a summation of a number of small regular obstacles used in this work, and then also the algorithm works well. However, the work may be extended towards the non-uniform obstacles also, where, initially the shape and size of the obstacles might be considered and thereby reducing the overall algorithmic time complexity. Another future direction of this present work might be to consider the non-Manhattan routing paths also viz. X or Y routing techniques too. Besides these the work may be extended to study the behaviour for multiple nets, and observation of other parameters such as congestion, signal integrity, considering thermal issues, and so on.

References

1. Li, L., Young, E.F.Y.: Obstacle-avoiding Rectilinear Steiner Tree Construction. In: Proceedings of International Conference on Computer-Aided Design (ICCAD), pp. 523–528 (2008)
2. Huang, T., Young, E.F.Y.: Obstacle-avoiding Rectilinear Steiner Minimum Tree Construction: An Optimal Approach. In: Proceedings of International Conference on Computer Aided Design (ICCAD), pp. 610–613 (2010)
3. Hu, Y., Feng, Z., Jing, T., Hong, X., Yang, Y., Yu, G., Hu, X., Yan, G.: FORst: A 3-step heuristic for obstacle-avoiding rectilinear Steiner minimum tree construction. Journal of Information & Computational Science 1(3), 107–116 (2004)

4. Liu, J., Zhao, Y., Shragowitz, E., Karypis, G.: A Polynomial Time Approximation Scheme for Rectilinear Steiner Minimum Tree Construction in the Presence of Obstacles. In: 9th International Conference on Electronics, Circuits and Systems, vol. 2, pp. 781–784 (2002)
5. Sapatnekar, S., Goplen, B.: Placement of 3D ICs with thermal and inter-layer via considerations. In: Design Automation Conference, pp. 626–631 (June 2007)
6. Kahng, A.B., Robins, G.: A New Class of Steiner Tree Heuristics with Good Performance: The Iterated 1-Steiner approach. In: International Conference on CAD (1990)
7. Xie, Y., Cong, J., Sapatnekar, S. (eds.): Three-Dimensional Integrated Circuit Design Series: Integrated Circuits and Systems. Springer (2009)
8. `http://www.vlsicad.eecs.umich.edu/BK/PDtools/`
9. Deng, Y., Maly, W.: Interconnect Characteristics of 2.5d system integration scheme. In: ACM International Symposium on Physical Design, pp. 171–175 (April 2001)
10. Shi, Y., Mesa, P., Yu, H., He, L.: Circuit-Simulated Obstacle-Aware Steiner Routing. ACM Transactions on Design Automation of Electronic Systems 12(3), Article 28 (August 2007)
11. Yan, J.-T., Ming-Ching, Zhi, J., Chen, W.: Obstacle-Aware Longest Path using Rectangular Pattern Detouring in Routing Grids. In: Asia and South Pacific Design Automation Conference, ASPDAC (2010)

Study of the EEG Signals of Human Brain for the Analysis of Emotions

Ashish R. Panat[1] and Anita S. Patil[2]

[1] Principal. Priyadarshani Indira College of Engineering, Nagpur, India
ashishpanat@gmail.com
[2] Cummins College of Engineering for Women, Pune, India
anita.patil@cumminscollege.in

Abstract. In this research, the emotions and the patterns of EEG signals of human brain will be studied. The aim of this research is to study the analysis of the changes in the brain signals in the domain of different emotions. The observations can be analysed for its utility in the diagnosis of psychosomatic disorders like anxiety and depression in economical way with higher precision.

Keywords: EEG, EDF format, Feature extraction, Image classifiers, Emotions, Psychosomatic disorders, Normal and excited brain.

1 Introduction

Brain is the organ that makes us human, giving people the capacity for art, language, moral judgments, and rational thoughts. It is also responsible for each individual's personality, memories, movements, and his perception about the world. It is one of the body's biggest organs, consisting of some 100 billion nerve cells that not only put together and highly coordinated physical actions but regulate our unconscious body processes, such as digestion and breathing.

Emotions play a powerful and significant role in everyday life of human beings. Impulsive emotions express an indication of psychosomatic disorders. These disorders can be reflected as the changes in the brain signals and images.

Anxiety disorder: The term anxiety disorder covers several different forms of abnormal and pathological fear and anxiety. It covers four aspects of experiences an individual may have: mental apprehension, physical tension, physical symptoms and dissociative anxiety [1]. Anxiety disorder is divided into three types viz, anxiety disorder, phobic disorder, and panic disorder; each has its own characteristics and symptoms. They also require different treatment. The emotions present in anxiety disorders range from simple nervousness to bouts of terror. The amygdala is central to the processing of fear and anxiety, and its function may be disrupted in anxiety disorders.

Depression: Depression is a state of low mood and aversion to activity that can affect a person's thoughts, behaviour, feelings and physical well-being [2]. Depressed people may feel sad, anxious, empty, worthless, guilty, irritable, or restless. They may lose

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 659–665.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

interest in activities that once were pleasurable, experience loss of appetite or overeating, or problems concentrating, remembering details or making decisions. They may even contemplate or attempt suicide. Insomnia, excessive sleeping, fatigue, loss of energy, or aches, pains or digestive problems that are resistant to treatment may be present. The signals of the brain in these situations can also be utilized to study the emotions which can lead to great help in diagnosis of psychosomatic disorders.

The research is conducted previously to analyse the emotions by looking at the physiological aspects like users' heart rate, skin conductance and pupil dilation.

2 Literature Survey

In [3], Researchers Z. Khalili et al. have worked on Emotion detection using EEG and peripherals signals as Galvanic skin resistance, Respiration, Blood pressure, and Temperature. From these inputs, common set of features such as Mean, standard deviation and minimum and maximum of the set of data are extracted.

In [4], their research is extended to study the improvement in the results of EEG by using correlation dimension. In [5], the researchers have explored on different modalities for emotion detection, such as, Visual (facial expression), Auditory (pitch, loudness, etc.), tactile (heart rate, skin conductivity etc.) and Brain signals (EEG). In [6], Researchers have studied Brain activation during judgments of Positive emotions: Pride and Joy. They have used fMRI images for this purpose.

Researchers Arman Savran et al. [7] have developed a technique for multimodal emotion detection using fNIRS, face video and EEG signal.

3 Block Diagram

Figure(1) shows the block diagram of the system to process the captured signals from Brain e.g. EEG in this study, extract the features after processing the signal, classify the processed signal and analyse it for emotion detection.

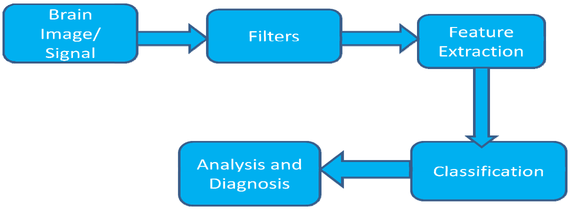


Fig. 1. Block Diagram of the system to analyse Brain signals

Initially, the EEG of the Brain is captured by the standard method recognized worldwide as International 10-20 system. In the Digital EEG system, these signals are first amplified and then digitized. The rate of digitization may vary from 100 Hz to 20 kHz, depending on the capacity of the system.

The EEG signal from the EEG machine is available in the EDF format (European Data Format). It is first converted into .Wav format which is suitable for processing. The signals are then filtered. The pass band of the filter depends on the frequency of the interest for that particular signal. After filtering the signal, the feature extraction process can be done.

4 Capturing the Signal

The fig.(2) shows the placement of the electrodes for capturing the EEG.

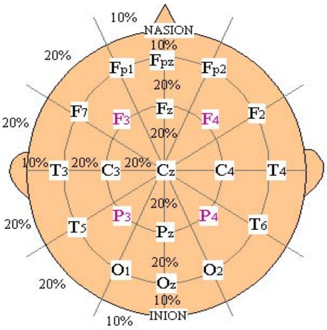


Fig. 2. Placement of electrodes for EEG [10]

As per the International Standard, known as international 10-20 system, 19 electrodes are connected to different locations on the scalp, which are salient points from the clinical point of view, plus one electrode is connected to ground. Whenever, the detailed study of EEG is intended in case of some patients, or for the research purpose, the number of electrodes may increase up to 256 also, where as in case of neonatal EEG, the number can be decreased.

5 Preprocessing

The EEG signal from the EEG machine is available in the EDF format (European Data Format). It is first converted into .Wav format which is suitable for processing. The fig (3) shows the flowchart of Pre-processing followed by feature extraction.

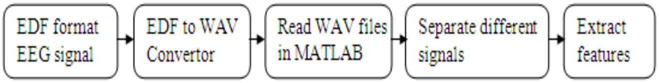


Fig. 3. Pre-processing and feature extraction

6 Feature Extraction

After pre-processing, the following features are extracted from this signal: Mean, Standard deviation, skewness, kurtosis, mean of absolute values of first difference of raw signals, mean of absolute values of first difference of normalized signal [3]. The images of these signals can be stored for further analysis. For feature extraction wavelet transform can be used.

Discrete wavelet Transform: The basic idea in the DWT for a one dimensional signal is as explained here. A signal is split into two parts, the high frequencies and the low frequencies. The edge components of the signal are largely confined to the high frequency component. The low frequency part is split again into two parts as high frequencies and low frequencies. This process is continued an arbitrary number of times, which is usually determined by the application at hand. Furthermore, from these DWT coefficients, the original signal can be reconstructed. This reconstruction process is called as inverse DWT (IDWT). The DWT and IDWT can be mathematically stated as follows.

Let

$$H(\omega) = \sum_k h_k e^{-jk\omega} \quad \text{and} \quad G(\omega) = \sum_k g_k e^{-jk\omega} \tag{1}$$

be a low pass and a high pass filter, respectively, which satisfy a certain condition for reconstruction.

A signal, x[n] can be decomposed recursively as

$$c_{j-1,k} = \sum_n h_{n-2k}\, c_{j,n} \tag{2}$$

$$d_{j-1,k} = \sum_n g_{n-2k}\, c_{j,n} \tag{3}$$

For j= J+1, J…, J_0 where $C_{j+1,}$ k = x[k], k ∈ Z;

J+1 is the high resolution level index and *J0* is the low resolution level index. The coefficients $C_{J0,\,k}$, $d_{J0,k}$, $d_{J0+1,k}$, …., $d_{J,\,k}$ are called the DWT of signal x[n] where $C_{J0,\,k}$ is the lowest resolution part of x[n] and $d_{J,,\,k}$ are the details of x[n] at various bands of frequencies. Furthermore, the signal x[n] can be reconstructed from its DWT coefficients recursively.

$$c_{j,n} = \sum_k h_{n-2k} c_{j-1,k} + \sum_k g_{n-2k} d_{j-1,k} \tag{4}$$

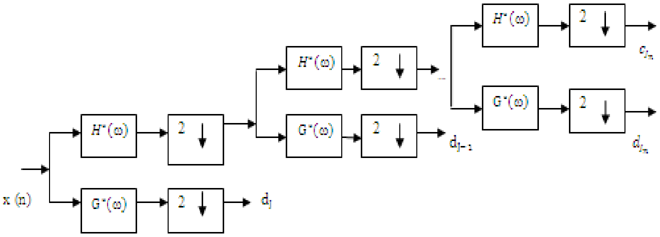


Fig. 4. Decomposition of signal by DWT

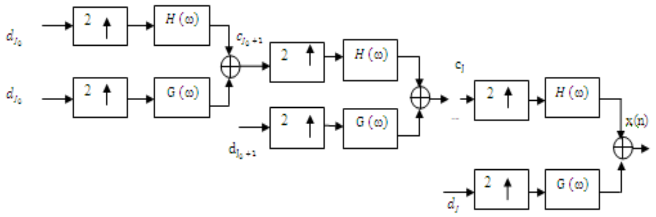


Fig. 5. Reconstruction of signal by IDWT

Then the classification of image is done by Image Classifiers.

Different techniques used by different researchers are, viz., Linear Discriminate Analysis (LDA) and *k*-th nearest neighbour (*k*-NN).

LDA: **Linear Discriminant Analysis (LDA)** and the related **Fisher's linear discriminant** are methods used in statistics, pattern recognition and machine learning. They are usually used to find a linear combination of features which characterizes or separates two or more classes of objects. The resulting combination may be used as a linear classifier, most of the times for dimensionality reduction before later classification. LDA is suitable when the measurements made on independent variables for each observation are continuous quantities [8].

k-NN: In pattern recognition, the ***k*-nearest neighbour algorithm** (*k*-NN) is a method commonly used for classifying objects based on closest training examples in the feature space. *k*-NN is a type of instance-based learning, also called as lazy learning where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is one of the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its *k*-nearest neighbours (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of its nearest neighbour [9].

7 Results

The following snap shots from MATLAB result windows show the conversion of EDF file (Fig.6) to .WAV file (Fig.7), band pass filter applied to theta wave (Fig.8), Delta wave (Fig.9).

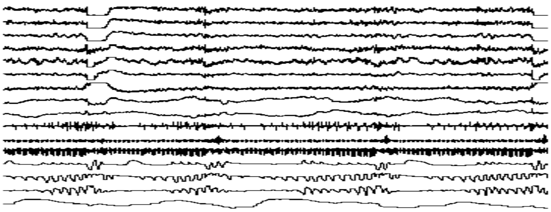


Fig. 6. Signal extracted from .EDF file

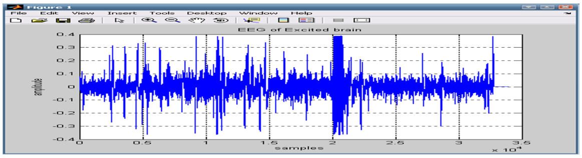
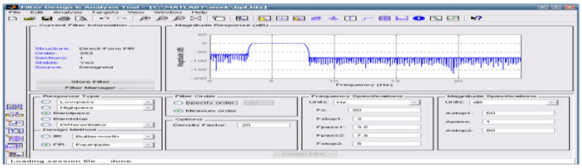


Fig. 7. Result of MATLAB code for EEG in .WAV format



Structure: Direct-Form FIR
Order: 263
Sections: 1
Stable: Yes
Source: Designed

Response Type: ● Bandpass
Design Method: ● FIR Equiripple
Filter Order: ● Minimum order
Density Factor: 20

Units: Hz
Fs: 50
Fstop1: 3
Fpass1: 3.5
Fpass2: 7.5
Fstop2: 8

Units: dB
Astop1: 60
Apass: 1
Astop2: 60

Loading session file ... done.

Fig. 8. Result of Band pass filter for Theta wave (using FDA tool)

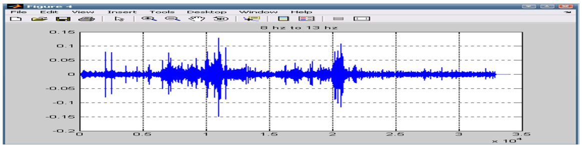


Fig. 9. Result for Delta wave (using FDA tool)

8 Conclusion

The study has proved the effective utility of economical and simple method of study of brain using EEG for diagnosis the two different emotion disorders viz., anxiety and depression. The EEG signal in EDF format is converted into .WAV format using EDF to WAV converter. The signal is then passed through the filters of different frequencies to separate alpha, beta, delta and theta waves. This technique has revealed the possibility of precise diagnosis of psychosomatic disorders in more simple and economical way.

References

1. Healy, D.: Psychiatric Drugs Explained, Section 5: Management of Anxiety. Elsevier Health Sciences, 136–137 (2008)
2. Kessler, R.C., Chiu, W.T., Demler, O., Merikangas, K.R., Walters, E.E.: Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Arch. Gen. Psychiatry 62(6), 617–627 (2005) PMC 2847357. PMID 15939839, doi:10.1001/archpsyc.62.6.617
3. Khalili, Z., Moradi, M.H.: Emotion detection using brain and peripheral signals. In: Proceedings of the CIBEC 2008, Biomedical Engineering Faculty, Amirkabir University of Technology, Tehran, Iran (2008) 978-1-4244-2695-9/08/$25.00 ©2008 IEEE

4. Khalili, Z., Moradi, M.H.: EEG Model and Location in Brain when at Emotion Recognition System Using Brain and Peripheral Signals using Correlation Dimension to Improve the Results of EEG. Biomedical Engineering Faculty, Amirkabir University of Technology, Tehran, Iran (2009) 978-1-4244-3553-1/09/$25.00 © 2009 IEEE
5. Special Issue on Multimodal affective Interaction : IEEE Transaction on Multimedia 12(6) (October 2010)
6. Takahashi, H., Matsuura, M., et al.: Brain Activation during Judgments of Positive Self-conscious emotion and Positive Basic Emotion: Pride and Joy. Cerebral Cortex 18, 898–903 (2008)
7. Savran, A., Ciftci, K., et al.: Emotion Detection in Loop from Brain signals and Facial Images. In: Enterface 2006, Dubrovnik, Croatia, July17 (2006)
8. EMcLachlan: Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience (2004)
9. Bremner, D., Demaine, E., Erickson, J., Iacono, J., Langerman, S., Morin, P., Toussaint, G.: Output-sensitive algorithms for computing nearest-neighbour decision boundaries. Discrete and Computational Geometry 33(4), 593–604 (2005)
10. `ftp://sigftp.cs.tut.fi/pub/eeg-data/data`
11. `http://www.cs.colostate.edu/eeg/eegSoftware.html`
12. `http://www.cs.colostate.edu/eeg/code/tutorial.html`
13. `http://serendip.brynmawr.edu`

A Dynamic Approach for Mining Generalised Sequential Patterns in Time Series Clinical Data Sets

M. Rasheeda Shameem[1], M. Razia Naseem[1], N.K. Subanivedhi[1], and R. Sethukkarasi[2]

[1] Department of Computer Science and Engineering,
R.M.K Engineering College,
Kavaraipettai, Tamil Nadu, India
`{ras.shameem,naseemrazia,nk.subanivedhi}@gmail.com`
[2] Department of Information Science and Technology,
Anna University,
Tamil Nadu, India
`sethumaaran@yahoo.co.in`

Abstract. Similarity based stream time series is gaining ever-increasing attention due to its importance in many applications such as financial data processing, network monitoring, Web click-stream analysis, sensor data mining, and anomaly detection. These applications require managing data streams, i.e., data composed of continuous, real-time sequence of items. We propose a technique for pattern matching within static patterns and stream time series clinical data sets. The main objective of our project is to ascertain hidden patterns between incoming time series clinical data sets and the set of predetermined clinical patterns. By considering the incoming image data at a particular timestamp, we construct a MultiScale Median model at multiple levels to adapt to the stream time series, characterized by frequent updates. Further, we employ a pruning algorithm, Segment Median Pruning on clinical Image data for pruning all candidate patterns. Experiments have been carried out on retinal disease data set known as Age Related Macula Degeneration (ARMD) and simulation results show that the system is efficient in processing image data sets for making efficient and accurate decision.

Index Term: Pattern match, stream time series, MultiScale Median.

1 Introduction

Time series data have a natural temporal ordering as it is a sequence of data points, measured typically at successive times spaced at uniform time intervals. There are two main goals of time series analysis[1]: identifying the nature of the phenomenon represented by the sequence of observations, and forecasting (predicting future values of the time series variable). Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, we can interpret and integrate it with other data (i.e., use it in our theory of the investigated phenomenon, e.g., seasonal commodity prices). Regardless of the

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 667–674.
springerlinkc.om © Springer-Verlag Berlin Heidelberg 2013

depth of our understanding and the validity of our interpretation (theory) of the phenomenon, we can extrapolate the identified pattern to predict future events.

Time series is said to be efficacious in clinical domain when compared to other models since the temporal nature of incoming data is taken into account. We are focusing our work mainly on a retinal disease, Age related macular degeneration (ARMD). It is the most common cause of vision loss in those aged over 50. It causes a gradual loss of central (but not peripheral) vision. The disease does not lead to complete blindness. Visual loss can occur depending on the type and severity of ARMD[2]. There are two main types of ARMD - wet and dry. 'Wet' ARMD is most severe. The disease begins with characteristic yellow deposits in the macula known as drusen. The drusen can be categorised as hard and soft drusen. Hard drusen have a well defined border, while soft drusen have boundaries that often blend into the background. In early stage, it is often difficult to detect the drusen. This serves as a motivation behind the proposed approach to provide technical support eliminating the necessity for manual assessment of ophthalmic images in diagnostic practices. In Fig 1 the circled area represents drusen.

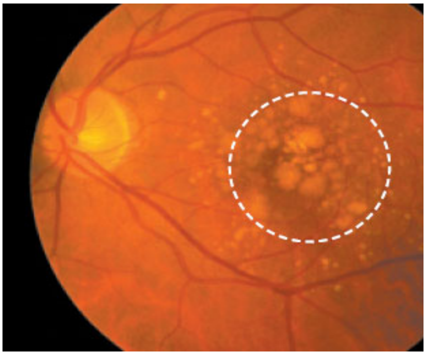


Fig. 1. Retinal image showing many drusens in the macula

We propose an efficient approach, MultiScale Median (MSM) representation to analyze image data streams and to facilitate similarity matching. We adopt pruning algorithm namely Segment Median Pruning on Clinical Image data(SMPCI) which reduces the search space by eliminating the false patterns, thus saving computational cost. Resulting patterns are compared with the predetermined patterns to retrieve candidates based on Lp-norms for p >= 1.

2 Related Work

This section overviews previous work on similarity search over archived time-series data. Many approaches have been proposed for performing similarity search.

Karam Gouda1 and Mosab Hassaan[3] proposed sequential pattern mining in very large databases using a new version of spade algorithm known as dSpade. Agarwal, Faloutsos and Arun Swami[4] proposed an indexing method for time sequences for processing similarity queries. In their work only the stronger coefficients of Discrete Fourier Transform (DFT) were used to map time sequences R* trees were used to efficiently answer similarity queries.

Ahmet Bulut and Ambuj [5] proposed a similarity search technique for online monitoring of data streams. A multi resolution indexing scheme used, provides high quality answers to aggregate monitoring queries, such as finding surprising levels of a data stream and similarity queries, such as finding interesting patterns. Yunyue Zhu and Dennis Shasha[6] proposed a technique for finding a song by analyzing humming part of the tune. Music was treated as the time series data. Improved Dynamic Time Warping (DTW) with the concept of envelope transforms was used for extending existing dimensionality reduction methods to DTW indexes.

A similar approach of automated diagnosis of ARMD was proposed [7] where the detection process involves discrete wavelet transform (DWT). We attempt to perform feature extraction adopting MSM instead of DWT.

3 Overview of the Proposed Project

Input to the system is retrieved from retinal image data set which will be viewed as time series image data sets. The first step of proposed system is preprocessing. Preprocessing is done to improve the quality of the image. It includes, extracting green component from the color fundus images followed by histogram equalization to improve the contrast and finally anisotropic diffusion to remove the noise. Next step is the MSM representation to facilitate feature extraction. Discrete Wavelet Transform (DWT) produces blurring and ringing noise near edge regions in images. MSM performs well when compared to DWT and requires less processing time. Our work assumes that trained pattern data sets are stored in a two dimensional grid index. Segment mean pruning is then adopted to generate matching candidate pattern with sliding window. Final step is the pattern matching which retrieves the matching pair.

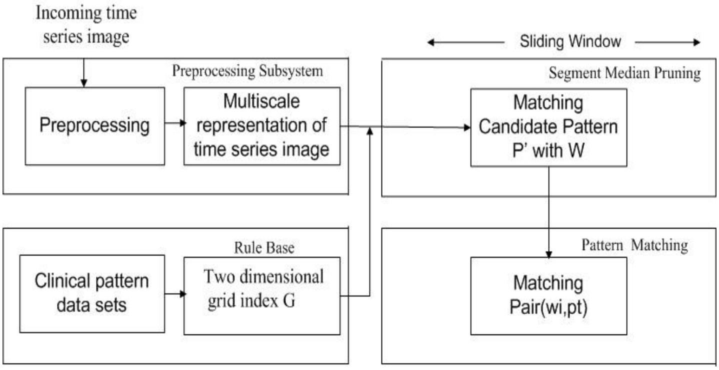

Incoming time series image
Sliding Window
Preprocessing Subsystem
Preprocessing
Multiscale representation of time series image
Segment Median Pruning
Matching Candidate Pattern P' with W
Rule Base
Clinical pattern data sets
Two dimensional grid index G
Pattern Matching
Matching Pair(wi,pt)

Fig. 2. The general framework of the system

4 Methodology

The approach consists of the following steps:

1. Preprocessing techniques are applied to the stream time series retinal images.
2. The processed image is represented using MSM approximation.
3. Segment mean pruning performs pattern matching over the MSMI representation.
4. Finally pattern matching is done to generate matching pair.

4.1 Preprocessing

Pre-processing is the first step in diagnosis of retinal images. The images are acquired using fundus camera. The quality of the retinal images however is heavily affected by factors that are difficult to control. The main purpose of preprocessing is to remove noise for the reliable feature extraction. Red and blue channels of the color retinal images are noisy or in low resolution. Green channel is therefore selected to offer better discriminatory power between the main retinal anatomy and retinal background. Therefore Green component is extracted after which, histogram equalization is done to enhance the contrast and improve the quality of the retinal image. Anisotropic diffusion, final step in preprocessing, aims to reduce image noise without removing significant parts of the image content, typically edges, lines or other details that are important for the interpretation of the image. [7].

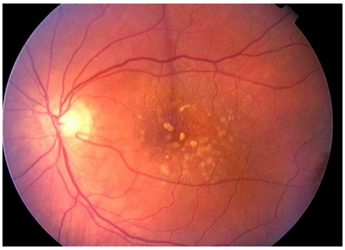


Fig. 3a. Original findus image

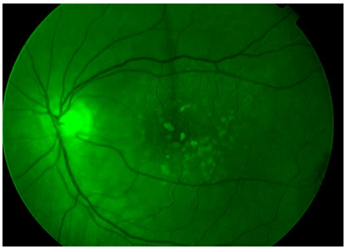


Fig. 3b. Green component extracted image

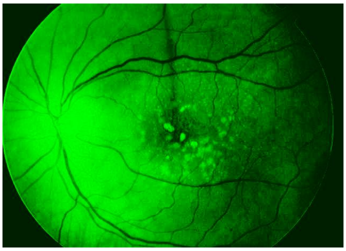


Fig. 3c. Histogram equalized image

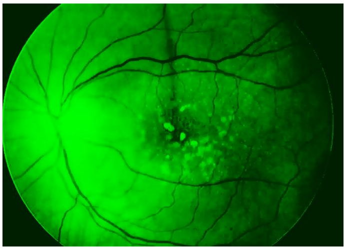


Fig. 3d. Anisotropic diffused image

4.2 MultiScale Median Representation on Clinical Image Data (MSMCI)

In this section, we represent the processed retinal images using MultiScale Median approach to enable the extraction of features for efficient ARMD disease diagnosis. The length of the processed time series images W are considered in the powers of 3 (i.e 3^l where l is the number of levels). Median values are computed considering the pixel values of each image on each level by which MSM of the time series are constructed from level 1 to level l. Level l computes the median values from the images. level l-1 computes from level l and so on up to level 1.[8]

4.3 Segment Median Pruning on Clinical Image Data (SMPCI)

SMPCI reduces the search space by performing pruning on the MSMCI representations. We assume that we have already calculated MSM approximations on any level j (between 1 and l), for every pattern in pattern set consisting of ARMD retinal images P and store it in a grid of dimension 2 along with their pattern identifiers. During the processing SMPCI a subset of (candidate) patterns from P that may match with W.

The algorithm is summarized as follows. It is driven by set of inputs –the Pattern set P, a sliding window W of length w = 3^l, a user defined similarity threshold ε and a 2 dimensional grid index G. It initially computes MSMCI approximation values of patterns in the pattern set and stores it in the grid index G with IDs. We start our pruning from the highest level of approximation, say from the root. Let l_{low2} be the highest level of approximation assign j to it. The pruning procedure repeats with approximations at different levels from j= l_{low} to l and until (G<> Ø). CP (i.e, a set of candidate patterns) are initialized to null. Comparison takes place between MSMCI values of patterns & Sliding Window at level j. If the L_2 norm distance between the pattern values in the grid and the approximation values of sliding window is within $\varepsilon/2^1$ the patterns are stored in CP. The algorithm returns the set of candidate patterns on its completion.

4.4 Pattern Matching

Here, we describe the method of pattern matching which will detect the presence of the ARMD by checking the actual distance between W and P'. Since we consider stream time series new image data can arrive at any time in the processing. To comply with the new arrival, the sliding window estimates the most recent w data values.

To summarize, the pattern matching algorithm involves the following. Incoming stream time series image sets, a set P of patterns and ε serves as the input. The algorithm works until no new data items are found. The sliding window is updated with the current w values, whenever new data item arrives at a particular timestamp. For every new data values, the MSMCI approximation values are calculated. The candidate patterns are pruned with SMPCI. If the real distance between the candidate pattern and the stream series is found to be within ε then the similar pairs (W_i; pt) are retrieved. The output returns matching pattern and stream series set.

4.5 Batch Processing

We have considered only one sliding window from a stream time series arriving at a time. We can optimize this for cases where multiple sliding windows are needed and data are buffered in a pool until enough are collected to minimize the cost of look up for such batch of data we can group sliding windows (say 3) and access the grid index 3 times. For instance, if there are 300 sliding windows we have to look up the gird index the same number of times i.e. 300. We group them into 3, thereby accessing grid only 3 times.

To illustrate this, consider a stream time series S, and a number of consecutive sliding windows $\{W_i, W_{i+1} \ldots .W_{i+m}\}$, which can be viewed as one group to look up GI once. The region that tightly bounds the second l_{min} level MSMs in GI, obtained from consecutive sliding windows $W_{j\ (i<j<i+m)}$ can be viewed as a minimum bounding rectangle. Since we want to retrieve those patterns pt in P whose distance to at least one Wj in bounding rectangle is within ε, we expand the rectangle by extending its length by ε on both sides along each dimension. The resulting rectangle is extended bounding rectangle. Any pattern that falls in the extended bounding rectangle is considered as a candidate that might match with the group of sliding windows [9].

5 Results and Discussions

For the processing of the fundus images, green component was selected since it appears less noisy and has finer detail than the other two primary colors, red and blue. Histogram equalization spreads out the frequency intensity values allowing for areas of lower contrast to gain a higher contrast. Anisotropic diffusion preserves strong edges and enhances the contrast of edges.

We consider the sliding window of length w=3^l. We group the images into three known as segments and compute median of each segment. The MSM values of the patterns are pre-computed and stored in a two dimensional grid. The SMPCI method prunes all candidate patterns by comparing the MSM values of patterns and sliding window at each level within $\varepsilon/2^l$. SMPCI retrieves the patterns efficiently by by comparing only a few levels of approximations thereby reducing the number of comparisons made between actual pattern set and time series data. Compared to DWT, MSM approximation can quickly detect similar patterns over stream time series data and thus requires less processing time.

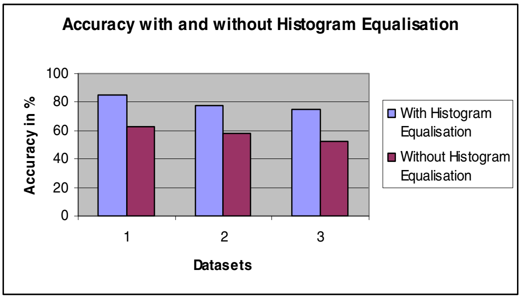


Fig. 4. Comparison of the accuracy of the proposed system with and without performig Histogram equalization in the pre-processing step

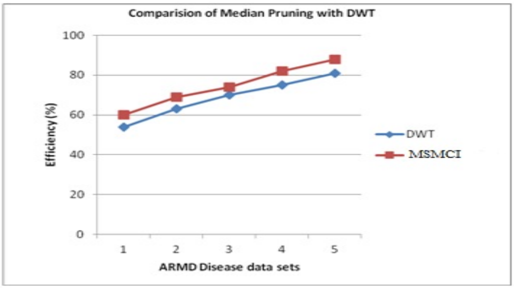


Fig. 5. Comparison of efficiency of the proposed system with DWT

6 Conclusion

The approach described can act as a diagnostic tool for accurate ARMD disease identification and can be domain expert competent. The results also suggest that the approach is time-efficient, which is essential for ophthalmologic applications. Diagnosis of other retinal diseases such as diabetic retinopathy employing the MSMCI and SMPCI technique could be done efficiently.

References

[1] Time series analysis, `http://www.statsoft.com/textbook/time-series-analysis`

[2] Age related macula degeneration, `http://www.patient.co.uk`

[3] Gouda, K., Hassaan, M.: Mining Sequential pattern in dense database. The International Journal of Database Management Systems (IJDMS) 3(1) (February 2011)

[4] Agrawal, R., Faloutsos, C., Swami, A.N.: Efficient Similarity Search in Sequence Databases. In: The Proceedings of the Fourth International Conference on Foundations of Data Organization and Algorithms, pp. 69–84 (1993)

[5] Bulut, A., Singh, A.K.: A Unified Framework for Monitoring Data Streams in Real Time. In: The Proceedings of 21st Internatioal Conference on Data Engineering (ICDE), pp. 44–55 (2005)

[6] Zhu, Y., Shasha, D.: Warping Indexes with Envelope Transforms for Query by Humming. In: The Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 181–192 (2003)

[7] Priya, R., Aruna, P.: Automated diagnosis of age related macular degeneration from color retinal fundus images. In: The Proceedings of the 3rd International Conference on Electronics Computer Technology (ICECT), April 8-10, vol. 2 (2011)

[8] Sethukkarasi, R., RajaLakshmi, D., Kannan, A.: Efficient and fast pattern matching in stream time series image data. In: The Proceedings of the 1st International Conference on Integrated Intelligence computing (ICIIC), August 5-7 (2010)

[9] Lian, X., Chen, L., Yu, J.X., Han, J., Ma, J.: Multiscale Representations for Fast Pattern Matching in Stream Time Series. The IEEE Transaction on Knowledge and Data Engineering 21(4), 568–581 (2009)

[10] Hussain, I., Ali, I., Zubair, M., Bibi, N.: International Conference on Information and Emerging Technologies (ICIET), June 14-16 (2010)

Motion Correction in Physical Action of Human Body Applied to Yogasana

Geetanjali Kale[1] and Varsha Patil[2]

[1] Department of Computer Engineering, PICT, Pune, India
[2] Vice Principal and HOD, Computer Engineering MCOERC, Nashik, India

Abstract. Here, we proposed framework for human motion analysis and efficient representation of motion as well as human body in the form of curve for motion correction. In our proposal Human body is represented by curves for which curve control points are identified, that may or may not be on body curve. Body motion is represented by motion trajectories for each control point. Control point motion is also represented by curves. As system proposes error correction in Yogasana, expert's motion body curves and control point motion trajectory paths and practitioner's motion are compared and error correction is suggested to practitioner. Framework can be applied for all scenes where standard requirements should meet, like sports, acting, physical exercise, and traditional dances.

Keywords: Human motion analysis, motion trajectory, Yogasana, Control points, motion correction.

1 Introduction

Motion related problems are extremely challenging but it has tremendous applications in interdisciplinary areas with scientific, commercial and social aspect. Some of the areas of motion are medical, entertainment industry, visual surveillance, precise analysis of athletic performance, art, content-based video retrieval, robotics, traffic monitoring, cloud and weather systems tracking, etc[1][2][11][4][9]. In motion the important area's attracting many researchers worldwide is human motion recognition, representation, segmentation, analysis, visualization, reproduction etc [9][10]. In this paper we concentrate on problem of motion correction in physical action by human body. This kind of approach has novel applications in design of affordable and efficient frameworks for sports training, to suggest corrections in movements for better results to practitioner, for auditing of scenes in movies, the correction required for correct moves in sports to improve the performance[2], distance learning for education involving physical motion of human body like different dances Japanese dance Niyon Bayo, Indian dance Bharatnatyam etc or physical exercise like Yogasana, Aerobics etc for fitness.

This paper concentrates on motion or action correction in Yogasana. As body posture can be represented by curve, here curve based approach is proposed for error correction.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 675–680.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

2 What Is Yogasana?

The present age of speed and competition has increased the stresses and strains resulting in an increasing prevalence of life style-related health problems such as hypertension, cardiac diseases, bronchial asthma, diabetes neurosis and depressive illness, low back pain [6]. With growing scientific evidence, Yogasana is emerging as an important health behavior-modifying practice to achieve states of health, demonstrated the beneficial effects of yogasana on health behavior in many life style-related problems [7]. Regular practice of Yogasana reduces the stress, which in resultant improves the performance of students in examination, many uses yoga for developing memory, intelligence and creativity [5].

The word yoga means "union" in Sanskrit. Asana is only one of the eight "limbs” (Ashtang in Sanskrit) of yoga, which concentrates on practice of specific breathing techniques and postures.

3 Proposed Approach for Error Correction

Yogasana can be defined by sequence of actions also known as ‘*Stiti*’. Each Stiti can be represented by curve made by human body to achieve that stiti and the curve is controlled by different control points (CP) on human body.

3.1 Body Curves and Sample Control Points for “Sun Salutation” Asana

Figure 1 shows few stiti’s in *‘Sun Salutation’* Asana. Here different stiti’s are Prarathanasan, Tadasan, Uttanasan, Chaturang Dandasan and each stiti make different body curves. Curve is controlled by different control points, which may or may not be on body.

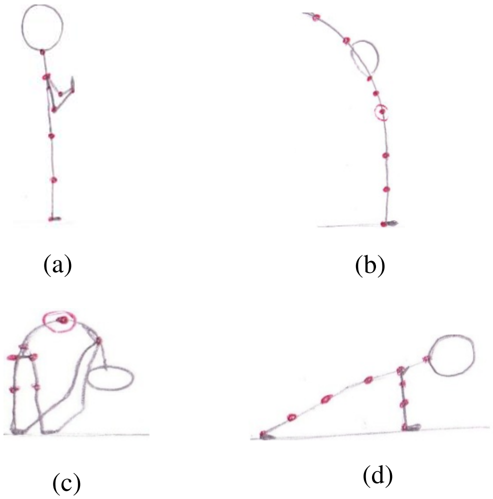


Fig. 1. Major steps (Stiti’s) in ‘*Sun Salutation’* with probable control points (a) Stiti 1- Prarthanasan, (b) Stiti 2- Tadasan, (c) Stiti 3- Uttanasan, (d) Stiti 4- Chaturanga Dandasan

So, the Asana is sequence of different curves with time. In Asana while transition of one Stiti to another Stiti velocity of control points is also important i.e. how much time is taken for transition of each control from one stiti to another.

System can be made intelligent from Yogasana done by expert's having different physical parameters. For proposing correction only stiti curves and motion sequence of CP trajectories required for particular asana are not sufficient as every human differs in physical attributes such as mass and body proportion [8]. Motion trajectories and allowed deviation varies according to physical characteristics and age of practitioner. Initially, we can consider only specific range of physical parameters and age.

3.2 Proposed Framework for Error Detection in Human Motion

Proposed framework is shown in figure 2. This framework applied for Yogasana. Here expert's body curves and control point trajectories needs to be calculated only once and store it permanently. Practitioner's body corves and control point motion trajectories needs to be calculated for each time Asana is done. According to physical parameters of practitioner's body, body curves and trajectory points templates are mapped and from that error detected and correction in Asana is suggested to practitioner if required.

3.3 Methodology / Algorithm

1. Asana A is sequence of different stiti's with respect to time

$$A = \{st_0, st_1, \dots\dots, st_n\}$$

where,

A is set of stitis/body curves for particular Asana. T is set relative time in stances for particular stiti.

$$\left.\begin{aligned} st_i &= \{cp_0, cp_1, \dots\dots, cp_m\} \\ \delta i &= \{\delta_0, \delta_1, \dots\dots, \delta_m\} \\ &\text{For all i, n} \le \text{i} \le 0 \end{aligned}\right\} \quad (2)$$

Where,

st_i is set of different control points representing body curve for i^{th} stiti.

δi Represents allowed deviation in Asana A for beginner in practice of Yogasna.

These kinds of templates can be prepared for each Asana for different physical parameters.

For each transition of body curve from one stiti to another stiti, information like type of motion, amount of motion and motion path should be maintained for each control point. It can be extracted from different key frames from video between two stiti's. Figure 3 shows motion path of control

1. point cp_2 for body motion from stiti 2 to stiti 3 of *'Sun Salutation'*.
2. From captured practitioner's video of Asana identify the key frames of required stiti and map that with the expert's Asana Stiti's control points,

$$\delta_i' = st_i - st_i' \quad (3)$$

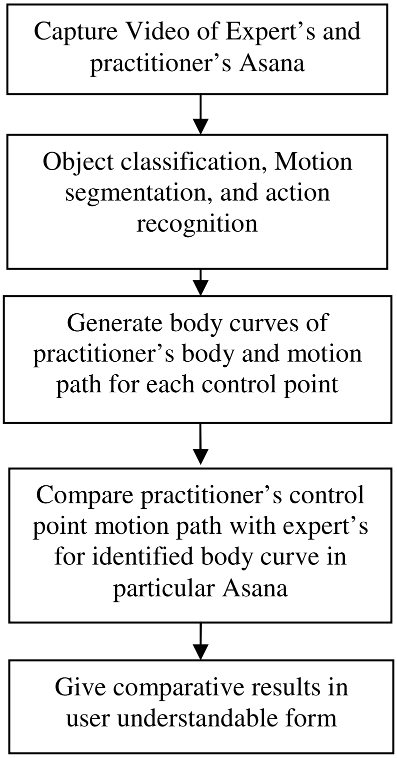

Capture Video of Expert's and practitioner's Asana
Object classification, Motion segmentation, and action recognition
Generate body curves of practitioner's body and motion path for each control point
Compare practitioner's control point motion path with expert's for identified body curve in particular Asana
Give comparative results in user understandable form

Fig. 2. Framework for Motion correction in Yogasana

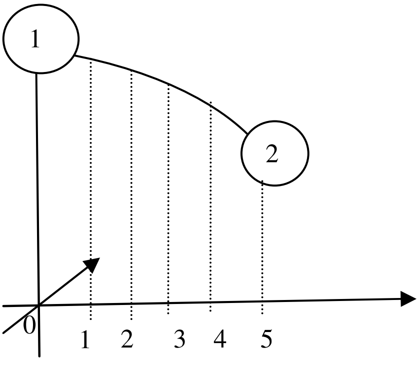


Fig. 3. Motion path of control point cp_2 from stiti1 to stiti2

δ_i' - Set of control points which gives difference between the body curves of expert and practitioner for i^{th} stiti of given Asana. Ideally it should be φ; practically it may vary with expertise level and physical parameters of practitioner.

Allowed deviation is given by δi as mentioned in (2), compare δi with δ_i' which gives set of control points where correction is required. Correction in Asana should be decided from physical parameter and expertise level of practitioner.

Beginner may have more deviation in body curve as compared to regular practitioner for similar physical parameters and age.beginner may have more deviation in body curve as compared to regular practitioner for similar physical parameters and age.Initially two levels can be maintained (a) Beginner (b) Regular Practitioner.

4 Conclusions

Human body and control point path representation using curves gives efficient representation and framework for error identification and correction. It has wide applications in sports, physical exercise, acting, traditional dances etc.

Still it has challenges like (i). human body curves vary with physique, age, gender, weight etc. (ii). Parameters of motion like center of gravity of each body segment, Moment of inertia are not considered. We are extending our work to consider few of the above mentioned parameters.

References

[1] Gavrila, D.M.: The visual Analysis of Human Movements: A Survey. Computer Vision and Image Understanding 73(1), 82–98 (1999)

[2] Wang, R., Leow, W.K., Leong, H.W.: 3D-2D Spatiotemporal Registration for Sports Motion Analysis. In: IEEE Conference on CVPR 2008, pp. 1–8 (2008)

[3] Lu, C., Ferrier, N.J.: Repetitive Motion Analysis: Segmentation and Event Classification. IEEE Transactions on Pattern Analysis And Machine Intelligence 26(2), 258–263 (2004)

[4] Naemura, M., Suzuki, M.: Extraction of rhythmical factors on dance actions thorough motion analysis. In: Proceedings of the Seventh IEEE Workshop on Applications of Computer Vision, WACV/MOTION 2005 (2005)

[5] Kauts, A., Principal, N.S.: Effect of yoga on academic performance in relation to stress. International Journal of Yoga 2, 39–43 (2009)

[6] Deshpande, S., Nagendra, H.R., Nagarathna, R.: A randomized control trial of the effect of yoga on Gunas (personality) and Health in normal healthy volunteers. International Journal of Yoga 1, 2–10 (2008)

[7] Williamsa, K.A., et al.: Effect of Iyengar yoga therapy for chronic low back pain; Venture, G., Ayusawa, K., Nakamura, Y.: Realtime Identification Software For Human Whole-Body Segment Parameters Using Motion Capture and Its Visualization Interface. In: IEEE 11th International Conference on Rehabilitation RoboticsKyoto International Conference Center, Japan, June 23-26 (2009)

[8] Goya, K., Zhang, X., Kitayama, K., Itaru: A Method for Automatic Detection of Crimes for Public Security by Using Motion Analysis. In: Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (2009)

[9] Williamsa, K.A., et al.: Effect of Iyengar yoga therapy for chronic low back pain. International Association for the Study of Pain 115, 107–117 (2005)

[10] Joon, J.S.: A preliminary study of Human Motion Based on Actor Physiques Using Motion Capture. In: Sixth International Conference on Computer Graphics, Imaging and Visualization, pp. 123–128 (2009)

[11] Azad, P., Asfour, T., Dillmann, R.: Toward an unified representation for imitation of Human Motion on Humanoids. In: IEEE International Conference on Robotics and Automation Roma, Italy, April 10-14 (2007)

Improvement the Bag of Words Image Representation Using Spatial Information

Mohammad Mehdi Farhangi, Mohsen Soryani, and Mahmood Fathy

Department of Computer Engineering, Iran University of Science and Technology

Abstract. Bag of visual words (BOW) model is an effective way to represent images in order to classify and detect their contents. However, this type of representation suffers from the fact that, it does not contain any spatial information. In this paper we propose a novel image representation which adds two types of spatial information. The first type which is the spatial locations of the words in the image is added using the spatial pyramid matching approach. The second type is the spatial relation between words. To explore this information a binary tree structure which models the is-a relationships in the vocabulary is constructed from the visual words. This approach is a simple and computationally effective way for modeling the spatial relations of the visual words which shows improvement on the visual classification performance. We evaluated our method on visual classification of two known data sets, namely 15 natural scenes and Caltech-101.

Keywords: BOW Representation, Spatial Information, N-gram Model, Spatial Pyramid Matching.

1 Introduction

As the acquiring and storing of images and multimedia data is becoming fast and easy, the databases of these data become very large. In this situation the necessity for developing methods to manage these databases becomes more and more important. Classifying images based on their content is one of these methods that finds the category of an image among several categories. However, this task is a challenging problem in the real world because we encounter a number of difficulties in the images where there exists occlusion, background clutter and lighting changes.

Many of the recent methods for classifying the images represent each image as sets of patches or regions, described by various descriptors. Based on this description, an image can be represented as a bag of visual words [1]. To achieve this representation, the first step is the extraction of the local patches from the image. Several methods proposed to extract local patches in the literature. While some researchers obtained local regions using regular grids which segment images by horizontal and vertical lines [2], others used various interest point detectors such as difference of Gaussian [3], Harris affine region detector [4] and Hessian matrix [5] to detect patches that contain local information of an image. After detecting the patches, a feature descriptor method like SIFT [3], SURF [5], etc is used to describe them. Previous studies have

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 681–690.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

shown that, the SIFT descriptor extracts robust features from the image which are invariant to affine transformations more than other descriptors [6]. After that, similar patches are clustered in the same groups and each of these groups is treated as a visual word. At this point, the vocabulary which consists of cluster centers is generated.

After the vocabulary construction, an image can be represented as a bag of visual words by assigning each local descriptor to one or several visual words with different weights [7]. Previous studies showed that by assigning the local descriptors to more than one word, the classification accuracy is increases [8].

Despite all of the successes in image classification based on the BOW, this type of representation does not consider the spatial information and this is because of the fact that the histogram representation naturally neglects the spatial location of visual words and spatial relations between them. One of the first attempts in order to utilize spatial information was proposed by Lazebnik et al. [9]. Their work was based on partitioning an image into increasingly finer grids. For each grid cell the frequency of visual words was computed. The BOWs from each cell were concatenated to each other and thus a representation of image which conveys the spatial location of visual words was obtained. In [10] a visual language model using training images was constructed. This model represents three kinds of relations including unigram, bigram and trigram between visual words and captures the proximity information of visual words. In [11] a new representation based on utilizing the informative adjacent word pairs were proposed. To find the informative word pairs, they measured the confidence that neighboring visual words are relevant. Visual words with high confidence were used to add to BOW representation.

In this paper we propose a new representation for images which adds the spatial information to bag of word representation. For this purpose we explore two types of spatial information. First, it is important to know where a certain visual word occurs in the image. For example a blue patch which is located above the image is probably representing a piece of sky while if this patch be in the bottom of the image, it may represent a part of a sea. Words adjacency is the second type of spatial information which although conveys important information about the content of the image, it is neglected in BOW model. For example a white patch can be part of a sheep, cloud or moon if it is surrounded by green grass, blue sky or dark area respectively. To consider this relation, we calculate number of times that each pair combination of words occurs in a certain neighborhood and construct the bag of N-grams inspired by Li et al. [10] and concatenate it to BOW representation.

The remaining sections of this paper are organized as follows. In section 2 we propose details of our method. Section 3 presents experimental results. And section 4 concludes the paper.

2 The Proposed Method

The new image representation which is called spatial bag of words (SPBOW) is constructed in two stages. In the first stage we use the spatial pyramid matching approach [9] and partition the image into fine sub regions and obtain the histogram of local features inside each sub regions. In the second stage the numbers of occurrences of visual word pairs are obtained and concatenated to the BOW representation as new features. The following subsections present these stages in details.

2.1 BOW Representation

In order to represent an image by bag of visual words, local patches are extracted from an image and every patch is described using SIFT descriptor. Since previous studies have shown that sampling on a regular grid outperforms other approaches like interest point detectors, we use the SIFT descriptor, sampled on a regular grid.

After that, each local descriptor of the image should be assigned to one or more visual words. If $\{r_1, r_2, \dots, r_n\}$ represents local descriptors in the image and $V = \{\omega_1, \omega_2, \dots, \omega_k\}$ represents the vocabulary, the hard histogram of visual words is computed as

$$\mathrm{HBOW}(\omega_j) = \sum_{i=1}^{n} \begin{cases} 1 & \text{if } \omega_j = \arg\min_{\omega \in W} \left(\mathrm{dist}(\omega_j, r_i)\right) \\ 0, & \text{otherwise} \end{cases} \qquad (1)$$

Where r_i is the i-th patch in the image and ω_j is the j-th word in the vocabulary.

2.2 Spatial Pyramid Matching

The BOW representation described above ignores some useful information of the image. For example in this representation there is no way to find out how many times a certain visual word takes place in a specific part of the image. To combine this information with BOW, we use the spatial pyramid matching proposed in [9] and partition the image into rectangular regions.

In details, pyramid matching works by placing a sequence of increasingly finer grids over the feature space and taking a weighted sum of the number of matches that occur at each level of resolution. At any fixed resolution, two points which fall into the same cell are matched. Matches found at finer resolutions are weighted more highly than matches found at coarser resolutions. More specifically, a sequence of grids is constructed at resolutions 0… L, such that the grid at level l has 2^l cells along each dimension, for a total of $D = 2^{dl}$ cells. Let H_X^l and H_Y^l denote the histograms of X and Y at this resolution, so that $H_X^l(i)$ and $H_Y^l(i)$ are the numbers of points from X and Y that fall into the i-th cell of the grid. Then the histogram intersection function finds the number of matches at level l.

$$I(H_X^l, H_Y^l) = \sum_{i=1}^{D} \min\left(H_X^l(i), H_Y^l(i)\right) \qquad (2)$$

In this equation, the number of new matches found at level l is given by $I^l - I^{l+1}$ for $l = 0, \dots L - 1$. The weight associated with level l is set to $\frac{1}{2^{L-l}}$, which is inversely proportional to the cell width at that level. Intuitively, since the matches found in larger cells involve dissimilar features, they should be weighted lower. So the following definition was obtained for the pyramid match kernel:

$$\kappa^l(X,Y) = I^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}\left(I^l - I^{l+1}\right) = \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l \qquad (3)$$

In order to combine this pyramid matching kernel with spatial location of words in the image, the elements of X and Y are used for representing the coordinates of a certain word in the image. Therefore, by placing an increasingly fine grid on this feature space, the spatial information is combined with BOW representation.

2.3 Spatial Relation Modeling

Although the relations between visual words in an image convey essential information about its content, this information is neglected in traditional BOW. In text categorization area the relations between words are obtained using the N-gram model and the conditional probability of word sequences are estimated using this model. However this relation is not considered in the image representation. One reason for neglecting this relation in previous studies is the fact that considering N-gram model for images consumes too much memory space which makes it impractical. To deal with this problem inspired by [12], we propose a method based on the visual ontology construction. An example of such ontology is shown in Fig. 1. The leaves of this tree are the visual words and the internal nodes are used for words adjacencies modeling. In details, after constructing the vocabulary which includes k visual words we employ the agglomerate clustering algorithm to hierarchically group word pairs. The leaves of this ontology are the visual words and the internal nodes are the ancestors of the words. We refer to these internal nodes as general words since they are constructed from two child nodes and contain features which are similar to the features of their child. We will use the internal nodes of this ontology at level l for word adjacencies modeling.

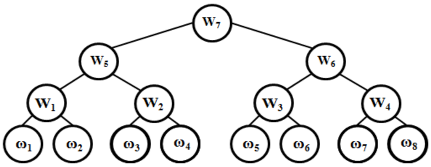


Fig. 1. An example of visual ontology. The leaves are visual words and represented by ω_i, the internal nodes are general words represented by W_j.

After defining the ontology, we use the general words to construct Bag of Bi-grams. We use general words for this purpose instead of visual words because if we directly use visual words the dimension of the features vector will be too high and can't be computed effectively. For example if the vocabulary consists of 200 words the number of Bi-grams will be more than 20000 which is very high and it is not suitable to consider it in BOW representation. In order to reduce the dimension of the feature vector, we can consider just 25 general words which are the ancestors of these words and obtain 325 Bi-grams. The diagram of this model is illustrated in Fig 2. For constructing the bag of Bi-grams we traverse the image from top-left to bottom-right and for each patch we consider the right, bottom and diagonal neighbors and assign each of them to one general word and count the number of general word pairs which are adjacent.

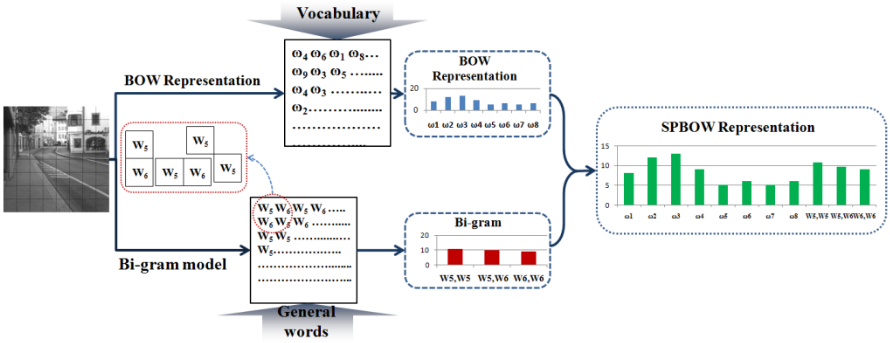


Fig. 2. Adding spatial information to BOW representation

Finally, we summarize our method for representing an image as follow:

1. Construct the visual ontology using agglomerative clustering of visual words and obtain the general words.
2. Calculate the frequencies of each individual word in the image. The histogram of occurrences of these words is referred to as BOW.
3. Count the number of times that every two general words are adjacent and concatenate these numbers to the BOW representation. We refer to this representation as SPBOW.

3 Experimental Results

In this section, we evaluate our proposed method for image classification on two data sets: 15 natural scene categories [9], and Caltech-101. Although these data sets contain color images, all the experiments are performed in grayscale. For experimental set up we follow Lazebnik et al. [9], and select randomly subsets of the data set to create train and test images. An SVM classifier with Laplacian radial basis function (LRBF) kernel is chosen to classify the images. To obtain image features, we extract SIFT features on a regular grid. The patches of this grid are 16*16 pixels and the sampling rate is set to 8 pixels. For constructing the vocabulary we employ the k-means algorithm on the features extracted from the training image and visual words are generated.

First experiments are performed on the 15 natural scenes which consists of 15 classes. Some samples of these dataset are shown in Fig. 3. We randomly select 100 images for training set and use the rest of the images in each class for testing. For all experiments a one level spatial pyramid is used. So, each image is partitioned to 2*2 sub images.

Table 1 shows the classification result of our method on this data set. For this experiment we use a vocabulary consisting of 256 words. The number of general words to serve as spatial relation modeling varies from 4 to 64. We see that when the number of general words is increased, the classification accuracy also increases. On the other hand, as the number of general words increases the algorithm becomes more complex because the number of Bi-grams becomes very high and therefore the algorithm needs more time and memory. So, it is not practical to use all of the general words and a

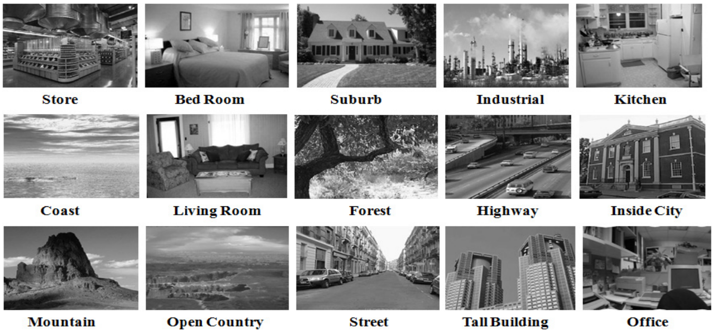


Fig. 3. Example images from the 15 natural scenes data set

tradeoff between the number of general words (algorithm complexity) and the classification accuracy is needed.

Table 1 shows that the classification accuracy doesn't change very much when we increase the number of general words from 16 to 32 and higher. For example when we use 64 general words the accuracy increases only 0.29 percent in comparison with the case we use 16 general words but we have to add almost 2000 value to feature vector in this case (There are 2080 and 136 pair combination for 64 and 16 general words respectively). Such behavior of the algorithm may be because the information content of the words adjacencies is limit. To illustrate this behavior, consider two white and blue patches occur in vicinity, showing a part of the sky. To realize that these two patches represent the sky, there is no need to quantize the blue color into several different blues and count the number of Bi-grams for every blue color. So, we can limit the number of general words to a predefined threshold. The value of this threshold depends on various parameters like number of classes, vocabulary size and the accuracy that we require.

Table 1. Classification results of the SPBOW representation

Number of general words	Number of Bi-grams	Classification accuracy
0	0	75.01 ± 0.4
4	10	75.86 ± 0.3
8	36	76.28 ± 0.4
16	136	76.35 ± 0.6
32	528	76.44 ± 0.5
64	2080	76.64 ± 0.3

Fig. 4 compares our method against BOW representation [2] and spatial pyramid matching (SPM) [9]. This figure plots the relationship between the classification accuracy and vocabulary size. In this experiment we used 16 general words for all vocabulary sizes. We see that our method which adds words adjacencies information to the BOW outperforms other methods which neglect this information. This supremacy can be seen for all vocabulary sizes but for small vocabularies this is more obvious. This behavior is because we use 16 general words for all vocabulary sizes. So, when we use small vocabularies the general words and visual words are more similar to each other in comparison with case we use larger vocabularies. For example when we use a vocabulary which consists of 16 visual words, the general words are the same as visual words. Furthermore, the number of visual words that each general word is a candidate for them increases when the size of vocabulary increases. For example, when the vocabulary consists of 512 words each of the 16 general words is a candidate for 32 of the visual words. In contrast each general word is a candidate for only 2 visual words when the size of the vocabulary is 32. So, as we model the spatial relation using general words, more information of visual words is neglected and we observe less improvement in classification accuracy for large vocabularies.

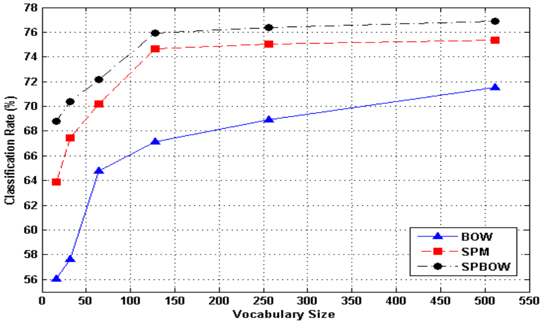


Fig. 4. Classification result on natural scene data set. The horizontal axis shows the vocabulary size and vertical axis represent the classification accuracy.

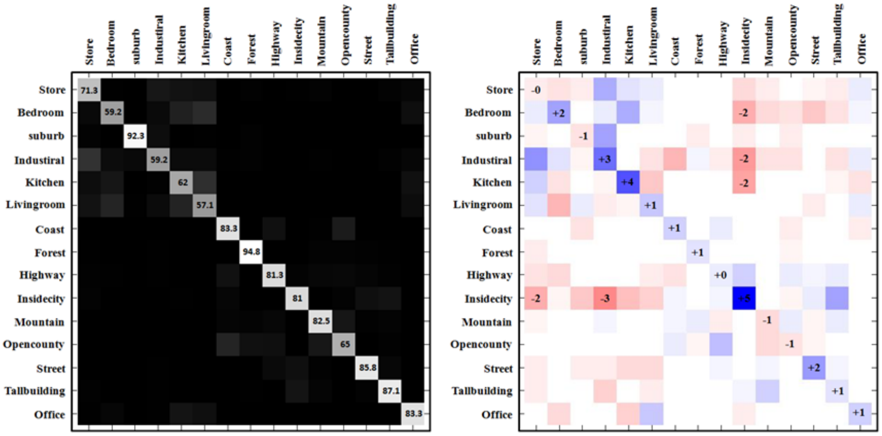


Fig. 5. (a) Confusion matrix of the 15 natural scene data set. The value at position (i,i) shows the classification rate for the class i. (b) Relative confusion matrix of natural scenes. The value at row i and column j which has been scaled, represents the difference between SPBOW and SPM to classify the images of class i as class j. We show positive and negative entries in blue and red respectively.

In Fig. 5 we show the confusion matrix of SPBOW representation and relative confusion matrix of 15 natural scene dataset. The relative confusion matrix illustrates the relation between confusion matrices of SPBOW and SPM representations. Every entry in this matrix denotes the absolute differences between entries in the confusion matrix of SPBOW and confusion matrix of SPM [9]. For this experiment the vocabulary size and number of general words are set to 256 and 16 respectively. We set up this experiment to observe the impact of words adjacency information on each class more clearly. The entries on the main diagonal of the relative matrix that shows the instances that correctly classified are mostly increased. As can be seen the classification accuracy of inside city, kitchen and industrial classes increase more than others. The non diagonal elements of this matrix show the misclassification rate and we see that the confusion declines for most of the class pairs. We clearly observe this improvement in the confusion between inside city as industrial, kitchen as inside city and industrial as inside city.

The second data set used for experiments is the Caltech-101. This data set consists of 101 objects and contains a broad range of objects. Fig. 6 shows some samples in this data set.

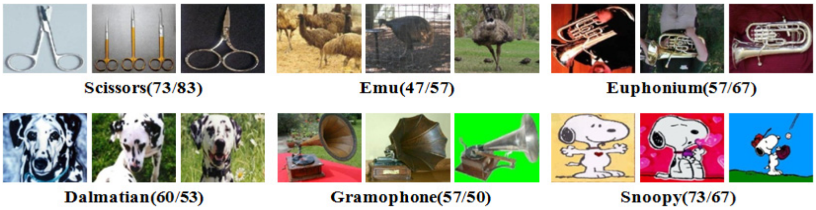


Fig. 6. Example images from Caltech-101 data set. Three top classes are those sample classes which our method has performed well compared to SPM and three bottom are samples which our method did not perform well. (SPM classification accuracy/SPBOW classification accuracy)

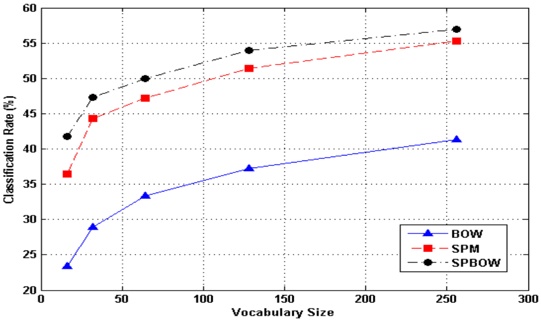


Fig. 7. Classification results on Caltech-101 data set. The horizontal axis shows the vocabulary size and the vertical axis represents the classification accuracy.

To construct train and test sets we randomly select 30 images per class for training and 30 images for testing. In Fig. 7 we show the classification accuracy of various image representation methods on Caltech-101 data set. In this experiment the number of general words used for words relation modeling is set to 16. We observe that spatial relations which are considered in our algorithm have positive effect on classification rate for all vocabulary sizes. In Fig. 6 some example classes are shown which our method has the best and worst performance on them.

4 Conclusion

In this paper we addressed the problem of neglecting the spatial information in BOW representation. For this purpose a new image representation based on modeling the words adjacencies using a tree structure was constructed and spatial relation between words were added to BOW. The experimental results on two known data sets showed that this new representation outperforms other representations and the spatial information plays an important role in detecting the content of the images. In this study the number of general words for modeling the relation between words was chosen based on the classification accuracy and algorithm complexity. However, an interesting future work is to find ways for selecting sub sets of the general words based on feature selection methods and using these sub sets in order to model the spatial relation.

References

[1] Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. ICCV (2003)

[2] Fei-Fei, L., Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2005)

[3] Lowe, K.D.: Distinctive Image Features from Scale-Invariant Keypoints. J. of Computer Vision 2(60), 91–110 (2004)

[4] Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. Alvey Vision Conf., pp. 147–151 (1988)

[5] Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)

[6] Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: Proc. CVPR 2003, Madison, WI, pp. 257–263 (June 2003)

[7] van Gemert, J.C., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M.: Visual word ambiguity. IEEE Trans. Pattern Analysis and Machine Intelligence 32(7), 1271–1283 (2010)

[8] Jiang, Y.G., Yang, J., Ngo, C.W.: Representation Of KeyPoint-Based Semantic Concept Detection: A Comprehensive Study. IEEE Trans. Multimedia 2(1), 42–53 (2010)

[9] Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2169–2178 (2006)

[10] Wu, L., Li, M., Li, Z., Ma, W.-Y., Yu, N.: Visual language modeling for image classification. In: ACM Multimedia Workshop on Multimedia Information Retrieval, pp. 115–124 (2007)
[11] Mei, L., Kweon, I., Hua, X.: Contextual Bag-of-Words for Visual Categorization. IEEE Trans. Circuits and Systems for Video Technology 21(4), 381–392 (2011)
[12] Jiang, Y.G., Ngo, C.W.: Bag-of-visual-words expansion using visual relatedness for video indexing. In: ACM Conf. on Research & Development on Information Retrieval, pp. 769–770 (2008)

Left Object Detection with Reduced False Positives in Real-Time

Aditya Piratla, Monotosh Das, and Jayalakshmi Surendran

Global R& D, Crompton Greaves Limited, Mumbai, India
aditya.piratla@gmail.com,
{monotosh.das,jayalakshmi.surendran}@cgglobal.com

Abstract. Identifying unattended objects in public places efficiently, is one of the major thrust areas of security. This paper proposes a real-time method to identify unattended or left objects in a region of interest that is under surveillance. This is a simple pixel based method for object detection which can be used for both indoor and outdoor environments. This method is robust to changes in illumination in case of high contrast foreground images, which is achieved through normalization. The false positives are eliminated through implementing a method following the ideas of codebook. The proposed method is tested on a live set up. It consists of an IP camera for video capture, an analytics server where the built in intelligence checks for the presence of any left object in the video and a user interface through which the concerned authority is intimated for any timely action. The entire communication in this system follows ONVIF (Open Network Video Interface Forum) standard. Apart from identifying unattended objects, it can be also used to keep designated areas clear of obstructions.

1 Introduction

Security is of vital concern in all public places these days. The significant increase in security measures in public places is the result of the terrorist attacks over the years. Sizeable part of the terrorist attacks is by leaving explosives in public places which are unidentified for a long time. Hence detection of left objects is an area of special interest to ensure security. Object detection even otherwise is an old but relevant problem in the field of image processing and computer vision. A number of literature is reported on this topic. Object detection indoors and outdoors need separate consideration. Object detection is mostly based on the background estimation and it has dynamic nature in outdoor scenarios. The characteristics intrinsic to the objects like the shape, color or size or the properties like motion can act as a cue for performing detection. When motion cue is used for this purpose, the underlying method separates the background and the foreground. Thus segmentation gains main thrust in these methods.

Video segmentation algorithms can be classified into three types: edge information based video segmentation, image segmentation based video segmentation and change detection based video segmentation. Edge information based algorithms, first apply Canny edge detector to find edge information of each frame and then keep tracking these edges. Image segmentation based algorithms first apply image segmentation algorithms, such as watershed transform [5] and colour segmentation on each frame to

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 691–698.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

separate a frame into many homogeneous regions [6]. Change detection based segmentation algorithms [2] threshold the frame difference to form change detection mask.

In training based technniques, for object classification, features extracted from images are compared with a visual dictionary. The best-matching features are learned by the classifier to determine the object class. In [3], the visual dictionary concept is extended with Interest Point Operators (IPOs). A fast algorithm for background modeling and subtraction is proposed in [10]. Sample background values at each pixel are quantized into codebooks which represent a compressed form of background model for a long image sequence.

Most of the video surveillance literature talks about analysis of video that is performed on the raw video. But literature is available in which the analysis is focused on compressed domain features of the encoded video that is served by the camera. In [4] a fast video segmentation algorithm for MPEG-4 camera systems is proposed. Object detection in compressed domain which brings out the possibility of avoiding the decoding of the video stream is discussed in [1]. Removal of noise in motion vectors that are vital in segmenting the objects in a scene is discussed in [7]. A moving object extraction technique for MPEG coded data directly is proposed in [8]. The inherent properties of surfaces and hence the illumination effects are also of importance in this field as the pixels values directly play a role in the entire procedure. Reference [9] talks about the way retina-and-cortex system (retinex) treat a color as a code for a three-part report from the retina, independent of the flux of radiant energy but correlated with the reflectance of objects. This paper largely talks about mondrian images.

This paper proposes an end to end video surveillance solution to detect any left object in real time. The algorithm for detection of objects is a simple pixel based method. These methods are generally highly illumination sensitive. The effect of illumination is reduced to a large extent through a normalization process. Whenever we consider outdoor scenes, one of the worst scenarios that we face with motion is the pixel difference due to tree leaf movement. The effects of tree leaf movement as well as falsely detected objects due to illumination change are eliminated in this work through an implementation adopting ideas of codebook. The novelty of this work is mainly on decreasing the false positives with this algorithm. This algorithm is suitable for real time applications since this does not require any learning phase. We address indoor as well as outdoor scenarios for object detection. The communication among different entities in the system follows ONVIF standard.

2 Proposed Method

In this paper, we present an algorithm to detect any left object in a given scene. The first step in the development is the background- foreground separation of the input frames. The background estimation in this system is done from the first few input frames itself. Any pixel that falls within the specified limits of the background pixel value will form the part of background. Else it is defined as a foreground pixel. To estimate the range of values for each background pixel, we find two different frames, namely, the minimum background frame and the maximum background frame represented as BG_{min} and

BG_{max} respectively. BG_{min} is formed through placing the minimum pixel values corresponding to every pixel position among the first say, N frames. Similary, the maximum values corresponding to every pixel position from initial N frames constitute BG_{max}. Let a pixel value in R, G or B plane of the i^{th} frame be represented by $I_i(x,y)$. Equation(1) shows the formation of BG_{min} and BG_{max}.

$$\begin{aligned} BG_{min}(x,y) &= min_{i=0}^{N}(I_i(x,y)) \\ BG_{max}(x,y) &= max_{i=0}^{N}(I_i(x,y)) \end{aligned} \tag{1}$$

This process is repeated for all three color planes.

Subsequently, illumination normalization is performed on the input image, background minimum image and background maximum image. The normalized image is multiplied by a constant value which reduces the dynamic range of pixel values and hence the effect of illumination changes to a large extent. The dynamic range of the pixels is reduced in all cases where the maximum pixel value of the frame is more than the constant of multiplication. Foreground extraction of the input image takes place subsequently. We check for the presence of foreground objects from the $(N+1)^{st}$ frame onwards with respect to BG_{min} and BG_{max} values allowing a tolerance ε. The Eq.(2) states the condition for a pixel $I(x,y)$ to be a part of the foreground.

$$\begin{aligned} &I(x,y) < (1-\varepsilon)BG_{min}(x,y) \qquad and \\ &I(x,y) > (1+\varepsilon)BG_{max}(x,y) \end{aligned} \tag{2}$$

If a pixel $I(x,y)$ lies within the background range specified as in Eq.(3), it is defined as a background pixel.

$$(1-\varepsilon)BG_{min}(x,y) \leq I(x,y) \leq (1+\varepsilon)BG_{max}(x,y) \tag{3}$$

After this stage all the closed contours of the foreground image are extracted. These represent the contours of the new objects with respect to the background. In spite of normalization procedure that is applied to the frames initially, there can be a number of false positives due to small movements like tree leaf movements and quantization errors. These errors are removed by implementing a method for removing extraneous contours using codebook. Any detected contour is at first represented by a bounding rectangle to it. While doing this, contours which have very small total length are neglected. If there are contours contained completely within another, only the outermost contour will be considered. Subsequent to this stage we use a codebook construction. The entries in the codebook correspond to the rectangles that are formed on each frame based on some pre-decided criteria.

The addtion and updation of codebook is explained in detail with the help of Fig. 1. Consider the first rectangle detected with diagonal co-ordinates (a,b) and (c,d). This is actually a bounding rectangle of one of the outermost contours that we get from the foreground image. Let us assume that the allowed threshold is th. We initialize the minimum and maximum limits for both x-co-ordinates and y co-ordinates. They are named as $a.min$, $a.max$, $b.min$, $b.max$, $c.min$, $c.max$, $d.min$ and $d.max$. Let us now

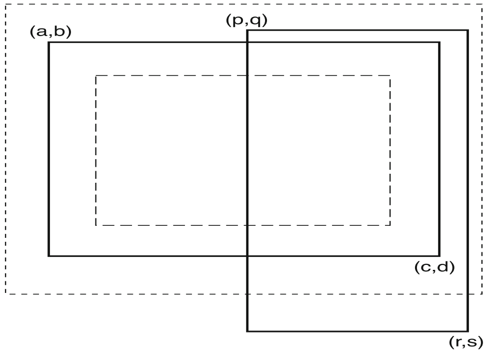


Fig. 1. Matching of Rectangles for Codebook Entry

consider only the co-ordinate (a,b) and see how the codebook entry is performed. The minimum and maximum limits for this coordinate is defined initially as in Eq.(4).

$$\begin{aligned} x.min &= a \\ x.max &= a \\ y.min &= b \\ y.max &= b \end{aligned} \tag{4}$$

With respect to the threshlod defined, the maximum and minimum values of x and y co-ordinates are defined. They are called 'learn threshold' values defined in Eq.(5).

$$\begin{aligned} x.learnth.min &= x.min - th \\ x.learnth.max &= x.max + th \\ y.learnth.min &= y.min - th \\ y.learnth.max &= y.max + th \end{aligned} \tag{5}$$

Similarly, thresholds are defined for co-ordinate (c,d) also. Let us now assume that a new rectangle appears with coordinates (p,q) and (r,s). We have to now decide if this new rectangle will become a part of the codebook or not. If

$$x.learnth.min \leq p \leq x.learnth.max \tag{6}$$

$$y.learnth.min \leq q \leq y.learnth.max \tag{7}$$

then, (p,q) qualifies to lie within the specified range. If the condition specified in Eq.(6) is satisfied, then initial values of *x.min* and *x.max* values are updated as follows. If

$$p \leq x.min \tag{8}$$

then

$$x.min = p \tag{9}$$

If

$$p \geq x.max \tag{10}$$

then

$$x.max = p \tag{11}$$

Similarly *y.min* and *y.max* values are also updated. Subsequent to this, the learn threshold values are also updated as per Eq.(5). A similar procedure is followed to verify if the co-ordinate (r,s) lies within the specified range of (c,d). If it lies within the specified limits, no new entry is made to the codebook but the existing entries are updated as per the above explanation.

Apart from the above mentioned parameters the codebook also contains three other parameters corresponding to every entry. Those are, the lifetime of that entry 'LT', the number of times it finds a match '*count*' and the number of consecutive frames that particular entry does not find a match with another codebook entry 'NA_t'. The first entry in the codebook will have the parameters LT, NA_t and *count* values set to 0. As frames are moved from one to another, the lifetime LT of the existing codebook entry is incremented by one. A comparison of the co-ordinates of the new entry with the existing entries are performed to see if it lies within a defined threshold of the older entries. If it finds a match, *count* is incremented by one. For every frame that does not find a single match for an entry in the codebook, NA_t is incremented by one. If a new addition is made to the codebook, the parameters, LT, NA_t and *count* are also entered for that entry. We have followed a definite strategy to display the object that is detected. As soon as an object is detected, an addition or updation to the codebook takes place with respect to the criteria mentioned in Eqs.(6)-(11). But the appearance of it is shown only if the value of *count* crosses a particular threshold. This allows elimination of false positives which occur instantaneously and also helps avoiding tree leaf and like movements in the scene. If a particular codebook entry does not find a match for a long time with respect to its lifetime, it is deleted from the codebook.

3 Experimental Setup

The experimental set up consists of one ONVIF compliant IP camera that captures video in CIF resolution, encodes it and streams it over RTSP. The decoded video from this stream is input to the analytics engine. Analytics engine detects new objects that appear in the scene and informs the server. The server reports all the output in real time to the UI (User Interface) that is highly interactive. The UI has provision to see the live view of the video as well as the recorded video. The UI allows the user to change the sensitivity of the algorithm which will allow the user or administrator of the system to change the threshold settings during the initial set up. Through the user interface, the camera can be also configured. The camera configurations include the encoder settings, frame rate, video resolution, brightness, contrast etc.

4 Results

This section shows results of the object detection along with illustrative diagrams. The results presented in this section include both indoor and outdoor scenarios. As a test sequence for indoor environment, video inside an office is used which does not have much of moving background. We have tested it for the case of object appearance with repect to the given scene. In Fig. 2a the background, with respect to which the object appearance is to be performed, is shown. Figure 2b and Fig. 2c show the input frame in which a new object is appeared and the corresponding foreground frame that the algorithm extracts respectively. Figure 2d shows the output of the algorithm where the new object that is appeared in the scene is highlighted. Figure 2c inludes every pixel that is not a part of the background and Fig. 2d displays only those objects which are having atleast a definite contour size and those which stay on the scene for a long duration in time as defined in the system.

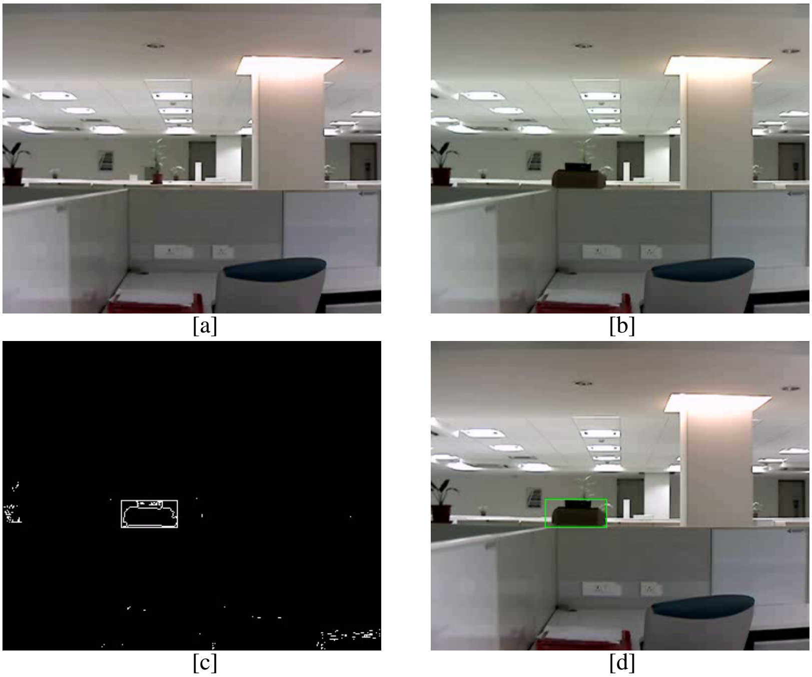


Fig. 2. Indoor Images: [a] Background [b] Input frame [c] Foreground [d] Output of the Proposed Method

Similarly, the background, input frame, extracted foreground and output frames of an outdoor scene are shown in Fig. 3. The foreground frame in this case shows the clear differences in the scene due to tree leaf movements. Moreover the video under consideration is captured from such a scene that includes both indoor and outdoor segments. The quantization noise effects at the edges present in the frame are also falsely extacted

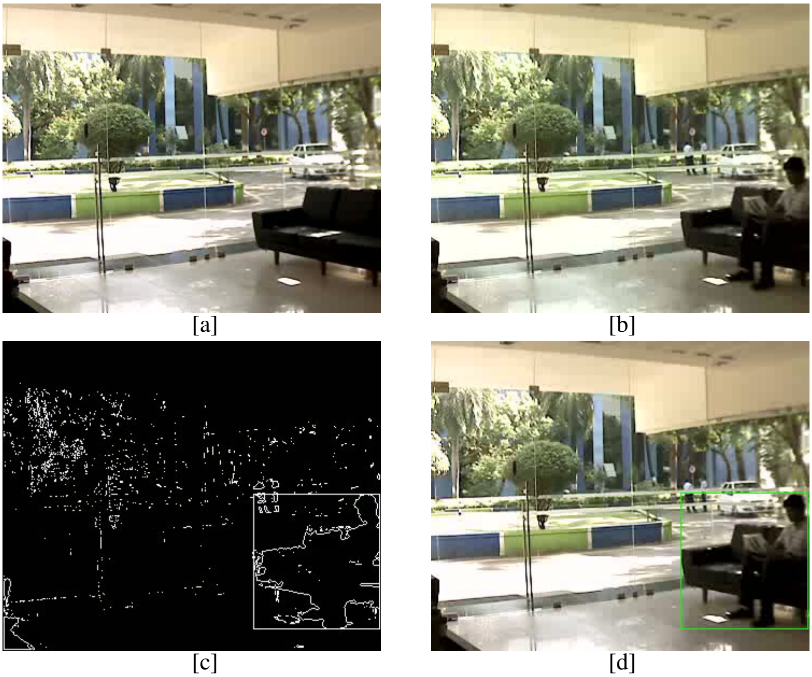


Fig. 3. Outdoor Images: [a] Background [b] Input frame [c] Foreground [d] Output of the Proposed Method

as foreground. The minimum contour size and the codebook implementation ensure the false positive reduction to a considerable level. From this result, the superiority of the algorithm is evident. The classification or identification of different objects will be the next step of our work. At present, the algorithm does not distinguish between human beings and other objects.

In case of outdoor scenario,the tree leaf movement causes a lot of false positives which are totally eliminated through our implementation. We have tested the algorithm with 100 hours of test clips captured from both indoor and outdoor environments. The response time for every frame is around 15 millisecond in our implementation.

5 Conclusion

This paper proposes a simple pixel based algorithm for detection of left objects in a given scene. The algorithm is capable of handling both indoor and outdoor scenarios which differ in their basic characteristics and hence prove to be robust. The foreground detection being a very basic method, wrongly classifies any change in background that results in treeleaf movement and illumination changes as foreground. But the codebook algorithm makes the results robust. The identification of objects as well as detection of objects that are occluded are not part of this work and will be addressed later.

References

1. Sabirin, H., Bäsey, G.: Video Surveillance of Today: Compressed Domain Object Detection. In: ONVIF Web Services Based System Component Communication and Standardized Data Storage and Export using VSAF – a Walkthrough. In Tech (2011)
2. Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B.: Image Change Detection Algorithms: A Systematic Survey. IEEE Trans. Image Processing 14(3), 294–303 (2005)
3. Wijnhoven, R., de With, P.H.N., Creusen, I.: Efficient Template Generation for Object Classification in Video Surveillance Survey. In: Proc. of 29th Symposium on Information Theory in the Benelux, pp. 255–262 (2008)
4. Beevi, Y., Natarajan, S.: An efficient Video Segmentation Algorithm with Real time Adaptive Threshold Technique. International Journal of Signal Processing, Image Processing and Pattern Recognition 2(4) (2009)
5. Moga, A.N., Gabbouj, M.: Parallel Image Component Labeling with Watershed Transformation. IEEE Trans. Pattern Analysis and Machine Intelligence 19, 441–450 (1997)
6. Wong, S.-S., Leow, W.K.: Color segmentation and figure-ground segregation of natural images. In: ICIP, vol. 2, pp. 120–123 (September 2000)
7. Rajkumar, R., SaiKrishna, D., Jayanth, A.: Survey on Motion Vector Filtering and Object Segmentation Methods in Compressed Domain. International Journal ofAdvancements in Technology 2(2) (2011)
8. Zeng, W., Gao, W., Zhao, D.: Automatic Moving Object Extraction in MPEG Video. In: Proc. of ISCAS 2003, vol. 2, pp. 524–527 (May 2003)
9. Land, E.H.: The Retinex Theory of Color Vision. Scientific American 237(6), 108–128 (1977)
10. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.: Background Modeling and Subtraction by Codebook Construction. In: Proc. ICIP, pp. 3061–3064 (2004)

Adaptive Fusion Based Hybrid Denoising Method for Texture Images

Preety D. Swami and Alok Jain

Samrat Ashok Technological Institute, Vidisha, India
preetydswami@yahoo.com, alokjain6@rediffmail.com

Abstract. This paper presents an efficient image denoising method by adaptively combining the features of wavelets and wave atom transforms. These transforms will be applied separately on the smooth areas of the image and the texture part of the image. The disintegration of the homogenous and nonhomogenous regions of noisy image is done by decomposing the noisy image into a noisy cartoon (smooth) image and a noisy texture image. Wavelets are good at denoising the smooth regions in an image and will be used to denoise the noisy cartoon image. Wave atoms better preserve the texture in an image hence is used to denoise the noisy texture image. The two images will be fused adaptively. For adaptive fusion different weights will be chosen for different areas in the image. Areas containing higher degree of texture will be allotted more weight, while the smoother regions will be weighed lightly. The information regarding the weights selection will be obtained from the variance map of the denoised texture image. Experimental results on standard test images provide better denoising results in terms of PSNR, SSIM, FOM and UQI. Texture is efficiently preserved and no unpleasant artifacts are observed.

1 Introduction

Many geometrical transforms have evolved as prominent tools for restoration of noisy images. Each transform has its certain area of expertise. Selection of the proper transformation tool is very important which actually depends on the kind of image to be denoised. All the natural images have some common characteristics [1]. Natural images basically comprise of homogenous regions, edges and texture. A wide range of methods belong to the class of wavelet transform based thresholding and shrinking of wavelet coefficients. Wavelet transform captures information about the image mostly in the low frequency regions where presence of noise is negligible. Failure of wavelets at the edges and in the texture regions have led to the generation of many geometric transforms which provide perfect reconstruction. Some of these transforms are curvelets [2], wedgelets [3], bandlets [4] and wave atoms [5]. Basic idea behind denoising using these transforms is to uniformly threshold the transform coefficients by deciding a suitable threshold. Another approach is to decide a different threshold for every scale of the transformed coefficients [6]. Further improvements can be done by employing a spatially adaptive threshold for each wavelet or curvelet coefficient rather than deciding a single threshold at each scale [7, 8].

N. Meghanathan et al. (Eds.): Advances in Computing \& Inform. Technology, AISC 177, pp. 699–707.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

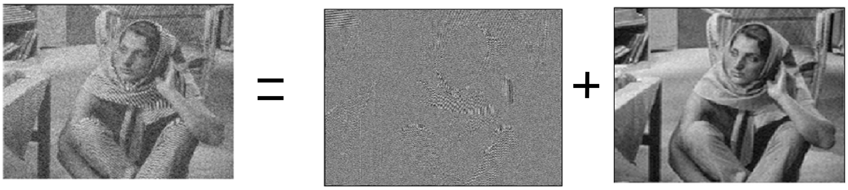
Newer denoising approach is to segment the image into edges and smooth areas and then to apply different models (Generalized Gaussian distribution and Gaussian distribution correspondingly) in these regions [9]. Some authors e.g. [10] and [11] have designed adaptive threshold functions in which thresholding is done using neural network approach. Considerable research has been done in the field of sparse and redundant representation modeling [12, 13] in which data characteristics are learnt either from the image itself or from some existing database and are then used as dictionaries.

In this paper, different transforms will be applied to denoise different regions of an image. Wavelets will be employed to denoise the smoother regions while wave atoms will be used to restore the textural regions. For the separation of smooth and the texture regions the image will be decomposed using curvelet transform. The denoised images will then be fused adaptively and the information regarding the weights needed for adaptive weight selection will be gathered by creating a variance map of the denoised texture image. The organization of the paper is as follows. In section 2, the proposed method is presented along with the introduction of the denoising methods used in the work. Experimental results are analyzed in Section 3. Finally, conclusion is given in Section 4.

2 Proposed Method

This work proposes a denoising method that employs the wavelet transform and the wave atom transform. The basis functions of these transforms are meant to perform efficiently in specific areas in an image. Wavelet transform gives excellent denoising results in the homogenous regions in an image but cannot restore the textural regions properly. Texture can be best denoised using wave atoms but artifacts are visible at the object boundaries due to periodization at the edges.

In the following discussion an attempt is made to present the superiority of wave atoms over wavelets to denoise texture in an image. For this purpose, the texture rich image of Barbara is chosen. Noisy image is then generated by adding white Gaussian noise having a standard deviation of 20. This noisy image in Fig. 1(a) is then separated into its textural part and homogenous (cartoon) part. Fig. 1(b) shows the noisy texture image and Fig. 1(c) shows the noisy cartoon image. Noisy texture image is denoised using wavelet transform and the image magnified around Barbara's face is shown in Fig. 2 (a). Same noisy texture image is denoised by wave atoms and the magnified version of the denoised image is plotted in Fig. 2(b). A comparison of Fig 2 (a) and Fig. 2 (b) shows better denoising of textures using wave atoms. This can be seen visually as well as by calculating the PSNR values for both the images. An improvement of 2.3 dB in PSNR value is obtained when the texture image is denoised using wave atoms. But the problem with wave atoms is the repetition of texture patterns along the object boundaries and edges. It is observed from Fig. 2(b) that the textural structure of Barbara's scarf is repeating and can be seen on the face. At the same time artifacts are largely visible in the homogenous regions where texture is not present. Same portion of wavelet denoised image can be seen in Fig. 2 (a). In this image the textural region are not visually pleasant but there are no repetitions of the texture regions as were the case with wave atoms.

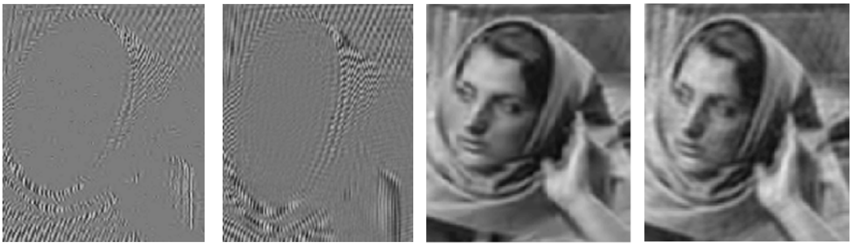


(a) Noisy Image

(b) Noisy Texture Image

(c) Noisy Cartoon Image

Fig. 1. Decomposition of noisy image into its textural constituent and cartoon constituent



(a)

(b)

(c)

(d)

Fig. 2. (a) Wavelet denoising of noisy texture image (b) Wave atom denoising of noisy texture image (c) Wavelet denoising of noisy cartoon image (d) Wave atom denoising of noisy cartoon image

Difference in denoising the cartoon part of the noisy cartoon image by wavelets and wave atoms is shown in Fig. 2(c) and Fig. 2(d). The noisy cartoon image of Fig. 1(c) is first denoised by wavelets and the magnified result is Fig. 2 (c). Same image of Fig. 1(c) is denoised using wave atoms and the magnified image is shown in Fig 2(d). Comparison of the two denoised figures reveals that the restored cartoon image is better in PSNR when denoising is done using wavelets. The restored cartoon image using wave atoms has a PSNR less by 1.47dB and many artifacts are observed in the image. In the proposed method the noisy cartoon image will be denoised using wavelets. The noisy texture image will be denoised using wave atoms. The two denoised images will be combined by adaptive fusion which will be based on the information obtained from a variance map created from the denoised texture image.

2.1 Decomposition of Noisy Image into Cartoon and Texture Image

In the proposed work, segmentation of noisy image into its cartoon part and textural part will be done using the curvelet transform. Different scales in the curvelet transform are actually different filters ranging from low pass to high pass [2]. Higher scales represent higher frequencies which correspond to the texture in an image. In this method the curvelet transform of noisy image is taken and the coefficients corresponding to the two highest scales are retained while the remaining lower scale coefficients are substituted to zeroes. This gives the noisy texture image. The noisy cartoon image is obtained by subtracting the noisy texture image from the noisy image. This decomposition is shown in Fig. 1.

2.2 Wavelet Denoising

The noisy cartoon image obtained from section 2.1 will be denoised using wavelets. The wavelet denoising method used in this work is inspired by the spatially adaptive Bayesian shrinkage of [7]. Denoising is based on estimating the probability that a given wavelet coefficient contains significant noise-free component. Each wavelet coefficient is shrunk by an amount which depends on its local surroundings as well as its global statistical properties.

2.3 Wave Atom Denoising

Natural images contain oscillatory patterns and textures. The wave atom transform is a new member in the family of oriented, multiscale transforms for image processing. Wave atom reconstruction is sparser for oscillatory patterns and textures as compared to wavelets and curvelets. To represent an oscillatory pattern f to some given accuracy and for some constant N (smaller or comparable to the number of pixels of the underlying image), $O(N^{3/2})$ curvelet coefficients or $O(N^2)$ wavelet coefficients are needed. Whereas to encode the same f to the same accuracy only $O(N)$ wave atom coefficients are required [5]. In this work the denoising algorithm used for wave atom denoising is hard thresholding of wave atom coefficients [14].

2.4 Proposed Adaptive Fusion

Simple addition of the two denoised images does not yield good results. The artifacts present in the smooth regions of denoised texture image tend to add to the denoised cartoon image and at the same time the repetition of texture on the face is largely visible. Thus, a fusion algorithm is proposed that overcomes the problems stated above.

For fusion of the two denoised images a variance map of the restored texture image is required. The variance map gives the information about the locations of homogenous regions and textures and edges in the original image. The variance map of an image is the collection of variances of each pixel of the image. This map is calculated by taking an odd-sized square window around a centre pixel of intensity x_i. The window is then made to slide around every pixel to give the complete variance map. Within the window the variance is calculated as

$$\sum_{i=1}^{n_p} (x_i - \mu)^2 \Big/ (n_p - 1) \tag{1}$$

where n_p, is the number of pixels inside the window, and μ is the mean map of the image in which each pixel is the mean of the pixel values inside the window, centered around the pixel whose variance is to be calculated. In this work a window of size 3*3 is employed to create the variance map of denoised texture image and is shown in Figure 3.

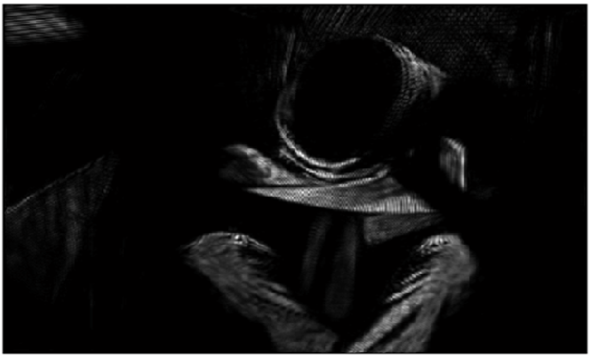


Fig. 3. Variance map of denoised texture image

From the variance map of some natural images it is observed that the highest range of variance values are associated with textures like that on Barbara's scarf. Midrange values of texture are found on the table cover and chair. Barbara's hair contains little texture. Thus more weight should be given to wave atom coefficients for texture regions like the scarf and less weight should be given to the smoother regions like that on Barbara's face. Overall variance map is thus classified into five regions of varying textures. This requires five different weight values which will be multiplied with the pixels of texture denoised image and the result will be added to the cartoon denoised image. These weights are calculated empirically for the image *Barbara* and tested on several standard images. For different groups, the weights *wt* depend upon the variance value of a particular pixel in the variance map of texture (VMT) image. These weights, *wt*, have been calculated as

$$
\begin{aligned}
&wt = 1.2 \quad if \ \mathrm{VMT} \geq 200 \\
&wt = 1.1 \quad if \ \mathrm{VMT} \geq 100 \ and \ VMT < 200 \\
&wt = 1.08 \quad if \ \mathrm{VMT} \geq 50 \ and \ VMT < 100 \\
&wt = 0.9 \quad if \ \mathrm{VMT} \geq 20 \ and \ VMT < 50 \\
&wt = 0.5 \quad if \ \ VMT < 20
\end{aligned}
$$

The final denoised image is now given as

*Combined Image = wt * Wave atom denoised texture image + Wavelet denoised cartoon image* (2)

3 Experiments and Results

Two standard images *Lena* and *Barbara* of size 512*512 are used for experimentation. Gaussian noise of levels 10, 20 and 30 is added to the original image and the denoising results are compared against four denoising methods in Table 1. The qualitative results are provided in Fig. 4. The different denoising methods are compared by using four measures: peak-signal-to-noise-ratio (PSNR), structural similarity index metrics (SSIM) [16], universal image quality index (UQI) [17] and figure of merit (FOM) [18]. These indices are defined as follows:

(i) *Peak-signal-to-noise-ratio (PSNR)*: This index indicates the noise suppression quality. Higher is the PSNR better is the noise suppression. PSNR is defined as:

$$PSNR = 10\log_{10}\left\{MAX^2 / MSE\right\}$$

where MAX is the maximum pixel value of image, MSE is the mean square error and is given by

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left\| I(i,j) - K(i,j)\right\|^2$$

where I is the original image, K restored image, m and n are the number of rows and columns respectively.

(ii) *Figure of merit (FOM)*: This is the standard index for measuring the edge displacement in the denoised image [18]. Maximum value of FOM is 1. FOM nearer to 1 indicates better edge preserving quality. By applying a particular edge operator on original and reconstructed images the edge plots I_d and I_r can be obtained. Then FOM is defined as:

$$FOM = \frac{1}{\max(n_d, n_r)}\sum_{k=1}^{n_d}\frac{1}{1+\gamma d_j^2}$$

where n_d and n_r are the number of pixels in I_d and I_r respectively, d_j is the Euclidean distance between j^{th} pixel of I_r and nearest pixel of I_d. γ is a scalar multiplier and equals 1/9.

(iii) *Universal image quality index (UQI)*: This index measures distortion which can be due to loss of correlation, luminance and contrast [17]. Maximum value of UQI is 1. Higher the UQI less is the distortion. For all the variables defined as before UQI is defined as:

$$UQI = \frac{4\sigma_{I,K}\bar{I}\times\bar{K}}{\left(\sigma_I^2+\sigma_K^2\right)\left(\bar{I}^2+\bar{K}^2\right)}$$

where $\sigma_{I.K} = \frac{1}{(n-1)}\sum_{k=1}^{n}\left(I_k-\bar{I}\right)\left(K_k-\bar{K}\right)$

in which $\bar{I}$, $\bar{K}$ are the mean and σ_I , σ_K are the variance of original and denoised images respectively.

(iv) *Structural similarity index metrics (SSIM)*: This index gives the visual perception quality of the denoised image [16]. Maximum value of SSIM is 1. If SSIM of denoised is closer to 1, denoised image is better. For all the variables defined as before SSIM is defined as:

$$SSIM = \frac{\left(2\sigma_{I,K} + c_2\right)\left(2\bar{I} \times \bar{K} + c_1\right)}{\left(\sigma_I^2 + \sigma_K^2 + c_2\right)\left(\bar{I}^2 + \bar{K}^2 + c_1\right)}$$

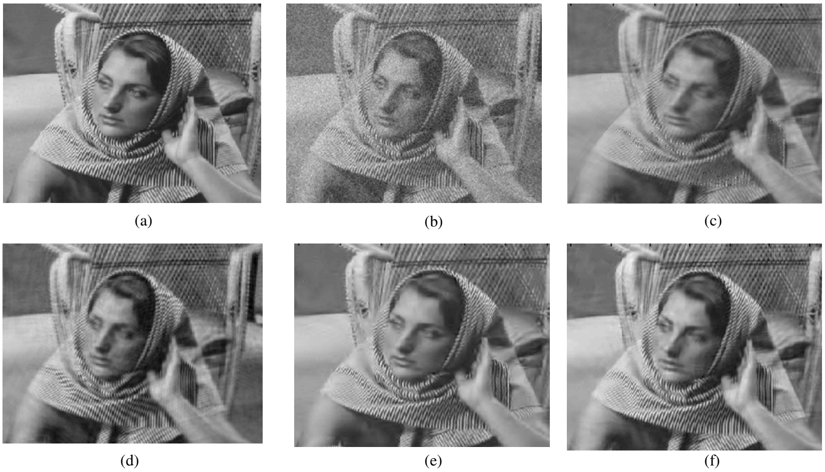
where $c_1 = (K_1 \times L)^2$ and $c_2 = (K_2 \times L)^2$

here K_1 and K_2 have the default values of 0.01 and 0.03 respectively. L is the dynamic range of pixel and for 8 bit image, its value is 255.

Table 1. Denoising Results of Different Methods Compared for Four Assessment Parameters

UPPER LEFT BLOCK: PSNR, UPPER RIGHT BLOCK: SSIM
LOWER LEFT BLOCK: UQI, LOWER RIGHT BLOCK: FOM

Image	Noise std. dev.	DCuT [2]		WA [5]		ProbShrink [7]		[15]		**Proposed**	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
		UQI	FOM	UQI	FOM	UQI	FOM	UQI	FOM	UQI	FOM
Lena	10	33.77	0.909	34.50	0.9547	35.24	0.9658	35.7	0.929	35.82	0.9750
		0.697	0.8348	0.7547	0.8541	0.7732	0.9142	0.735	0.9196	0.7734	0.9203
	20	31.08	0.859	30.97	0.9076	32.20	0.9309	32.49	0.886	32.52	0.9330
		0.597	0.7452	0.6412	0.7305	0.6836	0.7966	0.637	0.8429	0.6887	0.8536
	30	29.39	0.819	28.95	0.8622	30.33	0.8947	30.61	0.851	31.12	0.9000
		0.531	0.6604	0.5586	0.6336	0.6059	0.7181	0.571	0.7694	0.6093	0.7652
Barbara	10	29.18	0.875	32.56	0.9584	33.46	0.9685	33.5	0.933	33.61	0.9699
		0.754	0.8302	0.8192	0.9114	0.8409	0.9280	0.824	0.9598	0.8575	0.9603
	20	25.41	0.769	29.31	0.9140	29.53	0.9302	30.03	0.879	30.10	0.9333
		0.617	0.6275	0.7303	0.8360	0.7581	0.8530	0.736	0.9118	0.7597	0.9093
	30	24.35	0.711	27.52	0.8713	27.17	0.8816	28.07	0.831	28.13	0.8999
		0.543	0.5263	0.6630	0.7778	0.6718	0.7291	0.670	0.8643	0.6993	0.8597



(a) (b) (c)

(d) (e) (f)

Fig. 4. The denoising results of *Barbara* by different schemes. (a) Noiseless *Barbara* (b) Noisy *Barbara* (c) DCuT [2] (d) WA [5] (e) ProbShrink [7] (f) Proposed method.

4 Conclusion

The main contribution of this paper is to adaptively combine the features of wavelets and wave atoms for homogenous and texture regions in an image. Thus this method independently restores the attributes of both the transforms. It can be observed from the denoising results that for almost all images with various noise levels significant improvement in the values of PSNR is obtained. Improvements in the restored images are observed not only in PSNR values but also in the values of SSIM, FOM and UQI for images containing high quantity of texture. The uniqueness of this method is that it preserves the texture and at the same time removes the artifacts and the repetitions caused by wave atoms in the smoother areas. The proposed approach also tends to efficiently retain the edges, the structural similarity and the luminance and contrast correlation as can be seen from the results of FOM, UQI and SSIM values.

References

[1] Po, D.D.-Y., Do, M.N.: Directional Multiscale Modeling of Images Using the Contourlet Transform. IEEE Trans. Image Processing 15(6), 1610–1620 (2006)

[2] Starck, J.L., Candes, E.J., Donoho, D.L.: The Curvelet Transform for Image Denoising. IEEE Trans. Image Processing 11(6), 670–684 (2002)

[3] Donoho, D.: Wedgelets: Nearly Minimax Estimation of Edges. Ann. Statistics 27(3), 859–897 (1999)

[4] Le Pennec, E., Mallat, S.: Sparse Geometrical Image Approximation with Bandlets. IEEE Trans. Image Processing 14(4), 423–438 (2005)

[5] Demanet, L., Ying, L.: Wave Atoms and Sparsity of Oscillatory Patterns. Applied and Computational Harmonic Analysis 23(3), 368–387 (2007)

[6] Swami, P.D., Jain, A., Singhai, J.: A Multilevel Shrinkage Approach for Curvelet Denoising. In: IEEE Proc. of International Conference on Information and Multimedia Technology, Jeju Island, Korea, pp. 268–272 (2009)

[7] Pizurica, A., Philips, W.: Estimating the Probability of the Presence of a Signal of Interest in Multiresolution Single- and Multiband Image Denoising. IEEE Trans. Image Process. 15(3), 654–665 (2006)

[8] Tessens, L., Pizurica, A., Alecu, A., Munteanu, A., Philip, W.: Context Adaptive Image Denoising through Modeling of Curvelet Domain Statistics. Journal of Electronic Imaging 17(3), 033021-1—033021-17 (2008)

[9] Liu, J., Moulin, P.: Information-Theoretic Analysis of Interscale and Intrascale Dependencies between Image Wavelet Coefficients. IEEE Trans. Image Processing 10(11), 1647–1658 (2001)

[10] Nasri, M., Pour, H.N.: Image Denoising in the Wavelet Domain using a New Adaptive Thresholding Function. Elsevier J. Neurocomputing 72, 1012–1025 (2009)

[11] Bhutada, G.G., Anand, R.S., Saxena, S.C.: PSO-based Learning of Sub-band Adaptive Thresholding Function for Image Denoising. Signal Image and Video Processing (2010), doi:10.1007/s11760-010-0167-7

[12] Elad, M., Aharon, M.: Image Denoising via Sparse and Redundant Representations Over Learned Dictionaries. IEEE Trans. Image Processing 15(12), 3736–3745 (2006)

[13] Elad, M., Figueiredo, M.A.T., Ma, Y.: On the Role of Sparse and Redundant Representation in Image Processing. IEEE Proceedings - Special Issue on Applications of Sparse Representations and Compressive Sensing 98(6), 972–982 (2010)

[14] `http://www.waveatom.org/`

[15] Bhutada, G.G., Anand, R.S., Saxena, S.C.: Edge Preserved Image Enhancement using Adaptive Fusion of Images Denoised by Wavelet and Curvelet Transform. Digital Signal Processing 21, 118–129 (2011)

[16] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment from Error Visibility to Structural Similarity. IEEE Trans. Image Process. 13(4), 600–612 (2004)

[17] Wang, Z., Bovik, A.C.: Universal Image Quality Index. IEEE Signal Process. Lett. 9(3), 81–84 (2002)

[18] Pratt, W.K.: Digital Image Processing, 3rd edn. John Wiley and Sons (2006)

An Efficient De-noising Technique for Fingerprint Image Using Wavelet Transformation

Ashish Kumar Dass[1] and Rabindra Kumar Shial[2]

[1] Dept. of CSE, National Institute of Science & Technology, Berhampur, Odisha
ashishkumardass@yahoo.co.in
[2] Asst. Proff in CSE, National Institute of Science & Technology, Berhampur, Odisha
rkshial@yahoo.com

Abstract. Fingerprint acts as a vital role for user authentication as it is unique and not duplicated. For this reason fingerprint images are taken for different computer security purposes. Unfortunately reference fingerprints may get corrupted with noise during acquisition, transmission, or retrieval from storage media. Many image-processing algorithms such as pattern recognition need a clean fingerprint image to work effectively which in turn needs effective ways of de-noising such images. In this paper, we propose an adaptive method of image de-noising in the wavelet sub-band domain assuming the images to be contaminated with noise based on threshold estimation for each sub-band. Under this framework, the proposed technique estimates the threshold level by apply sub-band of each decomposition level. This paper entails the development of a new MATLAB function based on our algorithm. The experimental evaluation of our proposition reveals that our method removes noise more effectively than the in-built function provided by MATLAB.

Keywords: Wavelet Thresholding, Gaussian, Salt & Pepper noise, Fingerprint Image De-noise, Discrete Wavelet Transform.

1 Introduction

Finger print images have distinctiveness and persistence, which are highly desirable qualities for biometric applications and software security concerns. However, finger print images are generally of low contrast, due to skin conditions and application of incorrect finger pressure. Also, they inherently contain complex type of noise, originating from two distinct sources, such as the set of assorted devices involved in the acquisition, transmission, storage and display of the image and noise arising from the application of different types of quantization, reconstruction and enhancement algorithms. It is certain that every imaging method inherently involves noise. Many dots can be spotted in a Photograph of fingerprint taken with a digital camera or fingerprint reader under low lighting conditions or the machine hardware problem. Actually this type of noise is the uniform Gaussian noise. Appearance of dots is due to the real signals getting corrupted by noise (unwanted signals). On loss of reception or retrieve any Fingerprint image from the storage device random black and white

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 709–717.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

snow-like patterns can be seen on the Fingerprint images. This type of noise is called Salt & Pepper noise. The purpose of the de-noising algorithm is to remove such noise.

There are various methods to help restore an image from noisy distortions. Selecting the appropriate method plays a major role in getting the desired image to solve the de-noising problems in image analysis and pattern recognition. Generally, the de-noising techniques have been categorized into spatial and frequency domain techniques. The past experience has reveals that the wavelet technique is an efficient technique in comparison of others. In this paper a new shrinkage wavelet transformation method is proposed using the global threshold value, normalise it with all decomposed components and find out the rescaled threshold value. This method is an efficient technique compared to the MATLAB wavelet transformation and the various linear and non-linear spatial techniques. In this paper first the testing fingerprint image is noised with the Gaussian and Salt & Pepper noise differently. After that the proposed wavelet transformation is adapted in order to de-noise the fingerprint images, followed with the various other methods such as mean filter, median filter, library Matlab wavelet transformation techniques to de-noise the fingerprints image and lastly check which one is the best in terms of Pick signal to noise ratio(PSNR), Mean square error(MSE).

The paper is organized as follows. Section 2 relates to the existing work done in fingerprint de-noising whereas section 3 show brief introduction of wavelet transform and fingerprint de-noising. In section 4, new approach for fingerprint de-noising along with algorithm design. The proposed work is detailed in section 5 followed by conclusion in section 6.

2 Related Work

Maltoni D. Has proposed various methods and problems for fingerprint recognition. He has given idea how fingerprint get different noises with different stages of processing [1]. Louise has proposed fingerprint recognition for low quality images and emphasized upon ridge detection and Improved algorithms for enhancement of fingerprint images[2]. S.G.mallat described how to singularity detection using the wavelet transformation. Amra Graps uses the various wavelet technique as well as its importance from other de-noising techniques [4,8]. Rakesh has given the idea in order to utilise the wavelet transformation in fingerprint recognition [6]. Gornale S.S has given the idea to de-noise fingerprint using multi-resolution analysis through stationary wavelet transformation, which have the adaptive normalization based on block processing, are proposed. An orientation flow field of the ridges is computed for the fingerprint image. To accurately locate ridges, a ridge orientation based computation method is used [5]. But this method used the library Matlab function which is less efficient in order to de-noise. Zhen_bing Zhao has given a better idea for de-noising using wavelet transformation based on noise standard deviation estimation [3]. So the fingerprint image transformed by wavelet domain by an efficient way de-noise gives a better result and fulfils various authentication and pattern recognition methods.

3 Methodology

3.1 Wavelet Transform

Basically image de-noising techniques are fall into two basic categories namely spatial domain and frequency domain. Wavelet Transform (WT) is one of the frequency domain techniques emerged as very powerful tool and provide a vehicle for digital image processing applications.

A wavelet is a small wave with finite energy, which has its energy concentrated in time or space area to give ability for the analysis of time-varying phenomenon in other words it provides a time-frequency representation of the signal. Comparison of a wave with a wavelet is shown in below Figure1. Left graph is a Sine Wave with infinite energy and the right graph is a Wavelet with finite energy.

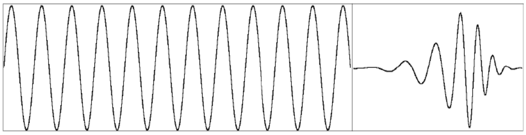


Fig. 1. Comparison of a wave and a wavelet

Fig.1 shows the comparison of wavelets with sine waves, which are the basis of Fourier analysis. Sinusoids do not have limited duration they extend from minus to plus infinity. Where sinusoids are smooth and predictable, wavelets tend to be irregular and asymmetric. Fourier analysis consists of breaking up a signal into sine waves of various frequencies. Similarly, wavelet analysis is the breaking up of a signal into shifted and scaled versions of the original (or mother) wavelet. Fig.1 shows that signals with sharp changes might be better analyzed with an irregular wavelet than with a smooth sinusoid. It also makes sense that local features can be described better with wavelets that have local extent. So wavelet has advantages in analyzing physical situations where the signal contains discontinuities and sharp spikes. Wavelets were developed independently in the fields of mathematics, quantum physics and electrical engineering. Wavelet Transform is used to split the signal into a bunch of signals and represents the same signal, but all corresponding to different frequency bands. The principle advantage is they provide frequency bands exists at what time intervals. Wavelet transform of any function f (t) represented as

$$\gamma(s,\tau) = \int f(t)\psi_{s,\tau}(t)dt \tag{1}$$

This equation shows how a function f(t) is decomposed into a set of basis functions called the wavelets. The variables s and, scale and translation, are the new dimensions after the wavelet transform.

Inverse wavelet transformation can be expressed as:

$$f(t) = \iint \gamma(s,\tau)\psi_{s,\tau}(t)d\tau ds \tag{2}$$

The wavelets are generated from a single basic wavelet ψ (t), the so-called mother wavelet, by scaling and translation:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right) \tag{3}$$

Applying wavelet transform on 1D signal, it can correctly detect the singularity in a signal. For images, the 2D scaling function φ (x, y) and mother wavelet ψ (x, y) is defined as tensor products of the following 1-D wavelets ψ (x), ψ (y) and scaling functions φ (x), φ (y).

Scaling function

$$\varphi (x, y) = \varphi (x) \times \varphi (y) \tag{4}$$

Vertical wavelets

$$\psi\, y (x, y) = \varphi (x) \times \psi (y) \tag{5}$$

Horizontal wavelets

$$\psi\, x (x, y) = \psi (x) \times \varphi (y) \tag{6}$$

Diagonal wavelets

$$\psi\, d (x, y) = \psi (x) \times \psi (y) \tag{7}$$

The use of wavelet transform on image shows that the transform can analyze singularities easily that are horizontal, vertical or diagonal.

3.2 *Wavelet* Thresholding

Image de-noising is used to remove the additive noise while retaining as much as possible the important features. Wavelet thresholding is an effective method which is achieved via thresholding. Wavelet thresholding procedure removes noise by thresholding only the wavelet coefficient of the details coefficients, by keeping the low-resolution coefficients unaltered. The plot of wavelet coefficients suggests that small coefficients are dominated by noise, while coefficients with a large absolute value carry more signal information than noise. Replacing noisy coefficients (small coefficients below a certain threshold value) and an inverse wavelet transform may lead to a reconstruction that has lesser noise. There are two thresholding methods frequently used: soft thresholding and hard thresholding.

The hard-thresholding T_H can be defined as

$$T_H = \begin{cases} x & \text{for } |x| \geq t \\ 0 & \text{in all other regions.} \end{cases} \tag{8}$$

Here t is the threshold value. A plot of T_H is shown in Figure.

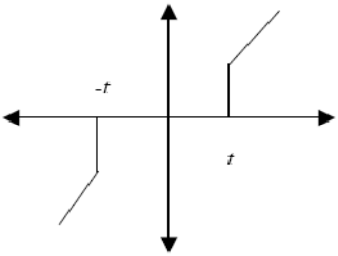


Fig. 2. Hard thresholding

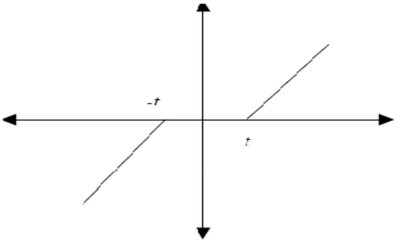


Fig. 3. Soft thresholding

Soft thresholding is where the coefficients with greater than the threshold are shrunk towards zero after comparing them to a threshold value. It is defined as follows in all other regions.

$$T_s = \begin{cases} \operatorname{sign}(x)(|x|-t) & \text{for } |x|>t \\ 0 & \text{in all other regions.} \end{cases} \tag{9}$$

In practice, it can be seen that the soft method is much better and yields more visually pleasant images. This is because the hard method is discontinuous and yields abrupt artifacts in the recovered images. Also, the soft method yields a smaller minimum mean squared error compared to hard form of thresholding. In this paper we have used the soft thresholding technique.

3.3 Fingerprint De-noising

A fingerprint image consists of non-ridge area, high quality ridge area, and low quality ridge area. It is well known that low quality ridge area in the fingerprint images would cause serious effects, which deteriorate the quality of the image. The Finger print image is corrupted with the Gaussian and Salt & Pepper noise. Many dots can be spotted in a Photograph of fingerprint taken with a digital camera or fingerprint reader under low lighting conditions or the machine hardware problem. Actually this type of noise is the uniform Gaussian noise. Appearance of dots is due to the real signals getting corrupted by noise (unwanted signals). On loss of reception or retrieve any Fingerprint image from the storage device random black and white snow-like patterns can be seen on the Fingerprint images. This type of noise is called Salt & Pepper noise. The resulting sub-image is extracted from the original fingerprint image with noise in the complex wavelet transform domain. Then, according to the characteristics of the sub-image data, the de-noised fingerprint is being used for further reference purposes.

4 Algorithm Design

Having fully analyzed the different condition characteristic of the useful signal and the noise in wavelet transformation domain, the above wavelet de-noising theory and corrected noise-estimate method are adopted to smooth the noise. This article proposed the image de-noising method based on the noise standard deviation estimation, normalize each detailed component, finding out the appropriate threshold value realizing steps are as follows:

Add the Gaussian noise and Salt & Pepper noise to the reference fingerprint Image.

Carry on the multi-scale wavelet decomposition to the observed image f (x, y) and obtain the low and the high frequency coefficients of each level.

Estimate the noise standard deviation σ by using the detail coefficients.

Determine the threshold value t by using the normalization of each level and producing the threshold value by global threshold method.

Use soft-threshold/hard-threshold function to make threshold processing to the each frequency coefficient, and obtained the estimate coefficient.

Realize de-noising and reconstruction by making wavelet inverse transformation to the low frequency coefficients and the processed high frequency coefficients.

5 Results and Discussions

This section shows the proposed method which consists different modules as shown in Fig. 4. In first module the test fingerprint Image is noised with Gaussian or Salt & Pepper noise. Then the noised image is De-noised by the proposed wavelet transformation which is described in the algorithm, after that apply multilevel wavelet

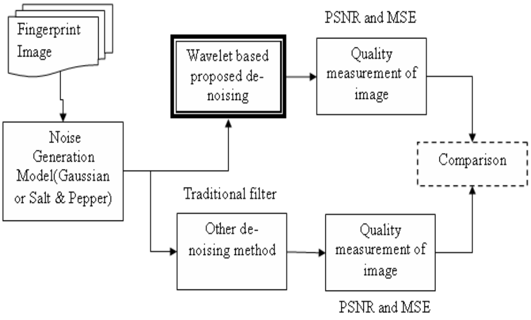
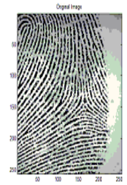

Fingerprint Image
Noise Generation Model(Gaussian or Salt & Pepper)
Wavelet based proposed de-noising
PSNR and MSE
Quality measurement of image
Comparison
Traditional filter
Other de-noising method
Quality measurement of image
PSNR and MSE

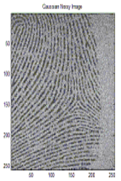
Fig. 4. Overall work layout

decomposition on the extracted Region of interest (ROI). At each level, the wavelet transform decompose the given image into three directional components, i.e. horizontal, diagonal and vertical detail sub bands.

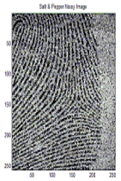
In next module the tested fingerprint Image is de-noised by mean, median and Matlab available wavelet function. Lastly all the de-noised algorithms are compared with the MSE and PSNR for quality measurement. We have use Matlab 7.0 to noised and de-noised the fingerprint image by proposed wavelet transformation, Library wavelet transformation, Mean and Median techniques. And different outputs of the programs are shown below.



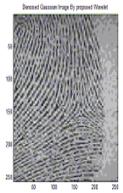
(original Image)



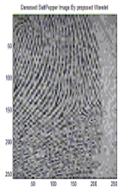
(Gaussian noised image)



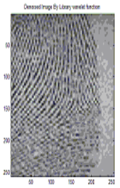
(Salt & pepper noised image)



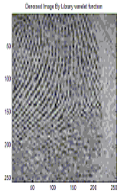
(propose wavelet Gaussian)



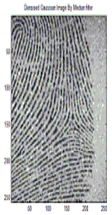
(Library wavelet Salt & pepper)



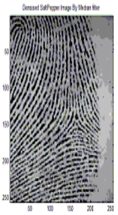
(Library wavelet Gaussian)



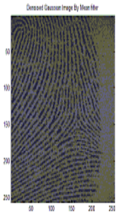
(propose wavelet Salt & pepper)



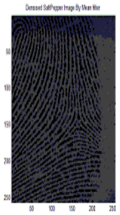
(Median Gaussian)



(Median Salt & Pepper)



(Mean Gaussian)



(Mean Salt & Pepper)

In this section deals with the comparison of the de-noising techniques .The Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE) of the output image is calculated which acts as a quantitative standard for comparison. The selection of the de-noising technique is application dependent. So, it is necessary to learn and compare de-noising techniques to select the technique that is apt for the application in which we are interested.

The Peak Signal to Noise Ratio (PSNR) is most commonly used as a measure of quality of reconstruction in image compression and image de-noising works [6]. It comes from mean square error (MSE). MSE of two images are defined as

$$\mathrm{MSE} = {}^{1}\!/_{mn} \left(\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I(i,j) - R(i,j))^2 \right)$$

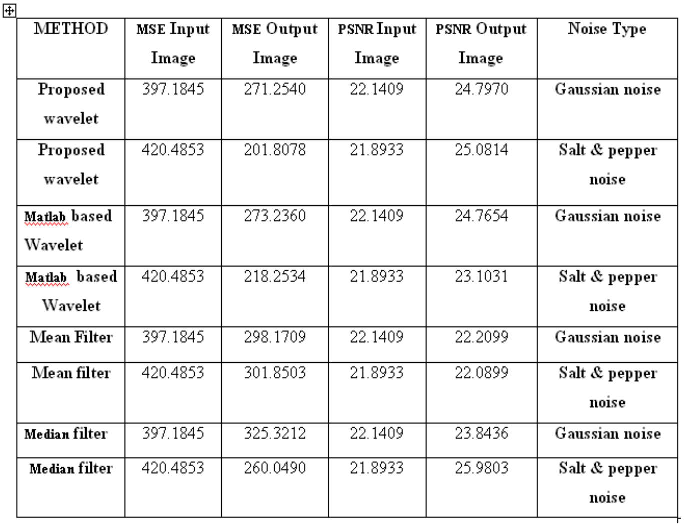
where *I* and *R* can be interpreted as input and reconstructed images respectively. *m* and *n* defines number of pixel in vertical and horizontal dimension of images I and R. Then the PSNR is de fined as

$$\mathrm{PSNR} = 10.\log_{10}\left(MAX_I^2/MSE\right) = 20.\log_{10}\left({}^{MAX_I}\!/_{\sqrt{MSE}} \right)$$

where MAX_I is the maximum pixel value of the image I. When the pixels are represented using 8 bits per sample "*grey scale*", MAX_I takes value "255". More generally, MAX_I is 2*B*-1 where *B* is a colour dept "*bit*" of an image.

Tables 1 shows the MSE and PSNR of the input and output images for all the filtering approach and wavelet transform approach.

Table 1. PSNR and MSE for Fingerprint.bmp as test Image



METHOD	MSE Input Image	MSE Output Image	PSNR Input Image	PSNR Output Image	Noise Type
Proposed wavelet	397.1845	271.2540	22.1409	24.7970	Gaussian noise
Proposed wavelet	420.4853	201.8078	21.8933	25.0814	Salt & pepper noise
Matlab based Wavelet	397.1845	273.2360	22.1409	24.7654	Gaussian noise
Matlab based Wavelet	420.4853	218.2534	21.8933	23.1031	Salt & pepper noise
Mean Filter	397.1845	298.1709	22.1409	22.2099	Gaussian noise
Mean filter	420.4853	301.8503	21.8933	22.0899	Salt & pepper noise
Median filter	397.1845	325.3212	22.1409	23.8436	Gaussian noise
Median filter	420.4853	260.0490	21.8933	25.9803	Salt & pepper noise

6 Conclusion and Future Work

In this paper we have seen the wavelet technique is better than the traditional mean and median spatial transformation techniques and the proposed wavelet function also de-noised the fingerprint better than the MATLAB wavelet function for Gaussian noised in terms of PSNR and MSE. If the noise is Salt & pepper type than by using Median filter gives better noise removal. The proposed method also nearly gives better quality as compared the median filter technique and better than other techniques. The de-noised fingerprint which we achieved, are more helpful for Automatic Fingerprint Recognition Systems or any pattern matching techniques. In future the work can be extended for other type of noises such as speckle noise, rician noise etc. and recover from the blurring effect of the fingerprint.

References

1. Maltoni, D., Maio, D., Jain, A.K., Prabahar, S.: Handbook of Fingerprint Recognition. Springer (2009)
2. Coetzee, L., Botha, E.C.: Fingerprint Recognition in Low Quality Images. Pattern Recognition 26(10), 1441–1460 (1993)
3. Zhao, Z.-B., Yuan, J.-S., Gao, Q., Kong, Y.-H.: Wavelet Image De-Noising Method Based On Noise Standard Deviation Estimation. In: Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition, Beijing, China, November 2-4 (2007)
4. Mallat, S.G., Hwang, W.L.: Singularity detection and processing with wavelets. IEEE Trans. Inform. Theory 38, 617–643 (1992)
5. Gornale, S.S., Humbe, V., Manza, R., Kale, K.V.: Fingerprint image de-noising using multi-resolution analysis (MRA) through stationary wavelet transform (SWT) method. International Journal of Knowledge Engineering 1(1), 5–14 (2010) ISSN: 0976–5816
6. Verma, R., Goel, A.: Wavelet Application in Fingerprint Recognition. International Journal of Soft Computing and Engineering (IJSCE) 1(4) (September 2011) ISSN: 2231-2307
7. Hrechak, A.K., McHugh, J.A.: Automatic Fingerprint Recognition using Structural Matching. Pattern Recognition 23, 893–904 (1990)
8. Graps, A.: An Introduction to Wavelets. IEEE Computational Science and Engineering 2(2) (Summer 1995)
9. Berry, J., Stoney, D.A.: The history and development of finger printing in advances in fingerprint technology, 2nd edn., pp. 1–40. CRC Press (2001)
10. Yang, R., Yin, L., Gabbouj, M., Astola, J., Neuvo, Y.: Optimal weighted median filters under structural constraints. IEEE Trans. Signal Processing 43, 591–604 (1995)
11. Ben Hamza, A., Luque, P., Martinez, J., Roman, R.: Removing noise and preserving details with relaxed median filters. J. Math. Imag. Vision 11(2), 161–177 (1999)
12. Jain, A.K.: Fundamentals of digital image processing. Prentice-Hall (1989)
13. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaption via wavelet shrinkage. Biometrika 81, 425–455 (1994)
14. Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association 90(432), 1200–1224 (2001)
15. Strela, V.: Denoising via block Wiener filtering in wavelet domain. In: 3rd European Congress of Mathematics, Barcelona. Birkhäuser (July 2000)
16. Zheng, J.-D., Gao, Y., Zhang, M.-Z.: Fingerprint Matching Algorithm Based on Similar Vector Triangle. In: Second International Congress on Image and Signal Processing, pp. 1–6 (2009)
17. Humbe, V., Gornale, S.S., Magar, G., Manza, R., Kale, K.V.: Fingerprint Image De-noising through Decimated and Un-decimated Wavelet Transforms (WT). In: International Conference on Future Computer and Communication, ICFCC 2009, pp. 500–504 (2009), doi:10.1109/ICFCC.2009.101
18. Technique on image denoising in wavelet transform domain CAI Hantian (South China University of Technology, Guangzhou 510641), doi: CNKI:SUN:GXJS.0.1998-06-002

Detection of Optic Disc by Line Filter Operator Approach in Retinal Images

R. Murugan[1] and Reeba Korah[2]

[1] Research Scholar, Centre for Research, Anna University of Technology, Chennai
`murugan.rmn@gmail.com`
[2] Professor, St. Joseph's College of Engineering, Chennai
`reeba26in@yahoo.co.in`

Abstract. The location of Optic Disc (OD) is of critical importance in retinal image analysis. This research paper carries out a new automated methodology to detect the optic disc (OD) in retinal images. OD detection helps the ophthalmologists to find whether the patient is affected by diabetic retinopathy or not. The proposed technique is to use line operator which gives higher percentage of detection than the already existing methods. The purpose of this project is to automatically detect the position of the OD in digital retinal fundus images. The method starts with converting the RGB image input into its LAB component. This image is smoothed using bilateral smoothing filter. Further, filtering is carried out using line operator. After which gray orientation and binary map orientation is carried out and then with the use of the resulting maximum image variation the area of the presence of the OD is found. The portions other than OD are blurred using 2D circular convolution. On applying mathematical steps like peak classification, concentric circles design and image difference calculation, OD is detected. The proposed method was evaluated using a subset of the STARE project's dataset and the success percentage was found to be 96%.

Keywords: Automated detection, Optic Disc, Line Filter Operator, Retinal Imaging.

1 Introduction

In ophthalmology, the automatic detection of optic disc may be of considerable interest for computer assisted diagnosis. Detecting and counting lesions in the human retina like microaneurysms and exudates is a time consuming task for ophthalmologists and prone to human error. That is why much effort has been done to detect lesions in the human retina automatically. Finding the main components in the fundus images helps in characterizing detected lesions and in identifying false positives. The detection of the optic disc is the first step in the early detection of the diabetic retinopathy.

This paper presents a line operator that is designed to locate the OD from retinal images accurately. Our proposed line operator is designed to capture the circular brightness structure associated with the OD. In particular, it evaluates the image variation along multiple oriented line segments and locates the OD based on the

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 719–728.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

orientation of the line segment with the maximum/minimum variation. Fig. 1(a & b) shows an example of retinal image in digital retinal images for vessel extraction (DRIVE) project's dataset.

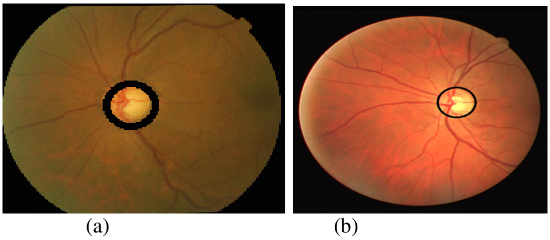

(a) (b)

Fig. 1. Circular brightness structure associated with the OD.(a & b) Example of retinal image in DRIVE project's dataset with OD labelled by a bold black circle.

The proposed method has several advantages. First, the designed line operator is tolerant to the retinal lesion and various types of imaging artifacts that most image-characteristics-based methods cannot handle properly. The tolerance to the imaging artifacts and retinal lesion can be explained by the proposed line operator that is designed to capture the unique circular brightness structure associated with the OD. Second, the designed line operator is stable and easy for implementation. It requires neither the retinal blood vessel nor the macula information.

The remaining part of the paper is organized as follows. Most of the available methods for automatic OD detection are reviewed in Section 2. Section 3 presents the proposed algorithm. The results are presented and discussed in Sections 4. Finally, conclusion and further work are found in Section 5.

2 OD detection Methods: A Literature Review

Automatic optic disc (OD) detection from retinal images is a very important task in ocular image analysis [1], [2] and computer-aided diagnosis of various types of eye diseases [3]–[5]. It is often a key step for the detection of other anatomical retinal structures, such as retinal blood vessels and macula [1], [6], [7], [8]. More importantly, it helps to establish a retinal coordinate system that can be used to determine the position of other retinal abnormalities, such as exudates, drusen, and hemorrhages [9], [10].

Some OD detection techniques have been reported in the literature. The early techniques make use of different types of OD specific image characteristics. In particular, some techniques search for the brightest regions [11], [12] or regions with the highest image variation [13], [14] resulting from the bright OD and the dark blood vessels within the OD. The limitation of these methods is that many retinal images suffer from various types of retinal lesion, such as drusen, exudates, microaneurysms, and hemorrhage, and imaging artifacts, such as haze, lashes, and uneven illumination (as illustrated in Figs. 9 and 10) that often produce brighter regions or regions with higher image variation compared to the OD.

Several OD detection techniques make use of anatomical structures among the OD, macula, and retinal blood vessels. For example, some methods are based on the anatomical structure that all major retinal blood vessels radiate from the OD [15]–[18]. Some other methods make use of the relative position between the OD and the macula that often varies within a small range [19], [20]. Compared with the image characteristics, the anatomical structures are more reliable under the presence of retinal lesion and imaging artifacts. However, the extraction of either retinal blood vessels or the macula is often a nontrivial task by itself.

Line operators have been used to locate linear structures from different types of images. For example, Zwiggelaar et al. used a line operator to detect linear structures from mammographic images [21], where a line strength is evaluated by the difference between the largest average image intensity along one oriented line segment and the average image intensity within a local neighbourhood window. Ricci and Perfetti [22] used a similar line operator to detect the linear structures that are associated with the retinal blood vessels.

3 Proposed Method

This section presents the proposed OD detection technique. In particular, we divide this section into four subsections, which deal with the retinal image pre-processing, designed line operator, OD detection, and discussion, respectively.

3.1 Retinal Image Pre-processing

Retinal images need to be pre-processed before the OD detection. As the proposed technique makes use of the circular brightness structure of the OD, the lightness component of a retinal image is first extracted. We use the lightness component within the LAB color space, where the OD detection usually performs the best [23]. For the retinal image in Fig. 1(a), Fig. 2(a) shows the corresponding lightness image.

The retinal image is then smoothed to enhance the circular brightness structure associated with the OD. We use a bilateral smoothing filter [24] that combines geometric closeness and photometric similarity as follows:

$$h(x) = k^{-1}(x) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi)c(\xi, x)s(f(\xi); f(x))d\xi \tag{1}$$

With the normalization factor

$$k(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c(\xi, x)s(f(\xi); f(x))d\xi \tag{2}$$

where f (x) denotes the retinal image under study. c(ξ, x) and s(f(ξ), f (x)) measure the geometric closeness and the photometric similarity between the neighborhood center x and a nearby point ξ. We set both c(ξ, x) and s(f(ξ), f(x)) by Gaussian functions.

The geometric spread σ_d and the photometric spread σ_r of the two Gaussian functions are typically set at 10 and 1 as reported in [24]. For the retinal image in Fig. 2(a), Fig. 2(b) shows the filtered retinal image.

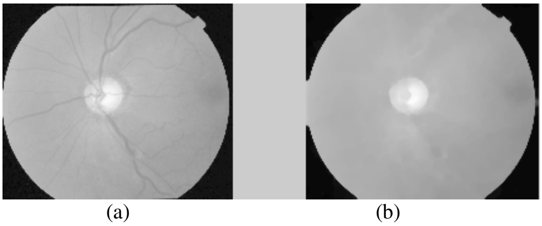

(a) (b)

Fig. 2. Retinal image pre-processing. (a) Lightness of the example retinal image in LAB color space. (b) Enhanced retinal image by bilateral smoothing where multiple crosses along a circle label the pixels to be used to illustrate the image variation along multiple oriented line segments.

3.2 Designed Line Operator

A line operator is designed to detect circular regions that have similar brightness structure as the OD. For each image pixel at (x, y), the line operator first determines n line segments L_i, i = 1. . . n of specific length p (i.e., number of pixels) at multiple specific orientations that center at (x, y). The image intensity along all oriented line segments can thus be denoted by a matrix $I(x, y)_{n\times p}$, where each matrix row stores the intensity of p image pixels along one specific line segment. The line operator that uses 20 oriented line segments and sets the line length p = 21, each line segment L_i at one specific orientation can be divided into two line segments $L_{i,1}$ and $L_{i,2}$ of the same length (p − 1)/2 by the image pixel under study (i.e., the black cell). The image variation along the oriented line segments can be estimated as follows:

$$D_i(x,y) = \| f_{\mathrm{mdn}}(I_{L_{i,1}}(x,y)) - f_{\mathrm{mdn}}(I_{L_{i,2}}(x,y)) \|, \quad i = 1,\ldots,n \tag{3}$$

Where $f_{mdn}(\cdot)$ denotes a median function. $f_{mdn}(IL_{i,1}(x, y))$ and $f_{mdn}(IL_{i,2}(x, y))$ return the median image intensity along $L_{i,1}$ and $L_{i,2}$, respectively. $D = [D_1(x, y)\ldots D_i(x, y)\ldots D_n(x, y)]$ is, therefore, a vector of dimension n that stores the image variations along n-oriented line segments.

The orientation of the line segment with the maximum/minimum variation has specific pattern that can be used to locate the OD accurately. For retinal image pixels, which are far away from the OD, the orientation of the line segment with the maximum/minimum variation is usually arbitrary, but for those around the OD, the image variation along Lc [labeled in Fig. 1(b)] usually reach the maximum, whereas that along Lt reaches the minimum. Fig. 4 shows the image variation vectors D(x, y) of eight pixels that are labeled by crosses along a circle shown in Fig. 2(b). Suppose that there is a Cartesian coordinate system centered at the OD, as shown in Fig. 2(b). For the retinal image pixels in quadrants I and III, the image variations along the 1st–10th [i.e., Lt in Fig. 1(b)] and the 11th–20th (i.e., Lc) line segments labeled in Fig. 3 reach the minimum and the maximum, respectively, as shown in Fig. 4. But for the retinal image pixels in quadrants II and IV, the image variations along the 1st–10th and the 11th–20th line segments instead reach the maximum and the minimum, respectively.

An orientation map can, therefore, be constructed based on the orientation of the line segment with the maximum (or minimum) variation as follows:

$$O(x, y) = \underset{i}{\operatorname{argmax}}\, D(x, y) \tag{4}$$

where D(x, y) denotes the image variation vector evaluated in (3). In addition, a binary orientation map can also be constructed by classifying the orientation of the line segment with the maximum variation into two categories as follows:

$$Q(x, y) = \begin{cases} -1, & \text{if } \underset{i}{\operatorname{argmax}}\, D(x, y) < \frac{n}{2} + 1 \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

where n refers to the number of the oriented line segments used in the line operator.

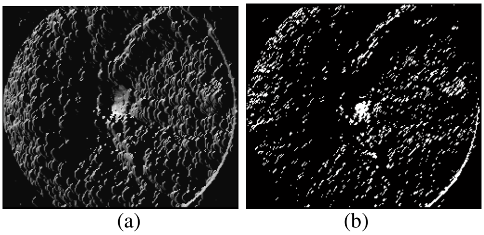

(a) (b)

Fig. 3. Orientation map of the retinal image in Fig. 2(b). (a) Gray orientation map that is determined by using (4). (b) Binary orientation map that is determined by using (5).

For the retinal image in Fig. 1(a), Fig. 3(a) and (b) shows the determined gray orientation map and binary orientation map, respectively. As shown in Fig. 3(a), for retinal image pixels in quadrants I and III around the OD, the orientation map is a bit dark because the orientation of the line segment with the maximum variation usually lies between 1 and (n/2) + 1. However, for retinal image pixels in quadrants II and IV, the orientation map is bright because the orientation of the line segment with the maximum variation usually lies between n/2 and n. The binary orientation map in Fig. 3(b) further verifies such orientation pattern. The OD will then be located by using the orientation map to be described in the following.

3.3 OD Detection

We use a line operator of 20 oriented line segments because line operators with more line segments have little effect on the orientation map. The line length p is set as follows:

$$p = k R \tag{6}$$

where R denote the radius of the central circular region of retinal images as illustrated in Fig. 1(a). Parameter k controls the line length, which usually lies between 1/10 and 1/5 based on the relative OD size within retinal images [25]. The use of R incorporates possible variations of the image resolution.

The specific pattern within the orientation map is captured by a 2-D circular convolution mask shown at the upper left corner of two peak images in Fig. 4.b. As shown in Fig. 4, the convolution mask can be divided into four quadrants, where the cells within quadrants I and III are set at–1, whereas those within quadrants II and IV are set at 1 based on the specific pattern within the orientation map. An orientation map can thus be converted into a peak image as follows:

$$P(x,y) = \sum_{x=x_0-m}^{x_0+m} \sum_{y=y_0-m}^{y_0+m} M(x,y)O(x,y) \tag{7}$$

where (x_0 y_0) denotes the position of the retinal image pixel under study. M(x, y) and O(x, y) refer to the value of the convolution mask and the orientation map at (x, y), respectively. Parameter m denotes the radius of the circular convolution mask that can be similarly set as p.

For the orientation maps in Fig. 5(a) and (b), Fig. 6(a) and (b) shows the determined peak images. As shown in Fig. 4, a peak is properly produced at the OD

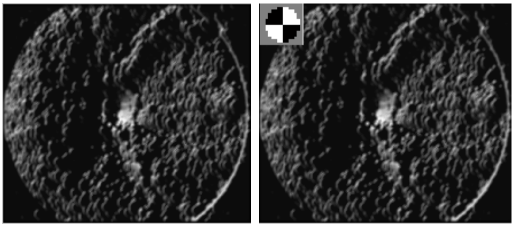

(a) (b)

Fig. 4. Peak images determined by a 2-D circular convolution mask shown in the upper left corner. (a) Peak image produced through the convolution of the gray orientation map in Fig. 3(a). (b) Peak image produced through the convolution of the binary orientation map in Fig. 3(b).

position. On the other hand, a peak is also produced at the macula center (i.e., fovea) that often has similar peak amplitude to the peak at the OD center. This can be explained by similar brightness variation structure around the macula, where the image variation along the line segment crossing the macula center reaches the maximum, whereas that along the orthogonal line segment [similar to Lc and Lt in Fig. 1(b)] reaches the minimum. The only difference is that the OD center is brighter than the surrounding pixels, whereas the macula center is darker.

We, therefore, first classify the peaks into an OD category and a macula category, respectively. The classification is based on the image difference between the retinal image pixels at the peak center and those surrounding the peak center. The image difference is evaluated by two concentric circles as follows:

$$\mathrm{Diff}(x,y) = \frac{1}{N_i}\sum_{d=0}^{R_1} I(d) - \frac{1}{N_o}\sum_{d=0}^{R_2} I(d) \tag{8}$$

where I refers to the retinal image under study and d denotes the distance between the peak and the surrounding retinal image pixels. R_1 and R_2 specify the radius of an inner concentric circle and an outer concentric circle where R_2 is set at 2 R_1. N_i and N_o denote the numbers of the retinal image pixels within the two concentric circles. In our system, we set R_2 at $(p - 1)/2$, where p is the length of the line operator. The peak can, therefore, be classified to the OD or macula category, if the image difference is positive or negative, respectively.

Finally, we detect the OD by a score that combines both the peak amplitude and the image intensity difference that by itself is also a strong indicator of the OD

$$S(x,y) = P(x,y)(\mathrm{Diff}(x,y) * (\mathrm{Diff}(x,y) > 0)) \tag{9}$$

Where P(x, y) denotes the normalized peak image. The symbol * denotes dot product and Diff(x, y) > 0 sets all retinal image pixels with a negative image difference to zero. The OD can, therefore, be detected by the peak in the OD category that has the

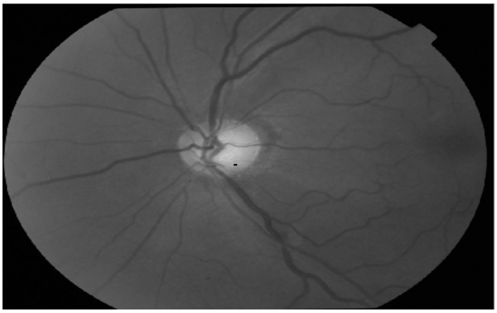
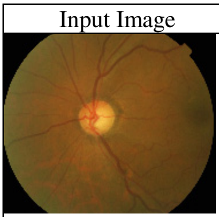
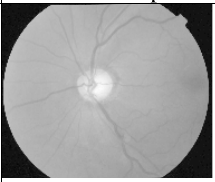
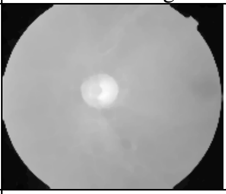
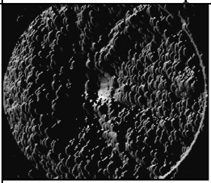
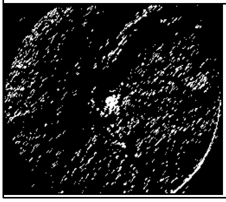
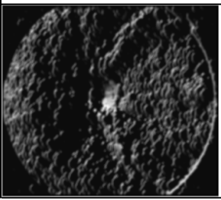
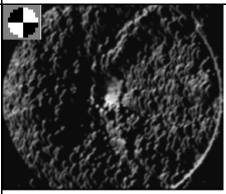
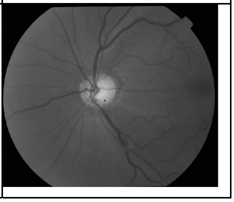


Fig. 5. OD detection. Score image by (9) for OD detection

maximum score. For the example retinal image in Fig. 1(a), Fig. 5(a) shows the score image determined by the peak image in Fig. 4(b). It should be noted that the image difference is evaluated only at the detected peaks in practical implementation. The score image in Fig. 5(a) (as well as in Fig. 5(b), 9, and 10) where the image difference is evaluated at every pixel is just for the illustration purpose.

4 Experimental Results

The proposed algorithm has been implemented in MATLAB. The corresponding results are follows.

Input Image	LAB - L Component	Smooth Image	Orientation map
			
Binary orientation map	Circular Convolution	Peak circular convolution mask	Detected Optic Disc
			

5 Conclusion and Future Work

The paper presented a simple method for OD detection using Line operator filter. The proposed approach achieved better results as reported. An extension for this study could be as follows.

The image segmentation is the foundation for the retinal fundus images. Efficient algorithms for segmenting the region of interest would be the main objective of future work. Optic disc can be detected easily after segmentation. A myriad of Optic Disc detection methods are available. The method which is compatible with the segmentation algorithm should be used. Diseases like Glaucoma and Diabetic Retinopathy are in rise and Optic Disc is an important indicator of these diseases. Hence the new methodologies to detect Optic Disc and the efficient use of already existing methods are the interest of future work.

Acknowledgments. The authors would like to acknowledge the faculty of Electronics and Communication Engineering department of Aalim Muhammed Salegh College of Engineering for their continuous support and encouragement and also acknowledge centre for research St.Joseph's College of Engineering, Chennai, Tamilnadu, India.

References

Akita, K., Kuga, H.: A computer method of understanding ocular fundus images. Pattern Recognit. 15(6), 431–443 (1982)

Patton, N., Aslam, T.M., MacGillivary, T., Deary, I.J., Dhillon, B., Eikelboom, R.H., Yogesan, K., Constable, I.J.: Retinal image analysis:Concepts, applications and potential. Prog. Retin. Eye Res. 25(1), 99–127 (2006)

Walter, T., Klein, J.C., Massin, P., Erginay, A.: A Contribution of Image Processing to the Diagnosis of Diabetic Retinopathy-Detection of Exudates in Color Fundus Images of the Human Retina. IEEE Trans. Med. Imag. 21(10), 1236–1243 (2002)

Chrastek, Wolf, M., Donath, K., Niemann, H., Paulus, D., Hothorn, T., Lausen, B., Lammer, R., Mardin, C.Y., Michelson, G.: Automated segmentation of the optic nerve head for diagnosis of glaucoma. Med. Image Anal. 9(4), 297–314 (2005)

Fleming, A.D., Goatman, K.A., Philip, S., Olson, J.A., Sharp, P.F.: Automatic detection of retinal anatomy to assist diabetic retinopathy screening. Phys. Med. Biol. 52(2), 331–345 (2007)

Pinz, A., Bernogger, S., Datlinger, P., Kruger, A.: Mapping the Human Retina. IEEE Trans. Med. Imag. 17(4), 606–619 (1998)

Tobin, K.W., Chaum, E., Govindasamy, V.P., Karnowski, T.P.: Detection of anatomic structures in human retinal imagery. IEEE Trans. Med. Imag. 26(12), 1729–1739 (2007)

Niemeijer, M., Abramoff, M.D., Ginneken, B.V.: Segmentation of the optic disc, macula and vascular arch in fundus photographs. IEEE Trans. Med. Imag. 26(1), 116–127 (2007)

Hsu, W., Pallawala, P.M.D.S., Lee, M.L., Eong, K.A.: The Role of Domain Knowledge in the Detection of Retinal Hard Exudates. Proc. Int. Conf. Comp. Vis. Pattern Recognit. 2, 246–251 (2001)

Sbeh, Z.B., Cohen, L.D., Mimoun, G., Coscas, G.: A new approach of geodesic reconstruction for drusen segmentation in eye fundus images. IEEE Trans. Med. Imag. 20(12), 1321–1333 (2002)

Walter, T., Klein, J.C.: Segmentation of color fundus images of the human retina: Detection of the optic disc and the vascular tree using morphological techniques. In: Proc. Int. Symp. Med. Data Anal., pp. 282–287 (2001)

Li, H., Chutatape, O.: Automatic location of optic disc in retinal images. In: Proc. Int. Conf. Image, vol. 2, pp. 837–840 (2001)

Sinthanayothina, C., Boycea, J.F., Cookb, H.L., Williamsonb, T.H.: Automated localisation of the optic disc, fovea, and retinal blood vessels from digital colour fundus images. Br. J. Ophthalmol. 83, 902–910 (1999)

Sekhar, S., Al-Nuaimy, W., Nandi, A.K.: Automated localisation of retinal optic disk using Hough transform. In: Proc. Int. Symp. Biomed. Imag.: Nano Macro, pp. 1577–1580 (2008)

Youssif, A., Ghalwash, A.Z., Ghoneim, A.: Optic disc detection from normalized digital fundus images by means of a vessels' directionmatched filter. IEEE Trans. Med. Imag. 27(1), 11–18 (2008)

Hoover, A., Goldbaum, M.: Locating the optic nerve in a netinal image using the fuzzy convergence of the blood vessels. IEEE Trans. Med. Imag. 22(8), 951–958 (2003)

Foracchia, M., Grisan, E., Ruggeri, A.: Detection of optic disc in retinal images by means of a geometrical model of vessel structure. IEEE Trans. Med. Imag. 23(10), 1189–1195 (2004)

Haar, F.: Automatic localization of the optic disc in digital colour images of the human retina. M.S. thesis, Utrecht University, Utrecht, The Netherlands (2005)

Li, H., Chutatape, O.: Automated feature extraction in color retinal images by a model based approach. IEEE Trans. Biomed. Eng. 51(2), 246–254 (2004)

Rovira, A.P., Trucco, E.: Robust optic disc location via combination of weak detectors. In: Proc. Int. Conf. IEEE Eng. Med. Bio. Soc, pp. 3542–3545 (2008)

Zwiggelaar, R., Astley, S.M., Boggis, C.R.M., Taylor, C.J.: Linear structures in mammographic images: Detection and classification. IEEE Trans. Med. Imag. 23(9), 1077–1086 (2004)

Ricci, E., Perfetti, R.: Retinal blood vessel segmentation using line operators and support vector classification. IEEE Trans. Med. Imag. 26(10), 1357–1365 (2007)

Osareh, A., Mirmehdi, M., Thomas, B., Markham, R.: Comparison of colour spaces for optic disc localisation in retinal images. In: Proc. Int. Conf. Pattern Recognit., vol. 1, pp. 743–746 (2002)

Tomasi, C., Manduchi, R.: Bilateral Filtering for Gray and Color Images. In: Proc. IEEE Int. Conf. Comp. Vis., pp. 839–846 (1998)

Tasman, W., Jaeger, E.A.: Duane's Ophthalmology, 15th edn. Lippincott Williams & Wilkins, Baltimore (2009)

Kauppi, T., Kalesnykiene, V., Kamarainen, J.K., Lensu, L., Sorri, I., Uusitalo, H., Kälviäinen, H., Pietilä, J.: DIARETDB0: Evaluation database and methodology for diabetic retinopathy algorithms. Tech. Rep., Lappeenranta Univ. Technol., Lappeenranta, Finland (2006)

Kauppi, T., Kalesnykiene, V., Kamarainen, J.K., Lensu, L., Sorri, I., Uusitalo, H., Klviinen, H., Pietil, J.: DIARETDB1 diabetic retinopathy database and evaluation protocol. Tech. Rep., Lappeenranta Univ. Technol., Lappeenranta, Finland (2007)

Staal, J.J., Abramoff, M.D., Niemeijer, M., Viergever, M.A., Ginneken, B.V.: Ridge based vessel segmentation in color images of the retina. IEEE Trans. Med. Imag. 23(4), 501–509 (2004)

An Extension of FFT Based Image Registration

P. Thangavel[1] and R. Kokila[2]

[1] Department of Computer Science, University of Madras, Chepauk, Chennai-600005, India
`thangavelp@yahoo.com`
[2] Department of Computer Science, University of Madras, Chepauk, Chennai-600005, India
`kok_oc25@yahoo.co.in`

Abstract. Image registration is considered as one of the most fundamental and crucial pre-processing task in image processing applications. It needs visual information from multiple images for comparison, integration or analysis. In this paper we present an extension of fast Fourier transform based image registration scheme. We have tested the proposed scheme with a number of selected images and found that the results are much better when compared to normal FFT method. The time complexity of our proposed method is of same order of the FFT based method [2].

1 Introduction

Image registration is an important tool in image processing. It is used to match two or more images of the same scene taken at different geometric viewpoint, different time or by a different image sensor. Image registration is widely used in different fields such as remote sensing for multispectral classification, environmental monitoring, change detection, image mosaicing, weather forecasting, creating super-resolution images and integrating information into geographic information systems(GIS). In medical applications it is used for combining data from different modalities such as computer tomography and magnetic resonance imaging(MRI) to obtain more complete information about a patient, monitoring tumor growth, treatment verification and comparison of the patient's data with anatomical atlases. Image registration is also used in cartography for map updating and in computer vision for target localization and automatic quality control.

Two images may differ from each other by its scale, rotation and translation that can be determined by using fast Fourier transform(FFT) method [2]. The Fourier method searches for the optimal match based on the information in the frequency domain. Because of this distinct features it differs from other registration strategies. Recently image registration using log polar mapping [3] and adaptive polar transform [8] were reported. Tzimiropoluos et al.,[9] have reported FFT based registration scheme with image gradients.

We present an extension of the phase correlation technique for automatic image registration, which is characterized by its insensitivity to translation, rotation, scaling and noise.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 729–737.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

2 Fourier Transform Based Image Registration

In this section we review the Fourier transform theory for image registration.

2.1 Translation

Let f_1 and f_2 be the two images with displacement (x_0, y_0):

$$f_2(x,y) = f_1(x - x_0, y - y_0) \tag{1}$$

The fourier transforms of f_1 and f_2 denoted by F_1 and F_2 are related as follows:

$$F_2(\omega_1, \omega_2) = F_1(\omega_1, \omega_2) * e^{-2\pi j(\omega_1 x_0 + \omega_2 y_0)} \tag{2}$$

The corresponding crosspower spectrum is:

$$\frac{F_2(\omega_1, \omega_2).F_1^*(\omega_1, \omega_2)}{\mid F_2(\omega_1, \omega_2).F_1^*(\omega_1, \omega_2) \mid} = e^{2\pi j(\omega_1 x_0 + \omega_2 y_0)} \tag{3}$$

Here F_1^* is the complex conjugate of F_1. The shift theorem guarantees that the phase of crosspower spectrum is equal to the phase difference between the images. The inverse Fourier transform of the crosspower spectrum produces impulse function that is approximately zero everywhere except at the point of displacement. The location of impulse is used to register the two images.

2.2 Rotation

Let f_2 be a translated and rotated image of f_1 with the translation (x_0, y_0) and rotation θ_0. Now f_2 can be written as follow as:

$$\begin{aligned} f_2(x,y) &= f_1(x \cos\theta_0 + y \sin\theta_0 - x_0, \\ &\qquad -x \sin\theta_0 + y \cos\theta_0 - y_0) \end{aligned} \tag{4}$$

Based on translation and rotation property of Fourier transform, the transforms of f_1 and f_2 are related by

$$\begin{aligned} F_2(\omega_1, \omega_2) &= F_1(\omega_1 \cos\theta_0 + \omega_2 \sin\theta_0, -\omega_1 \cos\theta_0 + \omega_2 \sin\theta_0) \\ &\quad * e^{-2\pi j(\omega_1 x_0 + \omega_2 y_0)} \end{aligned} \tag{5}$$

If M_1 and M_2 are the magnitude of F_1 and F_2 then the above equation becomes

$$\begin{aligned} M_2(\omega_1, \omega_2) &= M_1(\omega_1 \cos\theta_0 + \omega_2 \sin\theta_0, \\ &\qquad -\omega_1 \sin\theta_0 + \omega_2 \cos\theta_0) \end{aligned} \tag{6}$$

Rotation (θ_0) can be found using phase correlation by representing M_1 and M_2 in polar co-ordinates

$$M_1(\rho, \theta) = M_2(\rho, \theta - \theta_0) \tag{7}$$

2.3 Scale

When an image is scaled by scale factor a to an another image, the relation between the Fourier transform of these two images is expressed as follows:

$$F_2(\omega_1,\omega_2) = \frac{1}{a^2}F_1(\frac{\omega_1}{a},\frac{\omega_2}{a}) \tag{8}$$

The above equation can be rewritten in the logarithmic domain as follows :

$$F_2(\log\omega_1,\log\omega_2) = F_1(\log\omega_1 - \log a,\log\omega_2 - \log a) \tag{9}$$

From this we can obtain the scale factor a. In the above equation constant $\frac{1}{a^2}$ is ignored for simplicity.

2.4 Translation, Rotation and Scale

When translation, rotation, scaling are all present between the two given images, the corresponding magnitude spectrum can be represented as follows:

$$M_2(\rho,\theta) = M_1(\frac{\rho}{a},\theta - \theta_0) \tag{10}$$

The respective magnitudes in the log polar co-ordinate system are related as follow as:

$$M_2(\log\rho,\theta) = M_1(\log\rho - \log a,\theta - \theta_0) \tag{11}$$

Rotation and scaling can now be recovered using the Fourier phase shifting property. After recovery of these parameters, image f_2 can be wrapped to compensate for the rotation and scaling. Finally the standard phase correlation technique can be applied to cover the remaining translational offset between f_1 and f_2.

2.5 HighPass Filtering

The following highpass emphasis filter is multiplied with Fourier log magnitude spectra

$$H(\xi,\eta) = (1.0 - X(\xi,\eta)) * (2.0 - X(\xi,\eta)) \tag{12}$$

where

$$X(\xi,\eta) = [\cos(\pi\xi)\cos(\pi\eta)]\ and - 0.5 \leq \xi,\eta \leq 0.5 \tag{13}$$

2.6 LogPolar Transform

Logpolar Transform is used to convert an image from the cartesian co-ordinate to the polar co-ordinate. The mathematical expression of the mapping procedure is shown as

$$\rho = \log_{base}\sqrt{(x-x_c)^2 + (y-y_c)^2} \tag{14}$$

$$\theta = \tan^{-1}\frac{(y-y_c)}{(x-x_c)} \tag{15}$$

Here, (x_c,y_c) is the center pixel of the transformation in the cartesian co-ordinate, (x,y) denotes the sampling pixel in the cartesian co-ordinate and (ρ,θ) denotes the log radius and angular position in the log polar co-ordinate.

3 Proposed Method

The Fourier spectra of both images can computed by using two dimensional Fast Fourier Transform(FFT). The highpass filter used with the transfer function is multiplied with Fourier log magnitude spectrum . After this multiplication it is converted from the cartesian co-ordinate to the polar co-ordinate. The transformed image is phase correlated to find the translation. The translation will be either x or size of image - x. In determining the angle there is 180° ambiguity. This can be solved as follows: Let θ_0 be computed angle. First the translation is obtained by rotating the spectrum of source image by (θ_0), then rotate the spectrum of source image by $(180+\theta_0)$ and again compute the translation. If the peak value of the inverse Fourier transform of the crosspower spectrum is greater when the angle is θ_0 then the true angle of the rotation is θ_0. Otherwise $(180+\theta_0)$ is the angle of rotation.

If the obtained angle of rotation and scale is not able to register the source image with the target image, we then rotate the spectrum of target image in clockwise direction with 50°. Then perform the above process once again.

Now a new scale and rotation angle will be computed. Subtract the rotational angle from 50°. If the computed rotational angle is greater than 360° then subtract 360° from rotational angle ie (θ_0-360°). The computed rotation angle and scale factor is applied to the target image. Phase correlation technique is applied to the target image and source image to detect the translation.

3.1 Algorithm

1. Compute the forward Fast Fourier Transform on the source and target image and take absolute value of fast Fourier transform.
2. Multiply the absolute value of fast Fourier transform with the highpass filter.
3. Map the Fourier magnitude spectrum from the cartesian co-ordinates to the polar co-ordinates.
4. Calculate the scale factor and rotation angle by using the phase correlation technique on the log polar spectra of both the images.
5. Apply the scale factor and rotation angle to the target image and once again apply the phase correlation to detect the translation.
6. If the computed scale and rotation is not able to register the source image with the target image, then rotate the target image in clockwise direction with 50°. Then perform the above process.
7. rotation angle= θ_0-50°
   if $\theta_0 > 360°$
   rotation angle= θ_0-360°
8. Apply the scale factor and rotation angle to the target image as in step 5.

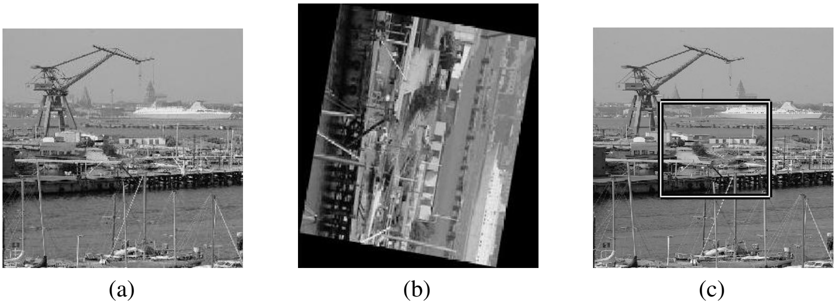
Matlab R2008a running on 3.4GHz Intel Pentium D machine is used to perform all the experiments. Depending upon the image size and the richness of texture content in the image its computation time for registering an image pair varies. In our experiment, the target image with size of 1035 * 1035 pixels are registered in approximately 236 seconds.

4 Experimental Results

To compute the effectiveness of the proposed image registration scheme, we have used seven different sets of images to test. Different rotation and scaling are applied to all images for the registration.

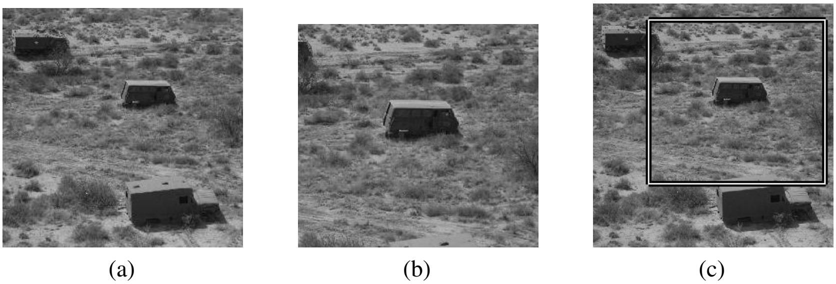
The kiel image (Fig.1(a)) is used as the source image. It is scaled and rotated with 1.6 and 260° respectively, and then used as target image (Fig.1(a)). Fig.1(c) shows the registration result using the proposed approach. The registration result is represented in the black and white rectangular area.

Figs.1-8 in which each figure consists of three subfigures. The subfigure on the left is the source image and the subfigure in the middle is the target image which is to be registered with the source image. The subfigure in the right is the registered image using proposed image registration method, in which the black and white rectangular area indicates the registration result.



(a) (b) (c)

Fig. 1. Image Registration result using the proposed approach on the kiel Image (a) Source Image (b) Target Image (scaled with 1.6 and rotated with 260°) (c) Registration result using the proposed approach (Target image is registered in the source image indicated in the black and white rectangular area)



(a) (b) (c)

Fig. 2. Image Registration result using the proposed approach on the Van Image (a) Source Image (b) Target Image (scaled with 3.4 and rotated with 360°) (c) Registration result using the proposed approach

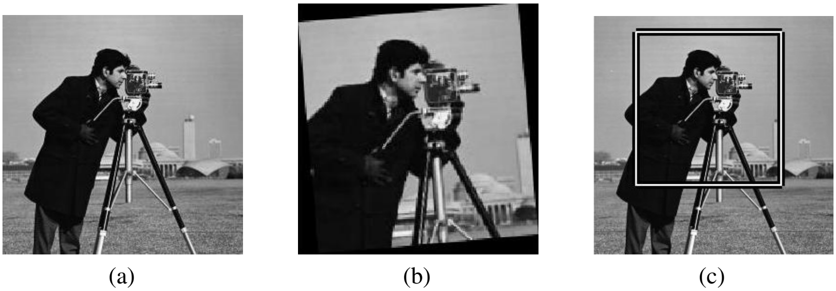

(a) (b) (c)

Fig. 3. Image Registration result using the proposed approach on the CameraMan Image (a) Source Image (b) Target Image (scaled with 2.23 and rotated with 5°) (c) Registration result using the proposed approach

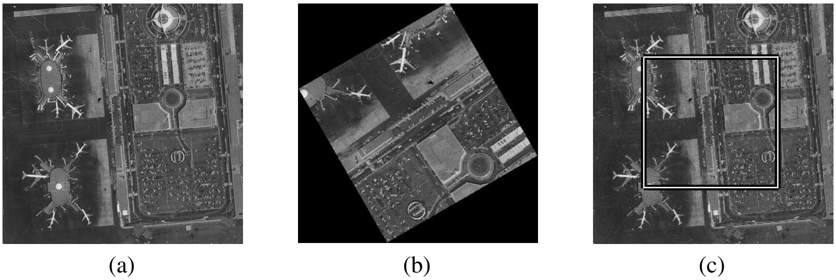

(a) (b) (c)

Fig. 4. Image Registration result using the proposed approach on the Airfield Image (a) Source Image (b) Target Image (scaled with 7.6 and rotated with 300°) (c) Registration result using the proposed approach

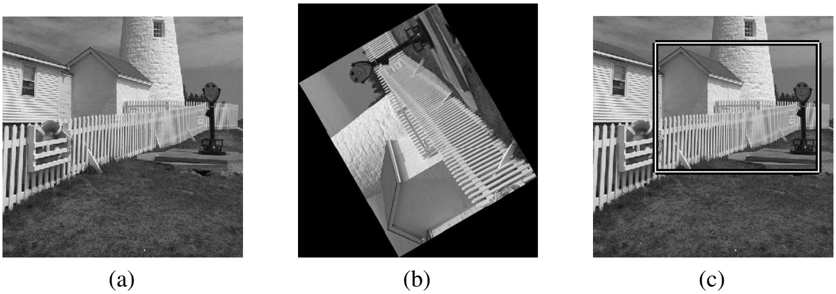
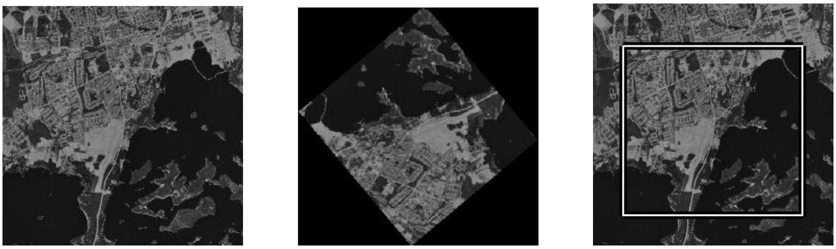

(a) (b) (c)

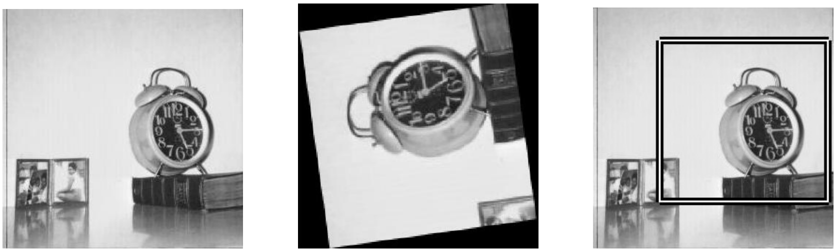
Fig. 5. Image Registration result using the proposed approach on the LightHouse Image (a) Source Image (b) Target Image (scaled with 0.99 and rotated with 120°) (c) Registration result using the proposed approach

As shown in the Table 1 we can see that the proposed method yields the robustness in scale, rotation and translation changes. The accuracy of the registration using proposed method is comparable with the existing methods.



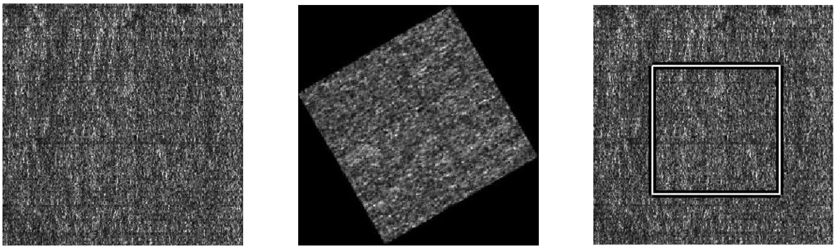
(a) (b) (c)

Fig. 6. Image Registration result using the proposed approach on the Land Satellite Image (a) Source Image (b) Target Image (scaled with 5 and rotated with 130°) (c) Registration result using the proposed approach



(a) (b) (c)

Fig. 7. Image Registration result using the proposed approach on the Clock Image (a) Source Image (b) Target Image (scaled with 2.5 and rotated with 98°) (c) Registration result using the proposed approach



(a) (b) (c)

Fig. 8. Image Registration result using the proposed approach on the Sonar Image (a) Source Image (b) Target Image (scaled with 5.9 and rotated with 120°) (c) Registration result using the proposed approach

Table 1. Scale factors and their corresponding Rotation recovered by the proposed method

Image	**Input values**		**FFT Method [2]**		**Proposed Method**	
	Scale	Rotation	Scale	Rotation	Scale	Rotation
Kiel	1.51	45.4°	1.5365	45.6046°	1.5086	45.5252°
	1.6	260°	1.5371	260.2174°	1.5984	259.8969°
	0.68	68°	1	−50°	0.6842	68.125°
	4.3	109.9°	1.9657	−48.7871°	4.305	109.4631°
	7.7	109.9°	1.1836	35.3594°	7.7006	109.9518°
	7.8	62°	1.9959	218.8963°	7.7891	62.0539°
	8	100°	1.1466	220°	7.9914	100.4948°
Van	2.89	9.6°	1	0°	2.8883	9.6097°
	3.4	360°	1.3372	129.7147°	3.3876	0.1107°
	5.3	75°	1.053	−49.8399°	5.3025	74.9726°
	7	180°	1.8227	39.6361°	7.0007	179.9341°
	1.5	260°	0.9894	308.7074°	1.4916	260.1319°
CameraMan	1	100°	0.9627	97.0313°	1	100.4688°
	1.2	125°	1.2092	125.1563°	1.2092	125.1563°
	1.2	225°	1	180°	1.1997	224.6667°
	2.23	5°	0.9587	183.1214°	2.2111	5.5378°
	2.9	135°	0.9665	92°	2.9936	134.9005°
	2.5	15°	0.7641	169.7938°	2.491	14.8313°
AirPort	5.8	154°	3.4171	−49.9454°	5.815	154.4957°
	7.6	300°	2.0379	41.7304°	7.6031	300.0249°
	0.7	63°	0.7022	62.9297°	0.7022	62.9297°
	3.9	10°	2.0874	130.6498°	3.8993	9.9418°
LightHouse	4.92	40.7°	1	130°	4.9447	39.9383°
	6.1	90°	1.9756	−49.7236°	6.0737	90.199°
	7.6	135°	1	130°	7.5894	135.0322°
	3.8	1.68°	0.9916	0.4436°	3.7874	1.9077°
	0.99	120°	0.8227	257.9688°	0.9892	120.1563°
Landsat	5	130°	0.8389	268.6737°	5.0022	130°
	4.9	333°	1	−50°	4.9137	333.0805°
	0.65	55°	0.64602	54.8438°	0.64602	54.8438°
	2.69	0°	0.8834	1.4063°	2.6776	0.1849°
Clock	3	290°	1	92.514°	2.9897	290.241°
	0.73	0°	0.7379	0°	0.7379	0°
	2.5	325°	1.0125	221.6071°	2.4926	325.0241°
	2.5	98°	1	98°	2.5157	97.6404°
	4.9	175°	3.0013	−50°	4.905	175.0858°
	5	140.69°	2.9976	130°	4.9983	140.5421°
Sonar	5.9	120°	1.9483	270.3362°	5.8854	120.1668°
	0.89	152°	0.89228	151.875°	0.89228	151.875°
	2.9	152°	1	180°	2.9385	150.9286°
	2.76	22°	1	180°	2.7668	22.3077°
	7.7	60°	1.9554	−48.626°	7.6356	60.2757°
	6.45	75°	0.8134	136.1503°	6.4327	74.7711°
	5.3	356°	1.2824	221.8519°	5.331	356.2609°

5 Conclusion

In this paper we have presented a fast and accurate image registration scheme based on FFT. The method is applied for a selected set of images to perform the image registration. The method succeeds in recovering scale factor upto 7.8 and rotation upto 360° with high accuracy than the FFT method [2].

Acknowledgements. This work is partially supported by University Grant Commission Major Research Project, Government of India.

References

1. Zitova, B., Flusser, J.: Image Registration Methods: A Survery. Image Vision Computing 21(11), 977–1000 (2003)
2. Reddy, B.S., Chatterji, B.N.: An FFT-Based Technique for Translation, Rotation and Scale-Invariant Image Registration. IEEE Trans. Image Processing 5(8), 1266–1271 (1996)
3. Zokai, S., Bean, C.P.: Image Registration using Log-Polar Mappings for Recovery of Large-Scale Similarity and Projection Transform. IEEE Trans. Image Processing 14(10), 1422–1434 (2005)
4. Lewis, J.P.: Fast normalised cross-correlation. In: Proceedings of Vision Interface, pp. 120–123 (1995)
5. Liu, H., Guo, B., Feng, Z.: Pseudopolar-Log-Polar Fourier Transform for Image Registration. IEEE Signal Processing Letters 13(1), 17–21 (2006)
6. Kuglin, C.D., Hines, D.C.: The Phase Correlation Image Alignment Method. In: Proceedings IEEE Conf. Cybernetics and Soc., pp. 163–165 (1975)
7. Casasent, D., Psaltis, D.: Position oriented and scale invariant optical correlation. Appln. Opt. 15, 1793–1799 (1976)
8. Matungaka, R., Zheng, Y.F., Ewing, R.L.: Image Registration Using Adaptive Polar Transform. IEEE Trans. On Image Processing 18(10), 2346–2354 (2009)
9. Tzimiropoluos, G., Argyriou, V., Zafeiriou, S., Stathaki, T.: Robust FFT-Based Scale-Invariant Image Registration with Image Gradients. IEEE Trans. on Pattern Analysis and Machine Intelligence 32(10), 1899–1906 (2010)

Scene Text Extraction from Videos Using Hybrid Approach

A. Thilagavathy, K. Aarthi, and A. Chilambuchelvan

Department of Computer Science and Engineering,
R.M.K Engineering College,
Kavaraipettai, Tamil Nadu, India
`atv.cse@rmkec.ac.in, aarthi.kme@gmail.com`

Abstract. With fast intensification of existing multimedia documents and mounting demand for information indexing and retrieval, much endeavor has been done on extracting the text from images and videos. The prime intention of the projected system is to spot and haul out the scene text from video. Extracting the scene text from video is demanding due to complex background, varying font size, different style, lower resolution and blurring, position, viewing angle and so on. In this paper we put forward a hybrid method where the two most well-liked text extraction techniques i.e. region based method and connected component (CC) based method comes together. Initially the video is split into frames and key frames obtained. Text region indicator (TRI) is being developed to compute the text prevailing confidence and candidate region by performing binarization. Artificial Neural network (ANN) is used as the classifier and Optical Character Recognition (OCR) is used for character verification. Text is grouped by constructing the minimum spanning tree with the use of bounding box distance.

Keywords: Caption text, Preprocessing, Scene text, Text extraction, Text grouping, and Video frame extraction.

1 Introduction

Text data present in images and video contain useful information for automatic annotation, indexing and structuring of images. According to Jung *et al[13]* , the text information extraction system (TIE) consists of four stages: text detection (finds the text region in the frame), text localization (groups the text region and generate bounding boxes), text extraction and enhancement (extract the text using some classifier and enhance it) and recognition (verify the extracted text with OCR).

Video consists of two types of text: Scene text and Caption text. *Scene text* is the text that in nature occurs in the area of capturing of video like text on banners, signs, container, CD cover, sign board, text on vehicle. It is also called graphics text. *Caption text or artificial text* on the other hand is the text that is artificially overlaid on the video/image such as the scores in sports videos, subtitles in news video, date and time in the video. It is also called superimposed text. However, the variations of text due to

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 739–748.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

differences in font, size, orientation, style, alignment, complex background, unknown layout makes the text extraction from video a challenging task.

The existing methods to detect and extract the text from video are classified into thresholding based and grouping based methods.

A. *Thresholding based method*

In this method, a global threshold is defined for the whole image and a local threshold is defined for a selected portion of the image. There are two types of thresholding based method:

1) *Histogram based thresholding:*

It is usually used for monochrome image. It counts the number of each pixel value in the histogram. Threshold is the value between the two peaks. The main disadvantage of this method is it doesn't work for images with complex background.

2) *Adaptive binarization techniques:*

It is used for grayscale image. Threshold is defined for the parts of the video. The most common method used is the Niblack's method which considers the mean and standard deviations over the radius of r.

3) *Entropy based method:*

This method is used only for the gray scaled image. It makes use of the entropy values of the different gray level distribution.

B. *Grouping based method*

This method groups the text pixel based on some criteria to extract the text. Grouping based method consists of the following types:

1) *Region based method:*

This method is based on texture analysis. Region based method can be either top-down and bottom-up. *Top-down* considers the whole image and moves to smaller parts of the image. Top-down approach considers the grayscale value. *Bottom-up* approach starts with the smaller parts of the image and then merges into a single image. The widely used bottom-up approaches are *connected component (CC)* based method and edge based methods.

2) *Learning based approach:*

It mainly makes use of the neural networks. Some of classifiers used are Multilayer perceptron (MLP), single layer perceptrons (SLP) etc.

3) *Clustering based approach:*

This method groups the text into clusters based on the color similarity. Some of the commonly used methods are K-means clustering, Gaussian mixture model (GMM), density based etc.

Extracting the scene text from video is demanding due to complex background, varying font size, different style, lower resolution and blurring, position, viewing angle and so on. In this paper we put forward a hybrid method where the two most well-liked text extraction techniques i.e. region based method and connected component (CC) based method comes together. Initially the video is split into frames and obtain the key frames. Text region indicator (TRI) is being developed to compute the text prevailing confidence and candidate region by performing binarization. Artificial Neural network (ANN) is used as the classifier and Optical Character Recognition (OCR) is used for character verification. Text is grouped by constructing the minimum spanning tree with the use of bounding box distance.

The paper is organized into the following sections. We discuss the related works in section 2, System overview in section 3 and Preprocessing in section 4, Connected component Analysis in section 5, Text grouping in section 6, Conclusion in section 7 and future contribution in section 8.

2 Related Work

Liu *et al* [1] proposed two text extracting methods: region based and connected component (CC) based method. The region based method is used for segmentation and CC for filtering the text and non-text components. In [2] Hu *et al* used a corner based approach to detect and extract the caption text from videos. It is based on the assumption that the text has a dense corner points. In [3] Weinman *et al* proposed the unified processing. The method involves the following information: appearance, language, similarity to other characters, and a lexicon.

Tsai *et al [4]* proposed the method for detection of signs from videos. It uses connected component analysis to detect the candidate region. A single detector is used for all the color instead of a separate detector for each color. The radial basis function network is used as the classifier. A rectification method was proposed to rectify a skewed road sign to its correct shape.

In [5] Nicolas *et al* implemented the Conditional Random Field (CRF) in the document analysis. It takes into account both the local and contextual feature. These features are extracted and feed as input to the Multilayer Perceptron (MLP). In [6] Chen and Yuille used Adaboost classifier where the weak classifiers are applied to train for a strong classifier in order to construct the fast text detector. This region identified by the classifier is given as input to the binarization and followed by CC analysis.

In [7] Kim *et al* presented an approach where they used SVM as a classifier and then perform the CAMSHIFT to identify the text regions. In [8] Lienhart *et al,* Complex-valued multilayer feed forward network is trained to detect text at an unchanging scale and position. In [9] Li et al makes use of the neural network as classifier and the extracted text is compared with the successive frames. It will identify the presence of the text in 16 X 16 windows only and SSD (Sum of Squared Difference) for frame similarity.

In [10] Zhong *et al* extract text from compressed video using the DCT. It applies horizontal thresholding to obtain the noise. In [11] Zhu *et al* employs Non-linear Niblacks method to perform the gray scale conversion and then fed into the classifier which is trained by Adaboost algorithm for filtering the text and non-text regions. In [12] Liu *et al* used edge detection algorithm to obtain the text color pixels. Connected component analysis is done to obtain the text confidence.

3 System Overview

The main objective of the proposed system is to extract the text from videos. The proposed system consists of the following stages: preprocessing, text extraction and text grouping. Video is split into frames based on the shots. Redundant frames are discarded by performing frame similarity which results in selection of key frames. The key frames containing the scene text.

In pre-processing stage, the text prevailing confidence is identified and its scale in the key frames. This stage identifies the region where the text is present i.e. candidate region. Then adaptive thresholding (binarization) is applied to identify the presence of text in the key frame. After the detection of text region, the connected component analysis is performed where both horizontal and vertical projection in the key frame is used to detect the text.

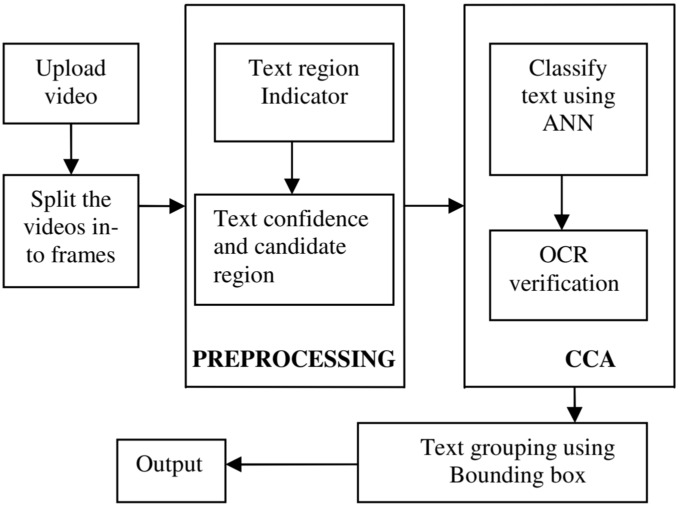

Upload video
Split the videos into frames
Text region Indicator
Text confidence and candidate region
PREPROCESSING
Classify text using ANN
OCR verification
CCA
Text grouping using Bounding box
Output

Fig. 1. Architecture diagram for text extraction process

In Connected Component Analysis (CCA) the CRF model is used to classify the candidate region into two classes: {text, non-text}. The Artificial neural network is trained to be a classifier to filter out the text and non-text components. The extracted text is passed to the OCR (Optical character recognition) for character confirmation. Then the texts are grouped into words and in turn into lines by using the horizontal and vertical bounding box distances by building minimum spanning tree.

4 Pre Processing

The first stage in Pre-processing is Video Frame Extraction. In this stage, the video containing the text is split into frames after reducing the rate of the video to 1 or 0.1 second. For the one frames per second, the size for the 320 x 240 frame is 75Kb and

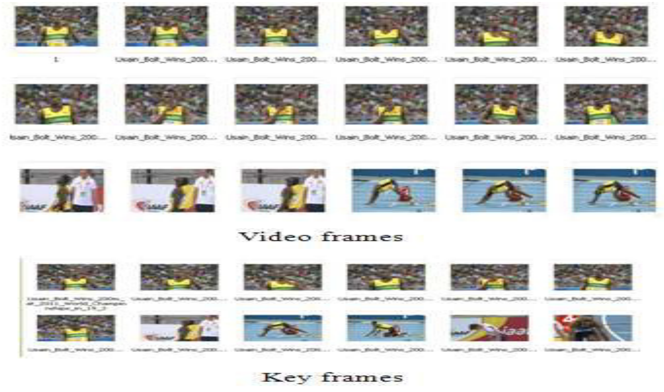


Fig. 2. Video splitting and key frame selection

the size per seconds' video: 4 MB size per minute video: 263 MB. Experimental results showed that when we consider this frame rate it will lead to the sizeable number of frames and so it will lead to redundant frames.

To overcome the temporal redundancy, edge comparisons between frames are done. Canny Edge detection is used for this purpose. The edges of a single frame are mapped with that of its neighbor ones to check for the frame similarity. When the inter frame space difference is high, it indicates that the frames are similar and we store only one frame and discard all the remaining frames. Thus by using this comparison we would be able to eliminate the redundant frames and will result in the distinct and unique frames (non- redundant video frames set). There is the need to choose the key frame from these non-redundant frames set. The key frames are those frames which contain the text in it. We make use of MODI (Microsoft Office Document Imaging) which identifies the frames containing text by discarding the frames containing special characters. The video frames that have only the non-special characters are

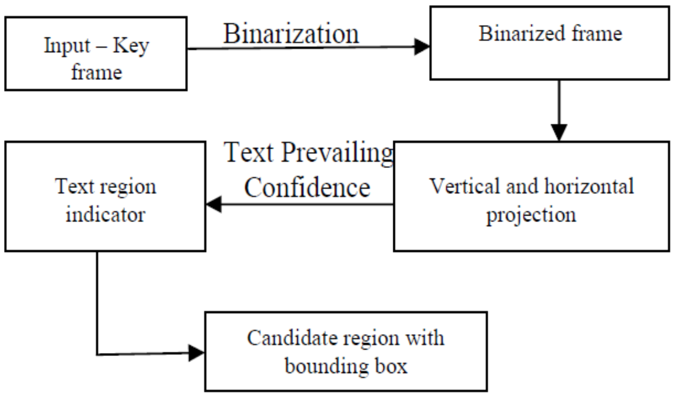

Input – Key frame
Binarization
Binarized frame
Vertical and horizontal projection
Text Prevailing Confidence
Text region indicator
Candidate region with bounding box

Fig. 3. Block diagram for text localization

stored. Thus we can filter out the non key frames easily and use only the key frames for further processing of the proposed method.

The next stage in preprocessing is to design the TRI. It is used to identify the candidate region. The transition map generation is used to differentiate the text from background. The text will have some kind of properties like vertical alignment, horizontal alignment, inter-character space, static motion, 2D motion, 3D motion to differentiate itself form the background.

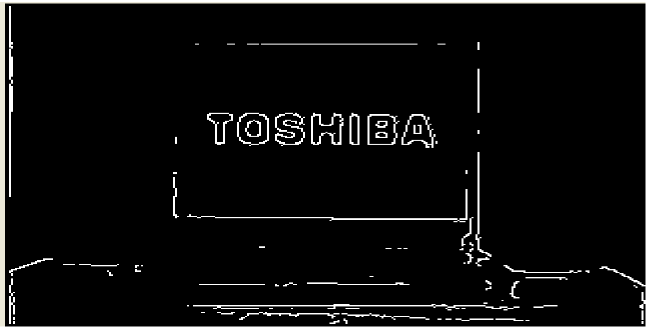


Fig. 4. Edge detection using Canny edge detector

The colored frame is converted into gray scale for binarization and then the adaptive thresholding is applied. The threshold value is based on the minimum size of text region. The Niblack binarization technique [14] is used which converts the gray scale image into binarized image. In Niblack's binarization algorithm the threshold value is selected according to the mean and standard deviation by sliding the small window over the key frame. Threshold value is calculated by using the maximum standard deviation of all the calculated windows.

To perform Connected Component, the successive pixels are examined by projecting the binarized image into two dimensions and a rectangular bounding box is generated

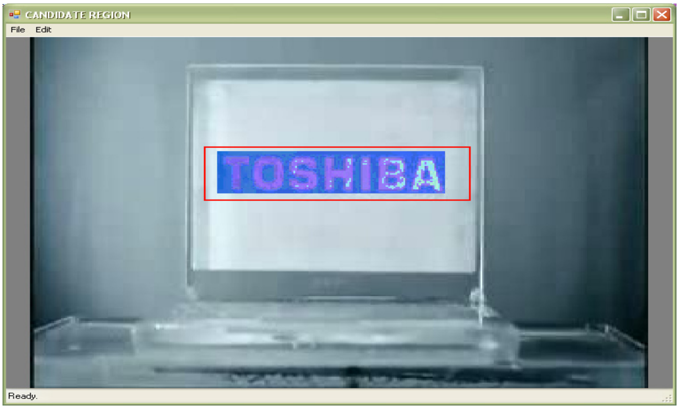


Fig. 5. A Key frames with the text highlighted (Candidate region)

by linking the four edges of the each character in the candidate region. In addition the Hidden Markov Model (HMM) is used to minimize the false detection.

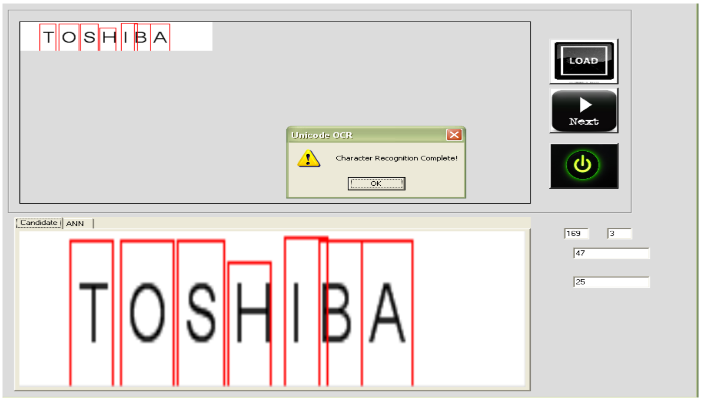
5 Connected Component Analysis

CRF model is used for classification of the candidate region into text and non-text. The CRF is a graphical model and depends on the Markovian property. The component locality graph is constructed by projecting the binarized image into two dimensions. The construction of graph assumes that the neighboring text has the same height and width. Euclidean distance is calculated between centroid of the two components. It also takes into consideration the height and width of the bounding box.

The Artificial Neural Network (ANN): Multi-layer Perceptron (MLP) is used as the classifier to categorize the text and non-text components. Training is done based on back propagation. The gradient descent technique is used to obtain the weights for the connection between the processing units. Training is through supervised learning where the input and the matching prototype are provided based on the learning rule. The hidden layer of the network contains the matrix value of the alphabets (both upper and lowercase) in the data array which is used for comparing the text from video frame. Thus it would classify the input and obtains the text. The extracted text is then passed on to the OCR (Optical character recognition) for the character confirmation.

6 Text Grouping

After text extraction, texts are grouped into words and then the words are grouped to lines by constructing the Minimum Spanning Tree (MST) using Kruskal algorithm. The tree is built on the basis of the bounding box distance between the texts: horizontally for the word and vertically for line partition. The spatial distance which is the distance between the bounding boxes is used for this purpose. The edge cuts are used as the partition in the tree construction which will lead to the text localization.

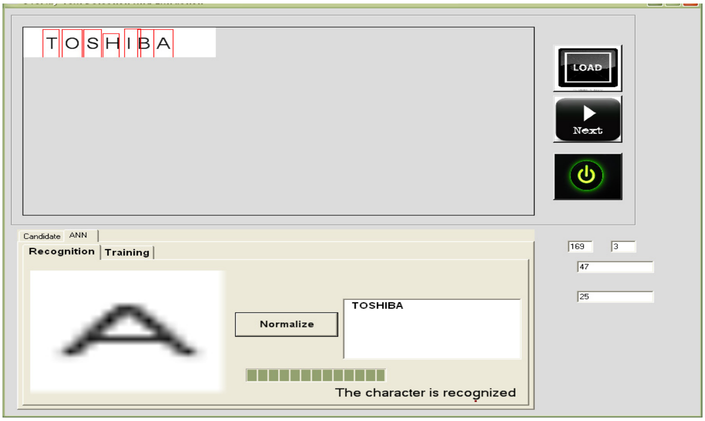


Character Recognition Complete!

169 3
47
25

(a)

Fig. 6. (a) Bounding box generation (b) Character recognition



TOSHIBA

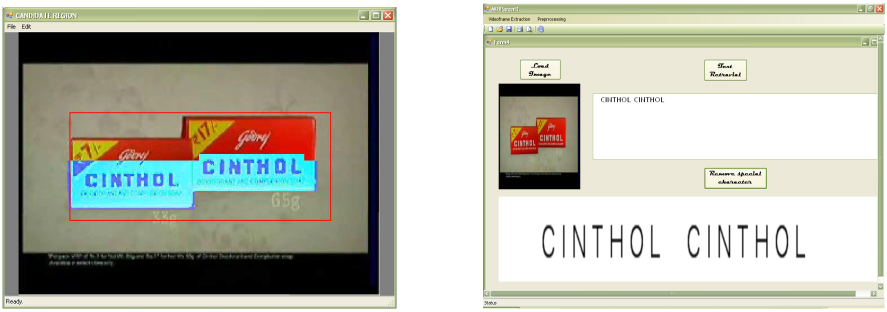
The character is recognized

(b)

Fig. 6. *(continued)*

7 Conclusion

In this paper we proposed a hybrid method where the two most well-liked text extraction techniques i.e. region based method and connected component (CC) based method comes together. The video is split into frames and key frames obtained. Text region indicator (TRI) developed to compute the text prevailing confidence and candidate region by performing binarization. Artificial Neural network (ANN) is used as the classifier and Optical Character Recognition (OCR) is used for character verification. Text is grouped by constructing the minimum spanning tree with the use of bounding box distance.

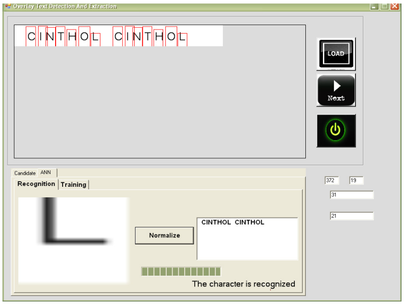


CINTHOL CINTHOL

CINTHOL CINTHOL

(a) (b)

Fig. 7. A example of text extraction process



CINTHOL CINTHOL

The character is recognized

372

19

31

21

(c)

Fig. 7. *(continued)*

8 Future Contribution

In this paper the main contribution is only on the English text extraction from the videos. In future, the work can be explored for multilingual languages.

References

[1] Pan, Y.-F., Hou, X., Liu, C.-L., Senior Member, IEEE: A Hybrid Approach to Detect and Localize Texts in Natural Scene Images. IEEE Transactions on Image Processing 20(3) (March 2011)

[2] Zhao, X., Lin, K.-H., Fu, Y., Member, IEEE, Hu, Y., Member, IEEE, Liu, Y., Member, IEEE, Huang, T.S., Life Fellow, IEEE: Text from Corners: A Novel Approach to Detect Text and Caption in Videos. IEEE Transactions on Image Processing 20(3) (March 2011)

[3] Weinman, J., Learned-Miller, E., Hanson, A.: Scene text recognition using similarity and a lexicon with sparse belief propagation. IEEE Trans. Pattern Anal. Mach. Intell. 31(10), 1733–1746 (2009)

[4] Tsai, L.W., Hsieh, J.W., Chuang, C.H., Tseng, Y.J., Fan, K.-C., Lee, C.C.: Road Sign Detection Using Eigen Colour

[5] Nicolas, S., Dardenne, J., Paquet, T., Heutte, L.: Document image segmentation using a 2-D conditional random field model. In: Proc. 9th Int. Conf. Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, pp. 407–411 (2007)

[6] Chen, X.R., Yuille, A.L.: Detecting and Reading Text in Natural Scenes. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2004), Washington, DC, pp. 366–373 (2004)

[7] Kim, K.I., Jung, K., Kim, J.H.: Texture-based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm. IEEE Transaction on. Pattern Anal. Mach. Intell. 25(12), 1631–1639 (2003)

[8] Lienhart, R., Member, IEEE, Wernicke, A.: Localizing and Segmenting Text in Images and Videos. IEEE Transactions on Circuits and Systems for Video Technology 12(4) (April 2002)

[9] Li, H.P., Doermann, D., Kia, O.: Automatic Text Detection and Tracking in Digital Video. IEEE Transaction on Image Processing 9, 147–156 (2000)

[10] Zhong, Y., Zhang, H., Jain, A.K., Fellow: Automatic Caption Localization in Compressed Video. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(4) (April 2000)

[11] Zhu, K.H., Qi, F.H., Jiang, R.J., Xu, L., Kimachi, M., Wu, Y., Aizawa, T.: Using Adaboost to Detect and Segment Characters From Natural Scenes. In: Proc. 1st Conf. Caramera Based Document Analysis and Recognition (CBDAR 2005), Seoul, South Korea, pp. 52–59 (2005)

[12] Liu, Y.X., Goto, S., Ikenaga, T.: A Contour-based Robust Algorithm for Text Detection in Color Images. IEICE Transaction. Inf. Syst. E89-D(3), 1221–1230 (2006)

[13] Jung, K., Kim, K.I., Jain, A.K.: Text information extraction in images and video: A survey. Pattern Recogn. 37(5), 977–997 (2004)

[14] Niblack, W.: An Introduction to Digital Image Processing. Strandberg Publishing, Birkeroed (1985)

MSB Based New Hybrid Image Compression Technique for Wireless Transmission

S.N. Muralikrishna, Meghana Ajith, and K.B. Ajitha Shenoy

Department of Master of Computer Application, Manipal Institute of Technology,
Manipal University, Manipal, India
{murali.sn,meghana.ajith,ajith.shenoy}@manipal.edu

Abstract. It is observed that Digital images requires a large amount of memory to store and when retrieved from the internet, can take a considerable amount of time to download. Our method enables us to compress image in such a way that the utilization of memory will be less. The Proposed MSB based hybrid method has good compression rate (more than sixty percentage of compression) and has linear time complexity i.e. $O(MN)$ where M and N denote number of pixels in horizontal and vertical directions. We have proved that the compressed image obtained after applying our algorithm has PSNR value greater than or equal to $32dB$ which is suitable for wireless transmission.

1 Introduction

There has been an astronomical increase in the usage of computers for a variety of tasks in recent years. One of the most common uses has been the storage, manipulation and transfer of digital images. The compressing of digital image by using files can be large and can also take up precious memory space on the computer's hard drive. If one considers a gray scale image of *256x256* pixels, it can have an approximate of *65,536* elements to store. It is observed that the downloading of these files from the internet can consume more bandwidth. Our method enables us to transmit details faster as we can compress the image efficiently.

Hybrid Image compression technique make use of both lossy and lossless image compression techniques. Examples are JPEG [1,2], JPEG2000 [3] etc. In our approach we make use of most significant bit for compression. Hence the first step is lossy. Then we apply Predictive [2,4] and Huffman [2,5] coding which are lossless. We will now give a quick description of these lossless techniques.

1.1 Huffman Coding [2,5]

Huffman encoding technique is one of the primary lossless encoding techniques. In this technique, given the characters that must be encoded, together with the probability of their occurrences, the Huffman coding algorithm determines the optimal code using the minimum number of bits. Hence, the length of the coded characters will differ. In text, the shortest code is assigned to those characters that occur most frequently. To determine a Huffman code, it is useful to construct a binary tree. The leaves of the

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 749–756.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

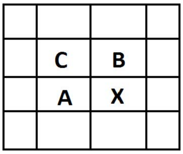


Fig. 1. Principle of the prediction in the lossless coding

tree represent characters that are to be encoded. Every node contains the occurrence probability of one of the characters belonging to this sub tree. 0 and 1 are assigned to the branches of the tree. The Huffman code for each character/symbol in the form of binary tree will be stored in a Huffman Table. The same Huffman table must be available for both encoding and decoding.

1.2 Predictive Coding [2,4]

As shown in the Figure 1, for each pixel X, one of eight possible predictors is selected. The selection criterion is a prediction that is as good as possible of the value of X from the already known adjacent samples A, B, C. The specified predictors are listed in Table 1.

Table 1. Predictors for lossless coding

Selection Value	Prediction
0	No Prediction
1	X=A
2	X=B
3	X=C
4	X=A+B-C
5	X=A+(B-C)/2
6	X=B+(A-C)/2
7	X=(A+B)/2

2 Our Approach

Consider an image with a pixel quantization of 8bits/pixel (i.e.8 bits for each Red, Green and Blue (RGB) components). Though we can use this method for any pixel quantization we are considering only 8 bits/pixel. We can represent each component of an image with a 2D array of M x N pixels where each pixel value ranges from 0-255. In our approach the pixels with 8 bit representation is converted to 5 bit representation by discarding the last three least significant bits. Now value of each pixel will be within the range of 0 to 31. This can be achieved with a simple bit shift operation. For example, if the pixel value of an image is 255, it can be represented as binary sequence 11111111, if we consider only 5 most significant bits then it is represented as 11111 i.e. 31 in decimal.

After this we apply lossless predictive coding for the sequence followed by Huffman encoding as an entropy encoding technique. Decompression for the lossless part is quite straight forward whereas for the lossy part we recover the pixel values by appending three zero bits as least significant bit. This also can be achieved by bit shift operation. For the same example, if the pixel value in binary is represented as 11111 after lossless decompression we append three zero bits at the least significant position, resulting in 11111000 i.e 248 in decimal. The maximum error in the pixel value can be at most 7 per pixel.

Algorithm 1. Compression Algorithm

```
Input : Uncompressed Image Components i.e. $M \times N$ integer matrix with 8 bit representation.
Output : Compressed Image Component.
1: For each pixel value consider most significant 5 bits and discard the last 3 bits. Now the new
   values are in the range from 0-31 (using 5 bits).
2: Apply Predictive coding (as described in Section 1.2). The number of chosen predictor,
   as well as the difference of the prediction to the actual value, is passed to the next step.
3: Apply Huffman Coding
```

Algorithm 2. Decompression Algorithm

```
Input : Compressed Image
Output : Uncompressed Image
1: Using Huffman table decode Huffman coded value.
2: Decode predictively coded data.
3: Add 3 zero bit at the end of each 5 bit value and make it 8 bit representation.
4:  Calculate the new pixel value in the range from 0 - 255
```

Theorem 1. *Time complexity of Algorithm-1: is $O(MN)$, where MxN denote number of pixels in the image.*

Proof. The first step in the algorithm is to discard last 3 least significant bits which takes O (MN) time as discarding 3 least significant bits from each 8 bit representation takes constant time. It is obvious that the next step (i.e. Predictive coding) in Algorithm-1 takes $O(MN)$ time. The last step is to find Huffman coding. The probability of occurrence of symbols can be found in O(MN) time. Since there are maximum of 32 distinct symbols, Huffman coding will take $O(32log32)$ time (since complexity of Huffman coding is $O(nlogn)$ where n denote number of distinct symbols) which is a constant. Hence the process of encoding entire data takes $O(MN)$ time .

In the following subsection we define two terms Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE) to prove that compressed image meets the requirements of wireless transmission quality.

2.1 PSNR Calculation [6,7]

Let us now prove that the compressed image obtained by applying our algorithm will meet the wireless transmission quality. To prove this we need to understand the concept of peak signal-to-noise ratio (PSNR). PSNR is an engineering term for the ratio between the maximum power of a signal and the power of the corrupting noise that affects the fidelity of its representation. This is because many signals have a very wide dynamic range. PSNR is expressed in terms of the logarithmic decibel scale. PSNR is most commonly used as a measure of quality of reconstruction of lossy compression codecs (e.g., for image compression). The signal in this case is the original data, and the noise is the error introduced by compression. When comparing compression codecs it is used as an approximation to human perception of reconstruction quality, therefore in some cases one reconstruction may appear to be closer to the original than another, even though it has a lower PSNR (a higher PSNR would normally indicate that the reconstruction is of higher quality). One has to be extremely careful with the range of validity of this metric; it is only conclusively valid when it is used to compare results from the same codec (or codec type) and same content. It is most easily defined via the mean square error (MSE) which for two *MxN* images I and K where I is uncompressed image and K is compressed image which is approximation of I is defined as:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 \tag{1}$$

PSNR is defined as follows $PSNR = 20.log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)$ where MAX_I is the maximum possible pixel value of the image. When the pixels are represented using 8 bits per sample, this is 255. Typical values for the PSNR in lossy image and video compression are between 30 and 50 dB. Acceptable values for wireless transmission quality loss are considered to be about 20 dB to 25 dB [6,7]

In our approach maximum possible difference between $|I(i,j) - K(i,j)|$ is 7. Therefore $[I(i,j) - K(i,j)]^2 \leq 49$. This implies that $MSE \leq 49$. Hence $PSNR \geq 20 * log10(255/7) = 31.23dB$ (since MAX_I is 255 in 8 bit representation). If we consider only 4 most significant bits $PSNR \geq 24.61dB$. So either we can consider 5 or 4 most significant bit based on the Image and quality that is needed. Hence our method can be used in wireless transmission.

In the next section we discuss the experimental results on benchmark images which show that our algorithm is maintaining a quality of the original image and has good compression rate.

3 Results

Our algorithm was implemented on Intel Core i3 with 2.2 GHz CPU. It was tested on benchmark test images [8] and we observed that our algorithm takes linear time and

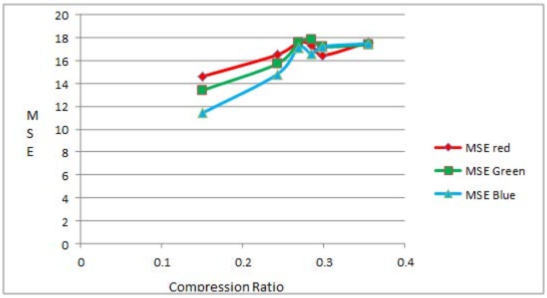


Fig. 2. Compression Ratio Vs MSE

the PSNR value was found to be greater than 35dB which is proved to be good for wireless transmission. This section follows with the experimental result. We calculated the Compression Ratio with the following formula [2,7]

$$\text{Compression Ratio} = \frac{\text{Compressed File Size}}{\text{Original File Size}} \tag{2}$$

A good compression makes use of less compression ratio and vive versa. We have considered six bench mark test images [8] for our experiment and the following tables and graphs summarizes the results: Table 3 summarizes the Mean Square Errors (MSE) for each component Red (R-MSE), Green (G-MSE) and Blue (B-MSE) of the different benchmark test images. The graph given in Figure 2 shows the relationship between Compression ratio and MSE. It shows that as Mean square error increases, compression ratio also increases. Table 4 summarizes the Peak Signal to Noise Ratio (PSNR) for each component Red (R_PSNR), Green (G_PSNR) and Blue (B_PSNR) of the different benchmark test images. Theoretically we have proved that by using our algorithm we can have PSNR values greater than or equal to 31 dB. Table 4 shows that by using our method we always get PSNR greater than 31 dB which is really good as per wireless transmission quality [6,7]. The graph given in Figure 3 shows the relationship between PSNR and compression ratio. This indicates that good compression gives high PSNR value and bad or less compression gives low PSNR values. Table 5 summarizes the Ex-

Table 2. Compressed File size, Space saving and Compression Ratio for different test images

File Name	M	N	Original File Size (in KB)	Compressed File Size (in KB)	Space Saving in percent	Compression Ratio
deer.ppm	2641	4043	31282	11118	64.46	0.355412
bridge.ppm	4049	2749	32610	9755	70.09	0.299141
big_tree.ppm	4550	6088	81154	23158	71.46	0.285359
big_building.ppm	5412	7216	114414	30799	73.08	0.269189
cathedral.ppm	3008	2000	17626	4283	75.70	0.242993
artificial.ppm	2048	3072	18433	2781	84.91	0.150871

Table 3. MSE values for different test images considered

File Name	R-MSE	G-MSE	B-MSE
deer.ppm	17.6207	17.3883	17.4932
bridge.ppm	16.4725	17.2295	17.237
big_tree.ppm	17.3572	17.8846	16.6048
big_building.ppm	17.5963	17.5689	17.1611
cathedral.ppm	16.5174	15.7315	14.8149
artificial.ppm	14.6178	13.4071	11.4269

Table 4. PSNR values for different test images considered

File Name	R_PSNR	G_PSNR	B_PSNR
deer.ppm	35.6706	35.7282	35.7021
bridge.ppm	35.9632	35.7681	35.7662
big_tree.ppm	35.736	35.606	35.9285
big_building.ppm	35.6766	35.6834	35.7854
cathedral.ppm	35.9514	36.1631	36.4238
artificial.ppm	36.482	36.8574	37.5515

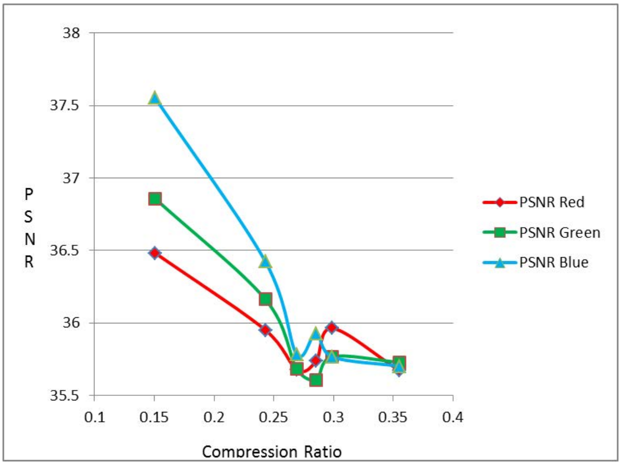


Fig. 3. Compression Ratio Vs PSNR

ecution time for compression of each test images. The graph given in Figure 4 is based on File Size and Execution time taken for compression. It shows that the Execution Time for our algorithm is Linear.

Table 5. Exectution Time noted for different test images

File Name	Original File size(KB)	Compressed File Size(KB)	Execution Time (in sec)
deer.ppm	31282	11118	0.46956
bridge.ppm	32610	9755	0.663543
big_tree.ppm	81154	23158	1.153356
big_building.ppm	114414	30799	1.508106
cathedral.ppm	17626	4283	0.277219
artificial.ppm	18433	2781	0.172618

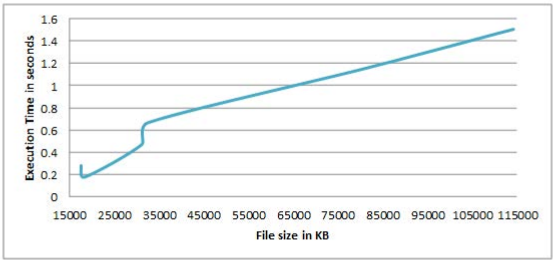


Fig. 4. File Size Vs. Execution Time

4 Conclusions

Even though there are many techniques for image compression we have introduced a new technique which makes use of the existing techniques. Our technique ensures a good quality image. We have also proved that the image that is obtained by using our technique meets the Wireless transmission quality. We have experimented our algorithm on several benchmark images and observed that our algorithm reduces the size of the image to $1/3^{rd}$ of the original size (More than 60% of reduction). As a future work we would like to apply some transformation techniques or any other compression techniques along with our algorithm in order to increase the compression rate. We can also think of considering luminance and chrominance component rather than RGB as luminance is more important than chrominance. We can think of assigning lesser bits to chrominance component to achieve better compression.

Acknowledgements. The authors would like to thank Manipal Institute of Technology, Manipal University, Manipal for encouraging us in pursuing this work.

References

1. Pennebaker, W.B., Mitchell, J.L.: JPEG Still Image Data Compression. Van Nostrand Reinhold, New York (1993)
2. Steinmetz, R., Nahrstedt, K.: Multimedia computing communication and applications. Printice hall PTR (1995)

3. Skodras, A., Christopoulos, C., Ebrahimi, T.: The JPEG 2000 still image compression standard. IEEE Signal Processing Magazine 18(5), 36–58 (2001)
4. Memon, Wu, X.: Recent developments in context based predictive techniques for lossless image compression. The Computer Journal 40, 127–136 (1997)
5. Bao, E., Li, W., Fan, D., Ma, X.: A study and implementation of the Huffman Algorithm Based on Condensed Huffman Table. Computer Science and Software Engineering 6, 42–45 (2008)
6. Saffor, A., Ramli, A.R., Ng, K.H.: A comparative study of image compression between JPEG and Wavelet. Malaysian Journal of Computer Science 14, 39–45 (2001)
7. PSNR, MSE, Compression Ratio, `http://www.wikipedia.org/wiki/`
8. Benchmark Test Images, `http://www.imagecompression.info`

Multi-temporal Satellite Image Analysis Using Unsupervised Techniques

C.S. Arvind[1], Ashoka Vanjare[2], S.N. Omkar[3,*] J. Senthilnath[2],
V. Mani[2], and P.G. Diwakar[3]

[1] Telibrahma Convergent Communication Pvt Limited, Bangalore
[2] Department of Aerospace Engineering, Indian Institute of Science, Bangalore
[3] Director, Earth Observation System, Indian Space Research Organisation, Bangalore
csarvind2000@gmail.com, omkar@aero.iisc.ernet.in

Abstract. This paper presents flood assessment using non-parametric techniques for multi-temporal time series MODIS (Moderate Resolution Imaging Spectro radiometer) satellite images. The unsupervised methods like mean shift algorithm and median cut are used for automatic extraction of water pixel from the image. The extracted results presents a comparative study of unsupervised image segmentation methods. The performance evaluation indices like root mean square error and receiver operating characteristics are used to study algorithm performance. The result reported in this paper provides useful information for multi-temporal time series image analysis which can be used for current and future research.

Keywords: MODIS satellite images, unsupervised image segmentation techniques, performance evaluation indices.

1 Introduction

Multi-temporal time series analysis of satellite images plays an important role to determine the land surface change detection [1]. The change detection study helps in surface analysis [2]. The NASA's MODIS satellite sensor has been considered as potential for multi-temporal image analysis because due to regular availability and open source. MODIS data provides excellent land and water discrimination along with wide area coverage [3]. The features like river, road network and vegetation are to be extracted and analysed, so this helps researchers to develop tools for analysing the surface changes occurred between different dates of imaging and it is useful in hydrological application such as flood assessment [4,5].

The researchers are continuously developing both supervised and unsupervised classification techniques for flood assessment. Rajiv Kumar Nath et al. [6] has worked on different remote sense data for flood assessment. The researchers have used supervised methods for flood application but performance limitation exists due to the extent and accuracy of the available and collected ground truth data. So in this context, several researchers are working towards unsupervised techniques.

---

[*] Corresponding author.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 757–765.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

R. Brakenridge et al. [3] have used unsupervised ISODATA method for flood risk analysis. An unsupervised or clustering technique automatically assigns each pixel to respective spectral clusters without manual intervention. It has a property of grouping individuals in the population and grouping of individuals is an outcome of partitioning of the data sets (i.e. mutually non-overlapping groups of the input datasets). But only a few unsupervised image segmentation methods like self organizing and K-means [] have been explored in flood assessment applications.

In this paper, unsupervised image segmentation methods like mean shift and median cut are used to extract non-linear features (river networks) from MODIS band-2 image [3] and the obtained results are verified with the ground truth data. Our investigation is in using unsupervised methods for low resolution MODIS satellite images in identifying water image pixels and thus identifying flooded places from satellite image. The results of image segmentation helps in separating water and non-water bodies so it will be useful in identifying flooded and non-flooded places from the extracted image.

Organization of the paper is as follows, in section -2, the study area description is presented; section-3 presents problem formulation, section-4 image processing methods are given. Section-5 gives the results and discussions. In the section-6, conclusion of this paper is presented.

2 Study Area

In this section, the study area chosen is Krishna and Tungabhadra rivers regions which flow in south India [] and area coverage is about 3, 13,568 sq mtrs2. During September-2009 rivers received heavy rainfall which caused river flooding so we have used MODIS (MOD09Q1) Terra Surface Reflectance 8-Day L3 Global 250 mtrs2 satellite images [3] because of wide coverage. Three different dated images like before (march-2009), during (September-2009) and after (November-2009) are considered in this study.

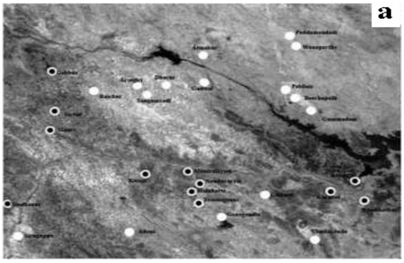


Fig. 1. Shows map of flooded (indicated by black dots within white dots) and non flooded places (indicated by white dots) which are used for flood assessment study.

3 Problem Formulation for Image Segmentation

This section explains problem formation for image segmentation. Segmentation problem involves the partitioning of a given image into a number of homogeneous segments and finally union of two neighbouring segments yields heterogeneous segments.

$$L = \{0, 1, 2, 3 \ldots\ldots\ldots\ldots L_m\}$$

be the set of the intensities of the image and

$$N_{mxn} = \{q = (z, w) \in S : | x - z | \leq \lfloor m/2 \rfloor, y - w \leq \lfloor n/2 \rfloor \quad (1)$$

$$S = \{(x, y)\} \quad 1 \leq x \leq N_c \quad 1 \leq y \leq N_r \quad (2)$$

are spatial co-ordinates of the pixel in N_r rows and N_c column image. The mxn neighbourhood of the pixel p=(x, y) over the S is given by

$$N_{mxn} = \{q = (z, w) \in S : | x - z | \leq \lfloor m/2 \rfloor, y - w \leq \lfloor n/2 \rfloor\} \quad (3)$$

Where: m and n are odd and $\lfloor . \rfloor$ Denotes the largest integer not greater that its argument.

The partition of an image is given by S

$\Delta_{k*}(S) = \{R_1, R_2 \ldots . R_{k*}\}$ for the natural number K^* such that

$$S = \bigcup_{k=1}^{k*} R_k$$

$$R_i \cap R_j = \varnothing$$

$\forall i, j \in \{i, 2 \ldots .. K*\}$ and $R_i \cap R_j = \varnothing \; \forall i, j \in \{i, 2 \ldots .. K*\} \; fori \neq j$

$for \cdot i \neq j$

$R_i \forall i \in \{1, 2 \ldots . K*\}$

$R_i \forall i \in \{1, 2 \ldots . K*\}$ is connected component.

$X(p)=C_m$ is a constant and C_m and C_n are not equal if R_m and R_n are adjacent.

The two regions are adjacent if they share a common boundary, i.e. if there is at least one pixel in one region, such that is 3x3 neighbourhoods contains at least one pixel belonging to the other region. According to the problem formulation the output of the image segmentation is represented by $\Delta_{k*}(S)$ and it is assumed that small pixel neighbourhoods contain either one (homogeneous) or two (heterogeneous) regions.

4 Image Processing Methods

Image filtering is applied to remove the clutters from the image. This noise reduction helps in preserving elongated river networks and thus helps to extract river features and group the similar water image pixels. Progressive median filter [7] is used to remove speckles.

4.1 Image Segmentation Methodology

Image segmentation is used to extract river networks by grouping the similar water image pixels. So to achieve this we have used non parametric mean shift and median cut methods.

Mean shift method (MS): MS is a popular non-parametric method based on kernel density estimation [8]. Initially both filtering and segmentation method is carried out on the image. This helps for identifying non-linear river network feature in the image. Initially arbitrary point is chosen in the feature space and move towards locally maximal density. The mean is shifted based on the weighted average and it is iteratively done to identify the similar pixels. Gaussian kernel is used for local point of convergence. Mean shift is carried-out in the two steps: a) Filtering and b) Segmentation.

In image filtering, the kernel density estimation is calculated using the similar image pixels which are given by:

$$\hat{f}(X) = \frac{1}{n h_i^d} \sum_{i=1}^{n} k \left\| \frac{(X - X_i)}{h_i} \right\|^2 \quad \ldots \tag{4}$$

The data points X_i, i=1, 2, 3.....n are in the d-dimensional space R^d, the kernel density estimation at the location x can is calculated using the bandwidth parameter h_i (where $h_i>0$). The kernel k is a spherically symmetrical kernel bounded which satisfies

$$K(x) = c_{k,d} k(\| x \|^2) > 0 where : \| x \| \leq 1 \tag{5}$$

where: The normalization constant $c_{k,d}$ definitely makes the k(x) integrates to 1 and k(x) is called the kernel profile. By assuming derivative of the kernel profile k(x) existed and g(x) =-k'(x) as the kernel profile.

The kernel G(x) is defined as $G(x) = c_{k,d} k(\| x \|^2)$

The gradient of equation is used to prove the property of kernel profile;

$$m_G(X) = c \frac{\nabla \hat{f}_k(x)}{\hat{f}(x)} \tag{6}$$

$where : m_G(X)$ is the mean shift vector, C is a positive constant and which gives the location x. Mean shift vector computed with kernel G is proportional to the normalized density gradient estimate obtained with the kernel K. The mean shift vector is defined as:

$$m_{h,G}(x) = \frac{\sum_{i=1}^{n} x_i g\left(\frac{\| x - x_i \|^2}{h}\right)}{\sum_{i=1}^{n} g\left(\frac{\| x - x_i \|^2}{h}\right)} - x \tag{7}$$

Mean shift vector thus points toward the direction of maximum increase in the density. The mean shift procedure is obtained by successive computation of the mean shift vector and translation of the kernel G(x) by the mean shift vector. Finally, it converges at a nearby point where the estimate has zero gradients and iterative equation is given by:

$$y_{i+1} = \frac{\sum_{i=1}^{n} x_i g(\| \frac{x - x_i}{h} \|^2)}{\sum_{i=1}^{n} g(\| \frac{x - x_i}{h} \|^2)} \qquad (8)$$

$$j = 1, 2, 3 \ldots . n$$

Initial position of the kernel is chosen as one of the data point's x_i. Usually, the local maxima / modes the density is the convergence points of the iterative procedure. The mean shift technique is based on unsupervised clustering. It is unsupervised as there is no information indicating the correct group, for example, for most image extraction problems there are no pixels marked as being part of a specific object. It iteratively shifts the mean of a pixel resulting in the pixel being drawn to a local point of convergence.

MEDIANCUT (MC): Median cut is a segmentation algorithm based on colour quantisation [9]. The colour quantization is a technique in computer graphics in order to find the best colour palette with the least differences between the original image and the quantized one. The colour quantisation is used as colour clustering algorithm of the satellite images in this paper. Pre-quantization precision, calculating the cutting position based on variance and searching reversely the colormap, significantly promotes both the speed and quality of the colour quantization. The median cut quantization algorithm is developed by Heckbert [9] and it is based on colour distribution of original image. The basic idea is to let each entry in the representative color set, Y, represent approximately the same number of pixels in the original image. Median Cut is carried out in three steps: a) representing image as cube of colours, b) sorting along axis and c) splitting based on median (carry out iteratively).

```
Median cut(image(x,y), level)
    {
            Image(x,y) : is the input image
            Level: box size (4, 6, 12, 32)
            //representing image as cube
            //create a colormap using image cube
            //sort along axis
            // splitting based on median (as reference)
            // adaptive partitioning
            //merging the similar groups
    }
```

Median cut algorithm constructs colour histogram based on the original image. The algorithm mainly performs a color reduction which reduces the number of possible of unique colors so storage and computation time is reduced. From the color histogram list L is constructed from non-zero entries. The RBG color value and corresponding histogram count is determined. Smallest box and largest box determined and sorted from smallest to largest using component as the sort key. The ordered list is spitted at based on median image pixels in order to create two sub lists. The entry is traversed until k such lists are formed. For each iteration, one of the previously constructed sub lists is selected for list splitting and the list is chosen with most pixels. Finally, representative set is created by computing the average color of each list.

5 Results and Discussions

In this section, two segmentation methods are used to extract the river network across the Krishna River from the March 2009 image and November image as shown in Fig. 3 and 4. The extracted region is overlaid on original image and verified with ground truth data. From figure-3, the extraction of the river network's can be seen and the same region is overlaid on the original image to verify the precision of extraction [11, 12] and it is measured using RMSE value.

We have used Root means square error (RMSE) parameter is used to verify the extracted image with ground truth image. RMSE is the statistical measure for varying magnitude quantity. The value near to zero is better extraction result.

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N} E_k^2} \tag{9}$$

Where: E_k is the difference between the ground truth data and algorithmically segmented image and N is the number of the pixels in the image

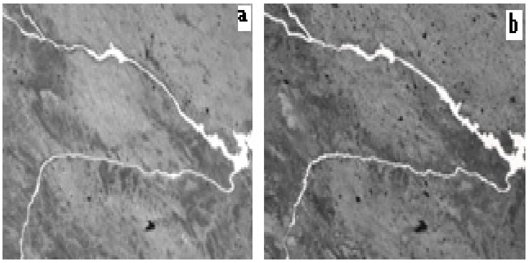


Fig. 2. Ground truth image for the a) March 2009 and b) November 2009 month

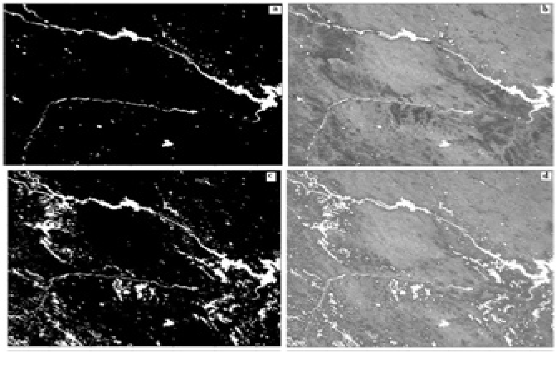


Fig. 3. (a) and (b) before flooded image March month using mean sift method for extraction and overlaying Fig-3.(c) and (d) median cut extraction and overlaid.

RMSE value between mean shift extracted March month image and ground truth image is 0.26.

RMSE between median cut extracted March month image and ground truth image is 0.37.

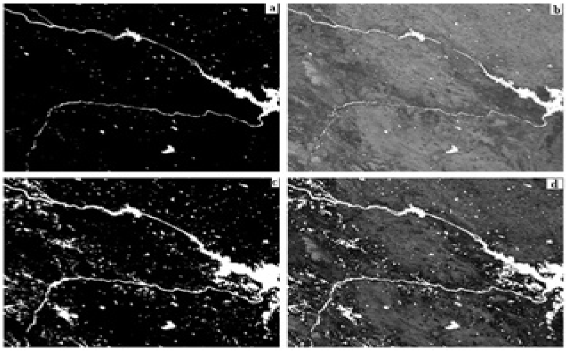


Fig. 4. (a) and (b) after flooded image November month using mean sift method for extraction and overlaying Fig-4.(c) and (d) median cut extraction and overlaid.

For the month of November month, mean sift extracted and ground truth is 0.27 and Median cut extracted and ground truth is 0.38

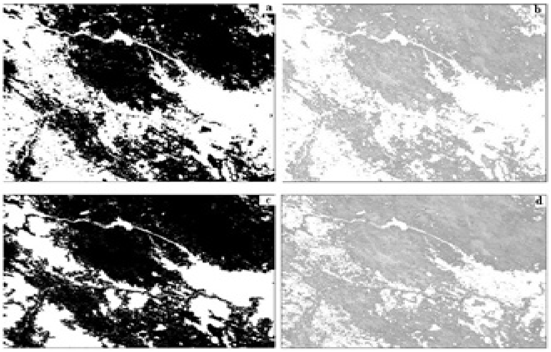


Fig. 5. (a) and (b) During flooded image September month using mean sift method for extraction and overlaying Fig-4.(c) and (d) median cut extraction and overlaid.

For the month of September 2009 image, mean shift and median cut are applied and extracted. But for during flooded image the extraction is verified using Receiver of characteristic parameters (ROC).

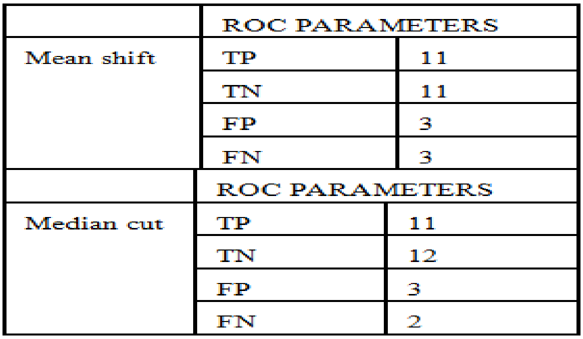
Receiver Operating Characteristics: ROC [10] consists of parameters like TP, TN, FP and FN which are calculated for the during flooded images for validation of the algorithmic performance [10].

In, during flooded images the ROC parameter is applied to locate the flooded places and distinguish flooded places from non-flooded places:

(a) Sensitivity or True Positive (TP) - A place which is positive according to the ground truth data and also according to the computed result is a True Positive (TP).
(b) Specificity or True Negative (TN) - A place which is negative according to the ground truth data and also according to the computed result is a True Negative (TN).
(c) False Positive (FP) - While the place which is positive according to the computed results but negative according to the ground truth data is a False Positive (FP).
(d) False negative (FN) - While the place which is negative according to the computed result but positive according to the ground truth data is a False Negative (FN).
The total number of flooded places – 12, Non-flooded places-16 and the total places verified -28.

These parameters are very helpful in order to determine the comparison of algorithms performance which is shown in the below table-1.

Table 1. Shows ROC parameter applied to September 2009 image in order to determine the flooded and non-flooded places



	ROC PARAMETERS	
Mean shift	TP	11
	TN	11
	FP	3
	FN	3
	ROC PARAMETERS	
Median cut	TP	11
	TN	12
	FP	3
	FN	2

6 Conclusion

We have used ROC parameter in order to identify the flooded cities. So from the table, mean shift performs better than median cut in identified the flooded places correctly. Also RMSE value of mean shift segmentation is less than median cut which proves a better extraction algorithm.

Acknowledgments. This work is supported by the Space Technology Cell, Indian Institute of Science and Indian Space Research Organization grant.

References

[1] Walkey, J.A.: Development of a change detection tool for image analysis. MS thesis. University of Wisconsin-Madison (1997)
[2] Bhavsar, P.D.: Review of remote sensing applications in hydrology and water resources management in India. Advances in Space Research 4(11), 193–200 (1984)

[3] Brakenridge, R., Anderson, E.: Modis-based flood detection, mapping and measurement: the potential for operational hydrological applications. NATO Science Series, 1, Volume 72, Transboundary Floods: Reducing Risks through Flood Management, 1, pp. 1-12

[4] Veronique, P., Zhou, Z., Songde, M.A.: A Framework for flood assessment using satellite images. IEEE International Geoscience and Remote Sensing, IGARSS 2, 822–824 (1998)

[5] Michael, S.A.: Multi-temporal Remote Sensing for mapping and monitoring Floods an approach involves validation of the KAFRIBA. Kafueu Flats. Zimba, Master Thesis (2007)

[6] Nath, R.K., Deb, S.K.: Water-Body Area Extraction from High Resolution Satellite Images-An Introduction, Review, and Comparison. International Journal of Image Processing (IJIP) 3(6), 265–384 (2010)

[7] Wang, Z., Zhang, D.: Progressive Switching Median Filter for the Removal of Impulse Noise from Highly Corrupted Images. IEEE Transactions on Circuits and Systems—ii: Analog and Digital Signal Processing 46(1) (January 1999)

[8] Comaniciu, D., Meer, P.: Mean shift analysis and applications. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision (1999)

[9] Heckbert, P.: Color image quantization for frame buffer display. In: Proceeding SIGGRAPH 1982 Proceedings of the 9th Annual Conference on Computer Graphics and Interactive Techniques. ACM SIGGRAPH Computer Graphics, vol. (16-3) (July 1982)

[10] Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874 (2006)

[11] Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Prentice Hall, Upper Saddle River

[12] Lillesand, T., Kiefer, R.W., Chipman, J.: Remote Sensing and Image Interpretation

Separable Discrete Hartley Transform Based Invisible Watermarking for Color Image Authentication (SDHTIWCIA)

J.K. Mandal and S.K. Ghosal

Department of Computer Science and Engineering,
Kalyani University, Kalyani,
West Bengal, India, 741235
`{jkm.cse,sudipta.ghosal}@gmail.com`
`http://www.klyuniv.ac.in, http://www.jkmandal.com/`

Abstract. In this paper a novel two-dimensional Separable Discrete Hartley Transform based invisible watermarking scheme has been proposed for color image authentication (SDHTIWCIA). Two dimensional SDHT is applied on each 2 × 2 sub-image block of the carrier image in row major order. Two bits are embedded in second, third and fourth frequency components of each 2 × 2 mask in transformed domain based on a secret key. Second and third bit position in each frequency coefficient has been chosen as embedding position. A delicate re-adjustment has incorporated in the first frequency component of each mask, to keep the quantum value positive in spatial domain without hampering the embedded bits. Inverse SDHT (ISDHT) is applied on each 2 × 2 mask as post embedding operation to produce the watermarked image. At the receiving end reverse operation is performed to extract the stream which is compared to the original stream for authentication. Experimental results conform that the proposed algorithm performs better than the Discrete Cosine Transform (DCT), Quaternion Fourier Transformation (QFT) and Spatio Chromatic DFT (SCDFT) based techniques.

Keywords: SDHTIWCIA, MSB, LSB, SDHT, ISDHT, DCT, QFT, SCDFT and Watermarked image.

1 Introduction

Watermarking is a technique of incorporating useful information into various digital media like image, audio etc. for ownership evidence, fingerprinting, authentication and integrity verification, content labeling and protection, and usage control. In our proposed scheme, we shall focus on separable discrete Hartley transform based invisible watermarking scheme for color image authentication.

The watermarking algorithm incorporates the watermark into the image, whereas the verification algorithm authenticates the image by determining the presence of the watermark and its actual data bits. Data can be embedded in both spatial and frequency domain. Frequency domain techniques are more suitable than spatial

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 767–776.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

domain techniques due to better security and robustness. The watermarking is performed in the cover (host) image through several frequency transformation approaches such as discrete cosine transform (DCT), discrete wavelet transform (DWT), discrete Fourier transform (DFT) etc. In frequency domain, hidden data are embedded into the frequency component of the transformed image pixels. To avoid severe distortion of the original image the midrange frequencies are best suitable for embedding to obtain a balance between imperceptibility and robustness. I. J. Cox et al. [1, 2] developed an algorithm to inserts watermarks into the frequency components and spread over all the pixels. DCT-based image authentication is developed by N. Ahmidi et al. [3] using just noticeable difference profile [4] to determine maximum amount of watermark signal that can be tolerated at each region in the image without degrading visual quality.

The Discrete Hartley Transformation [5] is used to convert the image from spatial domain to frequency domain. The frequency components values are used for embedding secret data. After embedding secret data, inverse Separable Discrete Hartley Transformation is applied to get back the embedded image into spatial domain. If carefully observe transformed and embedded pixel values, the pixel values are not preserved though embedded bits are intact, but, if we apply SDHT again, the frequency component values are not changed. The Hartley transform produces real output for a real input which can be designated as its own inverse. Thus it has computational advantages over the discrete Fourier transform, although analytic expressions are usually more complicated for the Hartley transform. The definition of SDHT is the difference of even and odd parts of the DFT.

The Separable Discrete Hartley Transform (SDHT) of spatial value f(x,y) for the image of size M x N is given in equation (1).

$$P_S(u,v) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} p(x,y)\mathrm{cas}(2\pi ux/N)\mathrm{cas}(2\pi vy/M) \tag{1}$$

Where, *u* varies from 0 to M-1 and *v* varies from 0 to N-1.

The variable *u* and *v* are the frequency variables corresponding to *x, y* and f(x,y) is intensity value of pixels in spatial domain. The sequence *cas* defined by:

$$cas(2\prod ux/N) = cos(2\prod ux/N) + sin(2\prod ux/N) \tag{2}$$

and

$$cas(2\prod vy/M) = cos(2\prod vy/M) + sin(2\prod vy/M) \tag{3}$$

Similarly, the inverse transformation to convert frequency component to the spatial domain value is defined in equation (2).

$$p(x,y) = \frac{1}{NM} \sum_{u=0}^{N-1} \sum_{v=0}^{M-1} P_S(u,v)\mathrm{cas}(2\pi ux/N)\mathrm{cas}(2\pi vy/M) \tag{4}$$

Where, *u* varies from 0 to M-1 and *v* from 0 to N-1.

The aim of SDHTIWCIA emphasizes on protection of secret information against unauthorized access. The proposed scheme exploits image authentication process by embedding the watermark data in both negative and positive frequency components along with the message digest MD (which is generated from watermark data) into the carrier image with a minimum change in visual pattern and improved security.

Problem motivation and formulation of transformation technique is given in section 2. Section 3 of the paper, deal with the proposed technique. Results, comparison and analysis are given in section 4. Conclusions are drawn in section 5. References are given at end.

2 Transformation Techniques

The formulations of image sub block of size 2 x 2 masks for four different image bytes in 2D-SDHT are as follows:

$$F\,(a_{0,0}) = \sum_{i=0}^{1}\sum_{j=0}^{1} a_{i,j} = c_{0,0}\ (say), \qquad F\,(a_{0,1}) = \sum_{i=0}^{1}\sum_{j=0}^{1} (-1)^{j}\ a_{i,j} = c_{0,1}\ (say),$$

$$F\,(a_{1,0}) = \sum_{i=0}^{1}\sum_{j=0}^{1} (-1)^{i}\ a_{i,j} = c_{1,0}\ (say), \qquad F\,(a_{1,1}) = \sum_{i=0}^{1}\sum_{j=0}^{1} (-1)^{i}\ (-1)^{j} a_{i,j} = c_{1,1}\ (say),$$

Where, $a_{0,0}$, $a_{0,1}$, $a_{1,0}$ and $a_{1,1}$ represent the spatial domain cover image bytes. Here $c_{0,0}$, $c_{0,1}$, $c_{1,0}$ and $c_{1,1}$ are all frequency component for $a_{0,0}$, $a_{0,1}$, $a_{1,0}$ and $a_{1,1}$ spatial domain values respectively. A fixed size of two bits is fabricated at the second and third bit position of LSB part of $c_{0,1}$, $c_{1,0}$ and $c_{1,1}$ based on a secret key. The first frequency component ($c_{0,0}$) is used as re-adjust phase to balance the quantum values between original and embedded data.

Similarly, by applying the inverse 2D-SDHT, the 2 x 2 transformed masks can be formulated as:

$$F\,(c_{0,0}) = \tfrac{1}{4}\sum_{i=0}^{1}\sum_{j=0}^{1} c_{i,j} = a_{0,0}\ (say), \qquad F\,(c_{0,1}) = \tfrac{1}{4}\sum_{i=0}^{1}\sum_{j=0}^{1} (-1)^{j}\ c_{i,j} = a_{0,1}\ (say),$$

$$F\,(c_{1,0}) = \tfrac{1}{4}\sum_{i=0}^{1}\sum_{j=0}^{1} (-1)^{i}\ c_{i,j} = a_{1,0}\ (say), \qquad F\,(c_{1,1}) = \tfrac{1}{4}\sum_{i=0}^{1}\sum_{j=0}^{1} (-1)^{i}\ (-1)^{j} c_{i,j} = a_{1,1}\ (say),$$

On re-adjustment, all inverse 2D-SDHT values are non negative and less than or equal to the maximum and greater than or equal to minimum possible value of a byte.

3 The Technique

In this paper a novel invisible watermarking scheme has been proposed for color image authentication (SDHTIWCIA) in frequency domain based on the two dimensional Separable Discrete Hartley Transform (SDHT). Initially, a 128 bit message digests (MD) and size of the watermark data is embedded using the proposed SDHTIWCIA scheme for authentication purpose. The SDHT is applied on 2 × 2 sub-image block for converting the spatial domain values to frequency components. This process is continued till the last sub-image block of the carrier/cover image in a row major order. In the proposed SDHTIWCIA scheme for each 2 × 2 sub-image block, the second and third bit position of the LSB part (i.e., LSB-2 and LSB-3) of each frequency components (except the first frequency component) are chosen for embedding authenticating bits. In each embedding, the authenticating watermark bits are embedded either in usual order or in reverse order to enhance the security of the hidden data. This can be accomplished by a secret key (K). The last two bits of each frequency component are kept unaltered. Now, If the modulo result of the summation of ASCII values of all the characters of key (K) by four (i.e., K % 4) matches with the numeric value produced by last two bits of the embedding frequency component then two bits from the watermark data are embedded in usual order else in the reverse order. Moreover, the first frequency component (excluding last two consecutive bits from LSB) has been used for re-adjustment of frequency components whenever it violates the basic principles of pixel representation in spatial domain like non-negative pixel value and a value less than or equal to 255 for eight bit representation. In the proposed technique, the value of frequency components does not become fractional as we are not changing the least two significant bits of the LSB. If the value becomes negative, then a multiple of four is added whereas if the value becomes greater than 255, an even multiple of four is deducted. Inverse Transform is applied on each 2 x 2 mask as post embedding to transformed embedded image (sometimes, re-adjusted as well) in frequency domain to convert back into spatial domain. Secret data is extracted in same manner from the watermarked image and the same is compared with original for authentication.

Consider the Baboon image as the cover/carrier image. 2D-SDHT is applied on the first 2 x 2 image block which consists of three sub-matrices namely R, G and B to convert it from spatial domain pixel value to frequency components value in transform domain.

$R_1=\{164,63,120,135\}$, $G_1=\{150,57,125,97\}$, $B_1=\{71,31,62,33)$

Applying 2D-SDHT the transformed frequency component values obtained as given below:

$F(R_1)=\{482,86,-28,116\}$, $F(G_1)=\{429,121,-15,65\}$, $F(B_1)=\{197,69,7,11\}$

Secret binary stream 101000010110000011 is embedded based on the secret key (K) “helloskf”. Hence,

$$(K \% 4) = ((104 + 101 + 108 + 108 + 111 + 115 + 107 + 102) \% 4) = 0$$

Now, based on the numeric value of last two consecutive bits of each frequency components (except the first) matches with the value produced by K % 4 then two bits from the authenticating watermark data are embedded in normal order into second and third bit position of second, third and fourth frequency component. Otherwise, two bits from the authenticating watermark bits are embedded by flipping the two bits i.e., in reverse order to second, third and fourth frequency component. Hence, the modified frequency components are:

$EF(R_1)=\{482,86,-24,112\}$, $EF(G_1)=\{429, 121, -11, 69\}$, $EF(B_1)=\{197,65,3,15\}$

Again, if we apply inverse SDHT values in pixel domain the regenerated pixel component values in spatial domain will be:

$F^{-1}(EF(R_1))=\{164,65,120,133\}$, $F^{-1}(EF(G_1))=\{150,57,123,95\}$, $F^{-1}(EF(B_1)) = \{71,29,60,33\}$

It is seen that the modified pixel values are non-fractional as, the last two bits of each frequency component are unaltered. Re-adjustment of the pixel values are not needed in this example as the spatial domain values are non-negative and not greater than 255.

The proposed scheme is described in the following sections namely, the Insertion, Re-adjustment and the Extraction. These are described in sec. 1, 2 and 3 respectively.

3.1 Insertion

Insertion is made at each transformed blocks of size 2 x 2 using two dimensional separable Discrete Hartley Transform. All the three channels of 2 x 2 masks in a 24 bit color image have been chosen for embedding two bits from the authenticating message in the second and third bit position of each transformed component except the first. The key (K) in each embedding frequency component specifies whether the watermark data is embedded in usual order or in reverse order. The authenticating message/image bits size is 1.5 * (m * n) – (MD + L) where MD and L are the message digest and dimension of the authenticating image respectively for the source image size of m x n bytes. The L and MD are used in extraction phase to extract the whole authenticating message\image and to authorize authenticating message/image.

Algorithm:

1. Obtain 128 bits message digest MD from the authenticating message/image.
2. Obtain the size of the authenticating message/image ((m + n) bits, where m bits for width and n bits for height).
3. Read authenticating message/image data do:
    - Read source image matrix of size 2 × 2 mask from image matrix in row major order and apply 2D-SDHT.
    - Extract two bits from authenticating message/image.
    - Check modulo result of the summation of ASCII values of all the characters of key (K) by four (i.e., K % 4). If the resultant value matches with the numeric value produced by last two consecutive bits of the embedding frequency component then two bits from the authenticating watermark data are embedded in usual order else in reverse order.
    - The authenticating message/image bits are embedded in second and third bit position within 2nd, 3rd and 4th frequency component values.
4. Apply inverse two dimensional SDHT using identical masks.
5. Apply re-adjust phase, if needed.
6. Repeat step 3 to step 5 for the whole authenticating message/image size, content and for message digest MD.
7. Stop.

3.2 Re-adjustment

In the proposed algorithm after embedding we have used inverse transformation (ISDHT) to obtain the embedded image in spatial domain. Applying inverse transform on identical mask with embedded data of the frequency component value which may change and can generate the following situation:

- The converted value may by negative (-ve).
- The converted value may be greater than the maximum value (i.e. 255).

The concept of re-adjust phase is to handle the above two serious problems by using the first frequency component of each 2 × 2 mask. In this phase if the converted value is negative (-ve) i.e. for case (i), the operation applied for each 2 x 2 mask is as follows:

$$F_B(0,0) = F_B(0,0) + I * 4 \tag{5}$$

Here, $F_B(0,0)$ is the first frequency component of the block number B and I is the multiple of four, takes values in the range, I = 1, 2, 3, …, n. That means, I is multiplied and incremented in each step till all the converted value in spatial domain value becomes positive.

For case (ii), if the converted value exceeds the maximum value of a byte (i.e., 255) in spatial domain, then the operation applied for each 2 x 2 mask is as follows:

$$F_B(0,0) = F_B(0,0) - J * 4 \tag{6}$$

Here, J is the even multiple of four, takes values in the range, J = 2, 4, 6,…, n when *n* is the positive even integer. That means J is multiplied and incremented in each step till all the converted value in spatial domain value becomes less than or equal to 255.

3.3 Extraction

The authenticated watermarked image is received in spatial domain. During decoding, the secret key (K) and watermarked image has been taken as the input and the authenticating message/image size, image content and message digest MD are extracted from it. All extraction is done in frequency domain from frequency component.

Algorithm:

1. Read watermarked image matrix of size 2 × 2 mask from image matrix in row major order and apply 2D-SDHT.
2. For each 2 x2 transformed mask do:
    - Extract two bits from each transformed frequency component except the first.
    - Check modulo result of the summation of ASCII values of all the characters of key (K) by four (i.e., K % 4). If the resultant value matches with the numeric value produced by last two consecutive bits of the

extracting frequency component then two bits of the authenticating watermark data are extracted in usual order else in reverse order.

- The authenticating watermark message/image bits are extracted from second and third bit position of 2nd, 3rd and 4th frequency component values.
- For each 8 (eight) bits extraction, it construct one alphabet/one primary (R/G/B) color image.

3. Repeat step 1 and step 2 to complete decoding as per the size of the authenticating message/image.
4. Obtain 128 bits message digest MD′ from the extracted authenticating message/image. Compare MD′ with extracted MD. If both are same then the image is authorized, else unauthorized.
5. Apply inverse SDHT using identical mask.
6. Stop

4 Results, Comparison and Analysis

This section represents the results, discussion and a comparative study of the proposed SDHTIWCIA scheme with the DCT, QFT based and Spatio-Chromatic DFT based watermarking methods in terms of payload capacity and visual interpretation, on the basis of peak signal to noise ratio (PSNR) analysis, bits per byte (BPB) and histogram analysis. Benchmark (PPM) images [6] are taken to formulate results and are shown in Fig-1. All cover images are 512 x 512 in dimension whereas the gold coin (i.e. the secret data) is embedded into the various source benchmark images. The experiment deals with ten different color images (i-x), where each pixel is represented by three intensity values RGB (Red, Green and Blue). Images are labeled as: (i) Lena, (ii) Baboon, (iii) Pepper, (iv) Airplane, (v) Splash, (vi) Earth, (vii) Sailboat, (viii) Foster City, (ix) San Diego, (x) Oakland. On embedding the watermark image that is the Gold-Coin image based on the secret key “helloskf”, the newly generated watermarked image produces a good visual clarity.

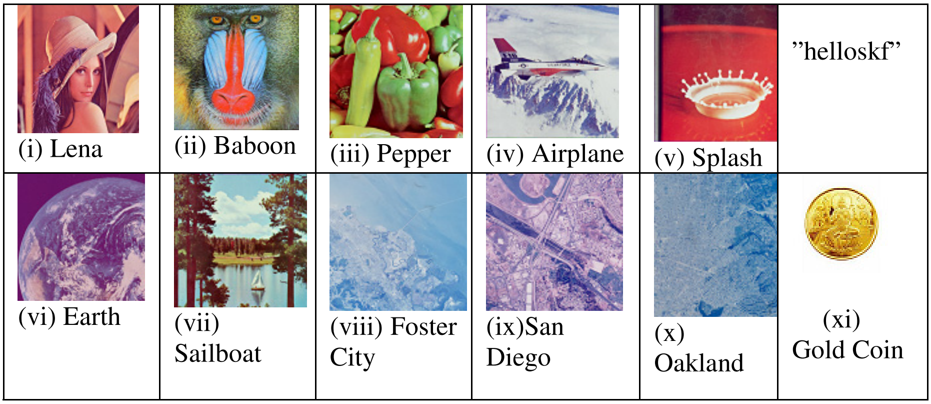

(i) Lena (ii) Baboon (iii) Pepper (iv) Airplane (v) Splash ”helloskf”
(vi) Earth (vii) Sailboat (viii) Foster City (ix)San Diego (x) Oakland (xi) Gold Coin

Fig. 1. Cover images (512 x 512), secret image (220 x 223) and that of Secret Key

In order to test the robustness, we have applied the technique on additional 25 PPM and 25 BMP color images of bit depth 24 bit, from which we can see that the technique can handle many kinds of visual and statistical attacks and it is quite difficult for the observer to detect the difference between the original and embedded image. From Table 1, we can identify the payload for carrier images which is 147456 bytes where the dimension of each original image is 512 x 512. After embedding the watermark data, the watermarked image is also retain a good visual clarity and produces value of 38 dB for peak to signal noise ratio in average cases. Moreover, the histogram analysis shows the changes made in the three images are more stable after embedding hidden bits. The table also shows that the bits embedded per byte (bpb) for each carrier image is 1.5. Fig-2 shows different states of modifications (before and after) of three different images viz. Lena, Baboon and Peppers.

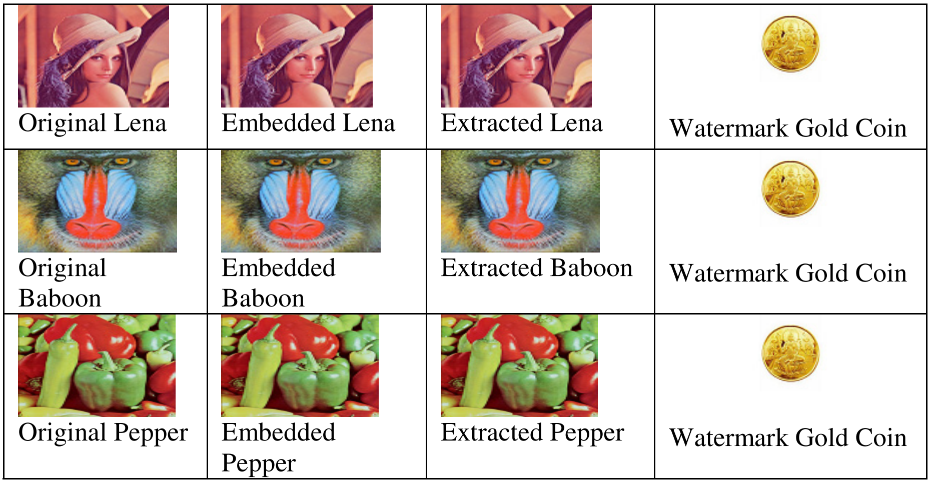

Original Lena
Embedded Lena
Extracted Lena
Watermark Gold Coin
Original Baboon
Embedded Baboon
Extracted Baboon
Watermark Gold Coin
Original Pepper
Embedded Pepper
Extracted Pepper
Watermark Gold Coin

Fig. 2. Cover, Watermarked, Extracted and Watermark Images using proposed SDHTIWCIA scheme

Also, a comparative study has been made among Discrete Cosine Transform (DCT), Quaternion Fourier Transformation (QFT) and Spatio Chromatic DFT (SCDFT) based scheme and our proposed SDHTIWCIA scheme based on the payload and the PSNR values. In the proposed scheme, the payload and PSNR is much more as compared to the SCDFT, QFT and SCDFT techniques, besides pertaining good visual clarity watermarked images. For the Lena image, the payload is more than 143616 bytes and PSNR enhancement are around 8 dB.

In Fig-3, the histogram analysis of Lena image is shown before and after embedding watermark data in an individual channel wise manner.

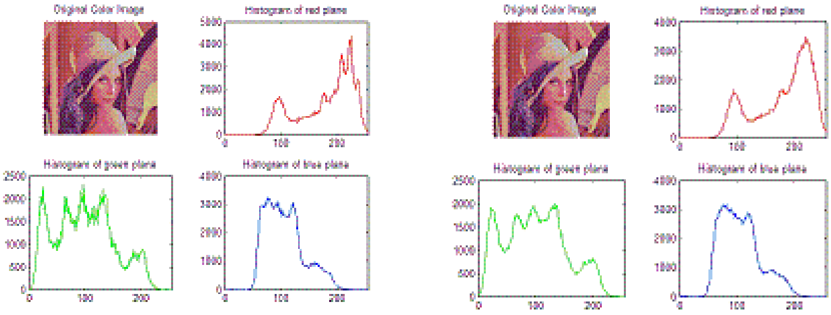
The histogram analysis shows the comparison results in terms of mean, standard deviation and median between original and watermarked 'Lena' image in a channelwise manner. The experimental results in Table 3 also ensures that the differences between two images is very minimal and tough to detect for the attacker.

Table 1. Results of embedding of 147234 bytes of information in each Image of dimension 512 x 512

Carrier Image	**Max. Payload (byte)**	**PSNR**	**BPB**
Lena	147456	37.95	1.5
Baboon	147456	38.57	1.5
Pepper	147456	37.76	1.5
Earth	147456	38.29	1.5
Sailboat	147456	38.23	1.5
Airplane	147456	37.40	1.5
Foster City	147456	38.08	1.5
Oakland	147456	38.30	1.5
San Diego	147456	38.55	1.5
Splash	147456	37.03	1.5
AVG	*147456*	*38.01*	*1.5*

Table 2. Comparison results of Payload Capacities and PSNR for Lena image in the existing technique namely SCDFT, QFT and DCT

Technique	**Payload(bytes)**	**PSNR(dB)**
SCDFT	3840	30.1024
QFT	3840	30.9283
DCT	3840	30.4046
SDHTIWCIA	*147456*	*37.95*



Histogram of Original Lena

Histogram of Watermarked Lena

Fig. 3. Comparisons Results of Histogram between of Original and Watermarked Lena Image

Table 3. Comparison results of Mean, Median and Standard Deviation of Original and Watermarked Lena Image

Image	Channel	Mean	Median	Standar Deviation
Original Lena	R	180.22	197	49.05
	G	99.05	97	52.88
	B	105.41	100	34.06
Watermarked Lena	R	180.20	196	49.21
	G	99.05	98	53.05
	B	105.11	101	34.32

5 Conclusion

The SDHTIWCIA scheme is an image authentication process in frequency domain to enhance the security compared to the existing algorithm. Authentication is done by embedding secret data in a carrier image. Using the technique a total of eighteen bits can be embedded in 2 x 2 image block. Experimental results conform that the proposed algorithm performs better than existing techniques.

Acknowledgments. The authors express gratitude to the Department of CSE, University of Kalyani and the PURSE scheme of DST, Govt. of India.

References

1. Cox, J., Kilian, J., Leighton, F.T., Shamoon, T.: Secure spread spectrum watermarking for images, audio and video. In: Proc. IEEE Int. Conf. Image Processing, September 16-19, vol. 111, pp. 243–246 (1996)
2. Cox, J., Kilian, J., Leighton, F.T., Shamoon, T.: Secure spread spectrum watermarking for multimedia. IEEE Trans. Image Processing 6(12), 1673–1687 (1997)
3. Ahmidi, N., Safabkhsh, R.: A novel DCT-based approach for secure color image watermarking. In: Proc. Int. Conf. Information technology: Coding and Computing, vol. 2, pp. 709–713 (April 2004)
4. Chou, H., Li, Y.C.: A perceptually tuned subband image coder based on the measure of just-noticeable distortion profile. IEEE Trans. Circuits Syst. Video Technology 5(6), 467–476 (1995)
5. Watson, A.B., Poirson, A.: Separable two dimensional discrete Hartley transform. J. Opt. Soc. Am. A. 3(12) (December 1986)
6. Weber, A.G.: The USC-SIPI Image Database: Version 5, Original release (October 1997), Signal and Image Processing Institute, University of Southern California, Department of Electrical Engineering, `http://sipi.usc.edu/database/`

Normalised Euclidean Distance Based Image Retrieval Using Coefficient Analysis

Nilofar Khan[1] and Wasim Khan[2]

[1] Department of Information Technology, University of RGPV Bhopal
Chameli Devi School of Engineering Indore, India
`nilofarwkhan@gmail.com`
[2] Department of Computer Science, University of RGPV Bhopal
Govt. Women's Polytechnic College Indore, India
`wasukhan1982@gmail.com`

Abstract. The article presented a novel method based on normalized Euclidean distance using application of discrete wavelet transform and bins intensity measurement, which is then coupled to a parameterized framework for content-based image retrieval. The discrete wavelet transform captures both frequency and location information and make image retrieval efficient. It further facilitates to incorporate recent research work on feature based coefficient distributions. We demonstrate the applicability of the proposed method in the context of color texture retrieval on different image databases and compare retrieval performance to a collection of state-of-the-art approaches in the area. Our experiment results on a large database further include a thorough analysis of computations of the main building blocks and runtime measurements of images**.**

Keywords: Content based image retrieval, discrete wavelet transform, Euclidean distance, Bins intensity measurement, Texture feature, Color feature.

1 Introduction

The fundamental challenge in image mining is to reveal out how low-level pixel representation enclosed in a raw image or image sequence can be processed to recognise high-level image objects and relationships [5] [27]. A picture is said to be worth a thousand words. If this statement is true, it is no wonder that computerised image retrieval is a challenging task. A key to a capable image retrieval system is how to extract and describe the image contents [1]. The current content-based image retrieval (CBIR) approach uses these image features to define the model of semblance between images [2]. Content-based image retrieval has been an active field of study for many years [2] [28]. Content Based Image Retrieval (CBIR) is an important research area for manipulating large multimedia databases and digital libraries [21]. High retrieval efficiency and less computational complexity are the desired characteristics of CBIR system [21]. CBIR finds applications in advertising, medicine, crime detection, entertainment, and digital libraries. Computational complexity and retrieval efficiency are the key objectives in the design of CBIR system [7] [18].

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 777–785.
springerlink.com

© Springer-Verlag Berlin Heidelberg 2013

Several attempts have been made in the recent past to improve the process of image retrieval, but a very few of them are information theoretic [8].During the last decade, a new image retrieval approach, called Content-Based Image Retrieval (CBIR), emerged. In this approach, the content of an image is described using low-level features such as color, texture, and shape. It is almost impossible to describe texture in words, because it is virtually a statistical and structural property [13] [29] [30].

Various texture representations have been investigated in pattern recognition and computer vision [8][20]. Despite their advantages over the traditional text-base image retrieval systems, CBIR systems face a major problem commonly referred to as the semantic gap, whereby the description of the images using the low-level features is unable to capture the semantic intended by the user in his/her queries. Therefore, CBIR systems produce a large amount of false positives in the retrieval process.

CBIR involves the following four parts in system realization: data collection, builds up feature database, search in the database, arrange the order and deal with the results of the retrieval.

1) Data collection Using the Internet spider program that can collect webs automatically to interview Internet and do the collection of the images on the web site, then it will go over all the other webs through the URL, repeating this process and collecting all the images it has reviewed into the server[4].

2) Build up feature database using index system program do analysis for the collected images and extract the feature information. Currently, the features that use widely involve low-level features such as color, texture and so on, the middle level features such as shape etc [4].

3) Search the Database The system extract the feature of image that waits for search when user input the image sample that need search, then the search engine will search the suited feature from the database and calculate the similar distance, then find several related webs and images with the minimum similar distance[4].

4) Process and index the results after researching Index the image obtained from searching due to the similarity of features, then return the retrieval images to the user and let the user select. If the user is not satisfied with the searching result, he can re-retrieval the image again, and searches database again [4]. The retrieval of content based image involves the following systems [9].Research in content- based image retrieval (CBIR) today is an extremely active discipline. There are already re vie w articles containing references to a large number of systems and description of the technology implemented. A more recent review reports a tremendous growth in publications on this topic. Applications of CBIR systems to medical domains already exist, although most of the systems currently available are based on radiological images [14].

Texture feature is a kind of visual characteristics that does not rely on color or intensity and reflects the intrinsic phenomenon of images. It is the total of all the intrinsic surface properties. That is why the texture features has been widely used in image retrieval [15].

The color histogram for an image is constructed by counting the number of pixels of each color [16][19]. In these studies the development of the extraction algorithms follows a similar progression (1) selection of a color space (2) quantization of the color space (3) computation of histograms [16][19].

CBIR systems can be based on many features, viz., texture, color, and shape and edge information. Texture contains important information about the structural arrangement of surfaces and their relationship to the surroundings [17].

A significant improvement is obtained by integrating the spatial distribution of the visual features since it captures better the contents of the images and reduces the number of false positives [10].

The exponential growth of image data that are being generated makes it imperative to use computers to save, retrieve and analyze images. Any pixel in the image can be described by three components in a certain colour space(for instance, red, green and blue components in RGB space or hue, saturation and value in HSV space), a histogram, i.e., the distribution of the number of pixels for each quantized bin, can be defined for each component [26]. By default the maximum number of bins one can obtain using the histogram function in MatLab is 256[16].The problem of image retrieval has been an active area of research since early 70's. In order to make the best use of information in images, we need to organize the images so as to allow efficient browsing, searching and retrieval. The basic two approaches for image retrieval are text-based and visual-base. Early image retrieval techniques were generally based on textual annotation of images rather than visual features. In other words, images were first annotated with text and then searched using a text-based approach from the traditional database management systems [2].Content-based image retrieval (CBIR) has been an active research topic in the last few years. Comparing to the traditional systems, which represent image contents only by keyword annotations, the CBIR systems perform retrieval based on the similarity defined in terms of visual features with more objectiveness[9][13]. Although some new methods, such as the relevant feedback, have been developed to improve the performance of CBIR systems, low-level features do still play an important role and in some sense be the bottleneck for the development and application of CBIR techniques[9][13]. A very basic issue in designing a CBIR system is to select the most effective image features to represent image contents [9][13]. Many objects in an image can be distinguished solely by their textures without any other information [22]. There is no universal definition of texture. Texture may consist of some basic primitives, and may also describe the structural arrangement of a region and the relationship of the surrounding regions [15] [22]. Content Based Image Retrieval (CBIR) is an important research area for manipulating large multimedia databases and digital libraries. High retrieval efficiency and less computational complexity are the desired characteristics of CBIR system [17]. The application area of CBIR is very vast and useful some applications are as follows:

- Art galleries and museum management,
- Architectural and engineering design,
- Interior design,
- Remote sensing and management of earth resources,
- Geographic information systems,
- Weather forecasting,
- Fabric and fashion design,
- Trademark and copyright database management.
- Law enforcement and criminal investigation.

2 Proposed Method for Image Retrieval

Content-based image retrieval (CBIR) is a new but widely adopted method for finding images from vast and un annotated image databases. In CBIR images are indexed on the basis of low-level features, such as color, texture, and shape that can automatically be derived from the visual content of the images[4] [14] [24].

Here we propose an efficient approach for image retrieval based on color and texture descriptor features. Similar to most CBIR systems, we need to index images by extracting their features in an offline process. We then submit a query image and find similar images to that query based on a matching criterion. We first start with feature extraction. Figure1 represents the proposed scheme architecture for this step. Firstly a database is prepared of different type of images. After this, analysis is performed on database. Analysis represents assessment of different descriptors used in this approach. Database is indexed according to values of different images. Finally the database is arranged on the basis of measures. When a user query is submitted for similarity matching the steps of analysis and feature selection is repeated as performed with image database. Now the value of query image is compared with the values of different images stored in database. As a result, the images having closest values compared to query image color and texture values are extracted from database.

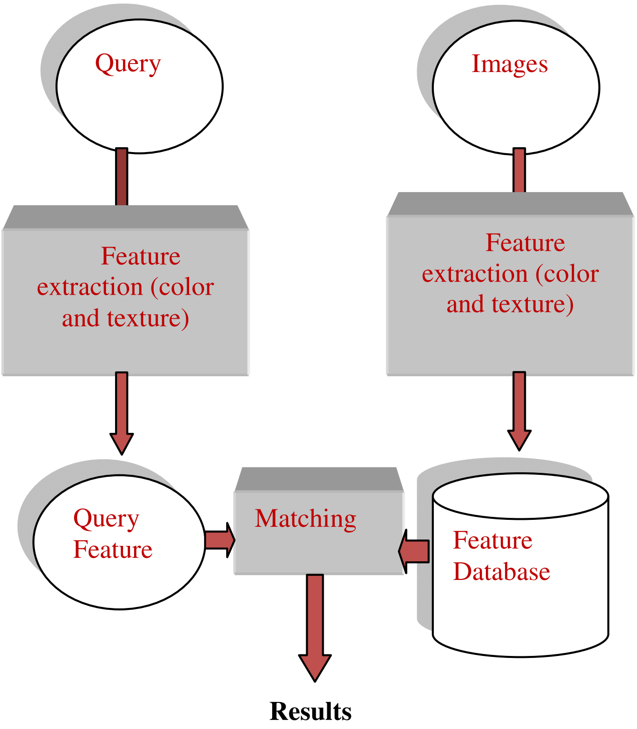

Query
Images
Feature extraction (color and texture)
Feature extraction (color and texture)
Query Feature
Matching
Feature Database
Results

Fig. 1. Proposed System Architecture

According to figure when a query image is submitted for image retrieval ,its color features are extracted and matching operation is performed between query image features and the image features stored in database .The result close to the query image is then retrieved from the database. First we load the database in the Matlab workspace after loading the database we resize the image for [128. 128] to get the similar size of images after that we Convert images from RGB to Gray and HSV for texture and color. Then we normalize the gray image for fixed mean [3]. After this we find the square mean value of this 10 coefficient that gives 10 signature of each image. After that we find out different signatures of hue, saturation and value .When a test image is loaded we apply the procedure 2-8 of algorithm to find signature of test image after that we determine the normalized Euclidean distance between query image signatures and database image signatures with indexing. The closest values are displayed on GUI as result [3].

Color histograms are frequently used to compare images. Examples of their use in multimedia applications include scene break detection and querying a database of images [3]. Color histograms are popular because they are trivial to compute, and tend to be robust against small changes in camera viewpoint [11]. In this paper, gray level variations are used to compute the texture feature of any image. For this purpose the color image is first converted in to gray level image. Then the values of wavelet coefficients are computed from gray level variations. According to combined color and texture features, images are extracted from the database.

Humans perceive color in an object through the nature of light reflected by that object. The characteristics generally used to distinguish one color from another are brightness, hue, and saturation [23] [25]. In computer vision, color represents a property of a point picture element or a "pixel" in a digitized image. For a given colored pixel, hue and saturation represent the chromaticity, while brightness, the intensity of the pixel. The purpose of color spaces or color models is to represent how color is to be perceived according to some standard [6].

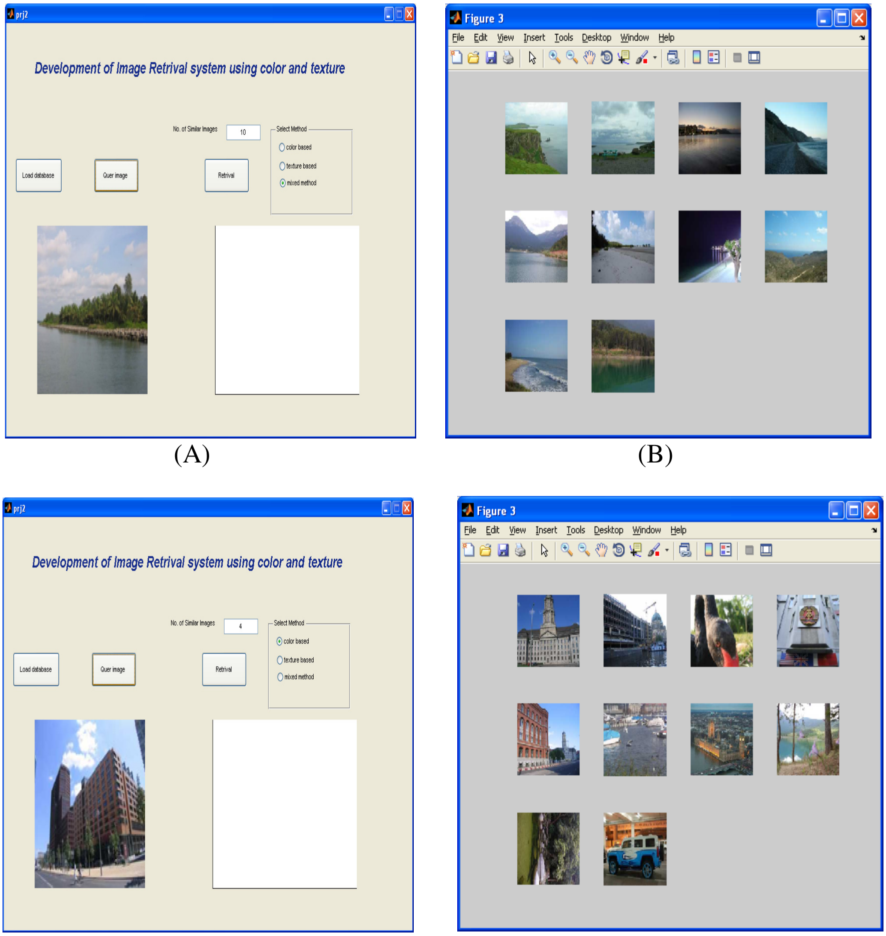
Images can be retrieved from the digital image database on the basis of colour, shape or texture. Among these texture is one of the most important features due to its existence in most real and artificial world images, which makes it under high interest not only for CBIR but also for many other applications in computer vision, medical imaging, remote sensing, and so on [12].

2.1 Algorithm for Scheme Used

Step 1: Load database in the Mat lab workspace.
Step 2: Resize the image for [128. 128].
Step 3: Convert image from RGB to HSV.
Step 4: Generate the Histogram of of hue, saturation and value.
Step 5: Generate 18 intensity value of hue, 3 for saturation and 3 for value.
Step 6: Convert image from RGB to Gray.
Step 7: Normalize the gray image for fixed mean.
Step 8: Apply the 3 level 2-D wavelet transform to find the 10 wavelet coefficient.
Step9: Find the square mean value of this 10 coefficient which give 10 intensity value of each image.
Step 10: Normalize the gray image for fixed mean.

Step 11: Combine the image feature.
Step 12: Store the intensity value of database images into the mat file.
Step 13: Load the test image.
Step 14: Apply the procedure 2-11 to find combine feature of test image.
Step 15: Determine the normalized Euclidean distance of intensity of test image with stored intensity of database.
Step 16: Sort obtained normalized Euclidean distance values to perform indexing to get the result.
Step 17: Display the result on GUI.

3 Experimental Results


Development of Image Retrival system using color and texture

No. of Similar Images 10

Select Method
○ color based
○ texture based
● mixed method

(A) (B)

Development of Image Retrival system using color and texture

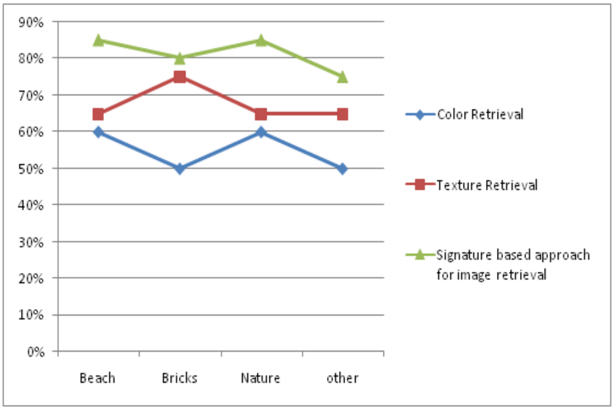
No. of Similar Images 4

Select Method
● color based
○ texture based
○ mixed method

(A) (B)

Fig. 3. (A) Query Image. (B) Result of Query Image

4 Comparison Chart of Schemes Used



5 Conclusions

The article represented a normalised Euclidean distance based method for image retrieval using coefficient analysis. For color retrieval, color histogram method is used. Color histogram counts the bins having same color intensity in image. In this paper HSV color model is used. For texture retrieval, discrete wavelet transform is used. From the wavelet family Haar transform is taken in to consideration. We then developed a mechanism for image retrieval based on these two image features with the help of MATLAB tool. We demonstrate the applicability of the proposed method in the context of color texture retrieval on different image databases and compare retrieval performance to a collection of state-of-the-art approaches in the area. When a query image is submitted, its texture and color value is compared with the texture and color value of different images stored in database. The images having closest value compared to query image are retrieved from database as result.

This work can be extended by taking other feature of image in to consideration. Some other parameter values can also be used for measurement.

References

[1] Chang, N.-S., Fu, K.-S.: Query by pictorial example. IEEE Transactions on Software Engineering 6(6), 519–524 (1980)

[2] Smeulders, A.W.M., Woming, S., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)

[3] Khan, W., Gupta, S.K.N., Khan, N.: A Proposed Method for Image Retrieval using Histogram values and Texture Descriptor Analysis. International Journal of Soft Computing and Engineering (IJSCE) I(II) (May 2011) ISSN: 2231-2307

[4] Singhai, N., Shandilya, S.K.: A Survey On: Content Based Image Retrieval Systems. International Journal of Computer Applications (0975 – 8887) 4(2) (July 2010)

[5] Zhang, J., Hsu, W., Li Lee, M.: An Information-Driven Framework for Image Mining. In: Mayr, H.C., Lazanský, J., Quirchmayr, G., Vogel, P. (eds.) DEXA 2001. LNCS, vol. 2113, pp. 232–242. Springer, Heidelberg (2001), doi:10.1007/3-540-44759-8_24

[6] Sheila Angeli Marcos, M., Soriano, M., Saloma, C.: Low-Level Color and Texture Feature Extraction of Coral Reef Components (2007)

[7] Srinivasa Rao, C., Srinivas Kumar, S., Chatterji, B.N.: Content Based Image Retrieval using Contour let Transform

[8] Das, A.: Entropy-Based Indexing On Color And Textur. In: Image Retrieval

[9] Yu, H., Li, M., Zhang, H.-J., Feng, J.: Color Texture Moments For Content Based Image Retrieval

[10] Bdesselam, A., Wang, H.H., Arayanan, K.: Spiral Bit-string Representation of Color for Image Retrieval

[11] Histogram Refinement for Content-Based Image Retrieval, Greg Pass Ramin Zabih

[12] Tamura's Texture Features

[13] Nirmal, S.: Proceedings of the 3rd National Conference; INDIA Com-2009 Computing For Nation Development, February 26-27, Bharti Vidhyapeet 's Institute of Computer Applications and management, New Delhi. Content Based Image Retrieval Techniques (2009)

[14] Ganeshwara Rao, N., Vijaya Kumar, V., Venkata Krishna, V.: Texture Based Image Indexing and Retrieval. IJCSNS International Journal of Computer Science and Network Security 9(5) (May 2009)

[15] Kavitha, C., Prabhakara Rao, B., Govardhan, A.: An Efficient Content Based Image Retrieval Using Color And Texture of Image Subblocks. International Journal of Engineering Science and Technology (IJEST) 3(2) (February 2011)

[16] Sharma, N., Rawat, P., Singh, J.: Efficient CBIR Using Color Histogram Processing. Signal & Image Processing: An International Journal (SIPIJ) 2(1) (March 2011)

[17] Reddy, P.V.N., Sataya Prasad, K.: Content Based Image Retrieval Using Local Derivative Patterns. 28(2) (June 30, 2011)

[18] Reddy, P.V.N., Satya Prasad, K.: Multiwavelet Based Texture Features for Content Based Image Retrieval. IJCST 2(1) (March 2011) ISSN : 2229 - 4333 (Print) | ISSN : 0976-8491(Online)

[19] Jeong, S.: Histogram-Based Color Image Retrieval. Psych221/EE362 Project Report (March 15, 2001)

[20] Long, F., Zhang, H., Feng, D.D.: Fundamentals Of content-Based image Retrieval

[21] Naresh Babu, K., Pothalaiah, S., Ashok Babu, K.: Image Retieval Color, Shape And Texture Features Using Content Based. International Journal of Engineering Science and Technology 2(9), 4278–4287 (2010)

[22] Rao, B., Prabhakara Rao, B., Govardhan, A.: Content Based Image Retrieval using Dominant Colorand Texture features. (IJCSIS) International Journal of Computer Science and Information Security 9(2) (February 2011)

[23] Kharate, G.K., Patil, V.H., Bhale, N.L.: Selection of Mother Wavelet For Image Compression on Basis of Nature of Image. Journal of Multimedia 2(6) (November 2007)

[24] Khokher, A., Talwar, R.: Image Retrieval: A State Of The Art Approach For Cbir. International Journal Of Engineering Science And Technology (IJEST)

[25] Karthikeyani, V., Duraiswamy, K., Kamalakkannan, P.: Conversion of Gray-scale image to Color Image with and without Texture Synthesis. IJCSNS International Journal of Computer Science and Network Security 7(4) (April 2007)

[26] Suhasini, P.S., Sri Rama Krishna, K., Murali Krishna, I.V.: CBIR Using Color Histogram Processing. Journal of Theoretical and Applied Information Technology

[27] Lakshmi Devasena, C., Sumathi, T., Hemalatha, M.: An Experiential Survey on Image Mining Tools,Techniques and Applications. International Journal on Computer Science and Engineering (IJCSE)

[28] Lawrence Zitnick, C., Kanade, T.: Content-Free Image Retrieval (May 2003)

[29] Grosky, W.I.: Image Retrieval - Existing Techniques, Content-Based (Cbir) Systems, `http://encyclopedia.jrank.org/articles/pages/6763/Image-Retrieval.html`

[30] Zhao, R., Grosky, W.I.: PART II:Content-Based Retrieval And Image Database techniques, `http://www.cs.sunysb.edu/~rzhao/publications/SemanticGap.pdf`

SoC Modeling for Video Coding with Superscalar Projection

S.K. Fairooz[1] and B.K. Madhavi[2]

[1] Research Scholar, Jawaharlal Nehru Technological University Hyderabad, A.P., India
`fairoozsk@gmail.com`
[2] Prof. ECE Department, Geethanjali College of Engineering and Technology,
Hyderabad, A.P., India

Abstract. This paper presents a concept for a better quality of service in scalable video streaming services with an improved display scales. The available funds will be administered by multiple connections with feedback to support video streaming applications and for improving the visualization of imaging samples. The scaling factor is achieved by increasing the level of visualization of the display unit. In this paper, modular approach to SoC design and implementation of scalable video coding algorithm for digital video source in a low overhead SoC design proposed for the scaling operation forcibly source environment.

Keywords: Scalar coding, resolution imaging, streaming video, SoC Design, Cross-Layer Design.

1 Introduction

With the increasing demands of high speed data and multimedia applications in wireless communications, the IEEE 802.16e-family [1-2] and the associated global interoperability SoC designed and built to support broadband wireless access (BWA) in the wireless metropolitan Area Network (WMAN). For wireless data and multimedia services for fixed subscriber stations is provided by IEEE 802.16-2004 standard is available [1] and for mobile broadband wireless access (MBWA) is provided by the IEEE 802.16e-2005 standard [2].The proposed conventional methods were observed to develop tools and to keep in mind its limitations. The optimization scheme described above can significantly improve the coding, but in the current scenario and future applications of these methods may get restricted. As resources such as bandwidth, performance, programming techniques are limited to certain minimum values. A layered structure, such as scalable video coding is used for space, time and quality (SNR) provide scalability.

The transport [4-7] Bit stream is better suited than the non-scalable bit stream when video packets over an error-prone channel with different bandwidth transfer. The encryption techniques improved by optimizing resources. First the required models can also be set locally for a higher - the quality of HR reconstruction. The second characteristic is the quality of the algorithm's efficiency for a practical solution

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 787–795.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

for image enhancement. Finally, the proposed method great flexibility in various aspects of the solution. In this paper, we have a cross-layer architecture scalar interpolation system .Low complexity resolution enhancement method that can be implemented in next generation systems with FPGA.

2 Resource Adaptive Communication Model

The end-to-end transmission of a streaming video in the soc system is depicted in Figure 1. As shown, the last mile wireless system is IEEE 802.16e/soc which consists of base stations terminals. The streaming monitoring and soc subsystem are interconnected by an IP-based backhaul network. The streaming service in this study is encoded by the scalable extension of H.264/AVC.

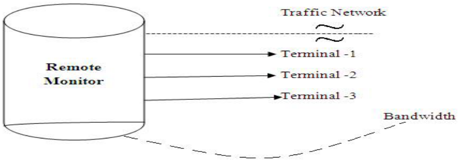


Fig. 1. End-to-end video streaming in SoC

In the proposed cross-layer design, the terminal will periodically report the average bandwidth to the streaming monitoring according to the residual terminal capacity and RF condition on the period of Report Period. The following subsections will describe the jobs of the streaming monitoring.

2.1 Streaming Controlling

In each of the Report Period, the terminal will request a target bit-rate from the traffic network (TN) streaming monitoring. The TN streaming monitoring then analyses the bitstream at the group of pictures (GOPs) that covers the current Report Period, as shown in Figure 2.

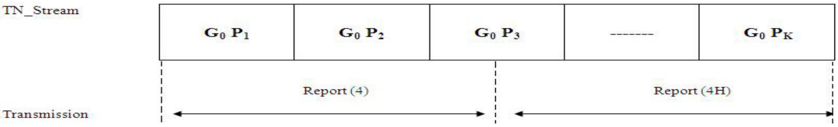


Fig. 2. Operational packet control in monitoring

In the TN bitstream, the data at lower spatial-SNR resolution is more important. And in each spatial-SNR layer, the lower temporal layer is more important. Therefore, according to the requested bit-rate from the terminal, the lower spatial-SNR layers are firstly extracted. At the spatial-SNR layer where the bit-rate cannot cover all the data, only the data at the lower temporal layers is extracted. With FGS (Fine Granular Scalability) coding, the data at the same temporal layers can be further truncated at any position to provide the exactly request bit-rate. The extracted data are then sent to the terminal.

It should be mentioned that some GOPs may belong to two Report Periods such as the GOP(x+2) in Figure 2. For such GOPs, the data that already sent in the first Report Period will not be sent again. However, if the second Report Period allows higher bandwidth, the remaining data in such GOPs will be transmitted. This makes the video quality smoother when the bandwidth changes frequently. Depending on the pre-load time of the related streaming service, the number of the overlapped GOPs between the Report Periods can be further extended to provide more smooth video quality. Further, this structure is also possible to enable the retransmission at the streaming monitoring with suitable pre-load time. To support multiple connections between the terminal and MS, the data sent to the terminal from the streaming monitoring is allocated into several connections according to the importance levels, as shown in Figure 3. The more important data is allocated to the connection that has more protection (i.e., higher transmission priority and MAC retransmission). To address the bandwidth fluctuation effect, in the proposed two-connection implementation, the monitoring allocates the more important data which occupies 80% of total data at the first connection, and put the remaining data at the second connection. This allows the terminal need only re-transmit the more important data when the actual transmission bandwidth is smaller than the expected bandwidth.

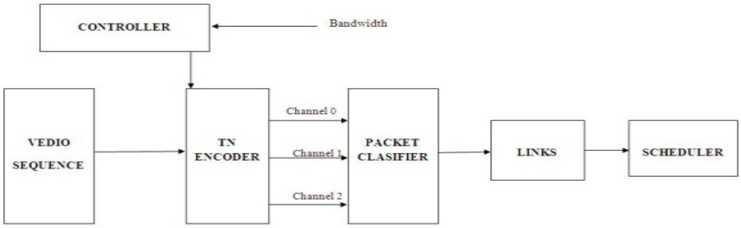


Fig. 3. Video streaming encoder on level selection

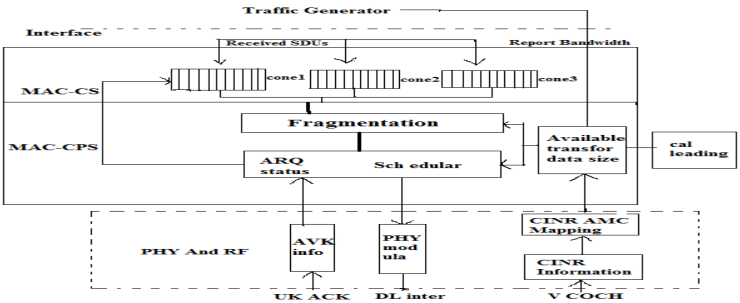


Fig. 4. Video MAC controller and decoder

2.2 Terminal Controlling

Upon receiving those packets in the soc subsystem, as shown in Figure 4, those packets will be first stored as MAC service data units (SDUs) in the queues. Then the MAC SDUs will be treated differently depending on the service types and MAC controls. The base station should collect the downlink (DL) channel condition and translate the information to relative available bandwidth. Besides, the base station will support multiple connections and have the capability of handling the connections by different means. The hardware architecture designed for the proposed algorithm is shown as the top level diagram in the figure below.

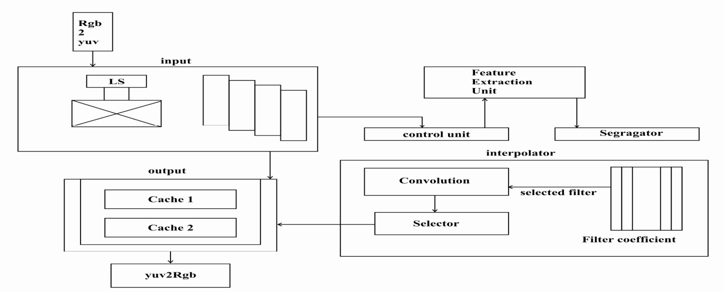


Fig. 5. Processing unit Block Diagram

In the above diagram there are different blocks, Modified RS algorithm is performed on the luminance (Y) path. Chrominance signal is interpolated by the pixel replication. At each input pixel arrival, the control unit reads a 4x1 pixel column from the Cache Buffer and provides the 3x3 and 5x5 pixel windows to he feature extraction and classification units. The feature extraction unit then extracts feature vector dimension of 8x1 which is then fed to the Segregator Unit for further processing. The purpose of Segregator is to perform distance calculation between prototypes of predetermined vectors and feature vector, the output is the index of the class having minimum distance of feature vector. Interpolator unit then uses this index to address filter coefficient LUT, selecting the appropriate filter for input pixel neighborhood. Constant coefficient multipliers are used form performing convolution and interpolated output pixels are then stored at output cache. Because of dynamic sample rate it is require having memories at the output to comply with the rate and order or data. The hardware implementation of the proposed scaling algorithm can be performed in two different ways:

Output - For each HD pixel coordinate, find the corresponding 5x5 low-resolution window, and perform feature extraction, classification, and interpolation on this neighborhood.

1. Input - For each SD pixel coordinate, find the corresponding four HD pixel coordinates. Since these HD pixels are to be interpolated from the same SD pixel neighborhood, perform feature extraction, and classification steps only for once.

Then perform interpolation for all four HD pixels corresponding to the SD pixel, and arrange the interpolated pixels in raster-scan order using cache buffers. Discard redundant pixels at the interpolation output if L_V or L_H (Scaling ratios, upper bounded by two) is a non-integer value.

Two main factors that can affect the implementation efficiency are the input/output data rate, and the target implementation platform. To provide an efficient implementation for different data rates, and different implementation platforms, the degree of resource sharing should be variable. The degree of resource sharing in feature extraction and classification units could be defined as:

D_r = Ne(# of elements in the feature vector)/N_p(# of processing paths in feature extraction) (1)

To achieve the desired data throughput with different resource sharing levels, core clock frequency, F_{clk_core}, must be related to the input pixel clock frequency, F_{Clk_in}, with the following formula: $F_{clk\_core} = [D_r]\, F_{Clk\_in}$ (2)

2.3 Operation of Control Unit

Control unit (CU) provides input data to the data-path blocks at appropriate timing and format The control unit:

1. Generates a sliding window to be used by FE and IN units, by shifting the 5 x 1 pixel column from the input memory unit into the 5 x 5 pixel window.

2. Generates the memory address and control signals for IM and OM blocks, and pipeline control signals for other blocks.

3 Generates the synchronization signals defined at the video standards (hsync, vsync, data enable).

4. Generates the Pv and PH values defined in equation below.

$$P_V = [Q_v\,(x_V/L_V - z_V) + \varepsilon\,] \tag{3}$$

$$P_H = [Q_H\,(x_H/L_H - z_H) + \varepsilon\,] \tag{4}$$

Where,
P_V =Vertical phase.
PH = Horizontal phase.
x_V = Vertical coordinates of high resolution pixel.
x_H = Horizontal coordinates of high resolution pixel.
ε = Very small number like 10^{-6}.
z_V = Vertical coordinates of low resolution pixel.
z_H = Horizontal coordinates of low resolution pixel.

5. For scaling ratios less than two, selects the pixels to be omitted using equation shown below.

$$z_V = [x_V/L_V]\ z_H = [x_H/L_H] \tag{5}$$

2.4 Operation of Input Memory

Input memory (IM) unit operates at input pixel clock frequency F_{clk_in}. The block consists of four cache buffers, to provide a 5 x 1 pixel column to the CU. The length of the line buffers is equal to the input video's horizontal size. Shown below is the block diagram of the Input Unit.

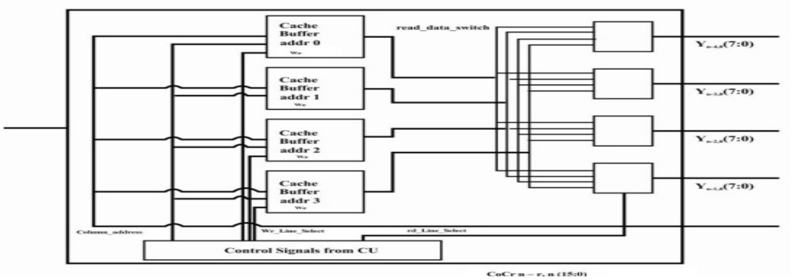


Fig. 6. Input Unit

2.5 Operation of Segregator

Segregator unit operates at core clock frequency CU, and generates four high resolution pixels using the selected F_{clk_core}. figure below shows the architecture of the feature extraction and context classification units. Parallel implementation where Dr = 1, will use 15 adders, 32 multipliers, and a serial-parallel implementation, where Dr= 4 reduces the number of adders/subtractors to four and the multipliers to eight, with a negligible increase in the number of pipeline registers.

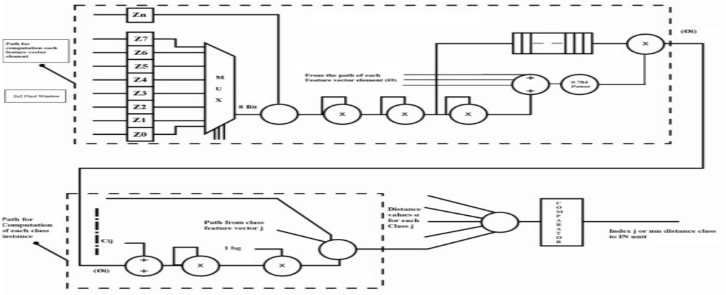

Zn
Z7
Z6
Z5
Z4
Z3
Z2
Z1
Z0
Path for computation each feature vector element
3x3 Pixel Window
M U X
8 Bit
From the path of each Feature vector element (Oi)
0.784 Power
(Oi)
Path for Computation of each class instance
Cij
Path from class feature vector j
1 bg
(Oi)
Distance values α for each Class j
COMPARATOR
Index j or mn distance class to IN unit

Fig. 7. Context Classification and Feature Extraction for scalar projection

2.6 Operation of Interpolation

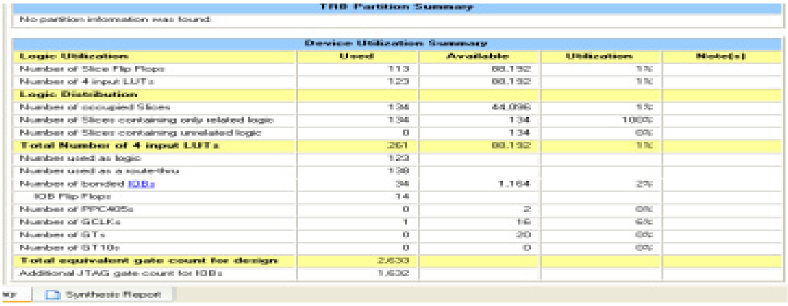
Interpolation unit (IN) unit operates at core clock frequency. 100 multiplications, and 96 additions required to perform 5 x 5 convolution for four HD pixels is resource shared with D_r= 4 to reduce the number of multipliers to 25, adders to 24. Therefore at each F_{clk_core} clock cycle, one convolution operation is performed, generating four HD pixels at every F_{clk-in}, clock cycle.

2.7 Output Unit

Output memory (OM) unit also operates at core clock frequency. The host machine which needs to communicate with the processor sends its data to the FIFO of the transmitter device and depending on the status of the control signal read or write operation is performed is performed and then the data is forwarded to the CRC generator where the FCS is generated and also to the frame builder where the frame is build and then the parallel data is converted to serial for transmission and send over the network. The data is transmitted over the network and at the receiving side the data is reconverted into parallel form and this data is fed into the FIFO and send to the frame reader where the is decomposed into different fields.

3 Simulation observations

For the functional evaluation of the designed system the developed system is designed using VHDL definition and simulated on the Active-HDL simulator for it's operational verification. For the testing of the designed system a test bench file is generated where two frame data is passed. This frame data is read by the video_codec file and divide into 8 X 8 Block. The block data is processed for noise estimation, motion estimation, and compression. Under recovery the data is decoded back for it's regeneration. For the real time Realization and resource utilization the developed design is synthesized using Xilinx synthesizer.



TRB Partition Summary

No partition information was found.

Device Utilization Summary

Logic Utilization	**Used**	**Available**	**Utilization**	**Note(s)**
Number of Slice Flip Flops	113	88,192	1%	
Number of 4 input LUTs	123	88,192	1%	
Logic Distribution				
Number of occupied Slices	134	44,096	1%	
Number of Slices containing only related logic	134	134	100%	
Number of Slices containing unrelated logic	0	134	0%	
Total Number of 4 input LUTs	261	88,192	1%	
Number used as logic	123			
Number used as a route-thru	138			
Number of bonded IOBs	34	1,164	2%	
IOB Flip Flops	14			
Number of PPC405s	0	2	0%	
Number of GCLKs	1	16	6%	
Number of GTs	0	20	0%	
Number of GT10s	0	0	0%	
Total equivalent gate count for design	2,633			
Additional JTAG gate count for IOBs	1,632			

Fig. 8. Summarized synthesis report for the developed estimation system

dlc=3bytes
Best master clock (BMC)= min (t_{stamp})
Default clock = ck1 (under asynchronous mode communication)
Frequency selection method = round robin

Total number of communicating nodes = 4
Total amount of data generated per node = 3 bytes
Total amount of expected data in processor = 12 bytes
Total time taken = 5580 ns (under non synchronous round robin based comm)
(Total time = processing time + comm Time)
Total time taken = 1445 ns (under 1588 synchronous mode comm)
Total time saved (ts)= 4135 ns
Total clock cycles saved = ts / BMC =4135 / 10 ≈ 413 cycles

The implemented of the video interface unit developed on VHDL and implemented onto the targeted FPGA (xcv300-bg432-6) for the real time realization. The observations of multiple frameset are carried out with the interface of a test bench and

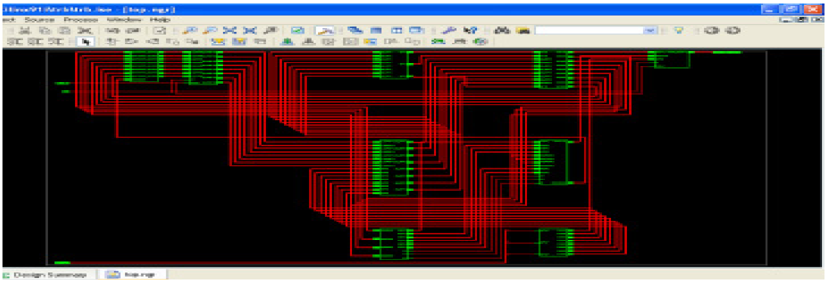


Fig. 9. RTL view of the implemented system using Xilinx synthesizer

are interfaced with Mat lab tool to observe the results. The original frame sequence is taken at a very low resolution with pixel representation of 150x250 size frame. These 5 frame sequences are passed to the developed system for pre-processing and the results obtained is shown below.

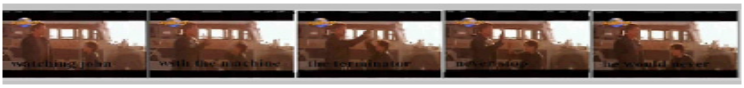


Fig. 10. Original image sequence considered

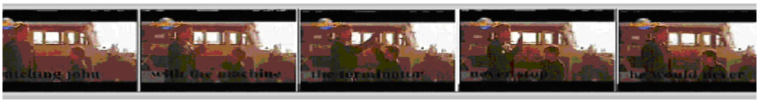


Fig. 11. Scaled image sequences at 1:2.5 ratio

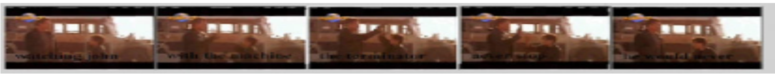


Fig. 12. Scaled image sequences at 1:2.5 ratio

The observation clearly illustrates the accuracy in retrieval in terms of visual quality as compared to the conventional Fourier based coding technique.

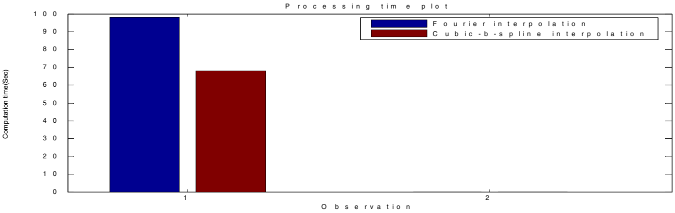


Fig. 13. Computation time taken for the two methods

The system developed is also evaluated for the computation time taken for the computation and projection of the frame sequence for interpolation. The total time taken for reading, processing and projecting is considered for the processing system.

4 Conclusions

In this paper, implementation of a dynamically reconfigurable video codec of a general structure by the concept of virtual codec and virtual tools is presented. The flexible structure of the proposed method provide a solution for the implementation of multiple profile MPEG-4 coding standard and hardware implementation for a classification based resolution enhancement method has not been presented previously. Proposed codec is in the reconfigurable codec structure mode, it is able to achieve better results than the standard approach for certain type of applications. By eliminating the need for soft interpolation and reducing the number of context classes. The computational complexity of resolution synthesis with negligible visual quality loss is lowered. The modified RS algorithm is simple enough to be implemented on low cost FPGA boards.

References

[1] Schoner, B., Villasenor, J., Molloy, S., Jain, R.: Techniques for FPGA Implementation of Video Compression Systems, Department of Electrical Engineering, University of California, Los Angeles, California

[2] Kulmala, A., Lehtoranta, O., Hamalainen, T.D., Ainen, M.H.: ScalableMPEG-4 Encoder on FPGAMultiprocessor SOC. Department of Information Technology, Institute of Digital and Computer Systems, Tamper university of Technology, P.O. Box 553, Korkeakoulunkatu 1, 33101 Tampere, Finland

[3] Kim, H., Altunbasak, Y.: Low-Complexity Macroblock Mode Selection For H.264/Avc Encoders. Center for Signal and Image Processing Georgia Institute of Technology, Atlanta

[4] Tian, D., Li, X., Al-Regib, G., Altunbasak, Y., Jackson, J.R.: Optimal Packet Scheduling for Wireless Video Streaming with Error-Prone Feedback. School of Electrical and Computer Engineering,Georgia Institute of Technology, Atlanta

[5] Cai, H., Zeng, B., Shen, G., Xiong, Z., Li, S.: Error-Resilient Unequal Error Protection of Fine Granularity Scalable Video Bitstreams, Microsoft Research Asia, Beijing

[6] Argyriou, A., Member: Error-Resilient Video Encoding and Transmission in Multi-Rate Wireless LANs. IEEE Transactions on Multimedia (August 2008)

[7] Cho, S., Pearlman, W.A.: Error Resilient Compression and Transmission of Scalable Video. Center for Digital Video and Media Research, Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, New York

[8] van Heesch, F., Klompenhouwer, M., de Haan, G.: Masking noise in upscaled video on large displays. Philips research laboratories, Eindhoven, The Netherlands

[9] Martinian, E., Behrens, A., Xin, J., Vetro, A.: View Synthesis for Multiview Video Compression, Mitsubishi Electric Research Labs, Broadway, Cambridge

[10] Chang, S.-F., Fellow IEEE, Vetro, A., Senior Member: Video Adaptation: Concepts, Technologies, and Open Issues. Proc of IEEE 93(1), 148–158 (2005)

Invisible Image Watermarking Using Z Transforms (IIWZT)

J.K. Mandal and Rakesh Kanji

Department of Computer Science and Engineering,
University of Kalyani,
Kalyani, Nadia-741235, West Bengal, India
`{jkm.cse,r.kanji.it55}@gmail.com`

Abstract. This paper presents an invisible image watermarking technique in frequency domain through Z transform, with a hiding capacity of 1.5 bpB (1.5 bits per byte). $Z(re^{j\omega})$ is a complex variable comes from Laplace transformation has two parameters, r denotes radius of Region of convergence and ω denotes angle respectively. In this technique (2 *2) sub matrix is taken from source image and converted into 1-diamensional (1*4) array which undergoes Z-transform with a is set of angular frequency(ω). 2 bits of secret message is embedded including the complex conjugate pair where multiple embedding is done. First co efficient is used for tuning purpose and not embedding for information. This process is repeated until exhaustion of secret image. Inverse Z transform is done at end to convert the image from frequency domain to spatial domain. Experimental result shows good PSNR and image fidelity which analytically suggest that the proposed scheme obtains better secrecy with improved fidelity.

Keywords: Frequency Domain Steganography, Invisible Watermark, peak signal to noise ratio (PSNR), mean square error (MSE) Z Transforms (ZT).

1 Introduction

Watermarking is a technique refers to embedding an authenticating object as a signature into an original object. A signature could be visible or invisible and object could be sound, image, video etc. Visible watermarking[1] refers to process of embedding signature into original object such that it can be seen and that of invisible watermarking refers to process that signature can't be seen[2,3,4,5].

This era of network world needs protection and security. Various sectors like military, bureaucrats, research organization, media deal with confidential information which demands high security. To overcome this kind of problem watermarking is very useful solution[6,7].

The proposed scheme refers to invisible image[8,9,10,11,12] watermarking which embeds bits of authenticating image as signature into frequency domain of the image. In 1947 the concept of Z is reintroduced by W. Haurewicz for solving linear,

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 797–804.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

constant coefficient difference equations. Later in 1952, Zedeh remodels this into sample data control group at Columbia University.

Z is a complex variable, converts discrete time domain signal to corresponding frequency domain. Z is polar representation having real and imaginary plane correspond to x axis and y axis respectively, means all Z points value is calculated by r(radius which is vector starting at origin) and ω (angle between radius vector and real plane) .

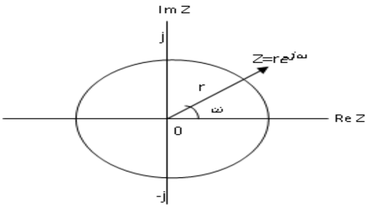


Fig. 1. The point Z=r$e^{j\omega}$ in complex Z plane

Analytically, Z=Re(Z)+Im(Z). It is seen that Fourier transform is the function F($e^{j\omega}$)which propagates with imaginary plane[8]. Fourier transform function assumes every function is converged within ROC=1, geometrically a circle having radius r=1,with a finite energy and stable as well. Z domain actually finds out the stable area in terms of circle of unstable signal or function that doesn't converge within r=1 or converge within a range of ROC, is seen when convolution of two different functions must be linear time invariant have different ROCs priory.

Z is a modern tool of digital signal processing could be designing any unstable signal's stability area in terms of region of convergence (ROC).

Forward Z transform expression in one dimension is given in equation 1,

$$F(Z)=\sum_{n=-\infty}^{\infty} r^{-n}\ g[n]\ e^{-jwn} \qquad (1)$$

g[n]-> is array having samples value taken in time domain. M is number of samples taken with equal time interval .n=0,1,2,3,4,5,……………….M-1

$$F(Z)=\sum_{n=0}^{M-1} r^{-n}\ g[n]\ e^{-jwn}$$

ω = anguler frquency (rad / second)

Inverse Z transform expression is given in equation 2,

$$g[n]=\frac{1}{2\pi j}\oint F(z)\ z^{n-1}dz \qquad (2)$$

Now, $z=re^{j\omega}$,So, $dz=jre^{j\omega}d\ \omega$

If we assume contour integration is in anticlock wise then ω will vary from –π to π is converted into definite integral of 2 π.However the ω could have varied -2 π to 0.

$$\text{So, } g[n]=\frac{1}{2\pi}\int_{-\pi}^{\pi} F(re^{j\omega})\ r^{n}e^{jwn}d\ \omega \qquad (3)$$

Equation (3) can be equivalently expressed in discrete representation as given in equation (4),

$$g[n]=\frac{r^n}{M}\sum_{k=0}^{M-1} F(z)\ e^{jwk}$$

2 The Technique

The scheme embeds data with r=1. The image pixel of secret messages is taken as row major order converting into binary (quantized level 0 to 7) and putting those bits into 1–dimensional array whose size equals to $8 \times M \times N$ is termed as secret array(M->row of secret image and N->column of secret image).

This technique takes (2×2) sub-cover image (source), transformed into coefficient values with a set of angular frequency $(\omega = [0,\frac{\pi}{2},\pi,\frac{3\pi}{2}])$. 2-bits from secret array is embedded into each coefficient of the transformed mask except first one. First coefficient of each mask in Z transform is kept for adjustment as there may be a possibility to generate negative values during inverse Z transform. This adjustment process is called tuning. The technique of insertion is discussed in section A that of extraction in B.

A. Insertion technique

The secret array(one dimensional) is taken in row major order, having 2×2 sub matrix taken each time from cover image ($M_1 \times N_1$) and transformed into one dimensional (1×4) matrix. Forward Z transform is applied with a set of angular frequency $(\omega = [0,\frac{\pi}{2},\pi,\frac{3\pi}{2}])$ and r=1.This gives four amplitudes, values corresponding to four angular frequencies, two real value out of which one is DC(frequency (ω)=0) value resieds into first position and two complex value makes jointly a conjugate pair. Excluding Dc value, real values are converted into binary form having quantized level 11(as highest value should be =1020) where 2 bits are embedded into second and third bit from LSB. The conjugate pair[(a-jb) and (a+jb)] is embedded in such a way that the net effect is nullified for proper embedding. To achieve this same secret bits are embedded at first and second bit position having quantized level 11. Embedded stego-frequency matrix is converted back into spatial domain through inverse Z transform. If any frequency coefficient value is found 0 on embedding it is converted as negative which is preferred than positive for better tuning. As 2×2 matrix is taken, number of coefficient is 4. For real value when data bits are getting embedded it must be the position which is multiple of sample number is 4(decimal). As inverse Z transform formula is rational number where denominator value is 4. Again two complex numbers jointly make conjugate pair its behavior should be taken together. So, identical bits are embedded in both conjugate components.

After embedding the mask is transferred into spatial domain. While transforming this into spatial domain two cases may arise.

Case-1: Algorithm will find out minimum value in (2×2) array, if any neative value is found then mask is taken back to frequency domain, the Dc value corresponding to first block of frequency matrix will be added with mod of that value multiplied by 4.

Case-2: Algorithm will find out maximum value. If it exceeds 255 and no negative value exists, then maximum value is subtracted from 255 and result is multiplied by 4. Taking mod of that, it is again subtracted from DC value. Now, the technique again finds out that this tuning resulted in giving any negative value, if so, inverse tuning is done as the final value which was subtracted from DC; will be added to DC results in restoring the previous values.

Insertion algorithm

Input: An $M_1 \times N_1$ source image and an $M \times N$ hidden image.
Output: An $M_1 \times N_1$ stego image.
Step1. Find out the size of hidden image in 8 bit representation.
Step2.Take a non overlapping window of 2×2 and make it 1×4 then apply forward Z transforms as Sliding manner on row major order of source image matrix .
Step3. Excluding 1st real coefficient in each window, embedding is done into other frequency coefficients for each bit of hidden image.
Step4. Inverse Z transform is done.
Step5. Tuning.
Step6. Repeated 1 to 5 steps till end of the hidden bits.
Step7. Stop.

B. Extraction technique

Extracting is done in reverse manner. The received stego image($M_1 \times N_1$) is transformed into one diamensional (1×4) form, applying forward Z transform it converted to frequency matrix from where the secret bits are taken out and is converted into binary form having quantized level 11, data being taken out from second and third position. From conjugate pair any one of complex number is converted into binary form having quantized level 11 and embedded information taken out from first and second position. This process is repeated until end of secret data.

Extraction algorithm

Input: An $M_1 \times N_1$ stego image.
Output: An $M \times N$ hidden image.
Step1. Take a non overlapping window of 2×2 and make it 1×4 then apply forward Z transform in sliding manner on row major order of source image matrix.
Step2. Excluding 1st real part in each window, extract the hidden bits from frequency coefficients.
Step3. Repeated 1 to 3 steps till all the hidden bits is extracted.
Step 4. Stop.

3 Schematic Diagram of IIWZT

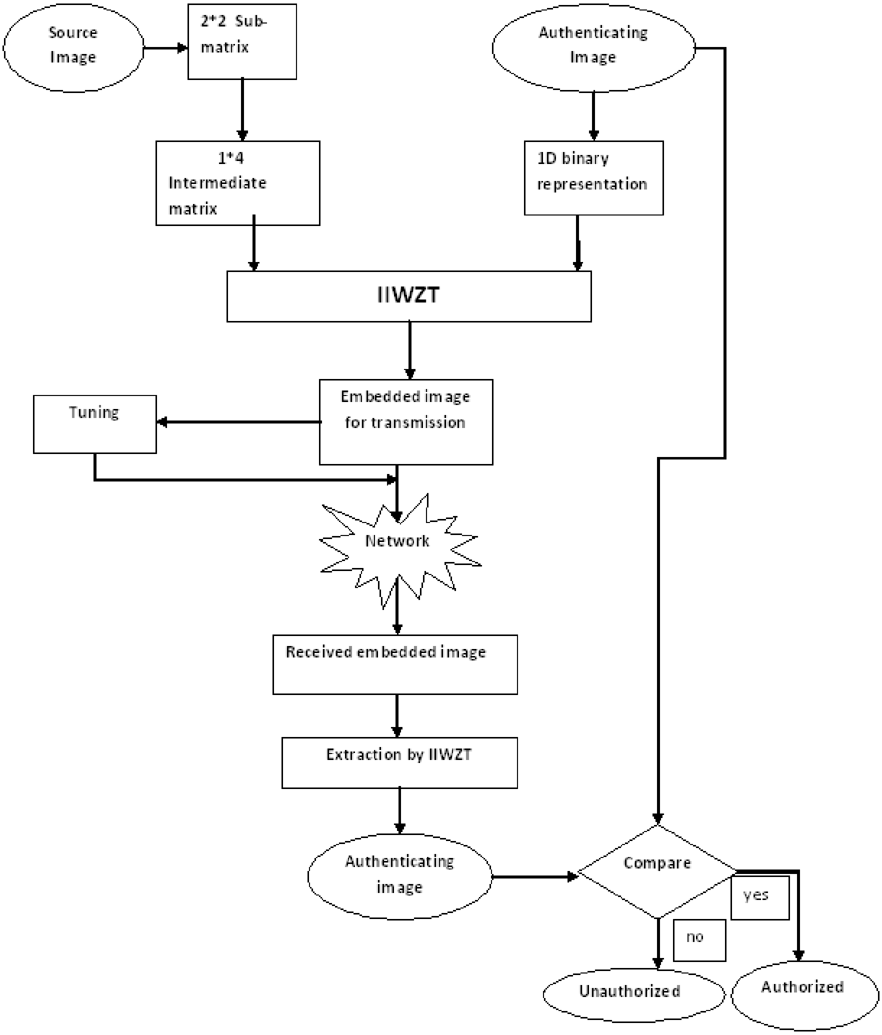


Fig. 2. Schematic diagram of IIWZT

4 Example

This process is described using a (2×2) matrix $\begin{bmatrix} 100 & 200 \\ 150 & 250 \end{bmatrix}$ with secret message stream="110010".

Step 1: The matrix is converted given matrix into (1×4)matrix ,[100 200 150 250]
Step2: Using Z transform, frequency matrix is obtained, $Z(2 \times 2) = \begin{bmatrix} 700 & -50-j50 \\ -200 & -50+j50 \end{bmatrix}$, it is achieved by transform 1D frequency matrix is converted into 2 cross 2 matrix.
Step 3: Embedding secret message stream=”110010”, gives stego frequency matrix as $Z(2 \times 2) = \begin{bmatrix} 700 & -48-j54 \\ -204 & -48+j54 \end{bmatrix}$.
Step 4: Inverse Z is applied, gives spatial matrix $\begin{bmatrix} 100 & 199 \\ 148 & 253 \end{bmatrix}$.
Step5: Forward Z transform is done and Taking out the secret bits.

5 Result, Analysis and Comparison

Analysis has been made on various images[13] using IIWZD technique given in table 1. Results are discussed in terms of visual interpretation, image fidelity, mean square error and peak signal to noise ratio(dB). Such cover image has dimensions (512×512). Each 2×2 non overlapping blocks is taken and goes embedding with 1.5bpB. Figure 3a shows the host images Aerial, Couple. Figure 3b shows embedded Aerial, Couple on embedding Coin image using IIWZD. Figure 3c is the authenticating image Coin(221× 221). From the table it is seen that the maximum value of the PSNR is 40.7516 and that of minimum value of the PSNR is 39.1764. The value of the PSNR is consistent for various images. The average PSNR for eight images is 40.448. The following formulas are used to calculate PSNR, MSE, IF.

$$\mathrm{MSE} = \frac{1}{\mathrm{MN}} * \sum_{\mathrm{m,n}} (\mathrm{I_1\ m,n} - \mathrm{I\ _2 m,n})^2$$

$$PSNR = 10\ \log\left(max\left(I_{m,n}{}^2\right)/MSE\right)$$

$$IF = 1 - \sum_{m,n} (I_{1_{m,n}} - I_{2_{m,n}})^2 / \sum_{m,n} I_{2_{m,n}}^2$$

Table 1. Results obtained for various Benchmark Images in IIWZD

Cover image (512× 512)	Mean square error (MSE)	Peak signal to noise ratio (PSNR)	Image fidelity(IF)
Aerial	5.7464	40.5368	0.9997
Elaine	5.5300	40.7035	0.9997
Stream and bridge	5.5254	40.7071	0.9996
Boat.512	5.4278	40.7845	0.9997
Couple	5.9770	40.3659	0.9996
Clock	7.8112	39.2035	0.9997
Pepper	5.5510	40.6870	0.9996
Tank	5.4983	40.7285	0.9995
Lena	5.6349	40.6218	0.9996
House	6.2850	40.1476	0.9997

Table 2. Comparison results obtained from RHDCTI and IIWZD

Capacity(bytes)	PSNR(dB)	METHOD
4096	38.71	RHDCTI-L_1
12288	32.97	RHDCTI-L_3
32732	28.20	RHDCTI-L_8
36721	27.60	RHDCTI-L_9
49152	40.45	IIWZT

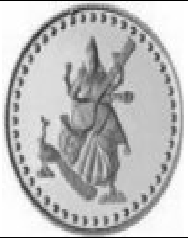
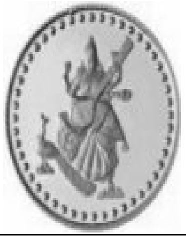
3.a.i cover image	3.b.i embedded image
3.a.ii cover image	3.b.ii embedded image
	
3.a.iii authenticating image	3.a.iii authenticating image

Fig. 3. Visual effect of embedding in IIWZD

Comparison of proposed technique has also been made with Chin-Chen et al method. Proposed method achieved higher average PSNR with almost identical or more hiding capacity in compared to RHDCTI[7] as shown in table 2.

6 Conclusion

The technique used Z transformation which provides extra parameter r(radius of region of convergence) for better security. Moreover standard PSNR and image fidelity analytically governs less distortion of cover image is likely to forgo intruder's smart eyes. Finally tuning value suggests there is scope for improving capacity of hidden image with good PSNR, which is further research in Z domain.

Acknowledgments. The authors express gratitude to the Department of Computer Science and Engineering, University of Kalyani and the PURSE scheme of DST, Govt. of India, under which the research has been carried out.

References

1. Yu, L., et al.: Improved Adaptive LSB steganography based on Chaos and Genetic Algorithm. EURASIP Journal on Advances in Signal Processing 2010, Article ID 876946 (2010)
2. Ghoshal, N., Mandal, J.K., et al.: Image Authentication by Hiding Large Volume of Data and Secure Message Transmission Technique using Mask (IAHLVDSMTTM). In: Proceedings of IEEE International Advanced Computing Conference IACC 2009, March 6-7. Thapar University, Patiala (2009) ISBN:978-981-08-2465-5
3. Ghoshal, N., Mandal, J.K., et al.: Masking based Data Hiding and Image Authentication Technique (MDHIAT). In: Proceedings of 16th International Conference of IEEE on Advanced Computing and Communications ADCOM 2008, Anna University, December 14-17 (2008) ISBN: 978-1-4244-2962-2
4. Ghoshal, N., Mandal, J.K.: A Novel Technique for Image Authentication in Frequency Domain using Discrete Fourier Transformation Technique (IAFDZTT). Malaysian Journal of Computer Science 21(1), 24–32 (2008) ISSN 0127-9094
5. Ghoshal, N., Mandal, J.K.: A Bit Level Image Authentication/Secrete Message Transmission Technique (BLIA/SMTT). Association for the Advancement of Modeling and Simulation Technique in Enterprises (AMSE), AMSE Journal of Signal Processing and Pattern Recognition 51(4), 1–13 (2008)
6. EL-Emam, N.N.: Hiding a large Amount of data with High Security Using Steganography Algorithm. Journal of Computer Science 3(4), 223–232 (2007) ISSN 1549-3636
7. Chang, C.-C., Lin, C.-C., Tseng, C.-S., Tai, W.-L.: Reverssible hiding in DCT based Compressed image (received July 19, 2005), `http://www.scincedirect.com` (received in revised form February 15, 2007; accepted February 15, 2007)
8. Mitra, S.K.: Digital Signal Processing, 3rd edn. TATA McGraw-Hill,
9. Dumitrescu, S., Xiaolin, W., Wang, Z.: Detection of LSB steganography via sample pair analysis. IEEE Trans. on Signal Processing 51(7), 1995–2007 (2003)
10. Chandramouli, R., Memon, N.: Analysis of LSB based image steganography techniques. In: Proc. of ICIP, Thissaloniki, Greece, pp. 1019–1022 (2001)
11. Wang, R.-Z., Lib, C.-F., Lin, J.-C.: Image hiding by optimal LSB substitution and Genetic algorithm. In: 2001 Pattern Recognition Society. Elsevier Science Ltd (2001)
12. Lin, C.Y., Chang, S.F.: A robust image authentication method surviving JPEG lossy compression. In: Proc. SPIE, San Jose, vol. 3312, pp. 296–307 (January 1998)
13. Weber, A.G.: The USC-SIPI Image Database: Version 5, Original release: October 1997, Signal and Image Processing Institute, University of Southern California, Department of Electrical Engineering, `http://sipi.usc.edu/database/` (last accessed on January 20, 2011)

Region Identification of Infected Rice Images Using the Concept of Fermi Energy

Santanu Phadikar[1], Jaya Sil[2], and Asit KumarDas[2]

[1] Department of Computer Science and Engineering,
West Bengal University of Technology, BF-142, Sector –I, Salt Lake, Kolkata 64
[2] Department of Computer Science and Technology,
Bengal Engineering and Science University, Shibpur, Howrah, India

Abstract. Automated disease detection using the features of infected regions of a diseased plant image is a growing field of research in precision agriculture. Usually, infected regions are identified by applying different threshold based segmentation techniques. However, due to various factors like non-uniform illumination or noises, these techniques fail to provide sufficient information for classifying diseases accurately. In the paper, a novel region identification method based on Fermi energy has been proposed to detect the infected portion of the diseased rice images. From the infected region, neighboring gray level dependence matrix (NGLDM) based texture features are extracted to classify different diseases of rice plants. Performance of the proposed method has been evaluated by comparing classification accuracy with other segmentation algorithms, demonstrating superior result.

Keywords: Fermi Energy, Disease Classification, Region Detection, Feature Extraction.

1 Introduction

Precision Farming is defined as information technology based farm management system to identify, analyze and manage variability within the fields for optimum profitability, sustainability and protection of land resources [1]. Precision farming aims at better decision making considering many aspect of crops, one of them is automated disease detection.

Rice is second largest crop produce in the World and the first largest crop in India. Therefore, accurate and timely diagnosis of rice plant diseases is absolutely necessary and may play significant role in country's economical growth. One of the crucial tasks is to detect the area of infection from the diseased plant images, on which the success of the automated system depends. Though a large number of segmentation techniques [3-6] are available for identifying the region of interest, no one can be used for all the applications due to great complexity for structuring visual information into meaningful regions. Thresholding is the simplest form of segmentation, or more general a two class clustering problem. Extensive works have already been carried out to introduce new and more robust thresholding techniques [4-12]. However, in most of the cases the threshold value is determined experimentally that depends on the quality

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 805–811.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

of the images. The energy function based segmentation methods detect boundaries in the images, as described by Kass [13]. The basic idea of the model is to locate sharp variations of the image intensity by minimizing energy functional. Caselle et. al. in their work [14] proposed a new intrinsic energy functional which uses information regarding the boundaries of the objects. Another energy based method uses energy functional to capture an object that exhibits high image gradients and a shape, compatible with the statistical model, best fit the segmented object [15]. Recently, an edge sensitive variational image thresholding method [2] explored locally adaptive image threshold technique by minimizing the variational energy functional to detect the binary images. However, all the existing energy based models are computationally expensive due to execution of different steps involved in the method.

In the paper, a novel method based on Fermi energy [16-17] has been proposed to identify the infected regions of the diseased rice plant images. Fermi energy is the highest possible energy of a particle in absolute temperature zero. Fermi energy (E_F) of an image has been considered as a threshold value to perform the segmentation. This method is computationally efficient because energy of different pixels and comparison with threshold are performed simultaneously. Once the infected portion is isolated, neighboring gray level dependence matrix (NGLDM) [18] based texture features are extracted. Information about homogeneity with respect to the distribution of entries have been calculated from the infected regions and used for classifying diseases. To compare the performance of the proposed method, infected regions are detected by applying the existing methods[19-20] and features are extracted to classify the diseases using different classifiers available in Weka Tool [21].

The paper is organized as follows: in section-2 Fermi energy based region extraction process has been discussed. Section 3 explains feature extraction process and, experimental results are given in section 4. Section 5 concludes the paper.

2 Fermi Energy Based Region Detection

The Fermi energy of a substance is attained when the temperature of a material is lowered to absolute zero [16-17], a simple concept developed based on Pauli Exclusion Principle. By this principle, only one electron can inhabit a given energy state. When the temperature is lowered to absolute zero, all electrons in the solid arrange themselves so that the total energy becomes minimum. As a result, they form a sea of energy states known as the Fermi Sea. The highest energy level of this sea is called the Fermi energy or Fermi level. At absolute zero no electrons have enough energy to occupy an energy level higher than the Fermi level.

Assume a three-dimensional cubical box, shown in Fig. 1 with side length *L*, an approximation for describing electrons in a metal. The states are now labeled by three quantum numbers n_x, n_y, and n_z. Calculation of energy of single particle is given in equation (1).

$$E_{n_x,n_y,n_z} = \frac{\hbar^2\pi^2}{2mL^2}\left(n_x^2 + n_y^2 + n_z^2\right) \tag{1}$$

Where n_x, n_y, n_z are positive integers, m is the mass of electron and $\hbar$ is plank's constant.

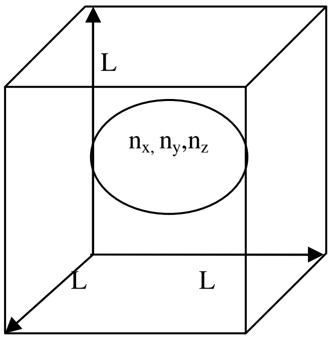


Fig. 1. Electron in a cube of length *L*

There are multiple states with the same energy, for example $E_{211} = E_{121} = E_{112}$. The ground state (zero temperature) for the electron corresponds to a distribution where exactly one electron in each energy state is placed, starting from the bottom, until all the electrons are accounted. In that state maximum energy state becomes the Fermi energy of object and it is represented by equation (2).

$$E_F = \frac{\hbar^2 \pi^2}{2mL^2} n_F^2 = \frac{\hbar^2 \pi^2}{2mL^2} \left(\frac{3N}{\pi}\right)^{2/3} \tag{2}$$

To compute the Fermi energy of an image, it is required to map the image corresponding to the absolute temperature zero. Therefore, it is assumed that the image will be in the ground state (absolute temperature zero) considering only one pixel with a specific colour value. Using equation (2) the Fermi energy is calculated where the values of '*N*', the number of particles in the system, the length '*L*' of the cube and mass '*m*' of a particle are to be known. Number of distinct color value in the image has been considered as number of particles '*N*'. Since any pixel can take a value between (0,0,0) to (255,255,255), the value of '*L*' is 256. Though, '*m*' is constant in case of electron, here mass '*m*' of a particular pixel is calculated by measuring probability of its presence in the image and its intensity value intensity value in the RGB plane using equation (3).

$$m_{r,g,b} = \frac{H_{r,g,b}}{p \times q}(r + g + b) \tag{3}$$

Where $H_{r,g,b}$ is the number of pixel with the RGB values *r, g, b* respectively and $p \times q$ is the size of image. Mass of the image '*m*' is the average mass of the pixels in the image, and used to calculate Fermi energy of image in equation (2)

Once the Fermi energy of the image is computed, it is treated as threshold value. Now the single pixel energy $\boldsymbol{E_n}$ is calculated using equation (1) where value of n_x, n_y, *and* n_z are corresponding to *R, G* and *B* value of the pixel respectively. If the value of $\boldsymbol{E_n} >= \boldsymbol{E_F}$ then it is treated as infected portion. The procedure is described below.

```
Algorithm: segment (image $IM$)
Input: A color image $IM$ of size $p \times q \times 3$
Output: Binary image $BW$ of size $p \times q$
Begin
  Initialize the matrix $H$ (256 ×256×256) to zero
  For i=1 to p
    For j=1 to q
          Increase the count of $H(r, g, b)$ by one where r, g and b represent the
          Red, Green and Blue component of the pixel of $IM(i,j)$
     End /* for loop j */
  End /* for loop i*/
  N= Number of nonzero elements in $H$ matrix.
  $m_{r,g,b}$ is computed using equation 3 for each color in the image
  Determine the mass of the image 'm' taking average of all $m_{r,g,b}$ .
  Compute the Fermi energy $E_F$ using equation (2).
  For i=1 to p
    For j=1 to q
      Compute $E_{r,g,b}$ of pixel $IM(i, j)$ with color values r, g, b using equation (1)
      If ( $E_F < E_{r,g,b}$) then
        BW(i, j)=1;
      Else
        BW (i, j) =0;
      End
    End
  End
  Return(BW)
End
```

3 Feature Extraction Process

Once the infected region is detected, features from the infected portion are extracted to identify the diseases. Shape and orientation of colours (textures) [22-24] in the infected portion may be used as features to identify the diseases. Here neighbouring gray level dependence matrix based texture features have been used for classification. This method computes textural information of images by evaluating gray level dependence between pixel and its adjacent pixels. In this texture feature extraction method, a matrix $C[C_I \times C_D+1]$ is built based on two user-given parameters d and T. C_I is the range of gray level of an image while C_D is the maximum number of neighbouring pixels with respect to another pixel within the block having distance d [18]. Each component of the matrix is represented by $C(k,l)$, where k is the gray value of pixel (x, y) while l is the number of neighbourhood pixels satisfying the following two criteria. First, this neighbouring pixel is within the block distance d from pixel (x, y), secondly the difference of gray value between the neighbouring pixel and pixel(x,y) is less than T, set as 10. Afterwards, statistical approaches are applied on the matrix to extract the following texture features.

a) Small number emphasis (SNE), measuring roughness of an image:

$$SNE = \frac{\sum_k \sum_l \frac{C(k,l)}{l^2}}{\sum_k \sum_l C(k,l)}$$

b) Large number emphasis (LNE), giving measurement of smoothness:

$$LNE = \frac{\sum_k \sum_l l^2 C(k,l)}{\sum_k \sum_l C(k,l)}$$

c) Second moment (SM), indicating homogeneity with respect to the entries contained in the matrix:

$$SM = \frac{\sum_k \sum_l C^2(k,l)}{\sum_k \sum_l C(k,l)}$$

d) Number of non-uniformity (NNU) :

$$NNU = \frac{\sum_l (\sum_k C(k,l))^2}{\sum_k \sum_l C(k,l)}$$

e) Entropy of the matrix (EM):

$$EM = \frac{\sum_k \sum_l C(k,l) \log C(k,l)}{\sum_k \sum_l C(k,l)}$$

4 Result and Discussion

The proposed method has been developed to detect the infected portion of the image from a diseased rice plant images. Three hundred sample images, 100 from each diseased class are acquired from the rice field of East Midnapur district of West Bengal, one of the major rice growing states in India. Images infected by three different diseases namely Leaf Blast [22-23], Leaf Brown Spot [22,24], Sheath Rot [22,24] have been considered for experiment. One infected sample images is shown in Fig 2. Infected portion of the plant image is detected using the proposed method as shown in Fig. 3.

Different features as described in the earlier section are extracted and classification accuracy obtained by PART, C4.5, Naïve bay's, SMO, bagging, boosting, MCS Classifier using "weka" tool [21]. Ten-fold cross-validations are carried out to obtain classification accuracy, as listed in Table-1.

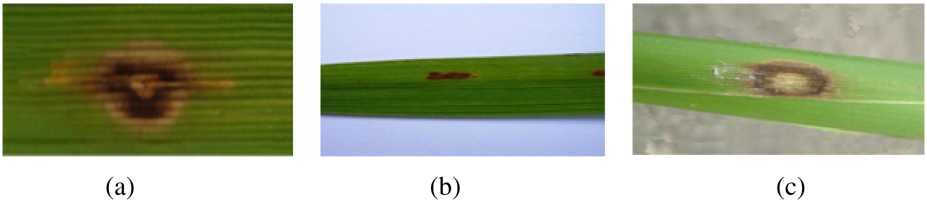

(a) (b) (c)

Fig. 2. Shows acquired images (a) image of rice leaf infected by blast disease; (b) rice leaf image infected by brown spot; (c) image of rice stem infected by sheath rot

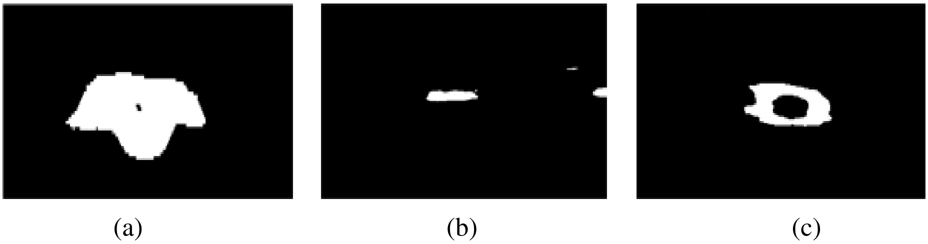

(a) (b) (c)

Fig. 3. Segmented images corresponding to Figure 2 using the proposed method

Table 1. Accuracy of the correct classification for different methods using different classifier

Classifier	**Proposed Method (%)**	**Otsu's method (%)**	**K-means method (%)**	**Energy Based method (%)**
J48	57.33	46.15	48.50	43.33
PART	50.33	47.49	46.15	47.67
MLP	55.67	**53.18**	55.52	53.33
Baye's Net	52.33	44.15	43.14	38.67
SMO	48	46.49	49.83	46
Boosting	52.67	45.42	42.47	36.33
Bagging	55.33	48.83	49.83	44.67
MCS	57.67	55.18	56.52	54
Average	53.67	48.36	49	45.5

5 Conclusions

The proposed method avoids selection of threshold value experimentally to segment the images. Complexity of the proposed method is less and may be used for online application. Results show that the proposed method gives better accuracy compare to other popular existing segmentation methods. The accuracy of the system may be increased by using other feature set. The roughness of the method may be evaluated by applying it other dataset.

References

1. Sing, A.K.: Advances in Data Analytical Techniques, vol. VI, pp. 165–174. Indian Agricultural Statistics Research Institute
2. Ray, N., Saha, B.N.: Edge Sensitive Variational Image Thresholding. In: ICIP, vol. VI, pp. 37–40 (2007)
3. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Pearson Education, New Delhi (2007)
4. Sankur, B., Sezgin, M.: Survey over image thresholding techniques and quantitative performance evaluation. Electron Imaging 13(1), 146–165 (2004)
5. Trier, O.D., Jain, A.K.: Goal-directed evaluation of binarization methods. IEEE Tran. Pattern Analysis and Machine Intelligence PAMI-17, 1191–1201 (1995)
6. Guo, R., Pandit, S.M.: Automatic threshold selection based on histogram modes and a discriminant criterion. Machine Vision and Applications 10, 331–338 (1998)
7. Cai, J., Liu, Z.Q.: A New Thresholding Algorithm Based on All-Pole Model. In: ICPR 1998, Int. Conf. on Pattern Recognition, Australia, pp. 34–36 (1998)
8. Cho, S., Haralick, R., Yi, S.: Improvement of Kittler and Illingworths's Minimum Error Thresholding. Pattern Recognition 22, 609–617 (1989)
9. Jawahar, C.V., Biswas, P.K., Ray, A.K.: Investigations on fuzzy thresholding based on fuzzy clustering. Pattern Recognition 30(10), 1605–1613 (1997)
10. Li, C.H., Tam, P.K.S.: An Iterative Algorithm for Minimum Cross-Entropy Thresholding. Pattern Recognition Letters 19, 771–776 (1998)
11. Savakis, A.: Adaptive document image thresholding using foreground and background clustering. In: ICIP 1998: Int. Conf. On Image Processing, Chicago (October 1998)
12. Cheng, S.C., Tsai, W.H.: A Neural Network Approach of the Moment-Preserving Technique and Its Application to Thresholding. IEEE Trans. Computers (42), 501–507 (1993)
13. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models: IJCV, pp. 321–331 (1987)
14. Caselles, V., Kimme, R.: Minimal Surfaces Based Object Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(4), 394–398 (1997)
15. Xavier, B., Pierre, V., Jean-Philippe, T.: A priori information in image segmentation: energy functionalbased oh' shape statistical model and image information. IEEE (2003)
16. Hummel, R.E.: Electronic Properties of Materials (6), 63–74 (2011)
17. Wei, G., Sammes, N.M.: An introduction to electronic and ionic materials (3), 41–44 (2000)
18. Chaoxin, Z., Da-Wen, S., Liyun, Z.: Recent applications of image texture for evaluation of food qualities—a review. Trends in Food Science & Technology 17, 113–128 (2006)
19. Otsu, N.: A Threshold Selection Method from Gray Level Histograms. IEEE Transaction on Systems, Man and Cybernetics 9, 62–66 (1979)
20. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press (1967)
21. WEKA, Machine Learning Software (2011), `http://www.cs.waikato.ac.nz/~ml/`
22. Ou, S.H.: Rice Diseases, England: Commonwealth Mycological Institute (1972)
23. Webster, R.K.: Rice blast disease identification guide. University of California, Davis (2000)
24. Rice Doctor, International Rice Research Institute, Philipines (2003), `http://www.irri.org`

Unsymmetrical Trimmed Midpoint as Detector for Salt and Pepper Noise Removal

K. Vasanth[1], V. Jawahar Senthil Kumar[2], and V. Elanangai[3]

[1] Research Scholar
Sathyabama University,
Chennai, Tamilnadu
[2] Research Guide, Anna University
Chennai, Tamilnadu
[3] Lecturer
Sathyabama University,
Chennai, Tamilnadu
`{Vasanthecek,elanangai123}@gmail.com, veerajawahar@annauniv.edu`

Abstract. A fixed 3x3 window, Unsymmetrical trimmed midpoint is used as a detector for the detection of fixed valued impulse noise is proposed for the increasing noise densities. The processed pixel is termed as noisy, if the absolute difference between processed pixels and unsymmetrical trimmed midpoint is greater than fixed threshold. Under high noise densities the processed pixel is noisy, so the median of the ordered array is found. The median is checked using the above procedure. If found true then the computed median is considered as noisy hence the corrupted pixel is replaced by the Unsymmetrical Trimmed midpoint of the current processing window. If median is not noisy then replace the median of the current processing window else if the pixel is termed uncorrupted, it is left unaltered. The proposed algorithm (PA) is tested on different varying detail images. The proposed algorithm is compared with the standard algorithms and found to give good results both qualitative and quantitatively for increasing noise densities. The proposed algorithm eliminates salt and pepper noise up to 80% and preserves edges up to 70%.

Keywords: Unsymmetrical trimmed midpoint filter, salt and pepper noise, threshold based midpoint filters.

1 Introduction

Images are often corrupted by Salt and Pepper noise due to faulty communication channels or transmissions of images over the channels. If a salt and pepper noise is corrupted in a Gray scale image, pixels corrupted by positive impulses appear as white dots and those corrupted by negative impulses appear as black dots. Filtering is a solution for the removal of noise. Two types of filters are used for noise reduction 1. Linear filters accomplish noise reduction with blurring. 2. Non-linear filters have a good edge and image detail preservation properties that are highly desirable for image filtering. Median filters are especially suitable for reducing "salt & pepper" noise.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 813–822.
springerlink.com

© Springer-Verlag Berlin Heidelberg 2013

Median filter is a spatial filtering operation, which uses a 2-D mask that is applied to each pixel in the input image. Median filtering preserves fine details of an image. Median filters are very efficient for smoothing of spiky noise. Median filter blurs the image for larger window size and insufficient noise suppression for small window sizes. [1]- [2]. The Median operation is applied uniformly over the entire image results in the modification of uncorrupted pixel. At high noise densities, the standard median filter is prone to edge jitter [2]. Subsequently, the effective removal of impulses is often at the expense of blurred and distorted features. Adaptive Median Filter (AMF) uses increasing window size based on number of noise corrupted pixels in the current processing window. AMF fairs well at low and medium noise densities [3], Due to the increasing window size AMF blurs the image at high noise densities. The pixels are decomposed based on various threshold levels and subjected to Boolean operation in Threshold decomposition filter (TDF). The need for complex Rank ordering technique is hence eliminated. This algorithm requires large threshold levels for operation and fails at higher noise densities. The center weighted median filters (CWF) uses a large weight to the central pixel, while processing this filter duplicates each input samples k_i times and choosing median from the considered array. This center weight parameter allows fine tuning of the processed pixel, at low and median noise densities the algorithm fared well but the performance declines at high noise densities [5]. The various median and its variant filters discussed so far operate uniformly over the entire image results in the changing the originality of uncorrupted pixel. The concept of filtering is that it should be applied only to corrupted pixels while leaving uncorrupted pixels unaltered. Therefore, a noise-detection process should differentiate between uncorrupted pixel and the corrupted pixel prior to applying nonlinear filtering is highly desirable. To overcome the drawback of the above filters switched median filters were introduced. These filters work on the basis of noise detection and correction. In Progressive Switched Median filter (PSMF) the decision is based on fixed threshold value and hence a procuring a robust decision is difficult. Hence at high noise densities the switched filters do not preserve any of the local detail of the image and hence fine details are not preserved properly [6]-[7]. The Detail preserving filter used an ordered minimum-maximum mean for impulse noise removal. The DPF filter removes noise at medium noise densities but fails to eliminate salt and pepper noise at high noise densities [8]. Hence a suitable impulse detection algorithm is a must for any good filter for better result. The paper sectioned as follows. The II Section deals with the Noise model for salt and pepper noise. The III Section deals with Proposed Algorithm and its elaborative methodology. Section IV exclusively deals with Exhaustive results and Discussion. Section V gives the conclusion.

2 Noise Model

Let the true image x belong to a proper function space $S(\Omega)$ on $\Omega = [0; 1]^2$, and the observed digital image y be a vector in $Rmxm$ indexed by $A = \{1,2,..m\}$ X $\{1,2,.m\}$.The image degradation can be modeled as $y = N(Hx)$, where $H : S(\Omega) \rightarrow Rmxm$ is a linear operator representing blurring, and $N : Rmxm \rightarrow Rmxm$ models the noise. Usually, $y = Hx + \sigma n$ where $\sigma n \in Rmxm$ is an additive zero-mean Gaussian noise with standard deviation $\sigma >= 0$. Outliers are modeled as impulse noise[11]. Then a realist model for our data is

$$y' = H.x.K + \sigma g \tag{1}$$

$$y = N(y') \tag{2}$$

Where *N* represents the impulse noise and K refers to speckle noise as given in equation 1 & 2. The noise model for salt & pepper noise is given as, If [0; 255] denote the dynamic range of *y'*, i.e., 0 <= y'ij <= 255 for all (*i,j*), then they are denoted by Salt-and-pepper noise: the gray level of *y* at pixel location (*i j*) is illustrated in the equation 3.

$$y_{ij} = \begin{cases} 0 & \text{with probability } p; \\ y'_{ij} & \text{with probability } 1 - p - q; \\ 255 & \text{with probability } q; \end{cases} \tag{3}$$

Where $s = p + q$ denotes the salt-and-pepper noise level [9].

3 Proposed Algorithm

A. Unsymmetrical Trimmed filters

The crux behind the above filter is to eliminate the outliers inside the current window and preserve edges at high noise densities. All the pixels of an image lie between the dynamic ranges [0,255] (8 bit image). Hence Rank order the pixels of the current window and trim (eliminate) on either side for 0 or 255. The elements that are not outliers survive the elimination process. The arithmetic mean of the untrimmed pixels from the above operation is the idea behind Un-symmetrical Trimmed Mean filters. Here the trimming of values is done Un-symmetrically based on the local pixel information.

B. Unsymmetrical Trimmed Midpoint Filters

Consider a 3X3 window from the corrupted image. Order the pixels of the current window in increasing order. Now perform trimming on either side of the ordered array for 0 or 255. The elements that are not eliminated are considered as non noisy candidates of the processed window. The midpoint of the untrimmed pixels from the above operation is the anatomy of Un-symmetrical Trimmed Midpoint filters.

C. Unsymmetrical Trimmed Median Filters

Consider a 3X3 window from the corrupted image. Arrange the pixels of the processing window in ascending order. Now perform trimming on either side of the ordered array for 0 or 255. The elements that are not eliminated are considered as non noisy candidates of the processed window. The Median of the untrimmed pixels from the above operation is the basic idea behind of Un-symmetrical Trimmed Median filters.

D. Proposed algorithm

The brief illustration of the proposed algorithm is as follows.

Step 1: Choose 2-D window of size 3x3. The processed pixel in current window is assumed as p_{xy}.

Step 2: sort the 2D window data in ascending order using snake like modified shear sorting which is given by S. now Convert sorted 2D array into 1D array. S_{med} is the median of the sorted array

Step 3: **Unsymmetrical trimmed Midpoint filter**

Initialize two counters, forward counter (F) and reverse counter (L) with 1 and 9 respectively. When a 0 or 255 are encountered inside the Sorted array (S), F is incremented by 1 or L is decremented by 1 respectively. The resulting array will be holding non noisy pixels of the current window. The midpoint of this array is termed as UTMP (unsymmetrical trimmed midpoint) .

Step 4: **Salt and pepper noise Detection using UTMPF**

Case (1): If the absolute difference between the processed pixel and unsymmetrical trimmed midpoint filter (UTMP) is greater than the fixed threshold (T) then pixel is considered as noisy. As illustrated in equation1

$$\text{If } \left| P(x,y)-\text{UTMP} \right| > T \tag{4}$$

Case (2): If the case 1 is true find the absolute difference between the median of and unsymmetrical trimmed midpoint filter (UTMP). Check the difference is greater than the fixed threshold (T1) then median is considered as noisy as illustrated in equation 2. Case 2 is done for high noise densities where the computed median is also noisy.

$$\text{If } \left| S_{med}-\text{UTMP} \right| > T1 \tag{5}$$

Step 5: **Salt and pepper noise Correction logic**

If the case1 $\left| P(x,y)-\text{UTMP} \right| > T$ is true then check for the second case2 $\left| S_{med}-\text{UTMP} \right| > T1$. If both the condition are true then processed pixel and computed median is noisy. Hence replace the corrupted pixel with median of Unsymmetrical trimmed midpoint. If condition 1 is true and condition 2 is false then corrupted pixel is replaced with the median of the sorted array. If both case 1 and case 2 fails then the pixel is termed as non noisy. The pixel is left unaltered.

E. Methodology of proposed work

The bigger matrix refers to image and values enclosed inside a rectangle is considered to be the current processing window. The element encircled refers to processed pixel. The above discussed methodology is illustrated as below.

Case (a): Initialize forward counter F=1 and reverse counter L=9. Convert the 2D array into 1D array and sort the converted array. F and L counter moves in forward and reverse directions respectively. When a 0 is detected F is incremented by 1 and when a 255 is detected L is decremented by 1.

$$\begin{pmatrix} 0 & 0 & 255 & 0 & 255 \\ 94 & 177 & 205 & 155 & 255 \\ 0 & 0 & \enclose{circle}{255} & 25 & 123 \\ 0 & 0 & 187 & 124 & 255 \\ 0 & 255 & 255 & 255 & 255 \end{pmatrix} \qquad \begin{pmatrix} 0 & 0 & 255 & 0 & 255 \\ 94 & 177 & 205 & 155 & 255 \\ 0 & 0 & \enclose{circle}{115} & 25 & 123 \\ 0 & 0 & 187 & 124 & 255 \\ 0 & 255 & 255 & 255 & 255 \end{pmatrix}$$

Corrupted image Segment Restored image Segment

Unsorted array: 177 0 0 205 255 187 155 25 124
Sorted array S_{xy} 0 0 25 124 155 177 187 205 255

Here the median S_{med} value is 155. The case (1) is illustrated as follows. Now check for the presence of 0 or 255 in the sorted array. Every time a 0 is detected F is incremented by 1 and if 255 is detected L is decremented by1. In the above example there is two 0 and one 255. Hence F is incremented by two times and L is decremented by one time. Now finally F is holding 3 and L is holding 8. Now the variable DET is assigned with the midpoint of the rank ordered unsymmetrical trimmed output i.e. corrupted pixel is replaced by Midpoint of the trimmed array i.e, (25+205)/5=115. i.e, DET=115. Now perform first step detection $|255\text{-}115| > 40$. This condition is true. The Second condition is checked $|155\text{-}115| > 20$ and the second condition is true. Hence the pixel and the computed median is considered as noisy. The corrupted pixel is replaced by midpoint of sorted array ie., output =115.

$$\begin{pmatrix} 0 & 0 & 255 & 0 & 255 \\ 94 & 0 & 0 & 125 & 125 \\ 0 & 185 & \enclose{circle}{0} & 255 & 255 \\ 0 & 255 & 255 & 255 & 255 \\ 0 & 255 & 255 & 255 & 255 \end{pmatrix} \qquad \begin{pmatrix} 0 & 0 & 255 & 0 & 255 \\ 94 & 0 & 0 & 125 & 125 \\ 0 & 185 & \enclose{circle}{130} & 255 & 255 \\ 0 & 255 & 255 & 255 & 255 \\ 0 & 255 & 255 & 255 & 255 \end{pmatrix}$$

Corrupted image Segment Restored image Segment

Case (b): Initialize forward counter F=1 and reverse counter

L=9. Convert the 2D array into 1D array and sort the converted array. When a 0 is detected F is incremented by 1 and when a 255 is detected L is decremented by 1.

Unsorted array: 0 185 255 0 0 255 125 255 255
Sorted array S_{xy} 0 0 0 125 130 255 255 255

Here the median S_{med} value is 130. The case (2) is illustrated as follows. Now check for the presence of 0 or 255 in the sorted array. Every time a 0 is detected F is incremented by 1 and if 255 is detected L is decremented by1. In the above example there is three 0 and three 255. Hence F is incremented by three times and L is decremented by three times. Now finally F is holding 4 and L is holding 6. Now the variable DET is assigned with the midpoint of the rank ordered unsymmetrical trimmed output i.e. (125+185)/2 = 127. i.e., DET=127. Now perform first step detection $|0\text{-}127| > 40$. This condition is true. The Second condition is checked $|130\text{-}127| > 20$ and the second condition is false. Hence the processed pixel is noisy and the computed median is considered as non noisy. Hence the corrupt pixel is replaced with median of the array i.e. 130 output=130.

$$\begin{pmatrix} 0 & 0 & 255 & 0 & 255 \\ 104 & \boxed{119} & 255 & 255 & 255 \\ 0 & 103 & 255 & 255 & 123 \\ 0 & 122 & 255 & 124 & 255 \\ 0 & 255 & 255 & 255 & 255 \end{pmatrix} \qquad \begin{pmatrix} 0 & 0 & 255 & 0 & 255 \\ 104 & \boxed{119} & 255 & 255 & 255 \\ 0 & 103 & 255 & 255 & 123 \\ 0 & 122 & 255 & 124 & 255 \\ 0 & 255 & 255 & 255 & 255 \end{pmatrix}$$

Corrupted image Segment Restored image Segment

Case (3): Un sorted Array 0 104 0 0 119 103 255 255 255
Sorted Array 0 0 0 103 104 119 255 255 255

Initialize F=1 and L=9. After sorting the current window in ascending order, the counters propagate in the 1D array resulting in holding F=4 and L=6. DET will hold trimmed midpoint (103+109)/2=106 ie, DET=106. Now perform impulse detection $|119\text{-}106| > 40$. This condition is false and hence processed pixel is considered as non noisy hence left unaltered.

4 Simulation Results and Discussions

The Quantitative performance of the proposed algorithm is evaluated based on Peak signal to noise ratio (PSNR) ,Mean Square Error (MSE) and Image Enhancement Factor (IEF) which is given in equations 4,5 ,6 respectively.

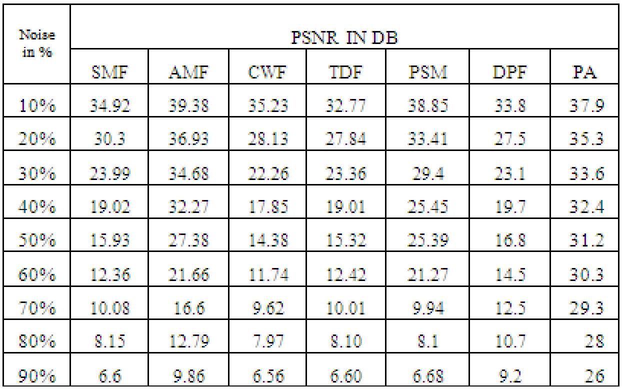
$$\text{PSNR} = 10\log_{10}\left(\frac{255^2}{MSE}\right) \tag{4}$$

$$\text{MSE} = \frac{\sum_i \sum_j (r_{ij} - x_{ij})2}{M \times N} \tag{5}$$

$$\text{IEF} = \frac{\left(\sum_i \sum_j (n_{ij} - r_{ij})2\right)}{\left(\sum_i \sum_j (x_{ij} - r_{ij})2\right)} \tag{6}$$

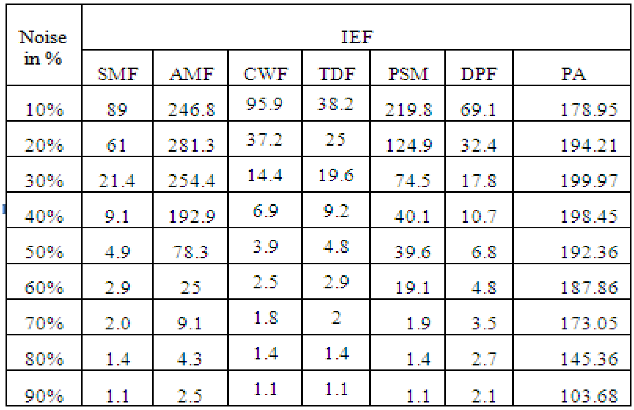
Where *r* refers to Original image, n gives the corrupted image *x* denotes restored image, M x N is the size of Processed image. The existing algorithms used for the comparison are SMF, AMF, CWF, TDF, PSMF, DPF, The qualitative performance of the proposed algorithm is tested on various images (Images are chosen as per the details of the image). Quantitative analysis is made by varying noise densities in steps of ten from 10% to 90% on images and comparisons are made in terms of PSNR, IEF, MSE. Results and graphs are given in Table 1, 2, 3 and figure 1, 2, 3 respectively. Figure 4 and 5 gives the qualitative performance of the proposed algorithm in terms of noise elimination and edge preservation. All the simulation is done in dual CPU E2140@1.6Ghz with 1GB RAM capacity. Better results were obtained when the pre-defined threshold T was between 20 and 40. And the second threshold T1 was between 15 and 30. From the table 1 we infer that for the proposed algorithm PSNR value is very high indicating how much the algorithm eliminates salt and pepper noise

Table 1. Performance of various algorithms at different noise densities for PSNR



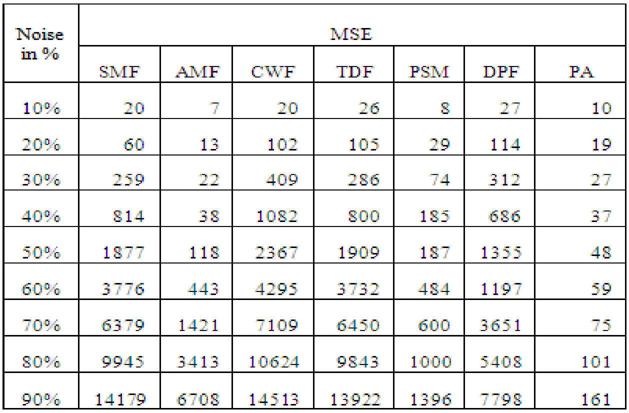
Noise in %	PSNR IN DB						
	SMF	AMF	CWF	TDF	PSM	DPF	PA
10%	34.92	39.38	35.23	32.77	38.85	33.8	37.9
20%	30.3	36.93	28.13	27.84	33.41	27.5	35.3
30%	23.99	34.68	22.26	23.36	29.4	23.1	33.6
40%	19.02	32.27	17.85	19.01	25.45	19.7	32.4
50%	15.93	27.38	14.38	15.32	25.39	16.8	31.2
60%	12.36	21.66	11.74	12.42	21.27	14.5	30.3
70%	10.08	16.6	9.62	10.01	9.94	12.5	29.3
80%	8.15	12.79	7.97	8.10	8.1	10.7	28
90%	6.6	9.86	6.56	6.60	6.68	9.2	26

Table 2. Performance of various algorithms at different noise densities FOR IEF



Noise in %	IEF						
	SMF	AMF	CWF	TDF	PSM	DPF	PA
10%	89	246.8	95.9	38.2	219.8	69.1	178.95
20%	61	281.3	37.2	25	124.9	32.4	194.21
30%	21.4	254.4	14.4	19.6	74.5	17.8	199.97
40%	9.1	192.9	6.9	9.2	40.1	10.7	198.45
50%	4.9	78.3	3.9	4.8	39.6	6.8	192.36
60%	2.9	25	2.5	2.9	19.1	4.8	187.86
70%	2.0	9.1	1.8	2	1.9	3.5	173.05
80%	1.4	4.3	1.4	1.4	1.4	2.7	145.36
90%	1.1	2.5	1.1	1.1	1.1	2.1	103.68

Table 3. Performance of various algorithms at different noise densities FOR MSE



Noise in %	MSE						
	SMF	AMF	CWF	TDF	PSM	DPF	PA
10%	20	7	20	26	8	27	10
20%	60	13	102	105	29	114	19
30%	259	22	409	286	74	312	27
40%	814	38	1082	800	185	686	37
50%	1877	118	2367	1909	187	1355	48
60%	3776	443	4295	3732	484	1197	59
70%	6379	1421	7109	6450	600	3651	75
80%	9945	3413	10624	9843	1000	5408	101
90%	14179	6708	14513	13922	1396	7798	161

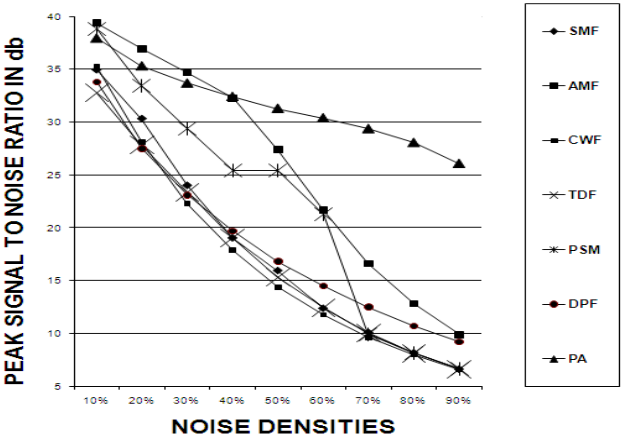


Fig. 1. Performance comparison of PSNR at various noise densities for low detail Lena image

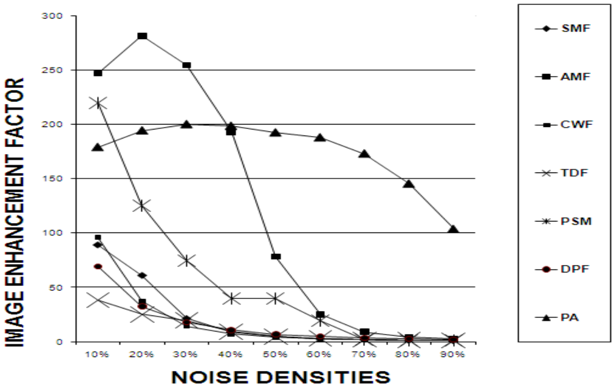


Fig. 2. Performance comparison of IEF at various noise densities for low detail Lena image

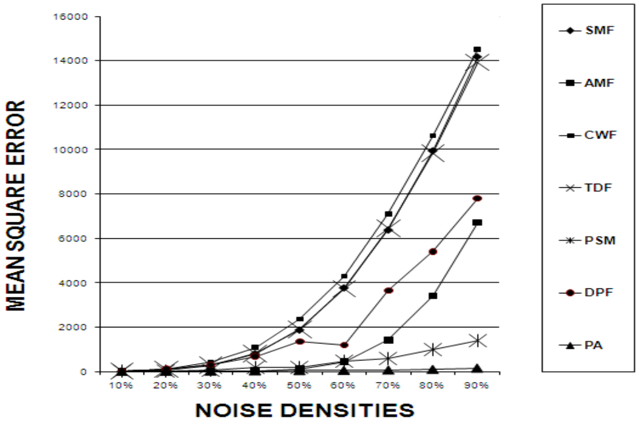


Fig. 3. Performance comparison of MSE at various noise densities for low detail Lena image

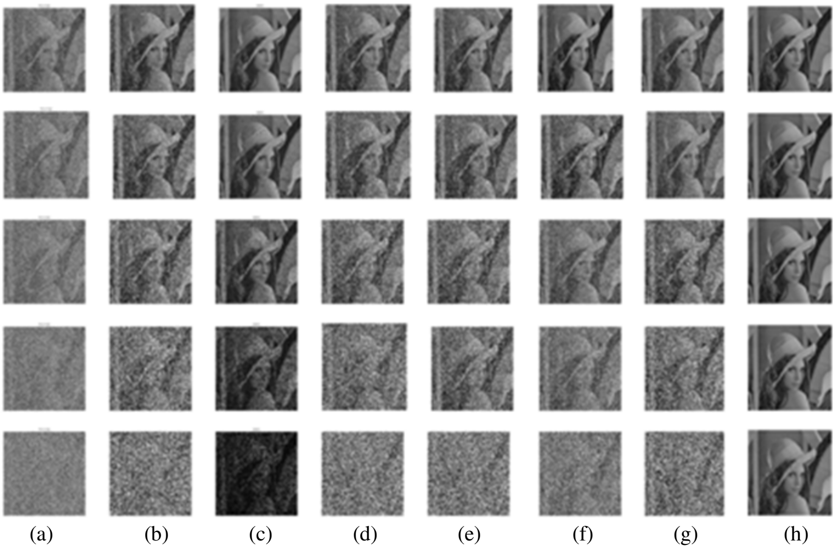


Fig. 4. Performance of various filters for Lena image corrupted by Salt and pepper noise from 50% to 90% in row1 to 5 respectively. Output of various filters in column 1 to 8 (a) fixed value salt and pepper noise (b) output of SMF (c) output of AMF (d) output of PSMF (e) output of DPF (f) output of MEANDET (g) output of MEDDET (h) output of PA.

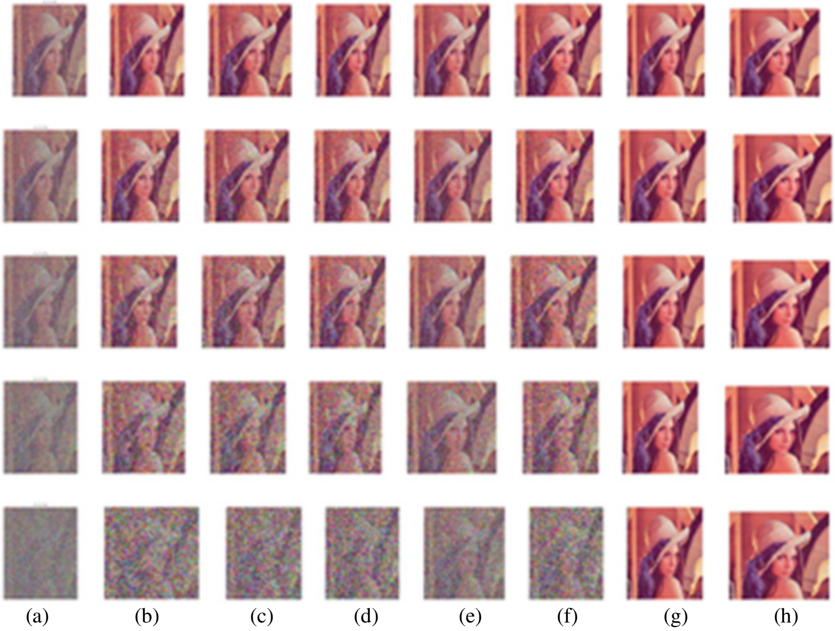


Fig. 5. Edge preservation of various filters for Lena image corrupted by Salt and pepper noise from 50% to 90% in row1 to 5 respectively. Output of various filters in column 1 to 8 (a) fixed value salt and pepper noise (b) output of SMF (c) output of MEAN DET (d) output of MED DET (e) output of PSMF (f) output of DPF (g) output of DBA (h) output of PA.

effectively even at high noise densities. Table 2 gives a high image enhancement factor, even at very high noise densities as high as 90%. From Table 3 we find the mean square error is also less for proposed algorithm at high noise densities. It is evident from figure 4 and 5 that the qualitative aspect of the proposed algorithm is very high against various algorithms for increasing noise densities (50% to 90%) for both Gray scale and color Lena (low detail image) and also edges were preserved and fine details are restored up to 70%. The good edge preservation is due to the fact that the obtained unsymmetrical midpoint is very close to median which generally has very good edge preservation capability. It was found that proposed algorithm preserved the global and local edges up to 70% of salt and pepper noise where filters such as SMF, AMF, PSMF, DPF fails. Hence none of the single level detector algorithm is able to detect and correct the short tailed noise at high noise densities. In this paper two values are considered for impulse detection hence make this algorithm more attractive. From last column of figure 5 we observe that the proposed algorithm preserves edges for increasing noise densities for color images. The proposed algorithm fairs less in comparison with adaptive median filter up to 30% of salt and pepper noise. From figure 1, 2, 3 it was found that the proposed algorithm has a good PSNR, high IEF and low MSE. The value of the threshold is updated based on the number of corrupted pixels inside the corrupted window.

5 Conclusion

A new 3x3 fixed window is proposed with two thresholds based Detector is proposed, which gives excellent noise suppression capabilities by preserving edges in images corrupted by salt and pepper noise as high as 70% for low detail image in both grayscale and color images. The Proposed algorithm outclasses many classical filters both in quantitative and qualitative aspects both for grayscale and color images.

References

[1] Pitas, I., Venetasanopoulos, A.N.: Non Linear Digital Filters: Principles and Applications. Kluwer, Boston (1990)

[2] Astola, J., Kuosmanen, P.: Fundamentals of non linear digital filtering. CRC Press (1997)

[3] Hwang, I.H., Hadded, R.A.: Adaptive Median filter: New algorithms and results. IEEE Transaction on image Processing 4(4), 499–502

[4] Zhang, H.: Generalized Threshold Decomposition. Journal of Electronics 14(1) (January 1997)

[5] Ko, S.-J., Lee, Y.-H.: Center weighted median filters and their applications to image enhancement. IEEE Trans. Circuits Syst. 38, 984–993 (1991)

[6] Ng, P.E., Ma, K.K.: A switching median filter with boundary discriminative noise detection for extremely corrupted images. IEEE Transactions on Image Processing 15(6), 1506–1516 (2006)

[7] Zhang, S., Karim, M.A.: A new impulse detector for switching median filters. IEEE Signal Processing Letters 9(11), 360–363 (2002)

[8] Naif Alajlan, K., Kamel, M., Jernigan, E.: Detail Preserving impulse noise removal. International Journal on Signal Processing: Image Communication 19, 993–1003 (2004)

[9] Bovik, A.: Handbook of Image and Video Processing. Academic Press (2000)

Identification of Handwritten Text in Machine Printed Document Images

Sandipan Banerjee

Dept. of Computer Science and Engineering,
National Institute of Technology, Durgapur, India
sandipan9008@gmail.com

Abstract. In our daily lives we come across many documents where both printed and handwritten text co-exist and sometimes intermingle. As the OCR techniques for processing the two are quite different it is necessary to classify and distinguish them first. In this paper, a scheme has been proposed by which handwritten, printed and "mixed" text regions in the same document image can be identified and demarcated from each other for Bangla, the second most popular Indian script. The proposed scheme has been established on the basis of the structural and statistical idiosyncrasies of printed and handwritten Bangla text.

Keywords: Document Processing, Optical Character Recognition, Bangla Script, Machine-printed and Handwritten Text, Indian Language.

1 Introduction

Document images are processed and analyzed successfully using the Optical Character Recognition (OCR) techniques for applications like natural language processing, text mining, human aid etc. Most of the OCR systems present in the market however are dedicated for machine printed texts only and some of the ones that do exist for handwritten ones are mainly for English and other Latin based scripts. For Indian languages however and especially Bangla, some work has been done in this field and they are mostly concerned with segmenting, extracting and recognizing individual characters from the given text.

But the whole equation changes when dealing with handwritten text as the recognition schemes for it are totally different from that of printed ones. In terms of preprocessing, segmentation, noise removal, feature selection and classification etc. handwritten text recognition is a different ball game totally. So, if a document page consists of both handwritten and printed text together in it, then it becomes very difficult to process it using OCR techniques and therefore the two types of text have to be separated and distinguished first before being fed to the OCR system. The separated printed and handwritten text can then be recognized using the standard OCR systems available today.

Documents with printed and handwritten text together are found quite frequently in our daily lives. Some of the common ones are question papers or feedback forms where the printed questions or queries are put inside a table and the answers are to be

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 823–831.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

provided by hand. Also, such documents can be found in printed text books where several lines are underlined, highlighted or annotated by hand for taking notes. Not only for proper processing but many historic documents, which are the last remaining prints of a text, if tainted with handwritings, can be cleaned up and the original text can be recovered and preserved by separating the handwritten text from it.

In this paper, the classification is done on the structural and statistical differences between machine-printed and handwritten text and a tainted text can be differentiated into any of the three categories: purely printed, purely handwritten or “mixed” i.e. where parts or whole of handwritten and printed text are very close or overlap on each other.

2 Pre-processing

The experiment was performed by using tainted documents where both the printed and handwritten text is in Bangla script. The collected document pages had a variety of handwritings from different people with different writing styles. The pages were then scanned and digitized to get the document images for working on. The gray-scale images were then binarized into a two tone image with pixels having ASCII value either as 0 or 255 where the former represents a black pixel and the later signifies a white space. The two-tone image thus obtained is then ready for the next stage of pre-processing.

The text matrix now obtained was labeled on the basis of its connected components using the 4-component Connected Component Labeling (CCL) technique in [8]. The connected components, as marked according to the algorithm, can be a single word, multiple joined words or an overlapping piece of intermingled text. The next step was to calculate the Bounding Box (BB) for each of the connected components. The Bounding Box is the minimum rectangle (or box) that can contain a connected component within it, like the figure below.

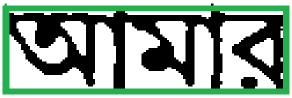


Fig. 1. Connected Components inside the Bounding Box

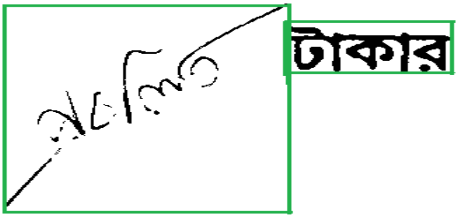


Fig. 2. Mutually overlapping Bounding Boxes

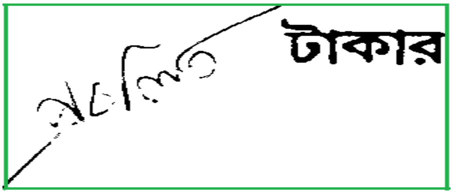


Fig. 3. Bounding Box Reconstruction for this “mixed” region

There may occur cases where the BB’s of two adjacent connected components overlap each other and create a problem in distinguishing one from the other as shown in Figure 2. In that case, we treated both the connected components as a single entity and reconstructed the BB which can contain both of them together like Figure 3. So, the text components were now properly distinguished according to their connected components and their corresponding Bounding Boxes were also updated for future use.

3 Classification Scheme

The classification strategy proposed here is a two stage classifier with the classification being done on the basis of the structural differences between the perfectly aligned and symmetric printed text and the coarse and skewed handwritings due to the human writing styles. The features are quite simple and easily detectable. The first stage classifier is based on three simple features that machine-printed text possesses and it separates the purely machine-printed text from the handwritten part of the document image. The second stage classifier discriminates the mixed regions of text, i.e. the BBs where machine printed text and handwritten ones are placed together, from the purely handwritten parts of text. The two stage classifier has been discussed below.

A. First Stage Classifier

The first stage classifier is fed the updated list of BBs which contain any one of the three: purely printed connected components, purely handwritten components or mixed components. The task of the classifier is to mark the BBs which contain purely printed text and separate them from the handwritten and mixed text. For doing this, it uses very basic and intrinsic features that are possessed by Bangla text, especially the machine printed ones. The features sought after by the extractor are as follows.

1. Feature One: Headline

The common idiosyncrasy found among most of the Bangla alphabets is the presence of the “matra” or the headline on the upper region. And when two or more such alphabets are placed together adjacently to form a word, their headlines join together to form an extended headline - basically a horizontal run. Now the probability that a given word will possess one or two headlines in it is quite high as most of the Bangla alphabets, 32 out of the 50 present, consist of a headline. Moreover, 11 among the 12 most used Bangla characters also possess a headline as well. On top of that, there are

41 characters that can start a word in Bangla and of them 30 are found with a headline. So it is obvious that the percentage of Bangla words that consist of at least one prominent headline is very high and is found out to be 99.4%.

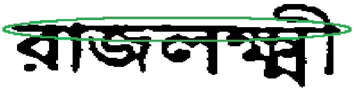


Fig. 4. Headline or Matra in a Printed Bangla Word

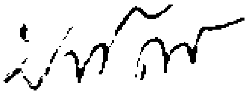


Fig. 5. Absence of a Headline in a purely Handwritten text

This special feature of Bangla script can be utilized as a distinguishing feature between machine printed and handwritten text. In machine printed text the headline region is denoted by a straight horizontal run in the row of pixels. So, in a word that possesses a headline, there is at least one long horizontal run in its pixel row. But in handwritten text, proper care is not taken by the individual while writing and in most cases, approximately 91%, the headline is not straight at all and therefore doesn't form a long horizontal run along its row of pixels. In some one off cases, some individuals do take proper care and write slowly and their headlines show a salient horizontal run in that case.

This variation between the headlines of handwritten and printed text has been taken as the first feature to be extracted in this scheme. A threshold T has been fixed and if the horizontal run of a word or text in a BB is found to be exceeding it, then it is likely to be a printed text and the BB is marked for being fed for the second feature extraction. And the BBs where such a horizontal run greater than T is not found, they are straight away kept to be fed to the second stage classifier. The value of T is taken as the average length of a Bangla alphabet of the printed text present in the document image that we are working on. So, the horizontal run in any text has to be at least longer than the average length of an alphabet for the text to be kept in contention for being a machine printed one. The BBs which satisfied this condition and were found to contain a long horizontal run were marked in the set BB_1 and the other set is named as BB_2.

2. Feature Two: Lowermost Point(s)

In Bangla, there are 39 consonants and 11 vowels. When a vowel is placed beside a consonant it usually takes a modified shape and is called a modifier thereafter. The modifier may be placed right, left, up or below the consonant which it modifies. So, the vowel 'উ' when placed adjacent to the consonant 'ক', it acts as a modifier and it changes its shape and the consonant now becomes 'কু'.

In machine printed Bangla text, the lower modifiers, if any, of the characters in a word lie on the same horizontal line, known as the lower line and the normal unperturbed characters of the word lie on another horizontal line, known as the base line. So basically the lowermost point of any character of a Bangla machine printed word lies in any one of the two horizontal lines mentioned above i.e. the base line or the lower line. But in handwritten text such care is not taken by the writer and the lowermost points of the characters of any word are distributed unevenly and therefore lie on more than two horizontal lines. And this feature has been used to further differentiate between printed and handwritten text.

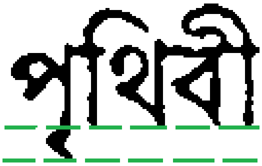


Fig. 6. Printed text having its lowermost points on two horizontal lines only (the base line and the lower line)

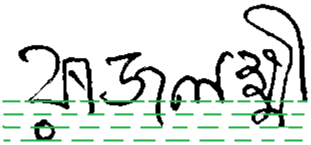


Fig. 7. Handwritten text with some of its lowermost points on more than two horizontal lines

The extracted and marked BBs in BB_1 after the first feature extraction i.e. the headline, are now to be tested on the basis of this feature and the printed text can therefore be found out. For mathematical determination and programmatic execution the method mentioned in [4] has been used. The lowermost points of the different components of any word is divided here into two sets B and L based on their proximity to the base line row B_r and L_r respectively. A component which has its lowermost point at the row R belongs to B if $| B_r - R | <= | L_r - R |$ and vice versa. All such positions for the lowermost points of a text or connected component is calculated and is placed in either of the two sets B, comprising of $b_1, b_2, b_3, ..., b_m$, where m is the number of lowermost points belonging to B, and L comprising of $l_1, l_2, l_3, ..., l_n$, where n is the number of lowermost points belonging to L. Now a feature called is the Character Lowermost Point Standard Deviation (CLPSD) is introduced and its value is calculated as:-

$$\text{CLPSD} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(b_i - b')^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(l_i - l')^2}$$

$$\text{where } b' = \frac{1}{m}\sum_{i=1}^{m} b_i \text{ and } l' = \frac{1}{n}\sum_{i=1}^{n} l_i$$

For machine printed text the lowermost points in B and L are almost all in the same row making the value of CLPSD very low i.e. the distribution is uniform. But for handwritten text, the distribution is usually uneven and therefore the individual row values in both B and L differ drastically from the mean value giving CLPSD a high value. The threshold used for the demarcation is kept as 0.1 of the mean height of the components in the text.

This feature helps to identify the set of BBs which contain printed text properly as in many cases where handwritten and printed text intermingle together, as in mixed regions, there can very well be a prominent headline due to the printed part of the text but the lowermost points of the components of such a BB are unevenly distributed giving it a very high value of CLPSD. For the BBs which have values of CLPSD less than the threshold they are identified as purely printed text and are separated out from the group and are put in the new set BB_3. The BBs which were found to have a high value of CLPSD are chalked out and mixed with BB_2 and we get a bigger set BB_4. This set is now fed to the second stage classifier.

B. Second Stage Classifier

Modern pens have this quality that any text written using them are quite uniform in terms of thickness and rarely varies if the writing style of the user doesn't vary in between. In case of any sudden punctuation or an extra effort on some components of the text we can get words which have components with non-uniform thickness, but that is quite rare. In case of machine printed text and especially those in Bangla, due to the structural feature of certain alphabets and due to the curvy nature of the script, variable thickness is very familiar. Normally where any curve stops and changes its direction or a loop is encountered; a sudden change in the thickness of the text is noted. This feature can be utilized to distinguish the components of BB_4 into purely handwritten and mixed text.

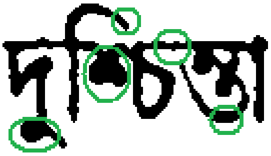


Fig. 8. The different thicknesses in the different components of a machine printed Bangla word

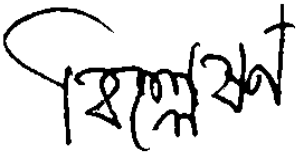


Fig. 9. A handwritten Bangla word having almost the same thickness for all its components

The text in each of the bounding boxes in the set BB_4 is now analyzed and their respective contours are found out. Then the resulting set of bounding boxes is fed to the classifier. The BBs where no sudden variation of the thickness of the contour is noted, those are marked as purely handwritten text and in the BBs where such a huge change of thickness is seen, they are marked as the mixed regions. Therefore we get two distinct set of BBs which are either purely handwritten or mixed regions of handwritten and printed text intermingled together. For noting the varying thickness (t) in any component we take the thickness in every position across a full run of a component and calculate the mean thickness (T_{mean}) of the component. If the mean thickness is lesser than a threshold value we conclude that the component is handwritten otherwise printed.

$$T_{mean} = \sum_{i=1}^{p} t_i$$

where $t_1, t_2, \ldots, t_p$ are the individual thicknesses for the p positions of the component.

The threshold for the variable thickness measurement is kept as 0.4 of the starting thickness for the connected component. That means for the thickness of the contour that we start with, an increase or decrease of only about 0.4 times of it is permissible. For free flowing handwritten text this condition is usually met as most of the times the writer writes the word in one complete go. But for machine printed text this varies heavily because of the presence of uneven zones of thickness like the headline, the modifiers, looping structures etc. So, there for printed text the threshold value is surpassed almost regularly and therefore we can separate it easily from the handwritten text. The initial thickness is recorded prior to every run across the contour of a component and is compared with the mean thickness found out in the end. If it is within the 0.4 bracket then it is marked as a handwritten text and if not then we conclude that it is a mixed region.

In the set BB_4 we were left with only purely handwritten text and mixed textual regions therefore the classifier now divides the set of BBs into two further sets: the BBs containing purely handwritten text, with a mean thickness lesser than or equal to the threshold, and the BBs which contain mixed regions of handwritten and machine printed text together and therefore has a mean thickness much greater than the threshold value. We mark the first of the two sets as BB_5 and the other one as BB_6.We already have a set of BBs which have purely machine printed text in BB_3, and now we get the two sets BB_5 and BB_6 as well marking the handwritten and mixed regions. Therefore the whole set of BBs has now been classified into three sets of purely printed, purely handwritten and mixed regions.

4 Results and Discussion

The scheme mentioned here was experimentally verified using a set of machine printed Bangla documents which were tainted by handwritings of different individuals. The set had over a total of 150 such samples of varying concentration of handwritten text. Most of the documents were pages from Bangla textbooks and question papers and some already scanned tainted documents were gleaned online. The printed text in them was of various styles and size. Each individual recorded three tainted pages of varying concentration and distribution of handwritten annotations:

one sparingly tainted with handwritten text and machine printed text in the ratio of 1:20, one with a relatively moderate concentration of handwritten text with some of them overlapping the printed lines and the third one being heavily tainted with almost 30% concentration of handwritten text. A total of 41 individuals recorded their handwritings for this purpose.

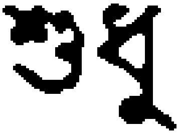


Fig. 10. A machine printed Bangla word without any headline

The algorithm was coded and implemented on the set of document images using the C programming language. Roughly, the accuracy of the proposed approach was found out to be 88%. The errors were mostly encountered when some of the machine printed texts were devoid of any headline as exhibited in Fig 10. But as discussed earlier such a case is quite rare and therefore should be exempted. Also, as expected most of the handwritten texts were devoid of any prominent headline region and therefore were discarded from being printed at the first go. Moreover, the thickness variation was also much more pronounced in case of printed word components and were therefore instrumental in demarcating between the purely handwritten and mixed regions of text. Most importantly, as the features used in the approach is independent of the style and font of the printed text so the scheme is independent of such considerations. For different documents however the thickness was different and therefore the initial thickness was calculated while parsing each component.

No. of pages	**No. of text boxes encountered**	**Correct Identification**	**Error**
152	17560	88%	12%

The proposed approach can be also implemented on documents having Devnagari, Assamese or Punjabi as their script as all of them are similar to Bangla. Even, after making some modifications on the classifier, it can be used to identify handwritten text in Latin based script like English as well. For future course of work, the next step would naturally be to formulate and implement the algorithm for identifying as well as separating the overlapping handwritten texts from the machine printed text in the mixed regions.

References

[1] Casey, R.G., Lecolinet, E.: A Survey of Method and Strategies in Character Segmentation. IEEE Transactions in Pattern Analysis and Machine Intelligence 18(7) (July 1996)

[2] Bishnu, A., Chaudhuri, B.B.: Segmentation of Bangla Handwritten text into characters by recursive contour following. In: Proceedings of the 5th International Conference on Document Analysis and Recognition, pp. 402–405 (1999)

[3] Pal, U., Chaudhuri, B.B.: Automatic Separation of Machine Printed and Handwritten Text Lines. In: Proceedings of the 5th International Conference on Document Analysis and Recognition, pp. 645–648 (1999)

[4] Pal, U., Chaudhuri, B.B.: Machine Printed and Handwritten Text Line Identification. Pattern Recognition Letters 22(3-4), 431–441 (2001)

[5] da Silva, L.F., Conci, A., Sanchez, A.: Automatic Discrimination between Printed and Handwritten Text in Documents. In: IEEE Xplore. Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI), pp. 261–267 (2009)

[6] Lemaitre, A., Chaudhuri, B.B., Couasnon, B.: Perceptive Vision for Headline Localization in Bangla Handwritten Text Recognition. In: Proceedings of the 9th International Conference on Document Analysis and Recognition (2007)

[7] Pal, U., Datta, S.: Segmentation of Bangla Unconstrained Handwritten Text. In: Proceedings of the 7th International Conference on Document Analysis and Recognition (2003)

[8] `http://en.wikipedia.org/wiki/Connected-component_labeling`

Optimization of Integration Weights for a Multibiometric System with Score Level Fusion

S.M. Anzar and P.S. Sathidevi

S.M. Anzar, Dept. of Electronics and Communication Engineering,
National Institute of Technology Calicut, India 673601
{p090004ec,sathi}@nitc.ac.in

Abstract. The effectiveness of a multibiometric system can be improved by weighting the scores obtained from the degraded modalities in an appropriate manner. In this paper, we propose an integration weight optimization scheme to determine the optimal weight factor for the complementary modalities, under different noise conditions. Instead of treating the weight estimation process from an algebraic point of view, an attempt is made to consider the same from the principles of linear programming techniques. The performance of the proposed technique is analysed in the context of fingerprint and voice biometrics using sum rule of fusion. The weight factor is optimized against the recognition accuracy. The optimizing parameter is estimated in the training/ validation phase using Leave-One-Out Cross Validation (LOOCV) technique. The proposed biometric solution can be be easily integrated into any multibiometric system with score level fusion. More over, it finds extremely useful in applications where there are less number of available training samples.

1 Introduction

In this paper, the primary focus is on the determination of an optimal integration weight (weighting factor) for the score level fusion of fingerprint and voice biometrics. Although the recognition accuracy of the voice biometrics is high in clean conditions, its performance tends to be significantly degraded under the presence of background noise. This is not desirable in real world applications. Integration weight depends on the reliability of the voice biometric since the feature vector varies with the amount of acoustic noise [1, 2]. At high SNR (Signal to Noise Ratio), the voice matcher outperforms the fingerprint matcher and the final decision heavily relies on the score values of the voice matcher. When the voice biometric is contaminated by noise, the fingerprint matcher outperforms the voice matcher. In this case the score values of the fingerprint matcher contribute more to the final decision. Therefore, it is crucial to estimate the optimal weight factor dynamically, to combine both the modalities for the maximum recognition accuracy. Previously, the integration weight estimation was treated from an algebraic point of view [2]. Here, our focus is to consider the same from the foundations of linear programming techniques. Intelligent fusion of information from the two modalities, requires an optimization approach. The estimation of optimal integration weight is important because it determines the amount of contribution of each modality

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 833–842.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

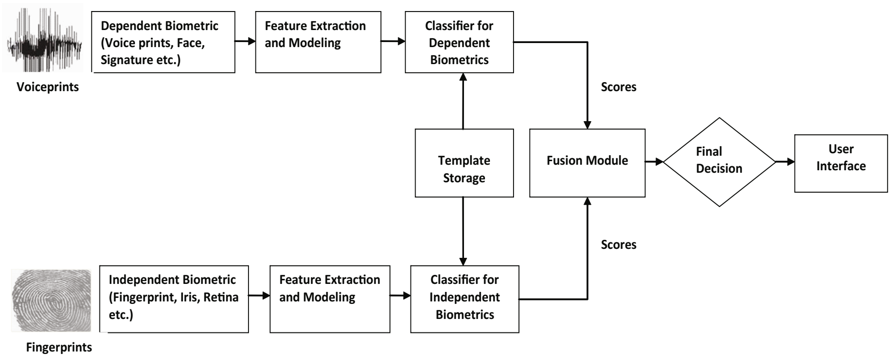

Voiceprints
Dependent Biometric (Voice prints, Face, Signature etc.)
Feature Extraction and Modeling
Classifier for Dependent Biometrics
Scores
Template Storage
Fusion Module
Final Decision
User Interface
Scores
Fingerprints
Independent Biometric (Fingerprint, Iris, Retina etc.)
Feature Extraction and Modeling
Classifier for Independent Biometrics

Fig. 1. Block diagram representing score level Fusion

towards the final decision. Otherwise, the system performs attenuating fusion. We are motivated by [1, 2 and 4] to develop a bimodal biometric system that is more robust to environmental and sensor noise. Fig. 1 shows the overall block diagram of the score level fusion strategy.

2 Proposal

As the inputs tend to be noisy, the sum rule of fusion is proved to be more effective [5]. But, this fusion strategy results in attenuating fusion, when the integration weights are not optimal. One can manually determine the integration weight, but this is a non-optimal approach which needs many iterations and often needs a lot of expertise [3]. In order to emphasize or de-emphasize the scores obtained from the unimodal systems, the integration weight factor must be adaptive and optimal. We can automate the estimation of the "best available weight factor", by formulating the problem using mathematical programming or optimization techniques. The proposed method systematically chooses the best weight factor β from a defined domain $(0 \leqslant \beta \leqslant 1)$ so as to maximize the objective function (recognition accuracy). The performance of the proposed scheme is compared with that of the baseline system (equal weight bimodal biometric system). To show the effectiveness of the proposed technique, we evaluated the performance of the system with a direct search optimization method (Grid Search) and a random search optimization method (Genetic Algorithm). Robust feature vectors were extracted from the two biometric traits. We considered minutiae based system that uses elastic matching algorithm for fingerprint matching [6]. MFCC features were extracted from the voice samples and Gaussian mixture model (GMM) was considered for representing the acoustic feature vectors [7]. Model with 16 MFCC feature vectors and 12 Gaussian mixtures were used. Score normalization is done to transform the match scores of the two matchers into a comparable (common) domain.

$$normalized{-}score = \frac{unnormalized\_score - min\_score}{max\_score - min\_score} \tag{1}$$

The optimal integration weight was obtained in the training/ validation stage. The normalized match scores from the two modalities were combined by the weighted sum rule to produce the final decision. Given the speaker scores $S^{(sc)}$ and the finger scores $S^{(fc)}$, the fused scores could be obtained by linearly combining the two scores.

$$S^{(fus)} = \beta S^{(sc)} + (1-\beta)\, S^{(fc)} \tag{2}$$

The weighting factor $\beta(0 \leq \beta \leq 1)$ determines how much each modality contributes to the final decision. Here, the focus is to optimize the weight factor β so as to get the highest recognition accuracy under all SNR conditions. The objective function is given by:

$$RecognitionAccuracy = -\frac{\sum diag(C_{Mat})}{\sum\sum(C_{Mat})} \times 100 \tag{3}$$

where C_{Mat} is the confusion matrix. Leave-one-out cross validation strategy was employed for the estimation of the optimal integration weight.

2.1 The One-Dimensional Grid Search

This methodology involves setting up of grids in the decision space and evaluating the values of the objective function at each grid point. The point which corresponds to the best value of the objective function is considered to be the optimum solution. The 1-dimensional grid search method can be formulated as the mapping f: $R^1 \longrightarrow R^1$ such that L $\leqslant \beta_1 \leqslant$ .....$\leqslant \beta_n \leqslant$ U, where $\beta_1, \ldots, \beta_n$ are the n test points [8]. This method determines the minimum of a real valued function based on the initial estimate of the location of the minimum point from the lower (L) and upper (U) bounds of the decision variable β. The number of test points n, in each iteration step, determines the rate of convergence of the algorithm.

2.2 Genetic Algorithm

Genetic Algorithm (GA) is a directed random search technique that is modelled on the natural evolution/ selection process towards the survival of the fittest. Individuals standing for possible solutions are often compared to chromosomes and represented by strings of binary numbers. In every generation, a new set of artificial individuals is created, using the information from the best of the old generation. The algorithm consists of initialization, evaluation, reproduction (selection), cross over and mutation. GA is expected to find the global minimum solution even in the case where the objective function has several extrema, including local extrema and saddle points [9]. The final solution gives the integration weight β for the score level fusion.

3 Simulation Results and Discussions

Finger images from the FVC2002 fingerprint database [10] and voice samples from ELSDSR database [11] have been employed for the experimentation. ELSDSR data base contains nine text independent speech samples of twenty three persons. So, finger images of twenty three different persons with nine impressions per finger were considered. Out of these nine samples per biometric, seven samples were used for training the individual classifiers and two samples were used for testing. As the fingerprint biometric is more robust, the performance of the system under varying noise conditions was not considered. The performance of the weak, voice biometric system under varying noise conditions was investigated by artificially degrading the test samples with additive white Gaussian noise. The performance of the system with varying SNR conditions is considered systematically from the feature extraction and model building stage to the testing stage. MFCC feature vectors of the order 12, 16 and 20 and the GMM with 12 and 16 mixtures were considered for the simulation studies, as they are widely used. Different model combinations were considered to select the best model that gives better recognition accuracy. The voice model with 16 MFCC feature vectors with 12 Gaussian mixtures gives improved recognition accuracy under normal operating conditions (10 db SNR to 20 db SNR). So, this model combination was considered for the subsequent analysis. The output of the different classifiers were consolidated into a single vector of scores using the sum rule of fusion. We explored the sum rule of fusion with equal and different weights to the two modalities.

3.1 Fusion with Equal Weights

In this case a constant value of $\beta = 0.5$ was assigned as an integration weight at all SNR levels. The score transformation could be achieved by the following equation.

$$S^{(fus)} = \frac{1}{n}\sum_{j=1}^{n} S_j \qquad ;n=2 \tag{4}$$

$$S^{(fus)} = \frac{1}{2}\left[S^{(sc)} + S^{(fc)}\right] \tag{5}$$

Table 1. Fusion using sum rule with equal weighting

No.	SNR in dB	Accuracy of Fingerprint Classifier	Accuracy of Voice Classifier	Combined Accuracy
1	20	95.6522	98.6956	95.6522
2	15	95.6522	91.4493	95.6522
3	10	95.6522	69.4203	95.6522
4	5	95.6522	33.1884	95.6522
5	0	95.6522	23.1884	90.0000
6	-5	95.6522	9.1304	77.6812
7	-10	95.6522	5.6521	66.8116

Table 2. Fusion with manually determined integration weight

No.	SNR in dB	Accuracy of Fingerprint Classifier	w_f	Accuracy of Voice Classifier	w_s	Combined Accuracy
1	20	95.6522	0.3	98.6956	0.7	100.0000
2	15	95.6522	0.35	91.4493	0.65	97.8261
3	10	95.6522	0.35	69.4203	0.65	97.2464
4	5	95.6522	0.65	33.1884	0.35	95.6522
5	0	95.6522	1	23.1884	0	95.6522
6	-5	95.6522	1	9.1304	0	95.6522
7	-10	95.6522	1	5.6521	0	95.6522

Table 3. Training/ Validation accuracy of individual classifiers

Modality		Voice							Fingerprints
SNR in dB	Set.No.	-10	-5	0	5	10	15	20	Clean
Accuracy	Val.1	9.3478	16.087	33.9130	50.6522	65.6522	87.1738	94.5652	95.6522
	Val.2	4.3478	6.7391	25.8696	44.7826	70.2174	86.5217	96.5218	91.3043
	Val.3	6.7391	20.0000	27.6087	42.1739	66.9565	84.5652	99.1304	91.3043
	Val.4	8.2609	12.8261	18.4783	33.9130	58.0435	87.8260	100.0000	86.9565
	Val.5	7.8261	20.2174	21.9565	30.2174	59.3478	76.0869	95.2174	60.8696
	Val.6	4.3478	4.5652	23.9131	51.5217	81.0870	95.2174	100.0000	86.9565
	Val.7	9.7825	13.2609	16.3044	45.6522	74.7826	95.4348	98.2609	95.6522
Average		7.2360	13.3851	24.0062	42.7019	68.0124	87.5465	97.6708	86.9565

where $S^{(fc)}$ - Finger Scores ; $S^{(sc)}$ - Speaker Scores. This technique will not favour one modality over another. More over the combined recognition accuracy may not be maximum always. The training accuracy of the bimodal system is shown in Table 1.

3.2 Manually Determined Integration Weight

This is a heuristic approach and often needs lot of expertise. Weighted average of the individual scores were considered here. The integration weights were randomly chosen on a trial and error basis for different SNR conditions and the weights that give higher accuracy were considered as the best choice . Sum rule of fusion with integration weights can be represented as follows.

$$S^{(fus)} = \frac{1}{n}\sum_{j=1}^{n} w_j S_j \qquad ; n = 2 \tag{6}$$

where $\sum_{j=1}^{n} w_j = 1$. This process required lot of time and effort and the results are not guaranteed to be optimal. The training accuracy of the manually determined integration weight is shown in Table 2. w_f and w_s are the individual weights given to the fingerprint and voice modality respectively. The result shows that for different noise conditions we need different weighting factors for both the modalities.

Table 4. Training/ Validation accuracy for integration weight optimization

Modality No.	SNR in dB	Accuracy of the Classifiers Fingerprints	Voice	Grid Search Combined Accuracy	β	Genetic Algorithm Combined Accuracy	β
1	20	86.9565	98.1366	100.0000	0.7071	100.0000	0.8087
2	15	86.9565	86.9565	95.6522	0.6928	95.0311	0.7387
3	10	86.9565	65.2174	92.5465	0.6500	92.0156	0.6605
4	5	86.9565	43.4783	87.5776	0.1255	87.5776	0.1255
5	0	86.9565	23.9131	86.9565	0.0000	86.9565	0.0000
6	-5	86.9565	13.0435	86.9565	0.0000	86.9565	0.0000
7	-10	86.9565	8.6957	86.9565	0.0000	86.9565	0.0000

Table 5. Testing accuracy with optimal integration weight

Modality No.	SNR in dB	Accuracy of the Classifiers Fingerprints	Voice	Grid Search Combined Accuracy	β	Genetic Algorithm Combined Accuracy	β
1	20	95.6522	98.6956	100.0000	0.7071	100.0000	0.8087
2	15	95.6522	91.4492	97.8261	0.6928	97.8261	0.7387
3	10	95.6522	69.4202	97.2464	0.6500	97.1015	0.6605
4	5	95.6522	33.1884	95.6522	0.1255	95.6522	0.1255
5	0	95.6522	23.1884	95.6522	0.0000	95.6522	0.0000
6	-5	95.6522	9.1304	95.6522	0.0000	95.6522	0.0000
7	-10	95.6522	5.6521	95.6522	0.0000	95.6522	0.0000

3.3 Leave-One-Out Cross Validation (LOOCV)

As the number of available biometric samples were limited, LOOCV strategy, was employed to fine tune the training/ validation phase and estimate the best optimal weight. LOOCV involves partitioning the data samples into complementary subsets and using a single observation from the data sample as validation data, and the remaining observations as training data. To reduce variability, multiple rounds of cross-validation were performed using different partitions, and the validation results were averaged over the rounds. Here, seven training samples from each modality were used for cross validation. So seven cross validation sets were considered from both modalities (Val.1 to Val.7). We choose each sample from the training set, for cross validation testing and the remaining 6 samples for cross validation training one at a time. The average training/validation accuracy of the individual classifiers is depicted in Table 3.

For instance, let us consider the estimation of the weight factor for 20 db. After performing the validation (testing), match score outputs from each validation set was stored separately. As we degraded the voice validation test samples with stationary noise, the match score values thus obtained exhibits a random behaviour. So, in order to arrive at a conclusive result, twenty separate experiments were performed with each validation data, so that each validation set possesses twenty voice matching scores. Score level fusion was performed with the matching scores thus obtained using equation 2, so as to

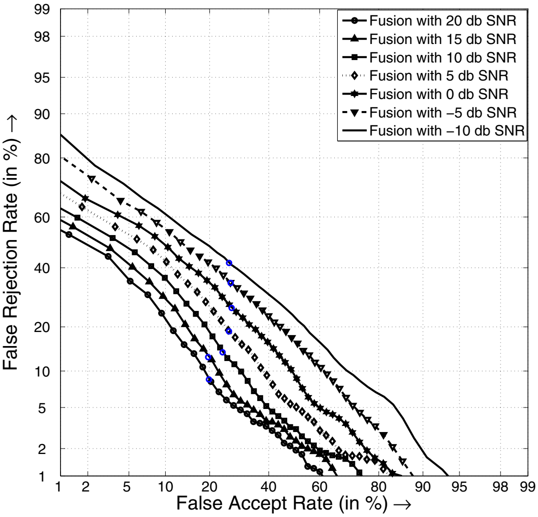


Fig. 2. Performance of the baseline system

maximize the objective function (recognition accuracy). That is, scores obtained from fingerprint validation set1 were fused with the scores obtained from all the voice validation sets. For each experiment, using grid search method and genetic algorithm, the algorithms returned the optimal β that maximizes the recognition accuracy. Hence, at the end of first round cross validation with fingerprint validation set1 (with all the validation sets from the voice modality), we had twenty fused scores along with twenty optimal β's from each set. In order to get a representative β, we considered the median values from each set. Finally we took the median of all these representative β's to get a more conclusive result. We preferred median values to the mean values, because of the fact that mean is to a great extent affected by the extreme values of the observation sets. More over the mode will not give a representative value for the real data sets. In such a case, median values gives better measure of the central tendency than the mean values and the mode. Similar experiments were repeated with all other finger validation sets on the voice validation scores. Thus multiple round of cross validation was performed to reduce the variability in the estimation process and the β's thus obtained were averaged over the rounds. Even though the computational complexity inherent in the leave-one-out strategy seems to be more, better tuning could be achieved by the process.

3.4 Training and Testing Performance

From the validation stage, we obtained optimal integration weights (β's) for different noise conditions from -10db to 20 db. The overall validation accuracy and the optimal integration weight β estimated for the various SNR conditions is shown in the Table 4

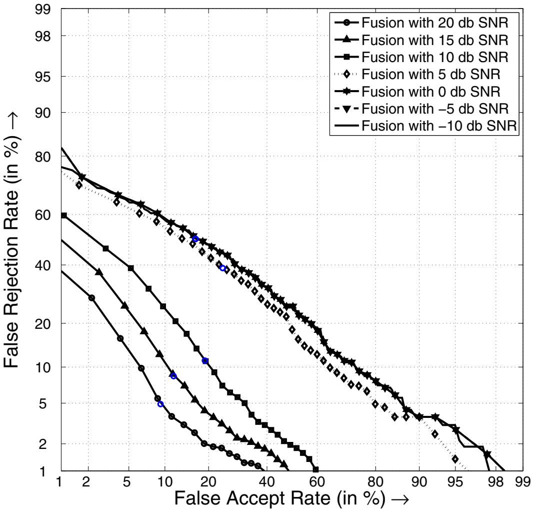


Fig. 3. DET performance curve for grid search based optimization

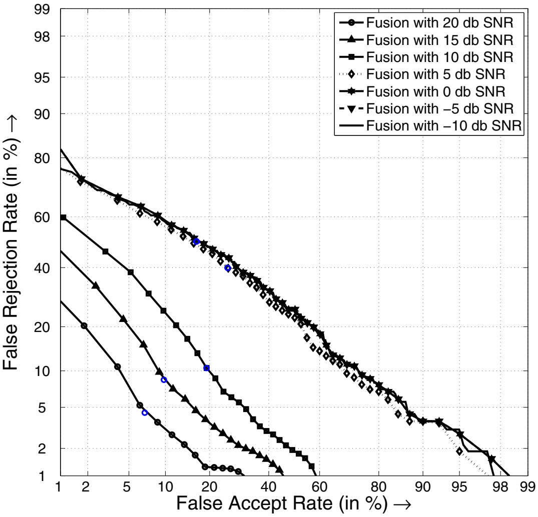


Fig. 4. DET performance curve for genetic algorithm based optimization

The β values thus estimated during the training/validation stage is used for testing. The overall testing accuracy is depicted in Table 5. The proposed method shows better performance than the base line system (sum rule of fusion method with equal weighting). With the base line method, the combined accuracy resulted in attenuating fusion for 20db, 0db, -5db and -10db and maintained the accuracy of the better unimodal system for all the other cases. The proposed method shows higher accuracy than any of the unimodal systems in the normal operating conditions and maintained the accuracy of the better unimodal ones for all adverse conditions. Further insight could be obtained from the DET plots [12]. Fig. 2., Fig. 3. and Fig. 4. show the DET plots for the baseline system and the proposed method respectively. It is clear from the figure that FAR (False Acceptance Rate) and FRR (False Rejection Rate) gets considerably reduced at the normal operating conditions for the proposed method. Even though the recognition accuracy remains same for the grid search and the genetic algorithm based method, lesser FAR and FRR could be achieved with genetic algorithm based method. This is evident from the respective DET plots.

4 Conclusion and Future Directions

Integration weight optimization technique is proposed here for improving multibiometric system performance under various noise conditions. We studied the performance of the proposed technique in the context of fingerprint and voice biometrics using score level fusion. The proposed method could successfully eliminate the attenuating fusion under various noise conditions. This method is straightforward without any theoretical complexities. Moreover, by estimating the optimal integration weight using linear programming and leave-one-out cross validation techniques, we could automate the process and make the system more robust to fluctuating inputs. Thus, the proposed method could very much reduce the computational complexity involved in the determination of the optimal integration weight. The proposed scheme could be easily integrated with any biometric modalities, where the accuracy and robustness depends on the quality of biometric samples at hand. This method finds extremely useful in applications where there are less number of training samples. One of the demerits of the proposal is that, at the extreme noise conditions the fusion module contributes zero weighting to the voice modality, to eliminate the attenuating fusion. Under these conditions, the overall performance depends solely on the fingerprint matcher. Our future concentration is to develop a multibiometric system that can improve the recognition accuracy at extreme noise conditions, than any of the unimodal systems, while maintaining reduced FAR (False Acceptance Rage) and FRR (False Rejection Rate).

References

1. Kekre, H.B., Bharadi, V.A.: Ageing Adaptation for Multimodal Biometrics using Adaptive Feature Set Update Algorithm. In: IEEE International Advance Computing Conference (IACC 2009), Patiala, India, March 6-7 (2009)

2. Toh, K.-A.: Fingerprint and speaker verification decisions fusion. In: International Conference on Image Analysis and Processing (ICIAP), Mantova, Italy, pp. 626–631 (September 2003)
3. Rajavel, R., Sathidevi, P.S.: Adaptive Reliability Measure and Optimum Integration Weight for Decision Fusion Audio-visual Speech Recognition. Springer J. Sign. Process. Syst. (February 2011)
4. Rajavel, R., Sathidevi, P.S.: The Effect of Reliability Measure on Integration Weight Estimation in Audio-Visual Speech Recognition. International Journal of Engineering Science and Technology 2(8) (2010)
5. Ross, A., Nandakumar, K., Jain, A.K.: Handbook of Multibiometrics. Springer, New York (2006)
6. Wuzhili: Finger Print Recognition, Honors Thesis (2002)
7. Reynolds, D.: Gaussian Mixture Models* MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA
8. Kim, J.: Iterated Grid Search Algorithm on Unimodal Criteria. Ph.D. Thesis, Blacksburg, Virginia (1997)
9. Yang, W.Y., Cao, W., Chung, T.-S., Morris, J.: Applied Numerical Methods using Matlab. Wiley, India (2007)
10. FVC 2002, the second International Competition for Fingerprint Verification Algorithms, FVC 2002 (2002), `http://bias.csr.unibo.it/fvc2002/`
11. Feng, L.: Speaker Recognition, Informatics and Mathematical Modelling, Technical University of Denmark, DTU (2004)
12. Martin, A., Doddington, G., Kamm, T., Ordowsk, M., Przybocki, M.: The DET Curve in Assessment of Detection Task Performance. In: Proc. Eurospeech 1997, Rhodes, pp. 1895–1898 (1997)

A Two Stage Combinational Approach for Image Encryption

H.S. Sharath Kumar, H.T. Panduranga, and S.K. Naveen Kumar

Department of Electronics, University of Mysore, P G Center,
Hemagangothri, Hassan, Karnataka
{sharath.kr83,nave12}@gmail.com, ht_pandu@yahoo.co.in

Abstract. Information security is becoming more important in the field of information processing and management. Due to the fast development in the communication, a wide verity of information are transmitting in multiple medias like telephone, mobile, television, satellite communication, optical communication. Since image is also a very important information and there is a need for security in transmitting the image in most of the applications like satellite image transmissions, military applications, medical application and teleconferencing etc. In this paper, we describe a method for image encryption which has two stages. In first stage, each pixel of an image is converted to its equivalent eight bit binary number and in that eight bit number, number of bits equal to the length of password are rotated and then reversed. In second stage, a carrier image generated from the same password is added on the resultant image of first stage to which results final encrypted image.

1 Introduction

With the fast development in communication and information technology, huge data is transmitted over a communication channel which needs security. There are many applications like information storage, information management, patient information security, satellite image security, confidential video conferencing, telemedicine, military information security and many other applications which require information security. Komal D Patel and Sonal Belani [1] have presented a survey on existing work which is used different techniques for image encryption and also given a general introduction about cryptography. There are several methods for image encryption with some advantages and disadvantages. Ismet Ozturk and Ibrahim Sogukpinaar [2] have discussed the analysis and comparison of image encryption algorithms. And they classify the image encryption methods in to three major types: position permutation, value transformation and visual transformation. Borko Furht et. al. [3] have discussed about some of the most important techniques for image encryption based on encrypting only certain parts of the image in order to reduce the amount of computation. Panduranga H T and Naveenkumar S K [4] have proposed an approach using bit reversal method in which the encryption process is divided into three stages. In first stage each pixel of an image is in the form of decimal number is converted in to its equivalent eight-bit binary number. In second stage this eight-bit binary number is arranged in reverse order. In third stage this new binary number is get converted in to its

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 843–849.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

equivalent decimal number. Panduranga H.T. et. al. [5] have proposed an approach for image encryption using dual carrier image which improves the efficiency of encryption. S. R. M. Prasanna et. al [6] have presented an image encryption method with magnitude and phase manipulation using carrier images. Here they used the concept of carrier images and one dimensional Discrete Fourier Transform for encryption purpose and it deals with private key cryptosystem, works in the frequency domain.

Organization of the paper is as follows: Section 2 explains the image encryption technique by using bits rotation and reversal methods based on password and also explains the creation of carrier image based on the same password. Results and discussions are explained in Section 3. This paper is concluded by providing the summery of the present work in section 4.

2 Proposed Technique

In this method, encryption process is divided into two stages. For two stage approach only one password is used as encryption key. In first stage, a password is taken along with input image. Value of each pixel of input image is converted into equivalent eight bit binary number. Now length of password is considered for bit rotation and reversal. i.e., Number of bits to be rotated to left and reversed will be decided by the length of password. Let L be the length of the password and L_R be the number of bits to be rotated to left and reversed (i.e. L_R is the effective length of password). The relation between L and L_R is represented by equation (1).

$$L_R = L \bmod 7 \tag{1}$$

where ‘7’ is the number of iterations required to reverse entire input byte.

For example, $P_{in}(i,j)$ is the value of pixel of an input image. $[B_1\ B_2\ B_3\ B_5\ B_6\ B_7\ B_8]$ is equivalent eight bit binary representation of $P_{in}(i,j)$.

i.e. $P_{in}(i,j) \xrightarrow{decimal\ to\ 8\ bit\ binary} [B_1B_2B_3B_4B_5B_6B_7B_8]$

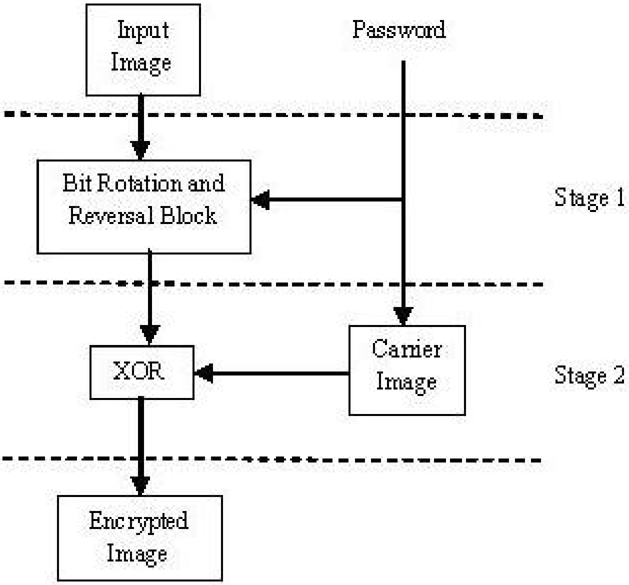
If L_R=5, five bits of input byte are rotated left to generate resultant byte as$[B_6B_7B_8B_1B_2B_3B_4B_5]$. After rotation, rotated five bits i.e. $B_1B_2B_3B_4B_5$, get reversed as $B_5B_4B_3B_2B_1$ and hence we get the resultant byte as shown below.

$$[B_6B_7B_8B_5B_4B_3B_2B_1] \xrightarrow{8\ bit\ binary\ to\ decimal} P_{out}(i,j)$$

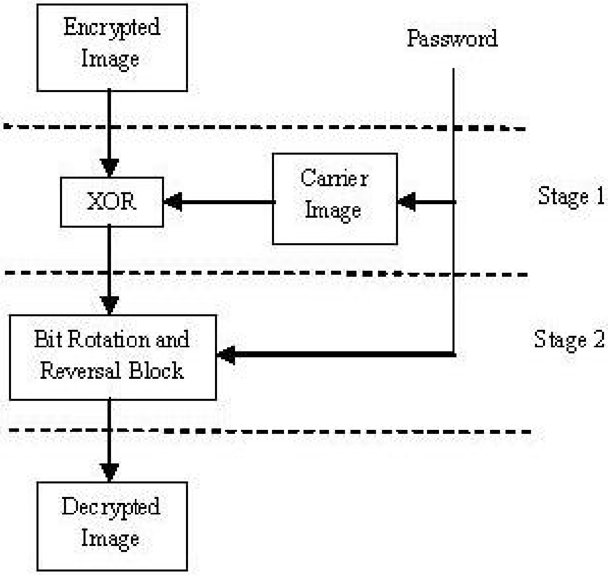
where $P_{out}(i,j)$is the value of output pixel of resultant image.

Since the weight of each pixel is responsible for its colour, this process change in weight of each pixel generates encrypted image. At the end of first stage, an encrypted image is generated, but this encrypted image can be decrypted by other passwords of same length as original password. To avoid this inconvenience second stage of encryption has designed.

At second stage, a carrier image is generated from the password which has entered in first stage. ASCII value of each element of password is arranged in a matrix form of size equal to the size of input image. Entering the ASCII values of password in



1(a).



1(b).

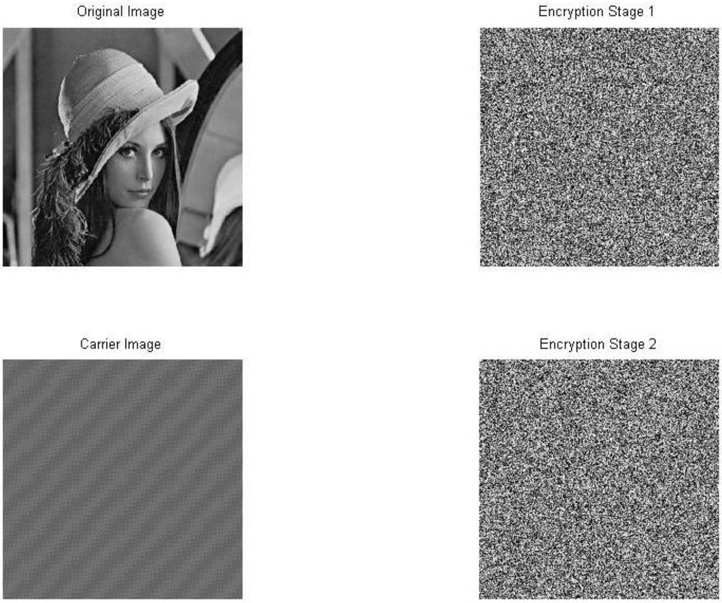
Fig. 1. (a). Block Diagram of Image Encryption Process. (b). Block Diagram of Image Decryption Process.

that matrix repeats till entire matrix is filled to generate a carrier image with the size equal to input image. *Bitwise-Exclusive-OR* operation is applied on the carrier image generated in this stage and encrypted image generated in first stage to generate final encrypted image.

Fig. (a, b). shows block diagram representation of Image Encryption and Decryption techniques respectively. As shown in Fig. 1(b). decryption process is carried out in two stages. In first stage, the carrier image generated by password undergoes *Bitwise-Exclusive-OR* operation with the encrypted image. Second stage of decryption process is similar to first stage of encryption process. But case of encryption, bits of input byte were rotated to left and then reversed, whereas in decryption process, bits of input byte are rotated to right and then reversed.

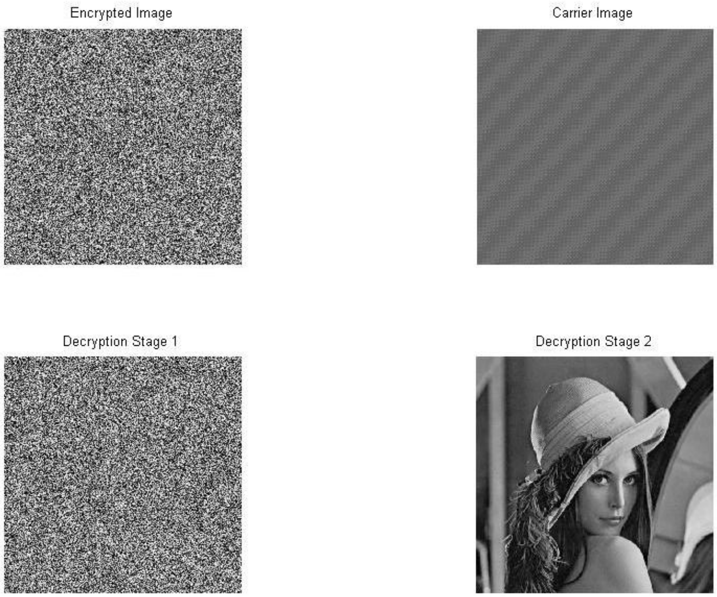
3 Results and Discussions

Here, the above mentioned technique is implemented for different images and passwords. We have observed that for good quality of encryption, minimum effective length of the password (L_R) should be four.

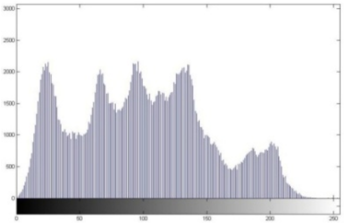


2(a)

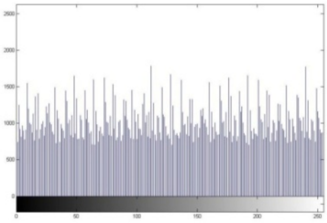
Fig. 2. (a). Encryption Process: Lena Image as input, Encrypted Image in first stage, Carrier Image generated by password “sharath”, Encrypted Image in second stage. (b). Decryption Process: Encrypted Image as input to Decryption process, Carrier image generated by password “sharath”, Decryption in first stage, Decryption In second stage. (c,d). Histograms of Input Image and final Encrypted image respectively.



2(b)



2(c)



2(d)

Fig. 2. (*continued*)

From Fig. 2(a). it can be observed that as the length of password "sharath" is seven, all the bytes in each pixel are reversed and hence encryption level is effective. Where as in Fig. 3. length of password is three, and hence encryption is not so effective.

Fig. 4. (a, b) is the result of implementation of encryption and decryption process for two passwords of same length. Since both passwords are of same length, at the final decryption stage image is partially decrypted though both passwords are different.

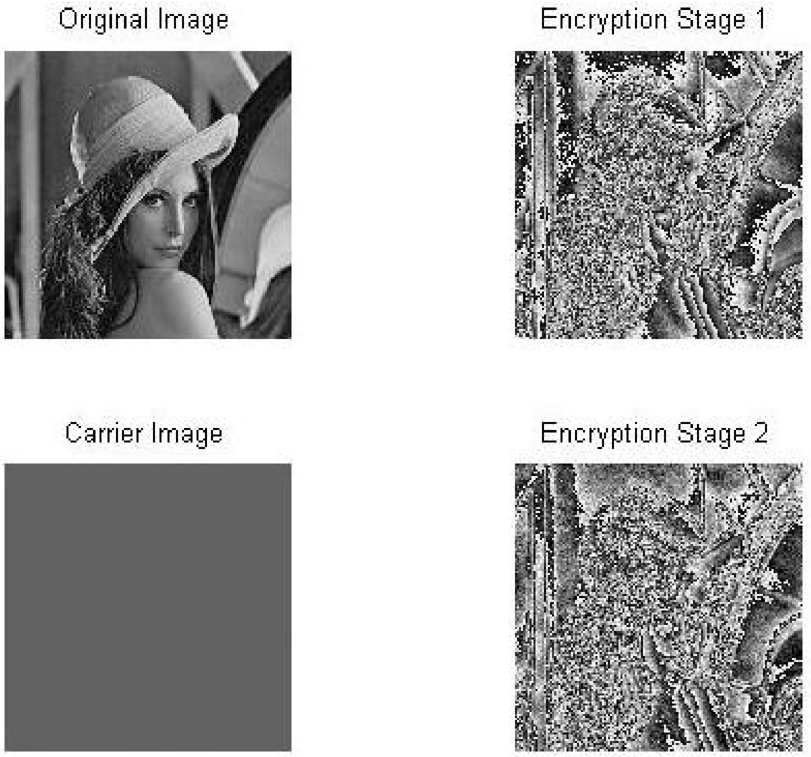


Fig. 3. Input Image, Encrypted Image at first stage, Carrier Image generated by password “abc”, Encrypted Image at second stage

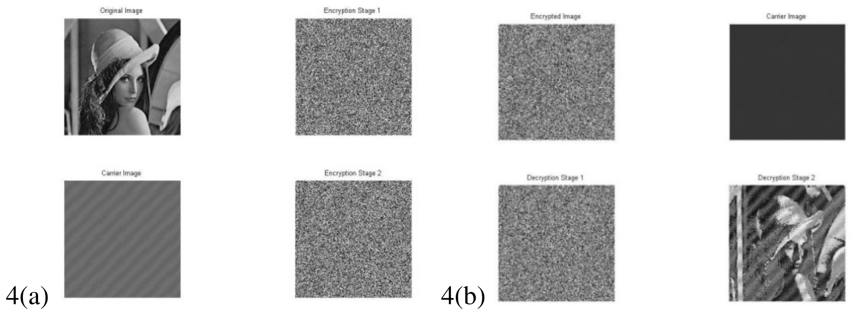


Fig. 4. (a). Encryption Process with password “sharath”. (b). Decryption process with password “1234567”.

4 Conclusion

In this paper, we explained the two stage image encryption technique which need only one password for both stages. In first stage based on password length, each pixel of image gets rotated and reversed. In second stage, a carrier image generated by the same password added with resultant image of first stage to generate final encrypted image.

Acknowledgment. The work discribed in this paper is supported by a grant from the *University Grants Commission,* New Delhi, India.

References

1. Patel, K.D., Belani, S.: Image encryption using different techniques: A review. International Journal of Emerging Technology and Advanced Engineering 1(1) (November 2011) ISSN 2250-2459
2. Ozturk, I., Sogukpinaar, I.: Analysis and Comparison of Image Encryption Algorithms. Transaction on Engineering, Computer and Technology 3, 38–42 (2004)
3. Furht, B., et al.: Multimedia Encryption and watermarking, Part II, pp. 79–120
4. Panduranga, H.T., Naveenkumar, S.K.: An image encryption approach using bit-reversal method. In: NCIMP 2010, pp. 181–183 (2010)
5. Panduranga, H.T., et al.: A New Image Encryption approach with dual carrier images. In: Proceedings of ICACT, pp. 478–481 (December 2008)
6. Prasanna, S.R.M., et al.: An Image Encryption Method with Magnitude and Phase Manipulation using Carrier Images. IJCS 1(2), 132–137 (2006)

Wavelet SIFT Feature Descriptors for Robust Face Recognition

Nerella Arun Mani Kumar and P.S. Sathidevi

Department of Electronics and Communication Engineering, NIT Calicut, Kerala, India
arunmanikumar@gmail.com, sathi@nitc.ac.in

Abstract. This paper presents a new robust face recognition technique based on the extraction and matching of Wavelet-SIFT features from individual face images. Here, Biorthogonal wavelet 4.4 is employed as the basis for Discrete Wavelet Transform of the images. Then, SIFT Face recognition method is applied on LL and HH sub band combination of images for recognition. The results obtained with the proposed method are compared with basic SIFT face recognition and classic appearance based face recognition technique (PCA) over three face databases: Nottingham database, Aberdeen database and Iranian database.

Keywords: Scale Invariant Feature Transform (SIFT), Wavelet Transform, Face Recognition, Principal Component Analysis (PCA).

1 Introduction

Face recognition is one of the most challenging research areas over last 40 years [11]. The most popular face recognition approaches make use of appearance-based projection methods, like PCA (Principal Component Analysis) [9], LDA (Linear Descriminant Analysis) [10] or ICA (Independent Component Analysis) [2]. However, the above algorithms are mostly sensitive to light, expression, pose etc. However, recent research interest on point detectors and invariant descriptors has opened a new alternative for model-based face recognition and authentication. In this paper, we propose an algorithm based on Wavelet-SIFT descriptors.

SIFT (Scale Invariant Feature Transform) descriptors were initially developed for object recognition purposes [8] and then, also used for others, like robot navigation [11], scene classification [7], etc. Later, SIFT features have also been used by different authors in the field of face recognition [3], with promising results. It has the advantage that feature descriptors are invariant to scale, rotation and affine transformation, etc. and has strong robustness for the occlusion problem. In this method, face image features are extracted using SIFT and then compared the test face image features with features of trainee database to recognize the face. To improve the accuracy and to reduce the time for recognition, we propose a new method based on Wavelet-SIFT descriptors for face images. Here, we apply Biorthogonal wavelet transform [6] to an image before it undergoes SIFT operation. Then we take the combination of LL and HH sub bands for SIFT operation.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 851–859.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

The paper is organized as follows. Section 2 describes SIFT. In section 3, proposed method is detailed. Experimental evaluation and results are presented in Section 4. Conclusion and future work are described in section 5.

2 Scale Invariant Feature Transform

Scale Invariant Feature Transform (SIFT) algorithm [5] has been proposed for extracting features that are invariant to rotation, scaling and partially invariant to changes in illumination and affine transformation of images to perform matching of different views of an object or scene. Steps for extracting SIFT features as follows.

Firstly, extreme values are detected at the different scales of the image, and are the keypoint candidates. Secondly, Taylor series and Hessian matrix are used to determine stable keypoints; thirdly, the gradient orientation is assigned to the keypoint by using its neighborhood pixels, and finally, keypoint descriptor is obtained.

2.1 Detection of Scale Space Extrema

The first stage of keypoint detection is to identify locations and scales that can be repeatedly assigned under differing views of the same object. The scale space of an image is defined as a function L(x,y,σ) that is produced from the convolution of a variable scale Gaussian function G(x,y,σ) with an input image I(x,y).

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{1}$$

Where, σ is scale of blurring.

The 2D Gaussian kernel is given by

$$G(x, y, \sigma) = \frac{1}{2\Pi \sigma^2} e^{-(x^2 + y^2) / 2\sigma^2} \tag{2}$$

To efficiently detect the stable keypoint locations, we use scale space extrema in the difference of gaussian (DOG) function convolved with the image, and is computed from the difference of two nearby scales separated by a constant multiplicative factor **k** ,

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{3}$$

In order to find the local extrema (maxima or minima) of D(x,y,σ), each sample point is compared to its eight neighbors in the current image and nine neighbors in the scale above and below the image. Local extrema is selected only if the current pixel is larger or smaller than the remaining pixels and is discarded otherwise.

2.2 Keypoint Localization

The location of keypoint is considered to filter the keypoints which are sensitive to noise or no edge effect in this process. The reference [4] shows that, according to Taylor quadratic expansion, DOG(x,y,σ) can delete the extreme points which have lower contrast and the value of Hessian vector and the ratio of determinant can reduce the edge effect.

2.3 Orientation Assignment

After the position and scale of the keypoint is determined, the next step is keypoint's direction, which can ensure the feature's rotation invariance. The direction is calculated by the image information of keypoint's neighborhood. For each extreme point, L(x, y) at this scale, the gradient magnitude m(x, y) and orientation θ(x, y) are computed using

$$m(x,y)=\sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2}$$

$$\theta(x,y)=\tan^{-1}((L(x,y+1)-L(x,y-1))/(L(x+1,y)-L(x-1,y))) \tag{4}$$

An orientation histogram is formed from the gradient orientations of sample points within a region around the keypoint. The orientation histogram has 36 bins covering the 360 degree range of orientations. Each sample added to the histogram is weighted by its gradient magnitude and by a Gaussian-weighted circular window with a σ that is 1.5 times that of the scale of the keypoint.

Peaks in the orientation histogram correspond to dominant directions of local gradients. The highest peak in the histogram is detected, and then any other local peak that is within 80% of the highest peak is used to create a keypoint with that orientation. Therefore, for locations with multiple peaks of similar magnitude, there will be multiple keypoints created at the same location and scale but at different orientations. Only about 15% of points are assigned multiple orientations, but these contribute significantly to the stability of matching. Finally, a parabola is fit to the 3 histogram values closest to each peak to interpolate the peak position for better accuracy.

2.4 Keypoint Descriptor

The descriptor for keypoint is highly distinctive and invariant to illumination and 3D view point changes. To generate a keypoint descriptor, consider a 16×16 window around the keypoint and divide it into sixteen 4×4 consecutive windows. Within each 4×4 window, gradient magnitudes and orientations are calculated. These orientations are put into an 8 bin histogram.

Any gradient orientation in the range 0-44 degrees add to the first bin. 45-89 add to the next bin. And the amount added to the bin depends not only on the magnitude of the gradient but also on the distance from the keypoint. This is done by using "Gaussian weighting function" and this generates a gradient. Multiply it with the magnitude of orientations and get weighted thing. Here the purpose of Gaussian window is to avoid sudden changes in the descriptor. Doing this for each 4×4 sub-window, for each window 16 pixels, we compute 16 random orientations into 8 predetermined bins. And totally, we have 4×4×8=16 (4×4 windows) ×8 (each window has 8 bins) =128 numbers, and normalize these numbers. These 128 numbers form the **"feature vector"** of corresponding keypoint.

2.5 Procedure for Face Recognition Using SIFT

a. Train a database and find the features for all images in the database using SIFT algorithm.
b. Input a new face image and find its features.
c. Now, compare the feature descriptors of new face image with feature descriptors of one image in Training database.
d. Calculate the number of nearest matches between two images.
e. Repeat steps c and d for each trainee image in the database and save corresponding number of matches.
f. The Image with more number of matches in trainee database is the recognized image.

The main idea of Face recognition using SIFT is that there is one point of the face which contains a very distinctive features of the subject, which could be found in the test image.

SIFT is a feature transform and extracts features from images. Normally, extracted features are more and also depend on nature of an image. If we have more number of such trainee images then recognizing a face for a test image will take more time for comparison. Time is also a crucial factor in recognition. To reduce time for recognition and to improve recognition accuracy we propose a new method, in which wavelet transform is applied to images before undergoing SIFT operation.

3 Proposed Method

Here, we proposed a method based on wavelet transform, applied on input images before extracting their feature points and descriptors.

The basic thought of wavelet transform [6] is using the same function by expanding and shifting to approach the original signal. The wavelet coefficients carry the time-frequency information in certain areas. It has good local characteristics both in time domain and frequency domain. It can maintain the fine structure of the original images in various resolutions and it is convenient to combine with human vision characteristics.

Compared with the orthogonal wavelet, biorthogonal wavelet has more obvious superiority in image processing because it balances the orthogonality and symmetry. In addition, the reconstructing signal of biorthogonal wavelet transform is suitable to embed watermark for its balance. Here we are using Biorthogonal wavelet 4.4.

Wavelet transform decomposes an image into one low frequency sub band (LL) and three high frequency sub bands (LH,HL,HH). In the proposed algorithm, LL and HH sub bands are used since the image information is preserved by LL component and edge information is preserved by HH component. The effect of facial expression could be attenuated by eliminating high frequency components [1].

3.1 Procedure for Face Recognition Using Wavelet-SIFT

a. For Trainee database, apply Biorthogonal wavelet 4.4 transform, then apply SIFT algorithm to LL, HH sub-band combination of each image and obtain corresponding feature descriptors.
b. For test image, apply Biorthogonal wavelet transform and obtain its feature descriptors by applying SIFT to LL, HH sub-band combination.
c. Now, compare the feature descriptors of new face image with feature descriptors of one image in Trainee database.
d. Calculate the number of nearest matches between two images using nearest neighbourhood criteria.
e. Repeat steps c and d from each trainee image in the database and save corresponding number of matches.
f. The image with more number of matches in trainee database is the recognized image.

Block diagram of the proposed method is shown in Fig. 1.

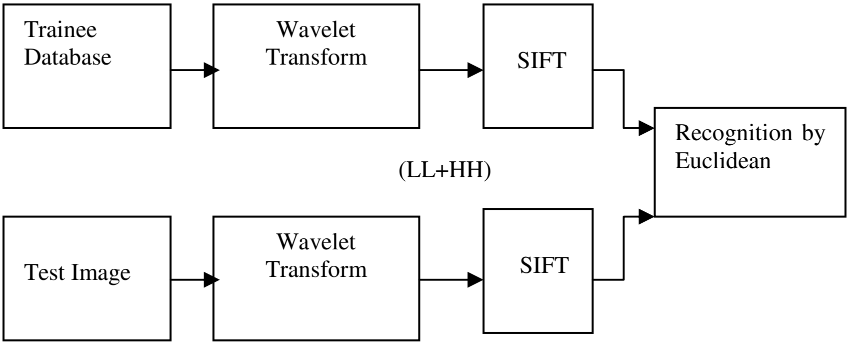


Fig. 1. Block diagram of Wavelet-SIFT method

4 Experimental Evaluation and Results

In this section, we present the experimental evaluation of our proposed method on different databases. Performance of our method is compared with SIFT based face recognition on accuracy and time for recognizing a single image. We run the Matlab code on Intel 3GHz core2dual processor.

Evaluation on Iranian database: Iranian database [12] contains 330 images of 33 subjects, each of 1200×900 dimensions. Each subject has 10 images varies in facial expressions, view point, illumination and affine changes. Here, we are taking different trainee databases depends on number of training samples (K) per subject.

Table 1 summarizes the comparison of recognition rate and time elapsed for recognizing single image with our method and with PCA and SIFT face recognition at different trainee image sizes (K).

Table 1. Results on Iranian database

Algorithm	% accuracy			Time (sec)		
	K=4	K=6	K=9	K=4	K=6	K=9
PCA	12.1	63.63	84.8	7.32	14.8	31.12
SIFT	12.1	81.81	90.9	450.8	780.5	1192.2
Proposed method	18.1	81.81	90.9	47.08	83.2	169.7

From Table 1, for K=4 accuracies of three methods are very less and our pro-posed method has little bit high accuracy than the other methods. Accuracies of three methods are increased with increasing trainee database. But this is not very significant in the case of PCA method. In all the cases, the proposed method takes less time for recognition compared to SIFT.

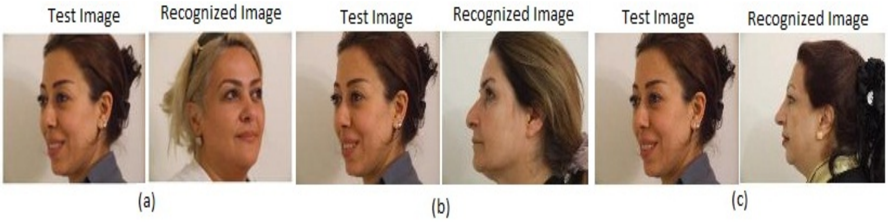


Fig. 2. Results on Iranian database for K=4 a) PCA b) SIFT c) Proposed method

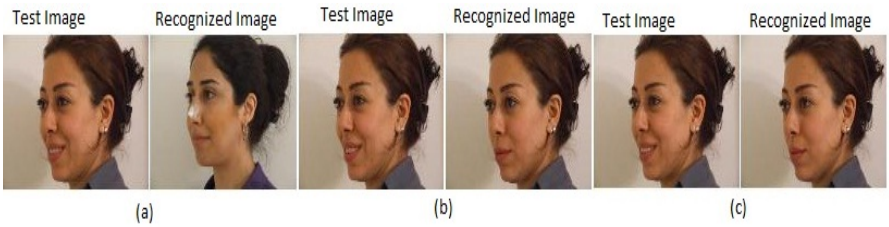


Fig. 3. Results on Iranian database for K=6 a) PCA b) SIFT c) Proposed method

Fig. 4. Results on Iranian database for k=9 a) PCA b) SIFT c) Proposed method

Figure 2 shows recognition results on Iranian database for K=4, where all methods fail to recognize the test image. For K=6, SIFT and proposed method recognizes the image but PCA fail to do, as shown in Fig.3. All the methods recognizes the test image, further increasing trainee data size to K=9 as shown in Fig.4.

Evaluation on Nottingham database: Nottingham database is one database in PICS database [13] which contains 497 images of 71 subjects, each of 288×384 dimensions. Each subject has 7 images varies in facial expression, affine changes and illumination changes. Here, we have taken 4 images per subject as trainee data, and other 3 as test images.

Table 2, summarizes the comparison of recognition rate and time elapsed for recognizing single image with the proposed method and PCA, SIFT face recognition techniques.

Table 2. Results on Nottingham database

Algorithm	% accuracy	Time (sec)
PCA	63.3	45.76
SIFT	97.18	200.78
Proposed method	98.59	85.26

From Table 2, we can observe that PCA takes less time compared to SIFT and proposed method, but accuracy is very less compared to other. Wavelet-SIFT takes less time compared to SIFT based face recognition.

Evaluation on Aberdeen database: Aberdeen database is another database in PICS database [13], which contains 252 images of 63 subjects, each of 400×560 dimensions. Each subject has 4 images varies in illumination and affine changes. We have taken 3 images per subject as trainee database and remaining image as test image.

Table 3 summarizes the comparison of recognition rate and time elapsed for recognizing single image with the proposed method and PCA, SIFT face recognition techniques.

From Table 3, accuracy of PCA is very less compared to other methods though time taken for recognition is less. Proposed method takes less time compared to SIFT method.

Table 3. Results on Aberdeen database

Algorithm	% accuracy	Time (sec)
PCA	40.32	14.13
SIFT	95.16	175.38
Proposed method	95.16	44.85

By comparing the above results with all data bases, it can be seen that PCA face recognition takes less time but the accuracy of the system is very low compared to SIFT and Wavelet-SIFT (proposed method). Our proposed method gives better results than SIFT face recognition in terms of Time. The accuracy of PCA, SIFT, Wavelet-SIFT increases with respect to trainee dataset.

5 Conclusion and Future Work

This paper presents a new approach for face recognition based on Wavelet-SIFT features. Experiments on different databases like Nottingham, Aberdeen and Iranian show that the proposed method gives better performance in terms of time and accuracy than PCA and SIFT based face recognition techniques. Though, PCA based face recognition takes less time for recognition, the accuracy is very low compared to the proposed method. Our method outperforms totally on SIFT based face recognition technique.

In this paper, we concentrated only on facial expressions and illumination changes of face images for face recognition. Our future work concentrates on extending this technique to face identification of distorted/ manipulated images.

References

1. Zhang, B.-L., Zhang, H., Ge, S.S.: Face recognition by applying wavelet subband representation and kernel associative memory. IEEE Transactions on Neural Networks 15(1), 166–177 (2004)
2. Bartlett, M.S., Movellan, J.R., Sejnowski, T.J.: Face recognition by independent component analysis. IEEE Transactions on Neural Networks 13(6), 1450–1464 (2002)
3. Bicego, M., Lagorio, A., Grosso, E., Tistarelli, M.: On the Use of SIFT Features for Face Authentication. In: Conference on Computer Vision and Pattern Recognition Workshop, CVPRW 2006, vol. 35, pp. 17–22 (June 2006)
4. Brown, M., Lowe, D.: Invariant Features from Interest Point Groups. Computer 332(6031), 253–262 (2002)
5. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Key-points. Int. J. Comput. Vision 60(2), 91–110 (2004)
6. Daubechies, I.: TenLectureonWavelets. CBMS, vol. 61. SIAM (1994)
7. Lucey, S., Matthews, I., Hu, C., Ambadar, Z., de la Torre, F., Cohn, J.: AAM derived face representations for robust facial action recognition. In: 7th International Conference on Automatic Face and Gesture Recognition, vol. 2-6, pp. 155–160 (April 2006)

8. Lowe, D.G.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
9. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cognitive Neuroscience 3(1), 71–86 (1991)
10. Belhumeur, P.N., Hespanha, O.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. Pattern Anal. Mach. Intell. 19(7), 711–720 (1997)
11. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Comput. Surv. 35(4), 399–458 (2003)
12. `http://pics.psych.stir.ac.uk/`
13. `http://pics.psych.stir.ac.uk/2D_face_sets.htm`

Image Denoising Based on Neutrosophic Wiener Filtering

J. Mohan[1], A.P. Thilaga Shri Chandra[2], V. Krishnaveni[1], and Yanhui Guo[3]

[1] Department of Electronics and Communication Engineering,
P.S.G. College of Technology, Coimbatore, India
jaimohan12@gmail.com, vk@ece.psgtech.ac.in
[2] Department of Electronics and Communication Engineering,
Sri Krishna College of Engineering and Technology, Coimbatore, India
thilagashrichandra@skcet.ac.in
[3] Department of Radiology, University of Michigan, Ann Arbor, USA
yanhuig@med.umich.edu

Abstract. This paper proposes an image denoising technique based on Neutrosophic Set approach of wiener filtering. A Neutrosophic Set (NS), a part of neutrosophy theory, studies the origin, nature, and scope of neutralities, as well as their interactions with different ideational spectra. Now, we apply the neutrosophic set into image domain and define some concepts and operators for image denoising. Here the image is transformed into NS domain, which is described using three membership sets: True (T), Indeterminacy (I) and False (F). The entropy of the neutrosophic set is defined and employed to evaluate the indeterminacy. The ω-wiener filtering operation is used on T and F to decrease the set indeterminacy and remove noise. We have conducted experiments on a variety of noisy images using different types of noises with different levels. The experimental results demonstrate that the proposed approach can remove noise automatically and effectively. Especially, it can process not only noisy images with different levels of noise, but also images with different kinds of noise well without knowing the type of the noise.

Keywords: Image denoising, Neutrosophic Set, Wiener filtering, Entropy, PSNR.

1 Introduction

An image is often corrupted by noise during its acquisition and transmission. Image denoising is used to remove the additive noise while retaining as much as possible the important signal features. Despite that a huge number of approaches has been proposed, denoise algorithms performance mainly depends on a suitable representation to describe the original image information and noise type.

The neutrosophic set approach had been applied into image processing such as image thresholding [1], image segmentation [2, 3] and image denoising based on neutrosophic median filtering [4]. In this article, the new image denoising technique based on neutrosophic set approach of wiener filtering has been proposed. First the image is transformed into NS domain, which is described using three membership

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 861–869.
springerlink.com
© Springer-Verlag Berlin Heidelberg 2013

sets: True (T), Indeterminacy (I) and False (F). The entropy of the neutrosophic set is defined and employed to evaluate the indeterminacy. The ω-wiener filtering operation is used on T and F to decrease the set indeterminacy and remove noise. After the filtering process, the noise will be removed. The image will be transformed from NS domain into normal image. We have conducted experiments on a variety of noisy images using different types of noises with different levels. The experimental results demonstrate that the proposed approach can remove noise automatically and effectively. Especially, it can process not only noisy images with different levels of noise, but also images with different kinds of noise well without knowing the type of the noise.

The rest of paper is organized as follows. In section 2 we formulate the neutrosophic wiener filter for image denoising. The results for the proposed filtering method are discussed in section 3. Finally some conclusions are drawn in section 4.

2 Neutrosophic Image Denoising

Neutrosophy, a branch of philosophy introduced in [5] as a generalization of dialectics, studies the origin, nature and scope of neutralities, as well as their interactions with different ideational spectra. Neutrosophy theory considers proposition, theory, event, concept or entity, <A> is in relation to its opposite <Anti-A> and the <Neut-A> which is neither <A> nor <Anti-A>. The neutrosophy is the basis of the neutrosophic logic, neutrosophic probability, and neutrosophic set and neutrosophic statistics [5]. In neutrosophic set, the indeterminacy is quantified explicitly and the truth-membership, indeterminacy-membership and falsity-membership are independent. The neutrosophic set is a general formal frame work which generalizes the concept of the classic set, fuzzy set [6], interval valued fuzzy set [7], intuitionistic fuzzy set [8], and interval valued intuitionistic fuzzy set [9], paraconsistent set, dialetheist set, paradoxist set and tautological set [5]. The definition of a neutrosophic set and its properties are described briefly.

2.1 Neutrosophic Set

Definition 1 (Neutrosophic Set): Let U be a Universe of discourse and a neutrosophic set A is included in U. An element x in set A is noted as $x(T, I, F)$. T, I, F are real standard and non standard sets of $]^{-}0, 1^{+}[$ with $\sup T = t\_\sup$, $\inf T = t\_\inf$, $\sup I = i\_\sup$, $\inf I = i\_\inf$, $\sup F = f\_\sup$, $\inf F = f\_\inf$ *and* $n\_\sup = t\_\sup + i\_\sup + f\_\sup$, $n\_\inf = t\_\inf + i\_\inf + f\_\inf$.

T, I and F are called the neutrosophic components. The element $x(T, I, F)$ belongs to A in the following way. It is $t\%$ true in the set, $i\%$ indeterminate in the set, and $f\%$ false in the set, where t varies in T, i varies in I and f varies in F.

2.2 Transform the Image into Neutrosophic Domain

Definition 2 (Neutrosophic image): Let U be a Universe of discourse and W is a set of U, which is composed by bright pixels. A neutrosophic image P_{NS} is characterized by three membership sets T, I, F. A pixel P in the image is described as $P(T, I, F)$ and belongs to W in the following way: It is t true in the set, i indeterminate in the set, and f false in the set, where t varies in T, i varies in I and f varies in F. Then the pixel $P(i, j)$ in the image domain is transformed into the neutrosophic set domain $P_{NS}(i, j) = \{T(i, j), I(i, j), F(i, j)\}$, where $T(i, j), I(i, j)$ and $F(i, j)$ are the probabilities belong to white pixels set, indeterminate set and non white pixels set respectively, which are defined as:

$$T(i, j) = \frac{\bar{g}(i, j) - \bar{g}_{\min}}{\bar{g}_{\max} - \bar{g}_{\min}} \tag{1}$$

$$\bar{g}(i, j) = \frac{1}{w \times w} \sum_{m=i-w/2}^{i+w/2} \sum_{n=j-w/2}^{j+w/2} g(m, n) \tag{2}$$

$$I(i, j) = \frac{\delta(i, j) - \delta_{\min}}{\delta_{\max} - \delta_{\min}} \tag{3}$$

$$\delta(i, j) = abs(g(i, j) - \bar{g}(i, j)) \tag{4}$$

$$F(i, j) = 1 - T(i, j) \tag{5}$$

where $\bar{g}(i, j)$ is the local mean value of the pixels of the window. $\delta(i, j)$ is the absolute value of difference between intensity $g(i, j)$ and its local mean value $\bar{g}(i, j)$.

2.3 Neutrosophic Image Entropy

For a gray image, the entropy is utilized to evaluate the distribution of the gray levels. If the entropy is maximum, the intensities have equal probability. If the entropy is small, the intensity distribution is non-uniform.

Definition 3 (Neutrosophic image entropy): Neutrosophic entropy of an image is defined as the summation of the entropies of three subsets T, I and F:

$$En_{NS} = En_T + En_I + En_F \tag{6}$$

$$En_T = -\sum_{i=\min\{T\}}^{\max\{T\}} p_T(i) \ln p_T(i) \tag{7}$$

$$En_I = -\sum_{i=\min\{I\}}^{\max\{I\}} p_I(i) \ln p_I(i) \tag{8}$$

$$En_F = -\sum_{i=\min\{F\}}^{\max\{F\}} p_F(i)\ \ln p_F(i) \tag{9}$$

where En_T, En_I and En_F are the entropies of sets T, I and F respectively. $p_T(i)$, $p_I(i)$ and $p_F(i)$ are the probabilities of elements in T, I and F respectively, whose values equal to i.

2.4 ω-wiener Filtering Operation

The values of $I(i, j)$ is employed to measure the indeterminate degree of element P_{NS} (i, j). To make the set I correlated with T and F, the changes in T and F influence the distribution of element in I and vary the entropy of I.

Definition 4 (ω- wiener filtering operation): A ω- wiener filtering operation *for* P_{NS}, $\hat{P}_{NS}(\omega)$ is defined as:

$$\hat{P}_{NS}(\omega) = P(\hat{T}(\omega), \hat{I}(\omega), \hat{F}(\omega)) \tag{10}$$

$$\hat{T}(\omega) = \begin{cases} T & I < \omega \\ \hat{T}_\omega & I \geq \omega \end{cases} \tag{11}$$

$$\hat{T}_\omega(i, j) = \underset{(m,n)\in S_{i,j}}{wiener}\{T(m,n)\} \tag{12}$$

$$\hat{F}(\omega) = \begin{cases} F & I < \omega \\ \hat{F}_\omega & I \geq \omega \end{cases} \tag{13}$$

$$\hat{F}_\omega(i, j) = \underset{(m,n)\in S_{i,j}}{wiener}\{F(m,n)\} \tag{14}$$

$$\hat{I}_\omega(i, j) = \frac{\delta_{\hat{T}}(i, j) - \delta_{\hat{T}\min}}{\delta_{\hat{T}\max} - \delta_{\hat{T}\min}} \tag{15}$$

$$\delta_{\hat{T}}(i, j) = abs(\hat{T}(i, j) - \overline{\hat{T}}(i, j)) \tag{16}$$

$$\overline{\hat{T}}(i, j) = \frac{1}{w \times w} \sum_{m=i-w/2}^{i+w/2} \sum_{n=j-w/2}^{j+w/2} \hat{T}(m,n) \tag{17}$$

where $\delta_{\hat{T}}(i, j)$ is the absolute value of difference between intensity $\hat{T}(i, j)$ and its local mean value $\overline{\hat{T}}(i, j)$ at (i, j) after ω- wiener filtering operation.

The summary of image denoising method based on neutrosophic set approach of wiener filtering is described as below:

Step 1: Transform the image into NS domain;

Step 2: Use ω - wiener filtering operation on the true subset T to obtain T_ω ;

Step 3: Compute the entropy of the indeterminate subset $\hat{I}_\omega$, $En_{\hat{I}_\omega}(i)$;

Step 4: if $\frac{En_{\hat{I}_\omega}(i+1) - En_{\hat{I}_\omega}(i)}{En_{\hat{I}_\omega}(i)} < \delta$, go to Step 5; Else $T = \hat{T}_\omega$, go to Step 2;

Step 5: Transform subset $\hat{T}_\omega$ from the neutrosophic domain into the gray level domain.

3 Results and Discussion

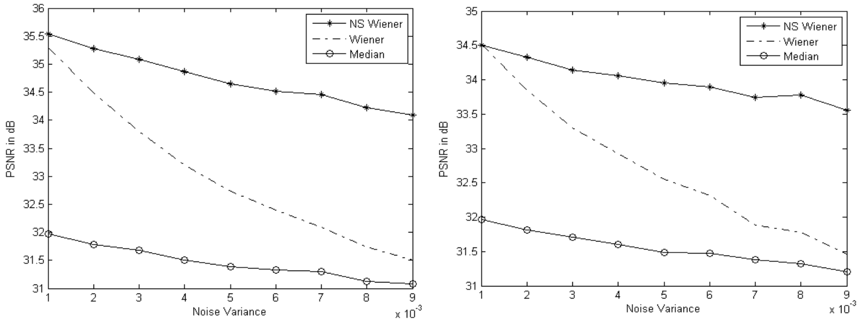
In the experiments, Lena and Cameraman images are used to evaluate the performance of the proposed method and compared with two filters such as median filter and wiener filter. Different kinds of noise such as Gaussian noise and salt and pepper noise with different levels are added into the image.

The performance of the denoising algorithm is measured by the peak signal to noise ratio (PSNR) in decibel (dB), which is computed using the following formula:

$$PSNR = -10 \log\left[\frac{\sum_{i=0}^{i=H-1}\sum_{j=0}^{j=W-1}\left(I(i,j) - I_d(i,j)\right)^2}{H \times W \times 255^2}\right] \tag{18}$$

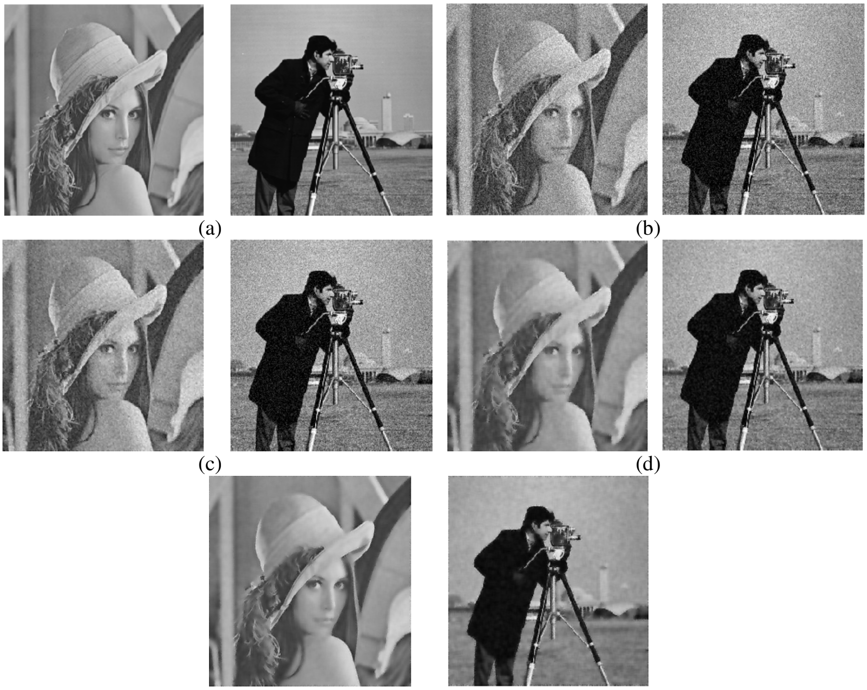
where $I(i, j)$ and $I_d(i, j)$ represent the intensities of pixels (i, j) in the original image and denoised image respectively. The higher the PSNR, the better the denoising algorithm is.

First, the proposed denoising algorithm is compared with median and wiener filter in removing Gaussian noise with different noise levels added to the Lena and cameraman images. The comparison of these filters based on PSNR with different noise levels is shown in Fig. 1a for Lena image and Fig. 1b for Cameraman image.



(a) (b)

Fig. 1. Comparison of Median, Wiener, and NS Wiener Filters for Gaussian noise based on the PSNR in dB values: (a) Lena (b) Cameraman.

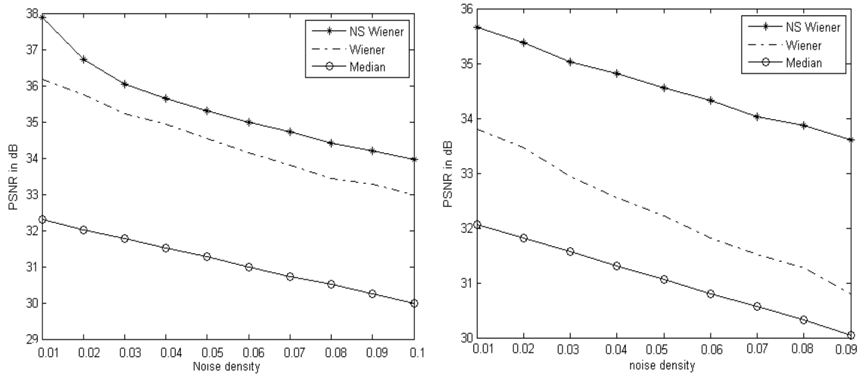


(a) (b)

(c) (d)

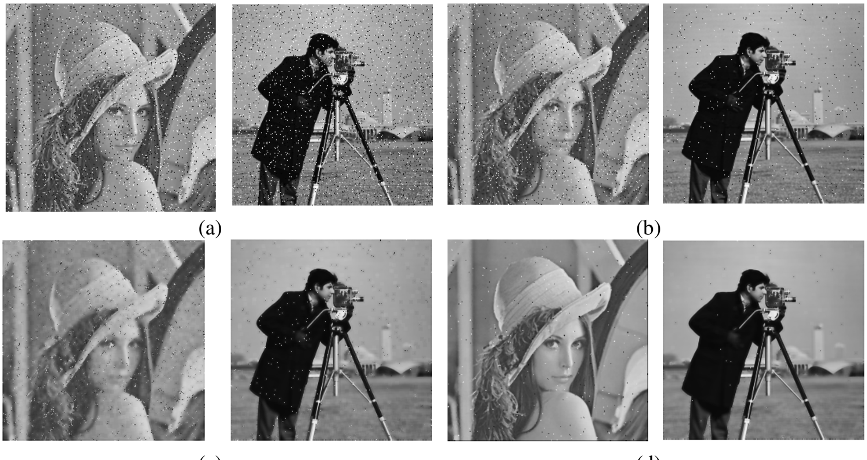
(e)

Fig. 2. a) Original Image, b) Noisy image (Gaussian noise): Lena (PSNR 29.57), Cameraman (PSNR 29.82) c) Denoised using Median Filter: Lena (PSNR 31.03), Cameraman (PSNR 31.53) d) Denoised using Wiener: Lena (PSNR 31.4), Cameraman (PSNR 33.01) e) Denoised using NS Wiener: Lena (PSNR 33.63), Cameraman (PSNR 33.59)



(a) (b)

Fig. 3. Comparison of Median, Wiener, and NS Wiener Filters for salt and pepper noise based on the PSNR in dB values: (a) Lena (b) Cameraman



(a) (b)

(c) (d)

Fig. 4. a) Noisy image (salt and pepper noise): Lena (PSNR 30.28), Cameraman (PSNR 30.32) b) Denoised using Median Filter: Lena (PSNR 32.24), Cameraman (PSNR 32.07) c) Denoised using Wiener: Lena (PSNR 35.28), Cameraman (PSNR 33.93) d) Denoised using NS Wiener: Lena (PSNR 36.11), Cameraman (PSNR 35.61)

Table 1. Comparison of Median, Wiener, NS Wiener Filters Based on PSNR in dB for Gaussian Noise

	PSNR (dB)					
	Median		**Wiener**		**NS wiener**	
Noise density	Lena	Camera man	Lena	Camera man	Lena	Camera man
0.001	31.97	31.96	35.29	34.51	**35.24**	**34.51**
0.002	31.79	31.81	34.49	33.85	**35.28**	**34.32**
0.003	31.68	31.7	33.8	33.29	**35.08**	**34.13**
0.004	31.51	31.61	33.2	32.92	**34.87**	**34.06**
0.005	31.39	31.48	32.74	32.55	**34.65**	**33.95**
0.006	31.32	31.43	32.39	32.31	**34.51**	**33.89**
0.007	31.29	31.37	32.09	31.89	**34.46**	**33.74**
0.008	31.12	31.32	31.74	31.78	**34.22**	**33.78**
0.009	31.09	31.21	31.51	31.46	**34.09**	**33.56**

From the comparison on visual and quantity measures in Fig. 2, conclude that the Neutrosophic approach of wiener filtering is not only removes the Gaussian noise and also achieves the high PSNR. Second, the proposed filter is employed to remove the other type of noise such as salt and pepper noise. The comparison of PSNR values of different filters are shown in Fig. 3a for Lena image and Fig. 3b for Cameraman

Table 2. Comparison of Median, Wiener, NS Wiener Filters Based on PSNR in dB for Salt and Pepper noise

	PSNR (dB)					
	Median		**Wiener**		**NS wiener**	
Noise density	Lena	Camera man	Lena	Camera man	Lena	Camera man
0.01	32.29	32.06	36.16	33.81	**37.88**	**35.66**
0.02	32.02	31.82	35.74	33.46	**36.73**	**35.38**
0.03	31.78	31.57	35.23	32.94	**36.04**	**35.03**
0.04	31.52	31.31	34.94	32.56	**35.66**	**34.81**
0.05	31.28	31.06	34.53	32.21	**35.31**	**34.55**
0.06	30.99	30.8	34.15	31.82	**34.99**	**34.32**
0.07	30.72	30.56	33.79	31.52	**34.71**	**34.03**
0.08	30.51	30.32	33.44	31.27	**34.41**	**33.88**
0.09	30.25	30.04	33.27	30.81	**34.21**	**33.61**

image. From Fig. 4, the proposed filter outperforms the other two filters. The PSNR values of the proposed filter, median and wiener filter for Lena and Cameraman images with Gaussian and Salt and pepper noise is tabulated in Table 1 and Table 2.

4 Conclusions

In this article, a novel image denoising technique based on neutrosophic set approach of wiener filtering has been proposed. The image is described as a NS set using three membership sets *T, I* and *F*. The entropy in neutrosophic image domain is defined and employed to evaluate the indetermination. The wiener filter is applied to reduce the set's indetermination and remove the noise in the image. The experimental results demonstrate that the proposed approach can remove noise automatically and effectively. Especially, it can process not only noisy images with different levels of noise, but also images with different kinds of noise well without knowing the type of the noise. The properties of neutrosophic image will achieve more applications in processing and computer vision.

References

[1] Cheng, H.D., Guo, Y.: A new neutrosophic approach to image thresholding. New Mathemetics and Natural Computation 4(3), 291–308 (2008)

[2] Guo, Y., Cheng, H.D.: New Neutrosophic approach to image segmentation. Pattern Recognition 42(5), 587–595 (2009)

[3] Zhang, M., Zhang, L., Cheng, H.D.: A Neutrosophic approach to image segmentation based on watershed method. Signal Processing 90(5), 1510–1517 (2010)

[4] Guo, Y., Cheng, H.D., Zhang, Y.: A New Neutrosophic approach to Image Denoising. New Mathemetics and Natural Computation 5(3), 653–662 (2009)

[5] Samarandache, F.: A unifying field in logics Neutrosophic logic. In: Neutrosophic Set, Neutrosophic Probability, 3rd edn. American Research Press (2003)

[6] Zadeh, L.A.: Fuzzy sets. Inform and Control 8, 338–353 (1965)

[7] Turksen, I.: Interval valued fuzzy sets based on normal forms. Fuzzy Sets and Systems 20, 191–210 (1986)

[8] Atanassov, K.: Intuitionistic fuzzy sets. Fuzzy Sets and Systems 20, 87–96 (1986)

[9] Atanassov, K.: More on Intuitionistic fuzzy sets. Fuzzy Sets and Systems 33, 37–46 (1989)

Modified Difference Expansion for Reversible Watermarking Using Fuzzy Logic Based Distortion Control

Hirak Kumar Maity[1], Santi P. Maity[2], and Debashis Maity[2]

[1] Dept. of ECE & EIE, College of Engineering and Management, Kolaghat
[2] Dept. of Information Technology, Bengal Engineering and Science University, Shibpur
`{Hirakmaity,debashis.cemk}@gmail.com, spmaity@yahoo.com`

Abstract. Digital watermarking using Difference Expansion (DE) is quite popular to embed reversibly the data followed by recovery of the original image. According to this algorithm, the least significant bit (LSB) of inter-pixel differences (between a pair of neighboring pixel) is used to embed data. It is seen that none of the DE works focuses on structural information retentions for the watermarked image at high embedding capacity. Moreover, security measure of the hidden data is not investigated under distortion constraint scenario. To this aim, a modification in DE is proposed that not only increase embedding space (hence, watermark payload) but also makes little change in structure and contrast comparison (imperceptibility) under similar luminance background. A simple fuzzy function is used to classify the image content into smooth, texture and edge region followed by adaptive distortion control. Modification also makes little change in relative entropy between the host and the watermarked data that leads to better security of the hidden data.

Keywords: Reversible watermarking, LSB modification, difference expansion, SSIM, KLD and distortion control.

1 Introduction

Digital watermarking is a method of embedding useful information into a digital work (especially for audio, image, or video) for the purpose of copy control, content authentication, distribution tracking, broadcast monitoring, etc [1]-[3]. This embedded information may be visible or invisible called as watermark and after embedding into the original data, the later one is called watermarked data. Distortion introduced by embedding the watermark is made small as much as possible so that the original and the watermarked image remain perceptually equivalent and imperceptible by human visualization. However, some typical data, for example, medical and military images, and some applications in legal domains do not allow even this small imperceptible distortion. This has led to an interest in lossless or reversible watermarking (RW), where the embedded watermark is not only extracted, but also perfect reconstruction of the host signal is possible from the watermarked image. RW provides authentication of the host image as well as can be used in covert communication. Performance

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 871–883.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

of a reversible watermarking algorithm is subjected to the following conflicting requirements:

Visible Quality: Visual quality degradation due to embedding should not be perceived by human eye and this distortion should be much less than the conventional watermarking scheme.

Payload Capacity: Data hiding capacity should be high enough to make it suitable for intended applications unlike the conventional watermarking scheme where low capacity may be acceptable.

Computational Complexity: The data hiding algorithm needs low mathematical complexity, hence, consequent implementation cost should be as low as possible.

Literature on reversible watermarking is quite rich by this time [1]-[9]. Among them, Tian's difference expansion (DE) scheme is quite popular one, where redundancy present in the image content is used to embed reversibly the payload (needed for exact recovery of the original image). This paper makes simple but effective modification of DE that improves data hiding capacity with much better visual and statistical invisibility of the hidden data. Increase in embedding space is shown geometrically with and without distortion control. To make watermark power adaptive, host image is partitioned into smooth, edge and texture regions using a simple fuzzy function of edge gradient obtained through sobel's operator. Then distortion control is set in different values based on these image characteristics. Performance improvement is not only shown through simulation results but also is explained through mathematical analysis.

The rest of the paper is organized as follows: Section 2 presents a brief literature review on RW, the limitations and scope of the present work. Section 3 introduces several mathematical models used in this work for watermarking performance assessment. Section 4 presents fuzzy logic based image partitioning for distortion control. Section 5 discussed with proposed algorithm and section 6 presents performance evaluation with discussion. Finally conclusions are drawn in Section 7 along with scope of future works.

2 Review of Related Works, Limitations and Scope of the Work

Several reversible image watermarking algorithms are reported using spatial domain and transform domain data embedding. Spatial domain technique embeds data by directly modifying pixel values of the host image, while transform domain technique embed data by modifying transform coefficients. Spatial domain methods are simpler to implement and provide higher hiding capacity but also suffer from lower robustness. On the other hand, transform domain technique offers imperceptibility in a better way with improved robustness against common signal processing approach. In literature, RW is classified into three major groups: 1) RW based on data compression approach, 2) RW based on difference expansion approach and 3) RW based on histogram bin shifting.

As mentioned earlier, Tian et al. [1], [3] first propose difference expansion (DE) algorithm. According to this scheme, the least significant bit (LSB) of inter-pixel differences (between a pair of neighboring pixel) is used to embed data. In principle, the redundancy present in digital images is used to achieve a high-capacity and low-distortion reversible watermarking. Later on, Alattar et al. [4] extended Tian's scheme by using DE of spatial and cross spectral triplets instead of pixel pairs which increase the hiding ability. Tian's algorithm allows embedding of one bit in every pair of pixels, whereas Alatter's algorithm [4] embeds two bits in every triplet. Soon after Alattar et al. [5] extended Tian's scheme in different way by using DE transform of quads and able to embed two bits in every quads. The amount of data that can be embedded into the host image depends largely on the characteristics of image. So this technique is not commonly acceptable. Subsequently, Alattar et al. [6] proposed another method based on generalized DE method with integer transform where more differences were available for expansion and require low cost to record overhead information. Wang et al. [7] developed a novel scheme based on vector map which reduces the size of the location map. Lin et al [8] had proposed location map free reversible data hiding technique based on DE algorithms. Later, Hu et al. [9] presented another DE scheme which demands efficient payload dependent overflow location which has good compressibility than earlier algorithms. It contains two types of overflow location: one from embedding and another from shifting.

It is seen that none of the DE works focuses on structural information retentions for the watermarked image at high embedding capacity. Majority of them had tried to explore the redundancy in digital images to achieve very high embedding capacity with low distortion. Moreover, there is no provision on distortion control analysis in DE algorithms [1], [3]. This work makes it adaptive based on different characteristics of the host image. We classify an image into smooth, edge and texture region based membership of simple fuzzy function for the magnitude of the edge gradient. Accordingly distortion control is set in different values in different regions. Furthermore, in the literature of DE; security measure of the hidden data is not investigated under distortion constraint scenario. To this aim, a simple yet effective modification in DE is suggested that not only increases embedding space but also makes little change in structure and contrast comparison under similar luminance perspective. Modification also makes a difference in relative entropy between the host and watermarked data that in turn leads to higher security for the hidden data.

3 Mathematical Models for Watermarking Assessment

This section briefly introduces different quantitative measures used in this work for quality assessment of watermarking methods. The measures include peak-signal-to noise-ratio (PSNR), structural similarity index metric (SSIM) for visual quality of the watermarked images, Kullback-Leibler distance (KLD) for security measure of the hidden data.

3.1 PSNR and MSE

The simplest, oldest and most widely used technique to quantify image/video signal quality is the mean squared error (MSE), computed by averaging the squared intensity differences of distorted (watermarked) and original (host) image pixels. Mathematically it is defined as:

$$\mathrm{MSE} = \frac{1}{\mathrm{MxN}} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (u - v)^2 \tag{1}$$

Where, two M×N images u and v, one of the images is considered a noisy (here watermarked image) approximation of the original one. In watermarking, PSNR is also the widely used measure to quantify visual distortion made by watermarking process as well as different attack operations. MSE is always useful to calculate PSNR in the following way:

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{255^2}{\mathrm{MSE}} \right) \tag{2}$$

Where, maximum pixel intensity value is 255 for an 8 bit gray scale image. Note that the PSNR does not contain the local content or structure of image or video signal. On the other hand, the mean SSIM (MSSIM) index has been shown to be more accurate for measuring the quality of images [10], as this measure consists of human visual characteristics as well as structure of the content. Although, more recent works incorporates entropy masking to include neighborhood contributions and statistical pattern that play important roles in human perception, we will consider MSSIM here.

3.2 SSIM and MSSIM

SSIM is consistent with characteristics and content of the data as well as visual environment. This method is based on the structural information of the image and provides a good measure for very different kinds of images, from natural scenes to medical images [10]. It compares local patterns of pixel intensities that have been normalized for luminance and contrast. SSIM index between two images u and v have three levels of comparisons and is defined as follows:

$$\mathrm{SSIM}(u, v) = [l(u, v)]^{\alpha} . [c(u, v)]^{\beta} . [s(u, v)]^{\gamma} \tag{3}$$

where, l, c and s are the luminance, contrast, structural comparisons between two images u and v, and are given by the expressions as follows:

$$l(u, v) = \frac{2\mu_u \mu_v + C_1}{\mu_u^2 + \mu_v^2 + C_1}, c(u, v) = \frac{2\sigma_u \sigma_v + C_2}{\sigma_u^2 + \sigma_v^2 + C_2} \text{ and } s(u, v) = \frac{\sigma_{uv} + C_3}{\sigma_u \sigma_v + C_3} \tag{4}$$

Where, C1, C2 and C3 are three constants. The symbols μ and σ are the local mean intensity and local standard deviation of an image. The symbol σ_{uv} is the local covariance coefficient between the images u and v. If $\alpha=\beta=\gamma=1$ and $c_3=c_2/2$ are considered, it can be rewritten as:

$$\mathrm{SSIM}(u, v) = \frac{(2\mu_u \mu_v + C_1)(2\sigma_{uv} + C_2)}{(\mu_u^2 + \mu_v^2 + C_1)(\sigma_u^2 + \sigma_v^2 + C_2)} \tag{5}$$

The overall value is obtained from the mean of the local SSIM, called MSSIM. The above mathematical forms reveal the fact that variance and covariance play an important role in SSIM measure and this issue should be taken care in watermarking work for better imperceptibility.

3.3 KLD

Kullback-Leibler distance (KLD) is one important security measure between two probability distribution function p_0 and p_1 those are closely related [11]. KLD is used widely in steganography for security measure of the hidden data. It is a natural distance function from a "true" probability distribution, p_1, to a "target" probability distribution, p_0. It can be interpreted as the expected extra message-length per datum due to using a code based on the wrong (target) distribution compared to, using a code based on the true distribution. It is frequently used as a measure of 'distance' from information theory viewpoint. If p_0 and p_1 are two probability densities, then KLD is defined as for continuous functions:

$$\mathrm{KLD}(p_1 \,\|\, p_0) = \int p_1(x) \log \frac{p_1(x)}{p_0(x)} dx \tag{6}$$

The log function has base 2. For discrete (not necessarily finite) probability distributions, the KLD is defined to be:

$$\mathrm{KLD}(p_1, p_0) = \sum_i p_{1i} \log_2 \left(\frac{p_{1i}}{p_{0i}} \right) \tag{7}$$

where, $p_1 = \{p_{11}, p_{12}, \ldots., p_{1n}\}$ and $p_0 = \{p_{01}, p_{02}, \ldots., p_{0n}\}$. As mentioned earlier, this is used here to quantify the security i.e. statistical invisibility of the hidden data.

4 Fuzzy Logic Based Image Partitioning and Distortion Control

It is seen that a natural images posses different characteristics like smooth, edge, texture region etc. and different regions are in a different way sensitive to embedding distortion. This needs adaptive watermark power control so that overall structure of the watermarked image is less affected. In automatic identification of image regions, it is difficult to find threshold gray values due to imprecise and uncertainties in gray values. This has led to the use of various soft computing tools like Neural Network (NN), Genetic Algorithms (GAs) and Fuzzy Logic (FL) to resolve uncertainties as well as adaptive and optimized solution. Here, we use a simple fuzzy function for identification of different image regions.

4.1 Image Partitioning Using Fuzzy Logic

We assume here that an image contains edge, smooth and texture regions. First, the image is portioned into three different regions and distortion is then controlled accordingly:

The gradient map of an image is generated by using Sobel's operator, that has two mask of same size, viz. one horizontal mask [-1 0 1;-2 0 2;-1 0 1] and one vertical mask [-1 -2 -1; 0 0 0; 1 2 1], for each pixel (P) with location (i, j), its gradient vector is defined as $\overline{M} = \{dx_{i,j}, dy_{i,j}\}$. Where $dx_{i,j}$ and $dy_{i,j}$ are obtained by vertical edge mask and horizontal edge mask, respectively. The magnitude of this vector is

$$g = \sqrt{(dx_{i,j}^2 + dy_{i,j}^2)} \tag{8}$$

Now we consider a fuzzy function, defined as:

$$func = \frac{1}{1+|g|} \tag{9}$$

It is clear from func (9), that the membership value of func varies from zero to one depending on g value varies from infinity to zero.

4.2 Error Distortion Control

The pixel intensity difference value between the host and the watermarked image is called error which causes visual distortion. To control or minimize this distortion we use a predefined error threshold (Δ). If (x, y) is the original pixel intensity value get modified to (x', y') after watermark embedding, the error is then $|x - x'|$ on x and on y is $|y - y'|$. For low data hiding bit-rate, distortion control is necessary in order to reduce the distortions introduced by the watermarking. According to this mechanism, any pixel pair will be transformed if it satisfies the following conditions:

$$|x - x'| < \Delta \text{ and } |y - y'| < \Delta \tag{10}$$

Using func (9), we partition image into smooth, edge, texture regions and set distortion as Δ_1, Δ_2 and Δ_3, respectively. It is assumed that overall embedding distortion using a global Δ would remain similar for Δ_1, Δ_2 and Δ_3 chosen adaptively in respective regions.

5 Proposed Modified DE Algorithm

This section presents the proposed watermarking scheme. For an 8 bit gray scale image, let x, y is a pixel pair bounded by its intensity value within the range of 0 and 255. Then the forward integer transform is given by:

$$x' = \left\lfloor \frac{x+y}{2} \right\rfloor, y' = \left\lfloor \frac{x-y}{2} \right\rfloor \tag{11}$$

where x' and y' are the forward transform pair and the symbol $\lfloor . \rfloor$ indicates floor function meaning “the greatest integer less than or equal to”. The inverse integer transform of (11) is given by:

$$x'' = x' + y', y'' = x' - y' \tag{12}$$

According to the general methodology of DE, we can embed a watermark bits (w) to the LSB position of y' using the following rule:

$$y'_1 = 2y' + w \tag{13}$$

where w can be 0 or 1 depending on the watermark bits. The watermarked pixel pairs will be represented by:

$$x''_1 = x' + y'_1 \, , \, y''_1 = x' - y'_1 \tag{14}$$

Using these transformations the original image is transferred to watermarked image. The next half of the work is related with watermark extraction to get back original image from watermarked image. At this point watermarked image is the input, by applying (11) to the marked pixel pairs, following pixel pair would be obtained.

$$x'_2 = \left\lfloor \frac{x''_1 + y''_1}{2} \right\rfloor, \, y'_2 = \left\lfloor \frac{x''_1 - y''_1}{2} \right\rfloor \tag{15}$$

Now watermark bit (b) is retrieved by discarding the LSB of y'_2 and the rest of the part (y''_2) will be considered for further operation.

$$y''_2 = \left\lfloor \frac{y'_2}{2} \right\rfloor \tag{16}$$

Retrieved watermark bits (b) will be same with embedded watermark (w) due to its reversibility. Now the original pair of pixel would be back by simply combing (12), (15) and (16).

$$x_2 = x'_2 + y''_2 \, , \, y_2 = x'_2 - y''_2 \tag{17}$$

According to the condition of reversibility the intensity value of the retrieved pixel should be same with its corresponding original one, i.e. $x_2 = x$ and $y_2 = y$. The whole retrieving/reconstruction process will be blind in nature i.e. without the use of the original image. In order to confined its applicability in 8 bit gray scale image, x' and y' should satisfy the conditions: $0 \leq (x'', y'') \leq 255$ to avoid overflow or underflow problem. Mathematically,

$$0 \leq (x' + y') \leq 255 \text{ and } 0 \leq (x' - y') \leq 255 \text{ or } y' \leq \min((255 - x'), x') \tag{18}$$

5.1 Condition of Changeability and Expandability

According to DE theorem, new embedded binary digits (b) will be placed into LSB position of the difference value y', i.e. expandable difference value after placing the embedded bit will be: $\ddot{y} = 2y' + b$. To avoid overflow or underflow problem, it should satisfy the following conditions:

$$\ddot{y} \leq \min((255 - x'), x') \text{ or } 2y' + b \leq \min((255 - x'), x') \tag{19}$$

where, b can be 0 or 1 and the above condition is called '*condition of expandability*'.

Similarly, without making overflow or underflow, any difference value will be changeable if and only if it satisfies condition (18) after embedding, i.e. the '*condition of changeability*' is defined as:

$$y' = 2\left\lfloor \frac{y'}{2} \right\rfloor + b \le \min((255 - x'), x')$$

$$\text{or } 2\left\lfloor \frac{2y'+b}{2} \right\rfloor + b \le \min((255 - x'), x')$$

$$\text{or } 2y' + b \le \min((255 - x'), x') \tag{20}$$

Note that *both the conditions* (condition of expandability (19) and condition of changeability (20)) *are same in this case*, so, it does not require to check for both the conditions separately, whereas in [2] these two conditions are different and needs separate checking. This in turn reduces mathematical complexity almost halved leading to reduction in implementation cost.

5.2 Data Embedding Process

The watermark embedding and retrieving process are very much similar as in [3] and can be summarized using the following steps:

1. First the original image is grouped into pair of non- overlapping pixel values. Any pair may contain two neighboring value or two pixel intensity value having a small difference. Pairing can be done horizontally or vertically or by any specific choice for the overall image or for any specific areas. If we make pair horizontally, then the total size of matrix that represents horizontal pairing, have the same number of column and half of the number of rows of its original. On the other hand, if we make pairing vertically, the size of matrix will be same number of rows and half of number of column of its original. Then forward integer transform (11) is applied to each pair and if it satisfies the condition (18), we will get the average value x' and difference value y' as shown in (11).

2. Now four disjoint set of difference values are formed as described below:

a. EZ: It stands for expandable zero, contains all expandable difference value (y') having zero value.

b. EN: It stands for expandable non zero, contains all expandable difference value (y') having non zeros value, i.e. $y' \notin$ EZ.

c. CN: It stands for changeable non zeros value, contains all changeable non zero values of y', i.e. $y' \notin$ (EZ $\cup$ EN).

d. NC: It stands for non-changeable values, contains all non-changeable values of y'.

3. For every difference value belonging to EZ group will be selected for data embedding. Depending on the payload size some selected value belongs to EN group will be selected for data embedding which is placed on subset EN1 and all non-selected value will be placed on subset EN2. For a difference value belongs to subset EZ $\cup$ EN1, bit 1is put into the location map and for subset EN2 $\cup$ CN $\cup$ NC bit 0 is put into the location map. From location map we can understand any value is expandable (if value=1) or not (if value=0) and finally location map has been compressed by using some lossless compression algorithms.

Original LSB's for difference values in EN2 $\cup$ CN are then collected and is kept it into an array c as a bit stream.

5. Now location map, original LSB's and payload for those difference values satisfy conditions (18) and (19) are embedded.

6. After embedding all bits, inverse transformation is performed to obtain watermarked image.

5.3 Data Recovery Process

The decoding and authentication process contains the following steps:

1. The watermarked image is grouped into pair of pixels in a same fashion used during embedding. Forward integer transform is performed to obtain difference and average values.
2. Next two disjoint set of difference value are created as below:
   a. CH: it stands for changeable difference value.
   b. NC: it stands for non-changeable difference value.
3. Collect LSB's of all difference value belongs to the subset CH, which form a bit stream b contains location map, original LSB's and payload.
4. Decode the location map by proper decompressing of the above bit stream and restore the original difference value.
5. Now apply reverse integer transform to get back the original image for authentication.

6 Experimental Results

This scheme presents performance evaluation of the proposed algorithm along with relative gain in imperceptivity, payload capacity and security compared to Tian's algorithms [3]. Although we have perform simulation over large number of gray scale images, due to space limitations, we report the same for two test images Lena and Boat as shown in Fig. 1. Fig. 1(a) and Fig. 1 (b) show the original image, Fig. 1 (c) and Fig. 1 (d) show the watermarked images with embedding capacity at 0.0549 bpp and 0.057 bpp, the corresponding PSNR values are 44.78 dB and 44.188 dB, respectively. Fig. 1 (e) and Fig. 1 (f) indicate the watermarked images with embedding capacity at 0.874 bpp, 0.8698 bpp and their corresponding PSNR values are 39.31 dB and 37.95 dB, respectively using over embedding. Numerical values have shown here obtained by using different sets of Δ values. Δ_1=3.5, Δ_2=11 and Δ_3=3.5 have been used for Fig. 1(c) and Fig.1 (d). Whereas Fig. 1(e) and Fig. 1(f) have been analyzed with Δ_1=11.7, Δ_2=36.7 and Δ_3=11.6.

By adapting over embedding we have achieved higher capacity up to 1.47 bpp with its respective PSNR and MSSIM values are 33.44dB and 0.34 for Lena images, where the payload size is 385320. Performing similar kind of analysis for Boat images, we have achieved the maximum capacity of 1.24 bpp with its respective PSNR, MSSIM and payload size values are 33.4 dB, 0.392 and 325982. Data hiding payload may be increased further, by more over embedding but the visual quality of the images will degrade. As boat image contains more details than Lena image, so it allows less embedding capacity for similar embedding distortion. The highest PSNR values we have achieved for Lena images is 44.78 dB with its embedding capacity 0.0549 bpp and for Boat image it is 44.18 dB with embedding capacity 0.057 bpp. So, from the above discussion we can conclude, our algorithm is convenient to use both for very high and very low embedding payload.

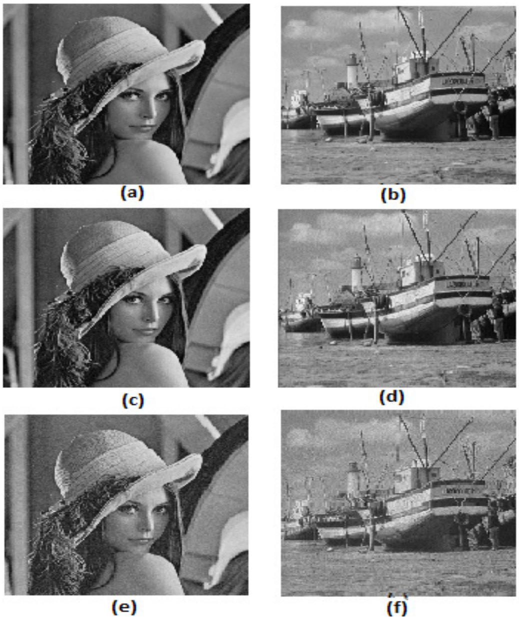


Fig. 1. Original Image (a) Lena (b) Boat; Watermarked Images: with Low capacity (c) 0.0549 bpp, 44.78dB, payload size (PS) 14404 (d) 0.057 bpp, 44.188 dB, PS 14970 and with high capacity using over embedding (e) 0.874 bpp, 39.31dB, PS 229063 (f) 0.8698 bpp, 37.95 dB, PS 228029 for Lena and Boat Image respectively.

We also have analyzed performance of the proposed algorithm with and without distortion control as shown in Fig. 2. For comparison we have shown our result with Tian's results under distortion control scenarios. DE with distortion control provides a rhombic shape. For a particular predefined error threshold (Δ), the larger rhombic area provides larger embedding space. So, from the Fig. 2 it is clear that our algorithm provide larger embedding space than Tian's one.

The watermarked image produced after transformation should not introduce visual artifacts. By taking the sum and difference of (10), we gets $x'' + y'' = x + y$ and $x'' - y'' = x - y$, which is exactly same with [3]. It indicates that our algorithm preserves both gray level averages and difference between transformed pair of pixels.

Furthermore to show the efficiency of our proposed algorithm, the detailed experimental results for Lena and Boat images are given in Fig. 3 and Fig. 4 respectively. In Fig. 3 and Fig. 4 first plot indicates RSNR (dB) vs. watermark payload (bpp), second plot indicates SSIM vs. capacity (bpp) and the last plot is a 3-D plot between

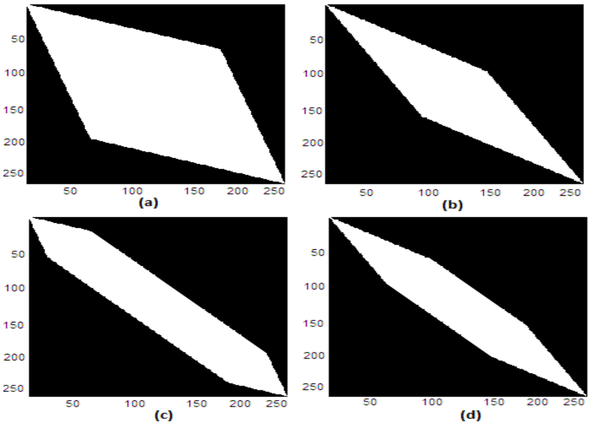


Fig. 2. Distortion control (DC) plot: (a) our algo without DC; (b) Tian's algo without DC; (c) Our algo with DC and (d) Tian's algo with DC

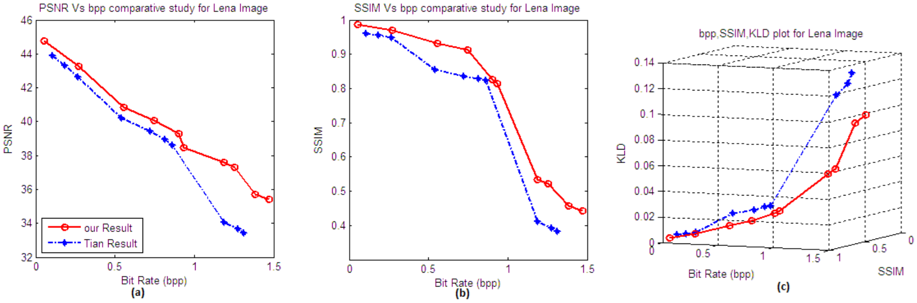


Fig. 3. Result for Lena Images: (a) PSNR Vs bpp; (b) SSIM Vs bpp; (c) KLD Vs SSIM Vs bpp

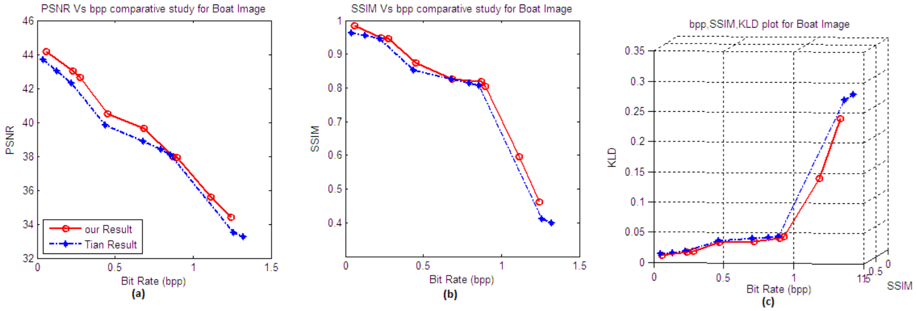


Fig. 4. Result for Boat Images: (a) PSNR Vs bpp; (b) SSIM Vs bpp; (c) KLD Vs SSIM Vs bpp

SSIM vs. bit rate (bpp) vs. KLD. From Fig. 3 and Fig. 4 it is clear that with embedding rate beyond 0.8 bpp, our results are superior to Tian's for Lena image, whereas it is not that much effective for Boat images as it contains more detail than Lena image. The lower value of KLD indicates that hidden data i.e. statistically invisible. Fig. 3(c) and Fig. 4(c) show the watermarked images (both Lena and Boat image) with different embedding capacity as security for the hidden data is higher compared to [3].

7 Conclusions and Scope of the Work

In this paper, we propose a modified DE algorithm for reversible watermarking. It has been shown geometrically as well as through simulation results that the proposed modification not only increases the watermark payload limit but also reduces visual distortion. Simulation results also show improvement in security of the hidden data compared to original DE method. As compare to Tian's scheme, we have reduced the difference component to embed watermark with more no of pixel pairs based on adaptive distortion control over the different image regions obtained using fuzzy function of edge magnitudes. The computational overhead has been reduced by merging two different conditions, 'condition of expandability' and 'condition of changeability'. Future work may be extended for the calculation of optimal distortion control (Δ's) values leads a given embedding distortion constrained scenario.

References

1. Tian, J.: Reversible Watermarking by Difference Expansion. In: Multimedia and Security Workshop at ACM Multimedia 2002, Juan-Les-Pins, France, December 6 (2002)
2. Feng, J.-B., Lin, I.-C., Tsai, C.-S., Chu, Y.-P.: Reversible Watermarking: Current States and Key Issues. International Journal of Network Security 2(3), 161–171 (2006)
3. Tian, J.: Reversible Data Embedding Using a Difference Expansion. IEEE Transactions on Circuit and Systems for Video Technology 13(8) (August 2003)
4. Alattar, A.M.: Reversible watermark using difference expansion of triplets. In: Proceedings of the IEEE International Conference on Image Processing, Catalonia, Spain, vol. 1, pp. 501–504 (September 2003)
5. Alattar, A.M.: Reversible watermark using difference expansion of quads. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, vol. 3, pp. 377–380 (May 2004)
6. Alattar, A.M.: Reversible Watermarking Using the Difference Expansion of a Generalized Integer Transform. IEEE Transaction on Image Processing 13(8), 1147–1156 (2004)
7. Wang, X.T., Shao, C.Y., Xu, X.G., Niu, X.M.: Reversible Data-Hiding Scheme for 2-D Vector Maps Based on Difference Expansion. IEEE Transaction on Information forensics and Security 2(3) (September 2007)
8. Lin, C.C., Yang, S.P., Hsueh, N.L.: Lossless Data Hiding Based on Difference Expansion without a Location Map. In: 2008 Congress on Image and Signal Processing, pp. 8–12 (2008)

9. Hu, Y., Lee, H.K., Li, J.: DE-Based Reversible Data Hiding With Improved Overflow Location Map. IEEE Transactions on Circuit and Systems for Video Technology 19(2) (February 2009)
10. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Asssesment: From Error Visibility to structural similarity. IEEE Transaction on Image Processing 13(4) (April 2004)
11. Georgiou, T.T., Lindquist, A.: Kullback–Leibler Approximation of Spectral Density Functions. IEEE Transactions on Information Theory 49(11) (November 2003)

A Novel Genetic Algorithm Based Data Embedding Technique in Frequency Domain Using Z Transform (ANGAFDZT)

J.K. Mandal[1], A. Khamrui[2], S. Chakraborty[2], P. Sur[2],
S.K. Datta[2], and I. RoyChoudhury[2]

[1] Dept. of Computer Science and Engineering, University of Kalyani, Kalyani, Nadia-741235
jkm.cse@gmail.com
[2] Dept. of Computer Science and Engineering,
Future Institute of Engineering and Management,
Sonarpur Station Road, Sonarpur, Kolkata-150, West Bengal
amritakhamrui@rediffmail.com, {sohamcha,priyashsur}@gmail.com,
{dattasaikat1,indranilrc29}@gmail.com

Abstract. In this paper a transformed domain based gray scale image authentication/data hiding technique using Z transform (ZT) termed as ANGAFDZT, has been proposed. Z-Transform is applied on 2 x 2 mask of the source image to transform into corresponding frequency domain. Four bits of the hidden image are embedded in each mask of the source image. Resulting image masks are taken as initial population. New Generation, Crossover and Mutation are applied on the initial population to obtain stego image. Genetic algorithm is used to enhance the security level. During the process of embedding, dimension of the hidden image followed by the content of the message/hidden image are embedded. Reverse process is followed during decoding. High PSNR obtained for various images compared to existing Chin-Chen Chang et al.[1] conform the quality of invisible watermark of ANGAFDZT.

Keywords: ANGAFDZT, Z-Transform, watermark, Genetic algorithm, PSNR, MSE, IF.

1 Introduction

Steganographic techniques embed secrete/authenticating information into various natural cover data like sound, images, logos etc. Embedded data is referred to as stego-data and it must be perceptually indistinguishable from its natural cover. Steganography includes the concealment of digital information within data. Generally, hidden information may be picture, video, sound file [6], [5]. A message may be hidden by using algorithms like invisible ink between the visible lines of innocuous documents to ensure the security which is a big concern in modern day image trafficking across the network. Security may be achieved by hiding information into images. Data hiding [4] in the image has become an important tool for image

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 885–893.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

authentication. Ownership verification and authentication are the major task for military people, research institute and scientists. Information security and image authentication has become very important to protect digital image document from unauthorized access [2]. Data hiding refers to the nearly invisible [3], [7], [12], [13], [14] embedding of information within a host data set as message, image or video. A classic example of steganography is that of a prisoner communicating with the outside world under the supervision of a warden. The data hiding represents a useful alternative to the construction of a hypermedia document or image, which is very less convenient to manipulate. The motive is to hide a message inside an image keeping its visible properties [8] as close as to the original. The most common methods to make these alteration is usage of the least-significant bit (LSB) developed through [8] masking, filtering and transformations on the source image[5]. Present proposal is an algorithm for secure message transmission through block based data hiding. Most of the works [11],[10], [1], [9] used minimum bits of the hidden image for embedding in spatial domain, but the proposed algorithm embed with a high payload in transformed domain with a bare minimum distortion of visual property.

Rest of the paper is organized as follows. Section 2 deals with the proposed technique. Results and comparisons are given in section 3. Concluding remarks are presented in section 4 and references are drawn at end.

2 The Technique

In the process of embedding a 2 x 2 mask is chosen from the host image in row major order. 2 x 2 mask is transformed into frequency domain using Z transform. Four bits of the authenticating message/image is embedded in each 2 x 2 transformed mask. Along with the hidden image, the dimensional values are also embedded into the real part of the host image mask on the second and third coefficient. First two rows contain the header informations. The dimension is extracted from the third row of each gray scale pgm image and the same is embedded into first four masks of the image as image information starts after fourth row of the image. The fourth row contains maximum pixel intensity. On the second coefficient, second and third LSB is chosen for embedding. On the third coefficient, third and fourth LSB is chosen for embedding. The position is chosen in such a way that there is no loss of precision after inverse Z transform. Identical embedding with same information is done in second and fourth coefficient to nullify the effect of algebraic addition of complex conjugate during reverse transformation. Embedded mask is transformed from frequency domain to spatial domain using inverse Z-Transform. 2×2 embedded image mask of size 32 bits are taken as initial population. Procedure of New Generation followed by Crossover and mutation has been applied on the initial population. For New Generation minimum coefficient of the mask is chosen. If the minimum is less than zero then subtract magnitude of minimum from each coefficient otherwise skip these step. The New Generation is done to keep the image fidelity nearer to original one and to avoid in generating negative pixel value during reverse transform. Crossover is performed onto the k bits by consecutive bitwise XOR operation on k steps, taking the MSB of the intermediate stream generated in each step. Mutation is performed between rightmost 2 bits of the consecutive two pixels of each mask as a result rightmost two bits of two consecutive pixels are swapped. Crossover and mutation is done in reversible manner. Genetic Algorithm is applied onto the embedded image to enhance the level of security.

The formula for Z- Transform is

$$X(z) = \sum_{m=0}^{\infty} x(m)\, r^{-m} e^{-j\omega m}$$

Limit is taken 0 to α as pixel value cannot be negative for an image. In the present implementation the value of r is taken as 1 and ω varies between 0<= ω <=2π. For a 2×2 sub image there are four pixel values in the mask and set of frequencies taken are: ω ={ 0, π/2, π, 3π/2} . The formula for forward Z- Transform is

$$Z(r\, e^{j\omega}) = \sum_{n=0}^{N-1} \sum_{\omega=0}^{P} g[n] e^{-j\omega n}$$

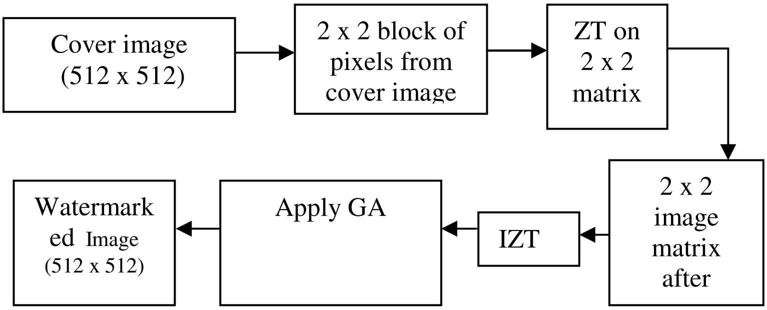


Fig. 1.1. The process to embed the Secret data into the source image

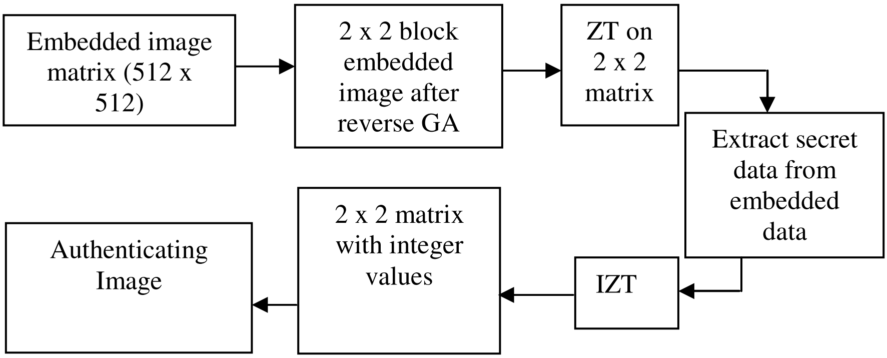

Embedded image matrix (512 x 512)
2 x 2 block embedded image after reverse GA
ZT on 2 x 2 matrix
Extract secret data from embedded data
IZT
2 x 2 matrix with integer values
Authenticating Image

Fig. 1.2. The process to extract Secret data from the watermarked image

Fig. 1. Schematic diagram of ANGAFDZT

and $e^{-j\omega n} = \cos\omega n - j\sin\omega n$, Where ω is frequency variables, varies from -∞ to +∞ and N is dimension of the matrix. The formula for inverse Z transform is

$$g[n] = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{\omega=0}^{p} F(\omega)e^{j\omega n}$$

And $e^{j\omega n} = \cos\omega n + j\sin\omega n$.

Schematic diagram of the technique is shown Figure1 of which Figure.1.1 shows process of encoding that of Figure.1.2 depicts the process of decoding. Algorithm of insertion and extraction are given in section 2.1 and 2.2 respectively. A complete example has also been illustrated in section 2.3.

2.1 Insertion Algorithm

The technique uses gray scale image of size p×q as input. Hidden image of size m×n is chosen. Four bits of hidden image is embedded in each mask of transformed coefficients in Z-domain followed by application of Genetic Algorithm.

Input : Host image of size p×q, hidden image of size m×n.
Output : Embedded image of size p×q.
Method : Insertion of hidden image bitwise into the gray scale image.

1. Obtain the size of the hidden image m×n
2. For each hidden message/image, read source image mask of size 2×2 in row major order. Apply Z-Transform onto the selected cover image mask (2×2) to obtain coefficients in transformed domain
3. Embed secret bits onto the second and third LSB position of the second coefficient of 2 x 2 mask. For embedding secret bits the dimension of the hidden image followed by the content is embedded. Dimension is extracted from the third line of the pgm format image and embedded onto the first four mask of the host image
4. Embed secret bits onto the third and fourth LSB position of the third coefficient of 2 x 2 mask
5. Copy second embedded coefficient onto fourth coefficient of the mask
6. Apply IZ-Transform to back the mask from Z domain to spatial domain
7. 2×2 embedded image mask of size 32 bits is taken as initial population in spatial domain. Perform New Generation operation on the initial population. Find out the minimum coefficient of the embedded image mask. If minimum is negative then subtract magnitude of minimum from each coefficient otherwise do nothing
8. Crossover is performed on the rightmost 3 bits from each byte of the New Generation is taken. A consecutive bitwise XOR is performed on it for the 3 steps. It will form a triangular form and first bit from each step is taken
9. Mutation is performed between rightmost 2 bits of the consecutive two pixels of each mask as a result rightmost two bits of two consecutive pixels are swapped

10. Repeat step 2 to 9 for the whole cover image
11. Stop.

2.2 Extraction Algorithm

The hidden image is received in spatial domain. The embedded image is taken as the input and the hidden message/ image size, content are extracted from it.

Input : Embedded image of size p×q.
Output : Host image of size p×q, hidden image of size m×n.
Method : Extract bits of hidden image from embedded image.

1. Reverse Mutation is performed on the rightmost 2 bits of two consecutive pixels of the each mask. For this rightmost two bits of two consecutive pixels are swapped
2. Reverse Crossover produce by consecutive bitwise XOR operation on the rightmost 3 bits of each byte in three steps. The first bit of each step is taken as the output
3. Read embedded image mask (of size 2×2) in row major order. Apply Z-Transform onto the embedded image mask to transform the embedded sub image from spatial to frequency domain so that four frequency components are regenerated
4. Extract the secret bits from the second coefficient of 2×2 mask on second and third LSB position. Replace hidden message/ image bit position in the block by '1'. For each eight extracted bits construct one image pixel of authenticating image
5. Extract the secret bits from the third coefficient of 2×2 mask on third and fourth LSB position. Replace hidden message/ image bit position in the block by '1'. For each eight extracted bits construct one image pixel of authenticating image
6. Repeat step 1 to 5 to regenerate hidden image as per size of the hidden image
7. Stop.

2.3 Example

Consider bits of Jet image (figure 2a) to be inserted into each mask of Lenna image (Figure 2c). Figure 2b shows pixels of Lenna image in spatial domain. Four bits of the Jet image is inserted into the Lenna image in 2×2 mask. Insertion is done in the second coefficient of each mask on second and third LSB bits and third coefficient on third and fourth LSB bits of the byte of Lenna. Resultant image after embedding is shown in Figure 2d in frequency domain and Figure 2e in spatial domain. Figure 2f shows New Generation. Figure 2g and 2h shows Crossover and Mutation.

01100001 Figure 2a: Bits of Jet Image	156 153 156 153 ↔ (Source sub image block) Figure 2b: Source Image Lenna	618 0 6 0 ↔ (Real Coefficients of transformed mask) Figure 2c: Source image Lenna after ZT	618 4 2 4 ↔ Figure 2d: Embedded Image Block
157 154 153 154 ↔ Stego sub image block after IZT Figure 2e: Stego Image Block after IZT	157 154 153 154 ↔ Stego sub image block after New Generation as no negative part is there Figure 2f: Stego Image Block after New Generation	158 154 153 154 ↔ Stego sub image block after Crossover Figure 2g: Stego Image Block after Crossover	158 154 154 153 ↔ Stego sub image block after Mutation Figure 2h: Stego Image Block after Mutation

Fig. 2. Encoding Process of ANGAFDZT

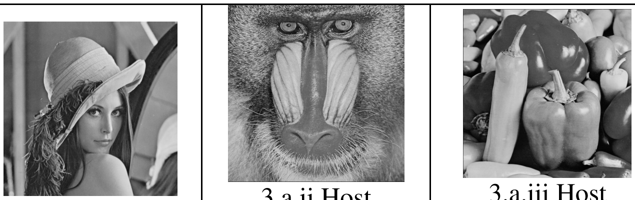
3 Result Comparison and Analysis

Extensive analysis has been made on various images using ANGAFDZT technique. This section represents the results, discussions in terms of visual interpretation and peak signal to noise ratio. Figure 3a shows the host images Lenna, Mandrill, Peppers. Figure 3b shows embedded Lenna, Mandrill, Peppers on embedding Jet image using ANGAFDZT. Figure 3c is the authenticating image Jet. Table I shows the PSNR value for each embedding against the source image. From the table it is seen that the maximum value of the PSNR is 40.973183 and that of minimum value of the PSNR is 40.701672. Table II shows the comparison of PSNR values of the proposed technique with the existing Chin-Chen Chang et al.[1]. In comparison with existing [1] it is seen that the proposed technique has better PSNR and capacity compared to the existing one. The following formula are used to calculate PSNR, MSE and IF (image fidelity).

$$PSNR = 10\ \log\left(max\left(\ I_{m,n}{}^{2}\right)/MSE\right)$$

$$MSE = \frac{1}{MN} * \sum_{m,n} (I_{1}\, m,n - I_{2} m,n)^{2}$$

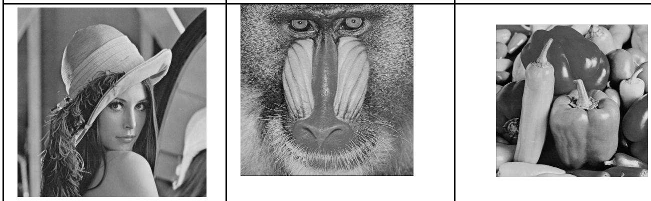
$$IF = 1 - \sum_{m,n} (I_{1_{m,n}} - I_{2_{m,n}})^{2} / \sum_{m,n} I_{2_{m,n}}^{2}$$



3.a.i. Host Lenna

3.a.ii.Host Mandrill

3.a.iii Host Peppers



3.b.i.Embedded Lenna

3.b.ii Embedded Mandrill

3.b.iii Embedded Peppers

3.c.i. Hidden Jet

Fig. 3. Visual Effect of Embedding in ANGAFDZT

Table 1. PSNR, MSE, IF values obtained for various images using ANGAFDZT

Host Image	***PSNR values***	***MSE Values***	***IF***
Lenna	40.917221	5.264515	0.999673
Mandrill	40.973183	5.197113	0.999721
Peppers	40.942974	5.233387	0.999706
Elaine	40.943676	5.232544	0.999747
Sailboat	40.927849	5.251648	0.999743
Boat	40.723259	5.504963	0.999710
Jet	40.701672	5.532394	0.999838

Table 2. Comparison of PSNR values between ANGAFDZT and existing[1]

Host Image	*PSNR values of ANGAFDZT*	*Capacity (bits) of ANGAFDZT*	*PSNR values of EXISTING [1]*	*Capacity (bits) of EXISTING [1]*
Lenna	40.917221	216000	30.34	36850
Mandrill	40.973183	216000	26.46	35402
Peppers	40.942974	216000	30.65	36804
Boat	40.723259	216000	29.75	36710
Jet	40.701672	216000	29.98	36817

4 Conclusion

The paper proposed a novel embedding approach termed as ANGAFDZT based on Z Transformation with genetic algorithm for gray scale images where large amount of information can be embedded. This paper shows that the proposed technique obtained better PSNR ratio for any message size than the existing approach Chin-Chen Chang et al.[1] as a result more data can be embedded with better visibility/quality. Effect of colors in images is the future scope of work as this paper works with gray scale images.

Acknowledgments. The authors express gratitude to the Department of Computer Science and Engineering, University of Kalyani and the PURSE scheme of DST, Govt. of India, under which the research has been carried out.

References

1. Chang, C.-C., et al.: Reversible hiding in DCT- based compressed images. Science Direct Information Science 177, 2768–2786 (2007)
2. Ghoshal, N., Mandal, J.K., et al.: Image Authentication by Hiding Large Volume of Data and Secure Message Transmission Technique using Mask (IAHLVDSMTTM). In: Proceedings of IEEE International Advanced Computing Conference, IACC 2009, Thapar University, Patiala, India, March 6-7 (2009) ISBN:978-981-08-2465-5
3. Ghoshal, N., Mandal, J.K., et al.: Masking based Data Hiding and Image Authentication Technique (MDHIAT). In: Proceedings of 16th International Conference of IEEE on Advanced Computing and Communications, ADCOM 2008, Anna University, December 14-17 (2008)

4. Ghoshal, N., Mandal, J.K.: A Novel Technique for Image Authentication in Frequency Domain using Discrete Fourier Transformation Technique (IAFDZTT). Malaysian Journal of Computer Science 21(1), 24–32 (2008) ISSN 0127-9094
5. Ghoshal, N., Mandal, J.K.: A Bit Level Image Authentication / Secrete Message Transmission Technique (BLIA/SMTT). Association for the Advancement of Modelling and Simulation Technique in Enterprises (AMSE), AMSE Journal of Signal Processing and Pattern Recognition 51(4), 1–13 (2008)
6. EL-Emam, N.N.: Hiding a large Amount of data with High Security Using Steganography Algorithm. Journal of Computer Science 3(4), 223–232 (2007) ISSN 1549-3636
7. Rechberger, C., Rijman, V., Sklavos, N.: The NIST cryptographic Workshop on Hash Functions. In: IEEE Security and Privacy, Austria, vol. 4, pp. 54–56 (January-February 2006)
8. Pavan, S., Gangadharpalli, S., Sridhar, V.: Multivariate entropy detector based hybrid image registration algorithm. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Philadelphia, Pennsylvania, USA, pp. 18–23 (March 2005)
9. Hashad, A.I., et al.: A Robust Steganography Technique using Discrete Cosine Transform Insertion. In: Enabling Technologies for the New Knowledge Society: ITI 3rd International Conference Information and Communication Technology (2005) ISBN:0-7803-9270-1
10. Moulin, P., O'Sullivan, J.A.: Information-theoretic analysis of information Hiding. IEEE Trans. on Info. Theory 49(3), 563–593 (2003)
11. Dumitrescu, S., Xiaolin, W., Wang, Z.: Detection of LSB steganography via sample pair analysis. IEEE Trans. on Signal processing 51(7), 1995–2007 (2003)
12. Chandramouli, R., Memon, N.: Analysis of LSB based image steganography techniques. In: Proc. of ICIP, Thissaloniki, Greece, pp. 1019–1022 (2001)
13. Wang, R.-Z., Lib, C.F., Lin, J.C.: Image hiding by optimal LSB substitution and Genetic algorithm. Pattern Recognition Society (2001)
14. Lin, C.Y., Chang, S.F.: A robust image authentication method surviving JPEG lossy compression. In: Proc. SPIE, San Jose, pp. 296–307 (January 1998)
15. Weber, A.G.: The USC-SIPI Image Database: Version 5, Original release: Signal and Image Processing Institute, University of Southern California. Department of Electrical Engineering (October 1997), `http://sipi.usc.edu/database/` (last accessed on January 20, 2011)

Speckle Noise Reduction Using Fourth Order Complex Diffusion Based Homomorphic Filter

Jyothisha J. Nair and V.K. Govindan

Department of Computer Science and Engineering,
National Institute of Technology Calicut, Calicut, 673 601, India
`{Jyothisha,vkg}@nitc.ac.in`

Abstract. Filtering out speckle noise is essential in many imaging applications. Speckle noise creates a grainy appearance that leads to the masking of diagnostically significant image features and consequent reduction in the accuracy of segmentation and pattern recognition algorithms. For low contrast images, speckle noise is multiplicative in nature. The approach suggested in this paper makes use of fourth order complex diffusion technique to perform homomorphic filtering for speckle noise reduction. Both quantitative and qualitative evaluation is carried out for different noise variances and found that the proposed approach out performs the existing methods in terms of root means square error (RMSE) value and peak signal to noise ratio (PSNR).

1 Introduction

One of the most important stages in image processing applications is the noise filtering. During image formation through an imaging device it may be subjected to various types of noise that may lead to degradation of the observed image. The noise may be additive or multiplicative in nature. Multiplicative noise, which is found in many real world signal processing applications are very difficult to be removed from the corrupted signal. This difficulty is mainly because of its multiplicative nature. As such noise will be amplified proportional to the pixel values; its presence will be amplified in the brighter areas and will be attenuated in the darker areas of the image. Hence, it will be difficult to develop statistical models for estimation and removal of the noise from the image which is corrupted by such multiplicative noise.

Image restoration techniques are based on modeling the degradation using a priori knowledge and applying the inverse process in order to restore the original image. Image restoration for improving the image quality may be employed as preprocessing techniques in a variety of applications such as pattern recognition, image processing and image compression. The presence of speckle noise in an imaging system reduces its resolution especially for low contrast images and suppression of speckle noise is an important consideration in the design of coherent imaging systems. For low contrast images, speckle noise is multiplicative in nature. Speckle noise creates a grainy appearance that can mask diagnostically significant image features and reduce the accuracy of segmentation and pattern recognition algorithms.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 895–903.
springerlink.com
© Springer-Verlag Berlin Heidelberg 2013

2 Related Works

Speckling is a common problem for different imaging modalities such as radio astronomy, synthetic aperture radar (SAR), ultrasound and laser imaging. Consequently, it promoted research and lead to many speckling reduction techniques. The most widely cited and applied filters in this category include Lee [1,2], Frost [3], Partial differential equation (PDE) based filters [4,5,6] and Complex diffusion based filters [7,8,9,10]. In Lee filter [1. 2], the output image is given as a linear function of central pixel and the average value of the pixels in the window. The Frost filter [3] output is based on an exponential filter kernel.

Partial Differential Equation (PDE) based methods [4,5,6], especially anisotropic diffusion (Perona-Malik), have proved to be particularly effective in pre-filtering images. Here, the basic idea is to deform an image, with a PDE and obtain the expected results as a solution to this equation. However, during the diffusion, the adjacent images in the stack are likely to diverge. 4th order PDEs and complex PDEs are suggested to solve this problem.

The concept of complex diffusion in image processing was introduced by Gilboa et al. [7] as an alternative to 2nd order anisotropic diffusion, which introduces blocky effects in images while processing. Several other authors also came up with modified complex diffusion techniques [8,9]. Fourth order non linear complex diffusion [10] is an improvement over the one proposed by Gilboa et al [7] in terms of edge preserving. Here also the fourth order complex diffusion technique is used for speckle noise reduction from digital images in log domain.

3 The Proposed Method

The speckle noise has complex amplitude which may be represented as given below:

$$a(x, y) = a_R(x, y) + ja_I(x, y)$$

where a_R and a_I are zero mean, independent Gaussian random variables for each (x,y) with some variance. The intensity field of speckle noise is given as

$$s(x, y) = |a(x, y)|^2 = a_R{}^2 + a_I{}^2$$

The image observation model [11] for such type of noise can be modeled as:

$$v(x, y) = u(x, y)s(x, y) + \eta(x, y)$$

where v(x, y) is the observed noisy image, u(x, y) is the original restored image, s(x, y) is the intensity of speckle noise and $\eta(x, y)$ is the detector noise which is additive in nature. Assuming the detector noise to be zero, the general observation model is:

$$v(x, y) = u(x, y)s(x, y) \tag{1}$$

In order to separate these two independent components and to facilitate their separate processing, we take logarithm transform on Eq. (1), thus

$$z(x,y) = \ln v(x,y) = \ln u(x,y) + \ln s(x,y)$$

If we process z(x, y) using a filter function h, we have

$$h(x,y) * z(x,y) = h(x,y) * \ln u(x,y) + h(x,y) * \ln s(x,y) \tag{2}$$

Here, '*' indicates convolution of the two function. Then Eq. (2) can be expressed in the form

$$Z(x,y) = U(x,y) + S(x,y)$$

Let g(x, y) be the desired enhanced image after filtering. As z(x, y) is the logarithm of v(x, y), we can yield g(x, y) using

$$g(x,y) = e^{Z(x,y)} = e^{U(x,y)} e^{S(x,y)}$$

$$g(x,y) = u_0(x,y) s_0(x,y)$$

Where $u_0(x,y) = e^{U(x,y)}$ and $s_0(x,y) = e^{S(x,y)}$

The enhancement approach using the foregoing concepts is summarized in Figure 1. This method is based on a special case of the homomorphic systems.

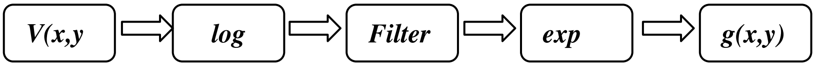


Fig. 1. The Flow chart of Homomorphic filtering

The homomorphic filtering approach [12] converts the multiplicative noise removal problem to additive noise removal problem which is easier to implement. The homomorphic filtering process refers to applying logarithm transformation on input image to convert the multiplicative noise to additive one, applying some inverse filter and finally taking the exponential of filter output to produce the restored image.

Here in this paper, fourth order complex diffusion based non-linear filter is proposed for the filtering part [10]. Complex diffusion is used as an efficient tool for denoising images by Gilboa [7] and is better compared to the ordinary PDEs. Proposed 4th order complex diffusion based homomorphic filtering method gives better performance with respect to Peak signal to noise ratio (PSNR). It is compared with the existing filters like 2nd order complex diffusion, Lee filter, Frost filter, PDE based filters and found superior.

The concept of complex diffusion in image processing was introduced by Gilboa et al. [7] as an alternative to 2nd order anisotropic diffusion, which introduces blocky effects in images while processing. This blocky effect is inherent in the nature of ordinary second order equations; it can be avoided by using complex diffusion. Complex diffusion is derived by combining the standard diffusion equation with the free Schrodinger equation. On application of the complex diffusion process, the real plane gives the low frequency components and the imaginary plane gives the high frequency components. The components in the real and imaginary plane are almost equivalent to that of the image convolved with a Gaussian and Laplacian of Gaussian (LOG) at various scales. Here, we are using a 4th order non linear complex diffusion [10], which is an improvement over its 2nd order counterpart in terms of preserving edges. The method is based on:

$$I^{t+1} = -\nabla^2 (c(\Im(I^t)))\nabla^2 I^t \tag{3}$$

where $\Im(.)$ takes the imaginary part, and

$$c(\Im(I^t)) = \frac{e^{i\theta}}{1+\left(\frac{\Im(I^t)}{k\theta}\right)^2}$$

with initial conditions $\Re(I^{t=0}) = I$ and $\Im(I^{t=0}) = 0$

k is the threshold parameter, θ is the phase angle and I is the original image. Finally the noise removed image is obtained by taking the exponentiation of the output obtained from the complex diffusion step.

4 Results and Discussions

The proposed method was implemented in MATLAB. For all the experiments, the parameters are chosen as k=0.08, θ=π/60, and the step size Δt=0.25. The proposed algorithm was applied on cameraman image for different amount of speckle noise variance and observed that it is better than existing methods like Lee filter, PDE based methods and even simple complex diffusion techniques.

To perform a quantitative comparison between the performances of the different filters, we computed some well-known speckle-reduction performance metrics. The first measure is the mean squared error (MSE), defined by Equation (4), where I_0 denotes the samples of the original image, I_f denotes the samples of the filtered image and M and N are the number of pixels in row and column directions, respectively.

$$MSE = \frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}\left[I_f(i,j) - I_o(i,j)\right]^2 \tag{4}$$

$$RMSE = \sqrt{MSE}$$

The Peak Signal to Noise Ratio (PSNR) is computed using the following equation

$$PSNR = 20\log_{10}\left[\frac{255}{RMSE}\right] \tag{5}$$

Figure 2, 3 and 4 shows the original, noisy and restored version of cameraman image for a noise variance of 0.02 and 0.004

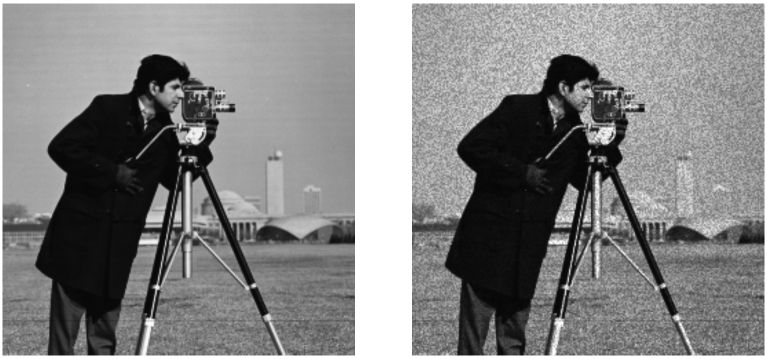


Fig. 2. Original Image, Speckled Image with variance 0.02

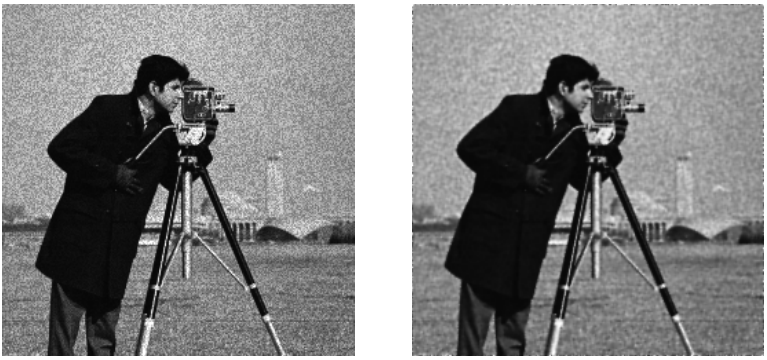


Fig. 3. Speckled Image with variance 0.02, Restored Image by the proposed approach

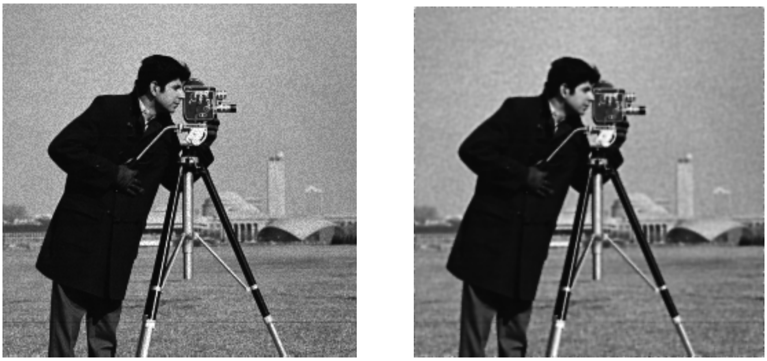


Fig. 4. Speckled Image with variance 0.004, Restored Image by the proposed approach

Table 1. Performance Comparison of proposed approach with other methods

Method	Speckle variance	RMSE	PSNR
Lee Filter	0.004	13.06	25.80
	0.02	14.31	25.01
	0.03	15.01	24.60
4th order PDE	0.004	14.68	24.79
	0.02	16.90	23.57
	0.03	18.1	21.11
Complex Diffusion	0.004	12.04	26.51
	0.02	13.87	25.28
	0.03	14.5	23.2
Homomorphic 4th order Complex Diffusion	0.004	9.81	28.28
	0.02	12.18	26.41
	0.03	13.68	25.40

The RMSE and PSNR values for the different speckle variances are given in the table 1. The proposed method converges to solution after 12 iterations for speckle variance 0.02 and for variance 0.04, it is 17 iterations. The performance of the proposed method was also compared with the existing methods like Lee, fourth order PDE, and simple complex diffusion. Figures 5 and 6 graphically display the PSNR values with respect to number of iterations for the two different variances of speckle noise.

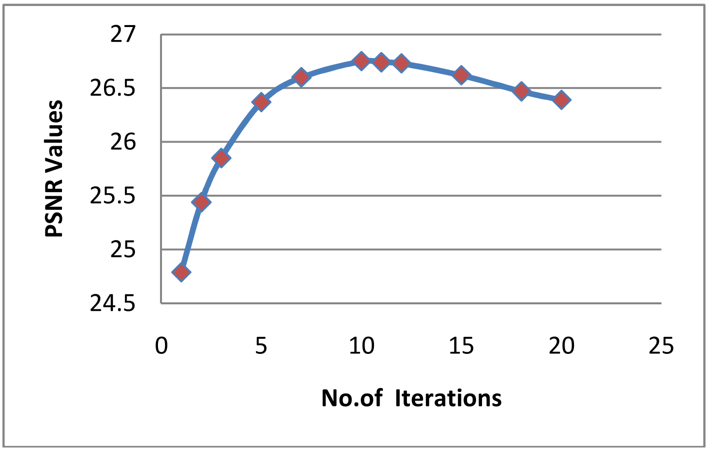


Fig. 5. No. of Iterations Vs PSNR for Speckle variance 0.02

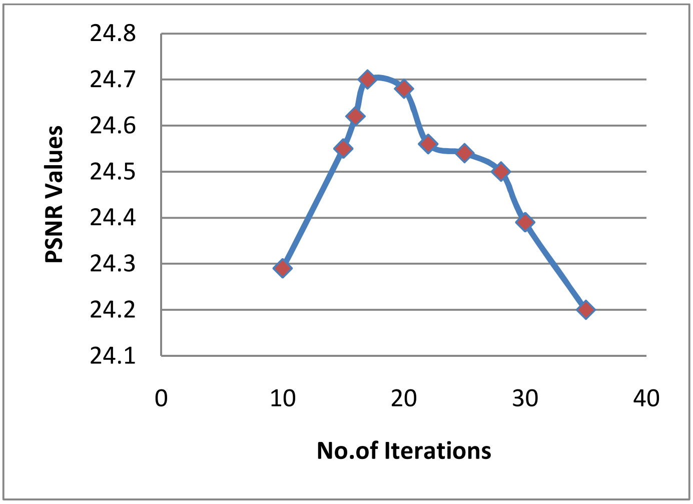


Fig. 6. No.of Iterations Vs. PSNR for Speckle variance 0.04

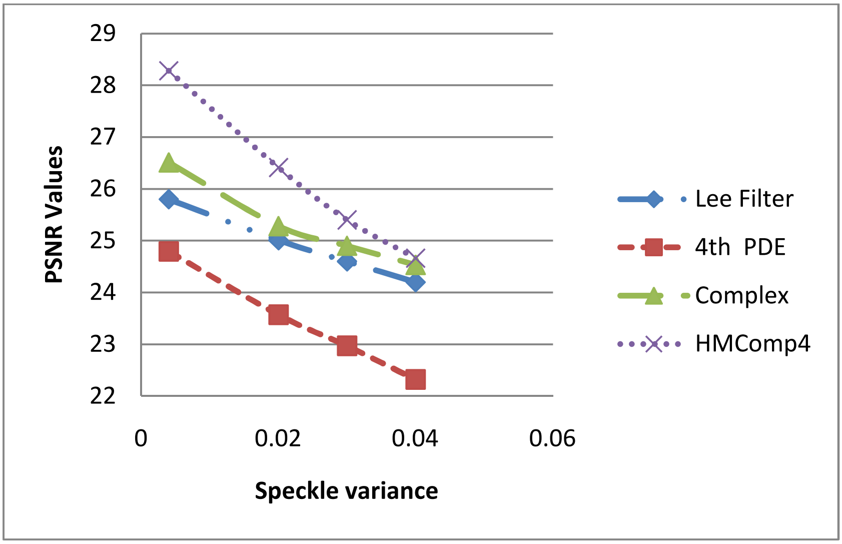


Fig. 7. PSNR computed against different noise sigma values

PSNR values and RMSE values for different methods are plotted in graphs and are given in fig 7 and 8. Experimental results demonstrate that the proposed approach provides the lowest RMSE and highest PSNR Value when compared to the existing approaches.

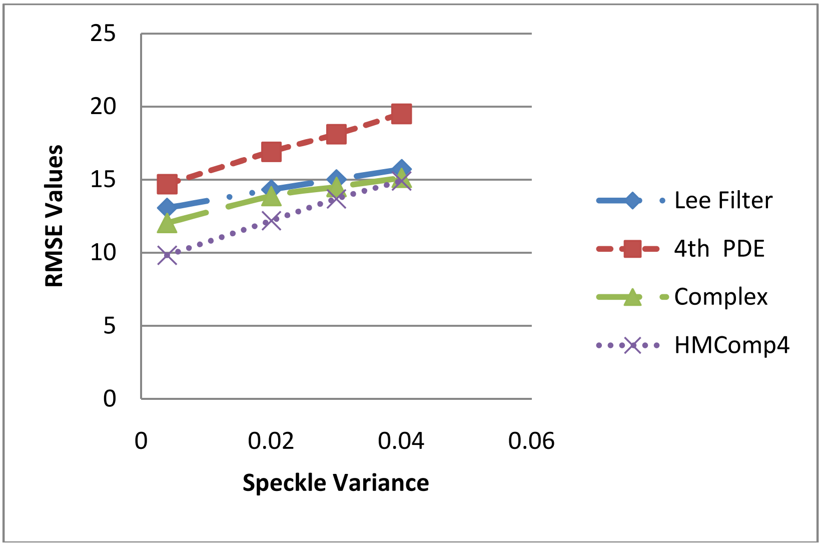


Fig. 8. RMSE computed against different noise sigma values

5 Conclusion

The presence of speckle noise in an imaging system reduces its resolution; especially for low contrast images and suppression of speckle noise is an important consideration in the design of coherent imaging systems. In this paper, a fourth order complex diffusion based homomorphic filtering is proposed for speckle noise reduction from digital images. Both quantitative and qualitative evaluations of the technique demonstrate the superior performance of the proposed approach when compared to the existing approaches.

References

1. Lee, J.S.: Speckle Analysis and Smoothing of Synthetic Aperture Radar Images. Computer Graphics and Image Processing 17, 24–32 (1981)
2. Lee, J.S.: Digital Image Smoothing and the Sigma Filter. Computer Vision. Graphics and Image Processing 24, 255–269 (1983)
3. Frost, V.S., Stiles, J.A., Josephine, A., Shanmugan, K.S., Holtzman, J.C.: A Model for Radar Images and Its Application to Adaptive Digital Filtering of Multiplicative Noise. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-4(2) (1982)
4. You, Y.L., Kaveh, M.: Fourth-order partial differential equations for noise removal. IEEE Trans. Image Process 9, 1723–1730 (2000)
5. Yu, Y., Acton, S.T.: Speckle reducing anisotropic diffusion. IEEE Trans. Image Process 11, 1260–1270 (2002)
6. Nadernejad, E., Koohi, H., Ssanpour, H.: PDEs-Based Method for Image Enhancement. Applied Mathematical Sciences 2(20), 981–993 (2008)

7. Gilboa, G., Sochen, N., Zeevi, Y.Y.: Image enhancement and de-noising by complex diffusion process. IEEE Trans. PAMI 25(8), 1020–1036 (2004)
8. Bernardes, R., Maduro, C., Serranho, P., Araújo, A., Barbeiro, S., Cunha-Vaz, J.: Improved adaptive complex diffusion despeckling filter. Opt. Express 18, 24048–24059 (2010)
9. Srivastava, R., Gupta, J.R.P., Parthasarthy, H.: Complex diffusion based speckle reduction from digital images. In: Proceeding of International Conference on Methods and Models in Computer Science ICM2CS 2009, pp. 1–6 (2009)
10. Rajan, J., Jeurissen, B., Sijbers, J., Kannan, K.: Denoising Magnetic Resonance Images Using Fourth Order Complex Diffusion. In: 13th International Machine Vision and Image Processing Conference IMVIP 2009, pp. 123–127 (2009)
11. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice-Hall, Inc., Upper Saddle River (1989)
12. Fan, C.-N., Zhang, F.-Y.: Homomorphic filtering based illumination normalization method for face recognition. Pattern Recognition Letters 32(10), 1468–1479 (2011)

A New Hybrid Approach for Denoising Medical Images

Deepa Bharathi[1] and Sumithra Manimegalai Govindan[2]

[1] KPR Institute of Engineering Technology, Coimbatore, Tamil Nadu, India
[2] BannariAmman Institute of Technology, Sathyamangalam, Tamil Nadu, India
`cool.deeps.143@gmail.com`, `mgsumithra@rediffmail.com`

Abstract. The most significant feature of diagnostic medical images is the removal of impulse noise which is commonly found in medical images and to make better image quality. In recent years, technological development has significantly improved in analyzing medical imaging. This paper proposes different hybrid filtering techniques for the removal of noise, by topological approach. The hybrid filters used here are hybrid median filter [hybrid min filter (H_1F) and hybrid max filter (H_2F)]. These filters are treated in terms of a finite set of certain estimation and neighborhood building operations. A set of such operations is suggested on the base of the analysis of a wide variety of nonlinear filters described in the literature. It is suggested from the simulation results that the proposed scheme yields better image quality after denoising. This approach is incorporated with spatial domain and frequency domain analysis. Results obtained by hybrid filtering technique are measured by the statistical quantity measures: Root Mean Square Error (RMSE) and Peak Signal-to-Noise Ratio (PSNR). Overall results indicate that the enhancement quality was performed well in proposed method when compared to other filtering techniques.

Keywords: Image enhancement, hybrid filter, PSNR, RMSE, denoising.

1 Introduction

In the early development of image processing, linear filters were the primary tools for image enhancement and restoration. Their mathematical simplicity and the existence of some desirable properties made them easy to design and implement. Moreover, linear filters offered satisfactory performance in many applications. However, they have poor performance in the presence of non additive noise and in situations where system nonlinearities or Gaussian statistics are encountered [1].

In image processing applications, linear filters tend to blur the edges and do not remove the existence of noise present in the image effectively. Previously, a number of schemes have been proposed for image denoising. Inherently noise removal from image introduces blurring in many cases. An adaptive standard recursive low pass filter is designed by Klaus Rank and Rolf Unbehauen by considering the three local image features like edge, spot and flats as adaptive regions with Gaussian noise [2]. Median filter has been introduced in [3]. Median filter now is broadly used in reducing noise and smoothing the images. In [4] topological median filter have used to improve conventional median filter. The better performance of the topological median

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 905–914.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

filters over conventional median filters is in maintaining edge sharpness. In [5], Yanchun proposed an algorithm for image denoising based on Average filter with maximization and minimization for the smoothness of the region, unidirectional median filter for edge region and median filter for the indefinite region. It was discovered that when the image is corrupted by both Gaussian and impulse noises, neither Average filter nor Median filter algorithm will obtain a result good enough to filter the noises because of their algorithm.

An improved adaptive median filtering method for denoising impulse noise reduction was carried out in [6]. An adaptive median filter (AMF) is the best filter to remove salt and pepper noise of image sensing [7]. The Computer Tomography images were denoised using curvelet and wavelet transforms in [8]. The objective of this study is to develop new hybrid filtering techniques and investigate their performance on medical images.

This work is organized as follows: Section 2 discusses types of noises involved in medical imaging. In Section 3 various existing filtering techniques for de-noising the medical images was discussed. Section 4 deals with proposed hybrid filtering techniques for de-noising the impulse noise in the medical images. In Section 5, both quantitative (RMSE and PSNR) and qualitative comparisons are provided. Section 6 puts forward the conclusion drawn by this paper.

2 Types of Noises

2.1 Salt and Pepper Noise

Salt and pepper noise is a form of noise typically seen on images. It represents itself as randomly occurring white and black pixels. A “spike” or impulse noise drives the intensity values of random pixels to either their maximum or minimum values. The resulting black and white flecks in the image resemble salt and pepper. This type of noise is also caused by errors in data transmission.

2.2 Speckle Noise

In medical literature, speckle noise is referred to as ‘texture’ and may possibly contain useful diagnostic information. Physicians generally have a preference for the original noisy images, more willingly, than the smoothed versions because the filter, even if they are more sophisticated, can destroy some relevant image details. In our work, we recommend hybrid filtering techniques for removing speckle noise in ultrasound images. The speckle noise model has the following form (* denotes multiplication). For each image pixel with intensity value f_{ij} ($1 \leq i \leq m$, $1 \leq j \leq n$ for an m x n image), the corresponding pixel of the noisy image g_{ij} is given by,

$$g_{i,j} = f_{i,j} + f_{i,j} * n_{i,j} \tag{1}$$

where, each noise value n is drawn from uniform distribution with mean 0 and variance σ^2

2.3 Gaussian Noise

Gaussian noise is statistical noise that has a probability density function of the normal distribution (also known as Gaussian distribution). Noise is modeled as additive white Gaussian noise (AWGN), where all the image pixels deviate from their original values following the Gaussian curve. That is, for each image pixel with intensity value fij ($1 \leq i \leq m$, $1 \leq j \leq n$ for an m x n image), the corresponding pixel of the noisy image gij is given by,

$$g_{i,j}= f_{i,j} + n_{i,j} \tag{2}$$

where, each noise value n is drawn from a zero -mean Gaussian distribution.

2.4 Poisson Noise

J = imnoise (I,'poisson') generates Poisson noise from the data instead of adding artificial noise to the data. If I is double precision, then input pixel values are interpreted as means of Poisson distributions scaled up by 1e12. For example, if an input pixel has the value 5.5e-12, then the corresponding output pixel will be generated from a Poisson distribution with mean of 5.5 and then scaled back down by 1e12. If I is single precision, the scale factor used is 1e6. If I is uint8 or uint16, then input pixel values are used directly without scaling.

3 Existing Filtering Techniques

In this section, we provide the definitions of some existing filters. The image processing function in a spatial domain can be expressed as

$$g(p)=\gamma(f(p)) \tag{3}$$

where γ is the transformation function, f (p) is the pixel value (intensity value or gray level value) of the point p(x, y) of input image, and g(p) is the pixel value of the corresponding point of the processed image.

3.1 Median Filter

The best-known order-statistic filter in digital image processing is the median filter. It is a useful tool for reducing salt-and- pepper noise in an image. The median filter [12] plays a key role in image processing and vision. In median filter, the pixel value of a point p is replaced by the median of pixel value of 8-neighborhood of a point ‘p’.

$$g(p)=median\{f(p), where\ p\in N_8(p)\} \tag{4}$$

The median filter is popular because of its demonstrated ability to reduce random impulsive noise without blurring edges as much as a comparable linear low pass filter. However, it often fails to perform well as linear filters in providing sufficient smoothing of non impulsive noise components such as additive Gaussian noise.

3.2 Center Weighted Median Filter

In uniform areas of the image the median and average filters will not differ by much. One way to improve the performance of the filter is give weights to the contribution of each pixel [9]. This is shown in the following figure where the origin is in the center of the template. The w's are the weights given to the pixel values of the corresponding neighborhood. If pixel p_i is in the neighborhood of pixel p and is associated with weight w_i, then w_i is the number of times p_i is repeated in the median calculation.

W(-1,1)	W(0,1)	W(1,1)
W(-1,0)	W(0,0)	W(1,0)
W(-1,-1)	W(0,-1)	W(1,-1)

Fig. 1. CWM Filter geometry

One common weighting is to weigh the pixels closest to the center with a higher value. These are called Center Weighted Median (CWM) filters. An extreme would be to let w (0, 0) only be greater than one. When w (0, 0) is very high, it is the identity filter. This can be useful in preserving features that correspond to single pixels. When w (0, 0) is low it is the ordinary median filter.

3.3 Hybrid Median Filter

Hybrid Median filter [10] is of nonlinear class that easily removes impulse noise while preserving edges. The hybrid median filter plays a key role in image processing and vision. In comparison with basic version of the median filter, hybrid one has better corner preserving characteristics. This filter is defined as

$$g(p) = avg\left\{\begin{matrix} avg\left\{\begin{matrix} f(p), p \in N_4(p) \\ f(p), p \in C_4(p) \end{matrix}\right. \\ f(p) \end{matrix}\right\} \tag{5}$$

4 Proposed Hybrid Filtering Techniques

In this section, we will provide the definition of proposed hybrid filters. These filters are yet to be applied by researchers to remove the Gaussian noise in the ultrasound medical images.

4.1 Hybrid Min Filter (H_1F)

Hybrid min filter plays a significant role in image processing and vision. Hybrid min filter is not a usual min filter. Min filter [11] recognizes the darkest pixels gray value and retains it by performing min operation. In min filter each output pixel value can be calculated by selecting minimum gray level value of $N_8(p)$. H_1F filter is used for removing the salt noise from the image. Salt noise has very high values in images. It is also proposed for Gaussian noise removal from the medical image. It is expressed as:

$$g(p) = \min \left\{ \begin{array}{l} median \left\{ \begin{array}{l} f(p),\ p \in L_3(p) \\ f(p),\ p \in R_3(p) \end{array} \right\} \\ f(p) \end{array} \right\} \tag{6}$$

In hybrid min filter, the pixel value of a point p is replaced by the minimum of median pixel value of LT neighbour's of a point 'p', median pixel value of RT neighbour's of a point 'p' and pixel value of 'p'.

4.2 Hybrid Max Filter (H_2F)

Hybrid max filter is not a usual max filter. Hybrid max filter plays a key role in image processing and vision. The brightest pixel gray level values are identified by max filter. In max filter [11] each output pixel value can be calculated by selecting maximum gray level value of N_8 (p). H_2F filter is used for removing the pepper noise from the image. It is also proposed for Gaussian noise removal from the medical image. It is expressed as:

$$g(p) = \max \left\{ \begin{array}{l} median \left\{ \begin{array}{l} f(p),\ p \in L_3(p) \\ f(p),\ p \in R_3(p) \end{array} \right\} \\ f(p) \end{array} \right\} \tag{7}$$

In hybrid max filter, the pixel value of a point p is replaced by the maximum of median pixel value of LT neighbours of a point 'p', median pixel value of RT neighbours of a point 'p' and pixel value of 'p'.

5 Experimental Results, Analysis and Discussions

The proposed hybrid filtering techniques have been implemented using MATLAB 10.0. The performance of various filtering techniques like Gaussian filter(GF),Median filter(MF),Weiner filter(WF),Trimmed average filter (TAF), Center-weight Median filter(CW-MF),Rank-order Median filter,(RO-MF),Hybrid minimum filter(H1F),Hybrid maximum filter,(H2F) is analyzed and discussed. The measurement of medical image enhancement is difficult and there is no unique algorithm available to measure enhancement of medical image. We use statistical tool to measure the enhancement of medical images. The Root Mean Square Error (RMSE) and Peak Signal-to-Noise Ratio (PSNR) are used to evaluate the enhancement of medical images.

$$RMSE = \sqrt{\frac{\sum (f(i,j) - g(i,j))^2}{mn}} \tag{8}$$

$$PSNR = 20 \log_{10}\left(\frac{255}{RMSE} \right) \tag{9}$$

Here f (i,j) is the original medical image with impulse noise , g(i,j) is an enhanced image and m and n are the total number of pixels in the horizontal and the vertical dimensions of the image. If the value of RMSE is low and value of PSNR is high then the enhancement approach is better. The original noisy image and filtered image of MRI brain cancer image, MRI Knee image, CT image of twins in the womb and the MRI brain image obtained by various filtering techniques are shown in the following Figures.

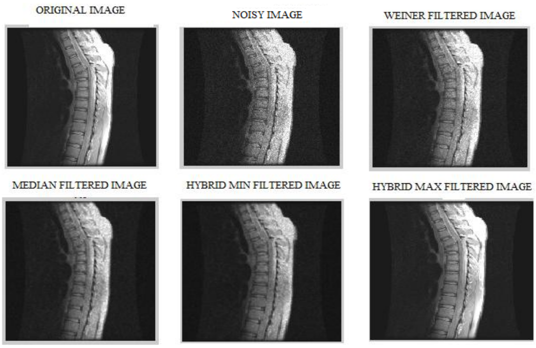


Fig. 2. Filtering analysis for MRI Knee image

Fig 2 represents the performance analysis of various filtering for the MRI Knee image corrupted by the speckle noise. It is understand from the result that denoising result is better in case of median filter than Weiner filter, but the proposed method yields better result than other filtering techniques. Comparative results were shown by both hybrid min filter and hybrid max filter.

Fig 3 indicates the performance of the various filtering techniques for the image corrupted by Gaussian noise. The performance of the proposed method is seen

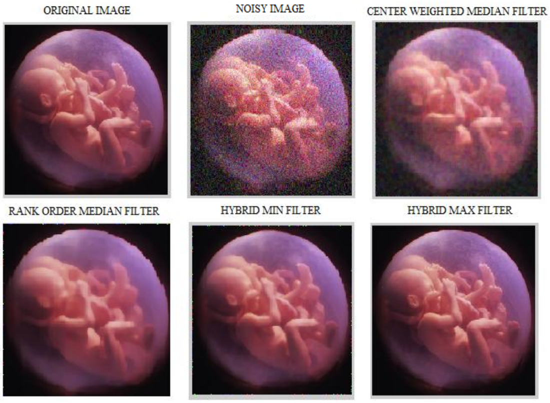


Fig. 3. Filtering analysis of CT image for twins in the womb

superior than the other filtering techniques. Next to proposed method, Rank-order median filter is showing fine denoising results.

The performance analysis for various filtering techniques for MRI Brain image corrupted by salt and pepper noise is shown in fig 4.Gaussian filter removes the noises present at the edge of image, while the trimmed average filter performs better than Gaussian filter. Comparative results are seen for hybrid min and max filter, thereby proposed method performs in a better manner than other.

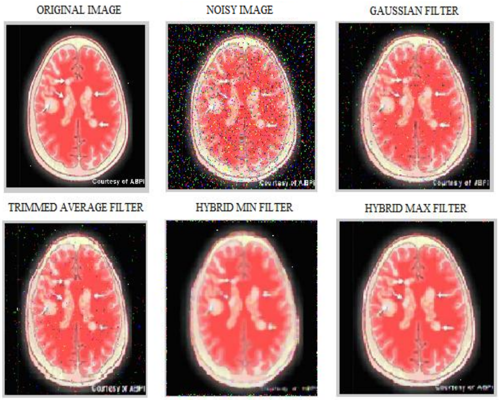


Fig. 4. Filtering analysis for MRI Brain image

Fig 5 indicates the performance of the various filtering techniques for the image corrupted by salt and pepper noise. It is observed from the results that the proposed method is showing enhanced performance than the other filtering techniques even for higher decibel(dB) of noise level.

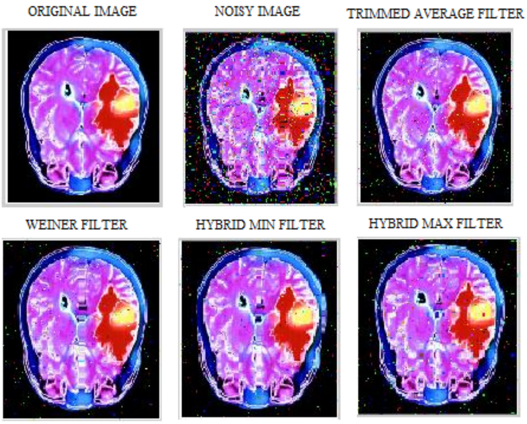


Fig. 5. Filtering analysis for MRI Brain cancer image

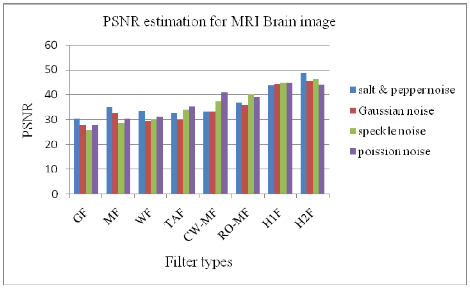


Fig. 6. Performance of PSNR estimation for MRI Brain image

The performance of PSNR estimation using different filtering methods for MRI Brain image affected by various noises is shown in fig 6.It is observed that for all types of noises, the proposed hybrid filter is performing in an enhanced manner than the others. In case of speckle noise PSNR value increases in linear manner for the filtering methods as considered in this paper.

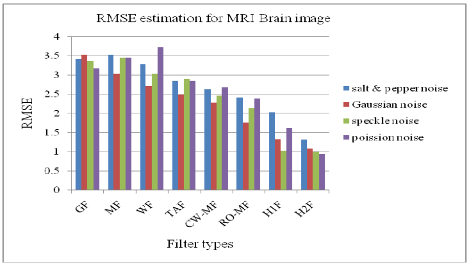


Fig. 7. Performance of RMSE estimation for MRI Brain image

It is inferred from the fig 7 that the mean square error for enhanced and noisy image is seen low in proposed method for all types of noises. The maximum error value is seen for Weiner filter for the poisson noise and for the median filter corrupted by the Gaussian noise. Comparative results is shown for the proposed hybrid filter for the speckle noise.

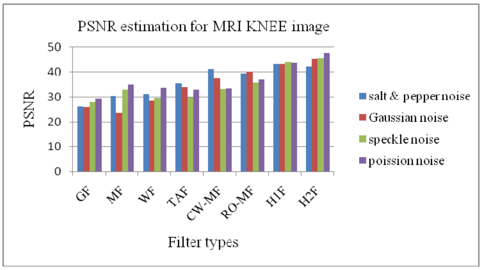


Fig. 8. Performance of PSNR estimation for MRI Knee image

It is seen from the fig 8 that, better image quality is obtained after denoising, while using hybrid max filter for all type of noises except salt and pepper noise and in case hybrid min filter is performing in an enhanced manner.

It is inferred from the fig 9 than the mean square error for enhanced and noisy image is seen low in proposed method for all types of noises. The maximum error value is seen for Gaussian filter for salt and pepper noise and for Weiner filter corrupted by the Gaussian noise. Comparative results is shown for the proposed hybrid filter for the speckle noise.

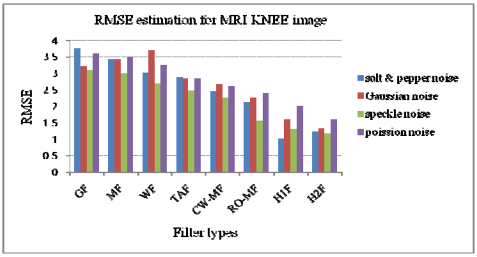


Fig. 9. Performance of RMSE estimation for MRI Knee image

Table 1. Performance estimation for the MRI brain image for the salt & pepper noise

EVALUATION PARAMETER	TYPES OF FILTERS							
	GF	MF	WF	TAF	CW – MF	RO-MF	H_1F	H_2F
PSNR	27.9	32	29.5	30	33.1	35.8	44.2	45.7
RMSE	3.12	3.5	3.2	2.8	2.6	2.4	2	1.6

Table 2. Performance estimation for the MRI KNEE image for Speckle noise

EVALUATION PARAMETER	TYPES OF FILTERS							
	GF	MF	WF	TAF	CW –MF	RO-MF	H_1F	H_2F
PSNR	27.9	33	29.5	30.1	33.2	35.8	44.2	45.7
RMSE	3.1	3.0	2.71	2.48	2.28	1.56	1.32	1.18

6 Conclusion

In this work, various hybrid filtering techniques for denoising medical images were introduced. To demonstrate the performance of the proposed techniques, the experiments have been conducted on blood cell , MRI Knee and Brain image and CT image for the twins in the womb, in- order to compare the proposed methods with many other well known techniques. The performance of noise removal by hybrid filtering techniques is measured using quantitative performance measures such as RMSE and PSNR. The experimental results indicate that the one of the proposed hybrid filter, Hybrid Max Filter performs significantly better than many other existing techniques and it gives the best results after successive iterations. The proposed filtering technique are computationally efficient than the other, since the running time of proposed method is 1.5ms, whereas the exsisting methods needs around 3.7ms.

References

[1] Pitas, I., Venetsanopoulos, A.N.: Nonlinear Digital Filters: Principles and Applications. Springer, NJ (1990)

[2] Rank, K., Unbehauen, R.: An Adaptive Recursive 2-D Filter for Removal of Gaussian Noise in Images. IEEE Transactions on Image Processing, 431–436 (1992)

[3] Tukey, J.W.: Nonlinear (nonsuperposable) methods for smoothing data. In: Proc. Congr. Rec. EASCOM 1974, pp. 673–681 (1974)

[4] Senel, H.G., Peters, R.A., Dawant, B.: Topological Median Filter. IEEE Trans on Image Processing 11(2), 89–104 (1992)

[5] Wang, Y., Liang, D., Ma, H., Wang, Y.: An Algorithm for Image Denoising Based on Mixed Filter. In: Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21-23, pp. 9690–9693 (2006)

[6] Juneja, M., Mohan, R.: An Improved Adaptive Median Filtering Method for Impulse Noise Detection. International Journal of Recent Trends in Engineering 1(1), 274–278 (2009)

[7] Al-amri, S.S., Kalyankar, N.V., Khamitkar, S.D.: A Comparative Study of Removal Noise from Remote Sensing Image. International Journal of Comp. Science. 7(1), 32–36 (2010)

[8] Sivakumar, R.: Denoising of Computer Tomography images using curvelet transform. ARPN Journal of Engineering and Applied Sciences 2(1), 21–26 (2007)

[9] `http://www.librow.com/articles/article-8`

[10] Gonzalez, R., Woods, R.: Digital Image Processing. Adison -Wesley, New York (1992)

Detection of Cerebral Aneurysm by Performing Thresholding-Spatial Filtering-Thresholding Operations on Digital Subtraction Angiogram

Jubin Mitra and Abhijit Chandra

Department of Electronics & Telecommunication Engineering
Bengal Engineering and Science University, Shibpur, Howrah, India
`jm61288@gmail.com`,
`abhijit922@yahoo.co.in`

Abstract. Cerebral aneurysm (CA) has been emerging as one of the life threatening diseases which have developed a deep concern amongst the neurologists in recent years. To be specific, it shows devastating characteristic due to the formation of abnormal bulging of artery in human brain followed by its rupture. Therefore detection of this abnormality prior to the rupture becomes inevitably essential to save our lives to a great extent. This paper throws enough light in detecting cerebral aneurysm of various sizes by combining the operations of spatial filtering and thresholding in an elegant way. A number of Digital Subtraction Angiogram (DSA) images, affected by cerebral aneurysm of various magnitudes, have been taken into consideration in this connection. Finally, the affected area has been marked with red colour to make it more prominent than the other parts of the image.

1 Introduction

Cerebral aneurysm (CA) is a cerebrovascular disorder resulting from the localised bulging of blood vessel. It may be of congenital type or may develop with age due to the weakness or certain injury in the wall of the blood vessels. CA is of saccular kind in nature located mostly at the bifurcation of the arteries, known as Circle of Willis [1]. Cerebral Aneurysm is also known as intracranial aneurysm in many articles.

This disorder of arteries in brain may take place at various locations. Extensive medical analysis have found that the frequent sites of occurrence of CA are Internal Carotid Artery (30-35%), Anterior Cerebral Artery (33-34%) and Middle Cerebral Artery (20%) [2]. Apart from having single Cerebral Aneurysm; multiple Cerebral Aneurysm does also exist which has an occurrence probability of 0.2 to 0.4. It is more predominant in women than in men. In USA, 10-12 billion people are affected with cerebral aneurysm, with rupture ratio 1:10,000 people per year.

Common causes of occurrence are due to infectious disease, hypertension, smoking and genetic abnormalities. Symptoms might range from nausea, severe headache, vision impairment, unconsciousness to no symptoms at all. If not treated at proper time, large Cerebral Aneurysm can lead to a bigger problem known as Subarachnoid Haemorrhage (SAH) which results due to the rupture of blood vessels [1]-[3].

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 915–921.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

Modern treatment method for cerebral aneurysm can be divided into non-invasive and invasive techniques. Non-invasive method includes Transcranial Doppler technique; whereas invasive techniques like CT scan, MRI or nuclear perfusion scanning are more popularly employed in this regard.

In recent past, few approaches were adopted by researchers towards the detection of cerebral aneurysm with the aids of image processing techniques. Segmentation of giant CA, composed of lumen and thrombus, has been carried out successfully by one multi-level object detection scheme based on Lattice Boltzmann Method (LBM) [4]. Experimental results have demonstrated the fact that with the proposed method both lumen and thrombus can be well segmented.

Another such method utilizes 2D Digital Subtraction Angiography (DSA) imaging technique, based on the calculation of Time to Peak (TTP) and Time Duration (TD) of flow of contrast agent in the blood vessels [5]. Combined TTP and TD algorithm was successful in detecting medium size aneurysm. However, the algorithm has certain weaknesses in the sense that without the help of the Multiscale Vessel Enhancement Filtering (MVEF), the quality of vessel segmentation is bad and may cause wrong detection of blood vessel segment. Moreover, the authors could not develop any suitable approach which can be useful in detecting small size aneurysm.

Our study of focus is on the detection of cerebral aneurysms of various sizes from 2D Cerebral Angiogram. The proposed method requires much less resource and is significantly faster than the previously proposed time based parametric technique [5]. Our approach uses simple yet elegant techniques like dual thresholding operation with one round of smoothing filtering to extract the region of interest. The resultant image is then superimposed over the original one to highlight and identify the aneurysm. A number of test images have been considered into our analysis to prove the supremacy of the proposed approach. Entire simulation work was performed by using MATLAB 7.0 software.

2 Theoretical Background

A. Cerebral Aneurysm (CA)

Aneurysm refers to a weak area in the wall of a blood vessel which results in bulging or ballooning out of the blood vessel. Aneurysm in the brain occurs if there is any weakened area in the wall of a blood vessel. Such an aneurysm which occurs in a blood vessel of the brain is called cerebral aneurysm (CA) [1]-[2]. It may be present from birth which is known as congenital aneurysm or it may be developed in later part of life due to a variety of reasons.

Aneurysms can be classified in terms of size and shape of the bulging. Berry aneurysm and giant berry aneurysm are more common in adults than children. The size of berry aneurysm can range from a few millimeters to over a centimeter. Giant berry aneurysms can reach well over 2 centimeters in size [2]. Amongst all aneurysms, multiple berry aneurysms are hereditary more often than other types of aneurysms. Apart from these two, large, medium and small size aneurysms can also be observed for many patients. The typical size of large size aneurysm is between 16 to 25 mm; whereas that of medium size aneurysm can vary between 6 to 15 mm. On the other hand, small size aneurysms may be as small as less than 5 mm [2]. Any factors like atherosclerosis, trauma and infection which can injure the blood vessel are mainly responsible for cerebral aneurysms.

Statistical measurement over a number of people has revealed that approximately 5% of the total population has some type of aneurysm in their brain, but only a small number of these aneurysms cause symptoms or rupture. Risk factors behind a ruptured cerebral aneurysm include family history of cerebral aneurysms and other medical problems such as high blood pressure, polycystic kidney disease, coarctation of the aorta and so on.

B. Digital Subtraction Angiography(DSA)

It is an improved method over traditional angiography to increase the visualization of the blood vessels drastically. Images are produced by subtracting pre-contrasting image from subsequent contrast enhancing image [3]. Image contrast is increased by administrating contrast medium gradually under the supervision of a radiologist.

C. Image Processing Techniques

This paper involves a very few image processing operations which seem to be essential in detecting cerebral aneurysm. These include Normalization, Global binary thresholding and Spatial averaging filtering.

a) Normalization:

In order to enhance the contrast of any digital image, normalization plays a very important role. More specifically, irrespective of the minimum and maximum gray level present in the original image; the contrast of the processed image is maximized through the process of normalization. If the gray level of the original image at location (x, y) is denoted by $f(x, y)$; then after normalizing the resultant gray value can be determined from the following equation [6]-[8]:

$$g(x,y) = \frac{f(x,y) - \min_{x,y} f(x,y)}{\max_{x,y} f(x,y) - \min_{x,y} f(x,y)} . \Delta \tag{1}$$

where Δ signifies the maximum possible gray value for the image under consideration.

b) Global binary thresholding:

Thresholding is a process of converting any gray image into binary image that will consist of extreme white and extreme black pixels only. For any given threshold λ, it can be outlined as [6]-[7], [9]:

$$g(x,y) = \begin{cases} \Delta & if\ f(x,y) < \lambda \\ 0 & if\ f(x,y) \geq \lambda \end{cases} \tag{2}$$

The term 'global' in binary thresholding simply implies that the value of the threshold λ remains fixed irrespective of the spatial location and gray value of the pixels.

c) Spatial averaging filtering:

Filtering is a process to extract some useful information from a given set of data depending upon the application. Averaging or smoothing spatial filtering finds out the average brightness of the entire image and thereby eliminates the minute details available in any image. As a matter of fact, this has the effect to blur the entire scene. The mathematical formulation of spatial averaging filtering takes the form [7]-[9]:

$$g(x,y) = \sum_{j=0}^{W} \sum_{i=0}^{W-1} f(x+i, y+j).m(i,j) \quad (3)$$

The parameter 'W' in (3) indicates the size of the filter mask 'm'.

3 Proposed Thresholding-Spatial Filtering-Thresholding (TSFT) Algorithm for Detecting Cerebral Aneurysm

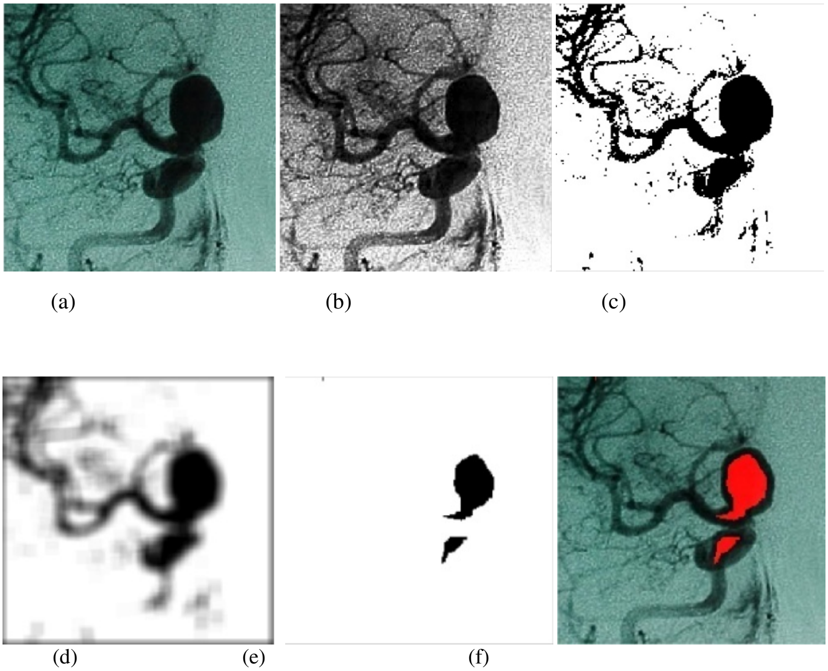
In this work, one novel algorithm has been proposed for the detection of cerebral aneurysm (CA) of various sizes. Simple yet useful point processing and spatial filtering operations have been combined wisely in achieving our goal. As a matter of fact, the name of the proposed algorithm is chosen as Thresholding-Spatial Filtering-Thresholding (TSFT). The entire algorithm has been implemented in MATLAB 7.0 software and summarized below:

Step 1 : Image acquisition through Digital Subtraction Angiography (DSA)
Step 2 : Conversion of the image into 8-bit gray-scale
Step 3 : Normalization of the gray-scale image
Step 4 : Segmentation of major vessels and abnormal outgrowth of vessel wall from the background through the operation of first binary thresholding
Step 5 : Scanning of averaging window over the binary image, as obtained in Step 4
Step 6 : Erosion of arteries and veins and interconnection of small vessels through the operation of second binary thresholding
Step 7 : Superposition of detected region over the original DSA image and high lighting the affected area with a red marker

As obvious from the above steps, the algorithm involves basic computations of image processing like normalization, global binary thresholding and average spatial filtering. The power of the algorithm lies in the sequential operations of first thresholding, spatial filtering and second thresholding which have made our proposed Thresholding-Spatial Filtering- Thresholding (TSFT) approach simple yet elegant than other complicated approaches, available in literature.

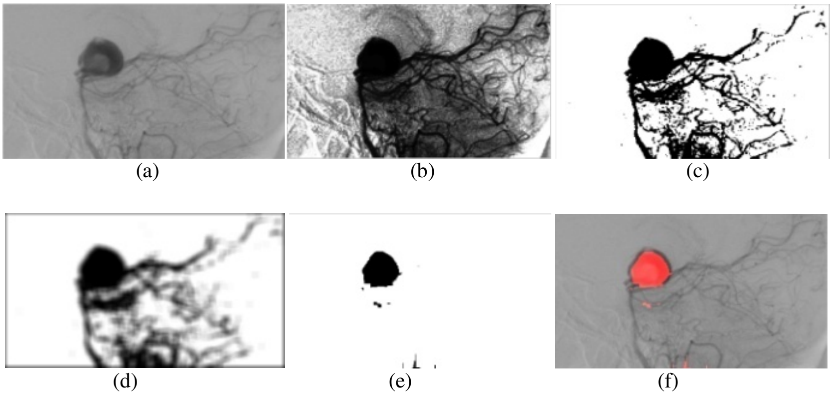
4 Simulation Results and Discussions

In order to show the competence of our proposed algorithm in detecting CA of various sizes, three such aneurysm affected DSA test images have been taken into consideration in this analysis. Our achievement in detecting aneurysm has been depicted in Fig. 1 through Fig. 3 below. In each of the figures, actual DSA image has been given in part (a). Resulting images after the process of normalization and first round of thresholding have been provided in part (b) and part (c) respectively. Effect of smoothing filter has been described in part (d), followed by second round of thresholding in part (e). Resultant superimposed image has been shown in part (f) which marks the affected area with a red color.


(a) (b) (c)

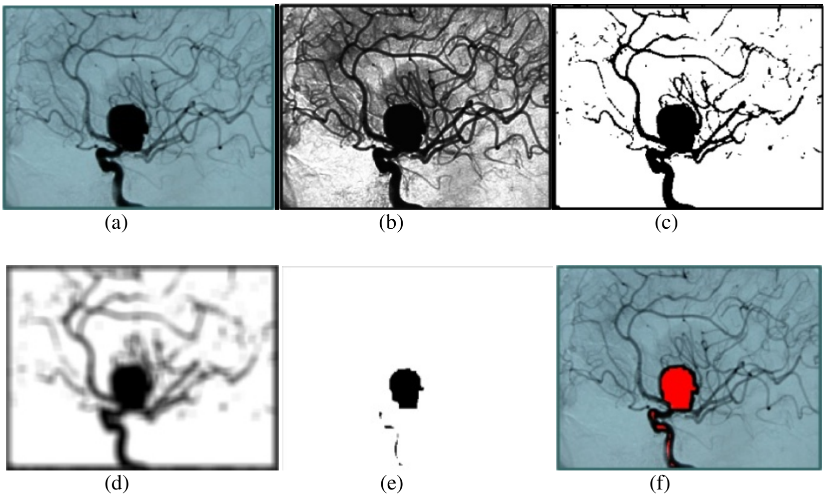
(d) (e) (f)

Fig. 1. Detection of Cerebral Aneurysm for 1st test image


(a) (b) (c)

(d) (e) (f)

Fig. 2. Detection of Cerebral Aneurysm for 2nd test image


(a) (b) (c)

(d) (e) (f)

Fig. 3. Detection of Cerebral Aneurysm for 3rd test image

Our algorithm requires three parameters which needs to be tuned in an efficient way to solve the purpose. The optimum set of those values has been listed in table 1.

Table 1. Optimum values of the parameters

Name of the parameters	Optimum value
First threshold (λ_1)	40
Size of filter mask (W)	10
Second threshold (λ_2)	1

As can be observed from the processed images, our proposed approach is capable enough in finding out the appearance and proper location of aneurysm. Moreover, it is very much efficient in highlighting the affected area with a higher degree of accuracy. As for example, the dark spot in part (a) and the red mark in part (f) cover approximately same area in the above images. Moreover for multiple CA, proposed approach recognizes both aneurysms, as identified in Fig. 1(f). The supremacy of TSFT algorithm lies in its simplicity in the sense that it incorporates very few fundamental point processing and spatial filtering approaches. This has made the proposed solution very much attractive in the field of medical imaging.

5 Conclusions

Cerebral aneurysm can successfully be detected by our proposed approach of smooth filtering sandwiched between two consecutive thresholding operations. However, proper selection of three user-defined parameters, namely first and second threshold values and the size of the filter mask are very much crucial and therefore needs special attention. Optimum selection of these values can lead to proper detection of artery, vein and even small aneurysm. Future research may be carried out in the direction of detecting small aneurysm and multiple aneurysms with a higher degree of precision by varying those parameters in an adaptive way.

References

[1] Gasparotti, R., Liserre, R.: Intracranial Aenurysms. European Radiology 15, 441–447 (2005)

[2] Wijdicks, E.F., Kallmes, D.F., Manno, E.M., et al.: Subarachnoid hemorrhage: neurointensive care and aneurysm repair. Mayo Clin. Proc. 80, 550–559 (2005)

[3] Strother, C.M., et al.: Parametric Color Coding for Digital Subtraction Angiography. American Journal of Neuroradiology 31, 919–924 (2010)

[4] Hunt, W.E., Hess, R.M.: Surgical Risk as Related to Time of Intervention in the Repair of Intracranial Aneurysms. Journal of Neurosurgery 28(1), 14–20 (1968)

[5] Wang, Y., Courbebaisse, G., Zhu, Y.M.: Segmentation of Giant Cerebral Aneurysms using a Multilevel Object Detection Scheme Based on Lattice Boltzmann Method. In: Proc. IEEE International Conference on Signal Processing, Xi'an, China, September 14-16 (2011)

[6] Zakaria, H., Kurniawan, A., Mengko, T.L.R., Santoso, O.S.: Detection of Cerebral Aneurysms by Using Time Based Parametric Color Coded of Cerebral Angiogram. In: Proc. International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, July 17-19 (2011)

[7] Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging 13(1), 146–165 (2004)

[8] Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn., pp. 595–611. Pearson Education, ISBN 81-7808-629-8

[9] McAndrew, A.: An Introduction to Digital Image Processing with MATLAB Notes for SCM 2511 Image Processing, pp. 1–264. School of Computer Science and Mathematics, Victoria University of Technology

[10] Pratt, W.K.: Digital Image Processing, 3rd edn. Wiley (2001)

Author Index

Aarthi, K. 739
Aarthi, R. 19
Abdulkadir, Ozcan 289
Abhishek, Kumar 561
Adeli, Ali 27
Ajith, Meghana 749
Ajitha Shenoy, K.B. 749
Alam, Mansaf 153
Ali, Mehreen 385
AlSudairi, Mohammed A.T. 253
Anzar, S.M. 833
Arora, Kumud 569
Arun Fera, M. 319
Arvind, C.S. 757
Asisha, Gottumukkala 215

Babukarthik, R.G. 479, 491
Bag, Anil Kumar 115
Balakrishnan, Kannan 355
Bandyopadhyay, Rajib 115
Banerjee, Sandipan 823
Bharathi, Deepa 905
Bhattacharyya, Nabarun 115
Bose, Shilpi 37

Chaki, Rituparna 469
Chakraborty, Abirlal 37
Chakraborty, S. 885
Chakraborty, Soubhik 107
Chakravorty, Sandeep 405
Chandra, Abhijit 915
Chattopadhyay, Samiran 37
Cheng, Xian-yi 57
Chilambuchelvan, A. 739
Chitra, S. 611
Chitrakala, S. 215

Dammavalam, Srinivasa Rao 145
Das, Arindam 649
Das, Chandra 37
Das, Monotosh 691
Das, Satrajit 649
Dass, Ashish Kumar 709
Datta, S.K. 885
De Sarkar, Ajanta 445
Dhavachelvan, P. 479, 491
Dinesh Acharya, U. 11
Diwakar, P.G. 757
Dubey, Gaurav 521
Durga Bhavani, S. 541

Elanangai, V. 813
Elloumi, Mourad 243

Fairooz, S.K. 787
Faizan Farooqui, Md. 623, 631
Farhangi, Mohammad Mehdi 681
Fathy, Mahmood 681

Ganesh Karthik, M. 595
Gantayat, S.S. 551
Garg, Poonam 569
Geetha Ramani, R. 433
George Amalarethinam, D.I. 395
Ghadirli, Hossein Movafegh 87
Ghosal, Prasun 649
Ghosal, S.K. 767
Govindan, Sumithra Manimegalai 905
Govindan, V.K. 895

Guo, Yanhui 861
Gupta, Kusum 279
Gupta, Savita 459

Haluška, Tomáš 273
Hamzeh, Ali 27
Hari, Vidya 175
Hima Bindu, G. 511

Ishwarya, C. 375

Jaganathan, Suresh 319
Jain, Alok 699
Jaiswal, Ajay 77
Javad Zomorodian, M. 27
Jawahar Senthil Kumar, V. 813
Jayaram Reddy, A. 581
Jeyarani, R. 375
Jhaveri, Maitri 195
Joby, Vivek 97
Josyula, Ramakrishna 135

Kale, Geetanjali 675
Kalpana, B. 611
Kamatchi, V. 365
Kammani, Abdullah 67
Kanji, Rakesh 797
Kanmani, S. 263, 341
Kapasi, Abbas 195
Khamrui, A. 885
Khan, Nilofar 777
Khan, Wasim 777
Kokila, R. 729
Kokilavani, T. 395
Korah, Reeba 719
Koundal, Deepika 459
Krishna Prasad, M.H.M. 145
Krishnaveni, V. 861
Kulkarni, P.J. 503
Kumar, Atul 639
Kumar, Prabhat 561
KumarDas, Asit 805

Lakshmeeswari, G. 511

Maddala, Seetha 145
Madhavi, B.K. 787
Madhavi Latha, M. 233
Maity, Debashis 871
Maity, Hirak Kumar 871
Maity, Santi P. 871
Majumder, Durjoy 161
Mandal, J.K. 767, 797, 885
Mandipati, Srivardhini 215
Mani, V. 757
Mani Kumar, Nerella Arun 851
Manjula Shenoy, K. 11
Manusha, G.V. 581
Maragathavalli, P. 263
Mitra, Jubin 915
Mohan, J. 861
Mohisin, G.S. 581
Mokaddem, Ahmed 243
Mukesh, Rajeswari 365
Mukharjee, Sourabh 329
Munivara Prasad, K. 595
Muqeem, Md. 623
Muralikrishna, S.N. 749
Murugan, R. 719
Muyeba, Maybin 531

Nagamalai, Dhinaharan 289
Nagarajan, G. 207
Nagaveni, N. 375
Nair, Jyothisha J. 895
Naveen Kumar, S.K. 843
Neela Madheswari, A. 175
Nithin Krishnan, N.U. 97
Nithya, S. 181
Nuli, Uday A. 503

Omkar, S.N. 757

Padmavathi, P. 299
Padmavathi, S. 19
Pal, Anita 639
Palanivel, K. 181
Panat, Ashish R. 659
Panduranga, H.T. 843
Pankaj, Dhanya S. 413
Pareek, Jyoti 195
Patil, Anita S. 659
Patil, Varsha 675
Paul, J.S. 355
Pauliček, Róbert 273
Phadikar, Santanu 805
Phalguna Krishna, E.S. 595
Pillai, Jyothi 531
Piratla, Aditya 691
Preethi Raj, S. 215

Qasim Rafiq, Md. 631

Rajakumar, 365
Raju, R. 479, 491
Rajya Lakshmi, D. 511
Ranichandra, C. 385
Ranjan, Rajeev 561
Rasheeda Shameem, M. 667
Rastgarpour, Maryam 87
Raw, R.S. 309
Ray, Amitava 405
Razia Naseem, M. 667
Rizwan Beg, Md. 623, 631
Roseline, R.A. 125
Roshan, Sahana 561
RoyChoudhury, I. 885

Sabitha, G. 299
Sadaf, Kishwar 153
Safeena, Rahmath 67
Saha, Suprativ 469
Sahukar, Latha 233
Salini, P. 341
Sankara Nayaki, K. 97
Sathidevi, P.S. 833, 851
Sen, Soumilla 445
Senthilnath, J. 757
Sethukkarasi, R. 667
Shabbiruddin, 405
Shanthi, S. 433
Sharath Kumar, H.S. 843
Sharma, Meena 77, 423
Sharma, Minakshi 329
Sharma, Sanjay Kumar 279
Sheeba, J.I. 299
Shen, Xue-hua 57
Shet, K.C. 11
Shi, Quan 57
Shial, Rabindra Kumar 709
Sil, Jaya 805
Sinaee, Mehrnoosh 27
Singh, Archana 521
Singh, N. 309
Singh, Niraj Kumar 107
Singh, Sukhwinder 459
Soman, K.P. 19
Soryani, Mohsen 681
Sreelakshmi, R. 97
Srinivas, Y. 511
Srinivasan, S. 375
Sriram Bharat, K. 385
Srivastava, Kamal Kumar 639
Subanivedhi, N.K. 667
Sumathi, P. 125
Sur, P. 885
Surendran, Jayalakshmi 691
Suvarna Vani, K. 541
Swami, Preety D. 699

Thangavel, P. 729
Thilaga Shri Chandra, A.P. 861
Thilagavathy, A. 739
Thyagarajan, K.K. 207
Toor, Vikas 309
Tripathy, B.K. 551, 581
Trivedi, Malhar 195
Tudu, Bipan 115

Vaidyanathan, Sundarapandian 1
Vanjare, Ashoka 757
Varghese, Abraham 355
Varghese, Reji Rajan 355
Vasanth, K. 813
Vasista, T.G.K. 253
Važan, Pavel 273
Vidhya, P. 225
Viginesh, S. 49
Vijay Krishnan, K. 49
Vijay Kumar, T.V. 521
Vijayraghavan, G. 49
Vishwakarma, Rajeev G. 423
Vivekanandan, K. 299
Vyas, O.P. 531

Wilscy, M. 413